and markedly influence social responses. It has been well known that under speech competition situations, vocally expressed emotions play an important role in modulating brain, behavioral, and body physiological responses to the attended voice. For example, some previous studies have shown that in dichotic-listening situations in which two voices are presented simultaneously at each ear, an utterance that is pronounced with an anger prosody activates both the superior temporal sulcus (STS; Grandjean et al., 2005; Sander et al., 2005) and the amygdala (Sander et al., 2005), relative to an utterance with an emotionally neutral prosody. Also, the utterance with an anger prosody leads to significant changes in both behavioral performance (e.g., reaction times to target articulation) and body physiological responses (e.g., skin conductance response [SCR]) (Aue, Cuny, Sander, & Grandjean, 2011). Note that in these studies meaningless utterances (pseudowords) spoken in different prosodies were used as the stimuli, and participants performed speech-content-irrelevant tasks (e.g., gender discrimination of a target speaker) without voluntarily processing the speech content. Thus, it is unclear whether emotional signals can be used as valid unmasking cues in speech recognition.

Dupuis and Pichora-Fuller (2014) approached this issue by examining listeners' recognition of speech content in a babble noise-masking condition. They found that words portrayed in negative emotion (fear or disgust) prosodies are recognized better than those for other emotions. Dupuis and Pichora-Fuller suggested that fear-prosody-enhanced recognition is caused by certain distinctive acoustical cues, such as the higher mean fundamental frequency, greater mean range of fundamental frequencies, and/or threatening information that captures auditory attention. Moreover, Iwashiro, Yahata, Kawamuro, Kasai, and Yamasue (2013) induced unpleasant emotion by using semantically negative words, and participants were asked to select the word heard at the attended ear from among four words presented visually on screen. Surprisingly, the reaction time was longer when a semantically negative word was presented at one of the ears in dichotic listening. This result suggests that when emotion is induced at the semantic level, it does not help but interferes with participants' performance.

Recognition of target speech in "cocktail-party" environments (with multiple talkers speaking simultaneously) largely depends on the facilitation of selective attention to the target speech (Schneider et al., 2007). Despite increasing evidence pointing to emotional modulation of auditory processing, it is still unclear whether introducing emotional information in target speech affects selective attention to the target speech against a highly complex background. A recent study by Mothes-Lasch, Becker, Miltner, and Straube (2016) has suggested that auditory background complexity is a crucial factor that modulates neural responses to emotional prosody. Specifically, the middle superior temporal region (mSTR) responds more strongly to emotionally negative prosody than to neutral prosody only under conditions of low auditory complexity (when a target voice is presented with a single animal sound), but not of high auditory complexity (when a target voice is presented with five animal sounds). Also, the prosody effect in the mSTR is not affected by attention. On the other hand, the prosody effect in the amygdala or in anterior superior temporal cortex is gated by attention but not by background complexity. Clearly, how the speech prosody, voice emotion, attention to target, and background signals are integrated in the brain is still largely unknown. Thus, establishing a new task paradigm is needed in order to specifically investigate how emotional processing affects a listener's selective attention to target speech, thereby facilitating recognition of that target speech against speech-on-speech informational masking.

It has been shown that two masking talkers cause the greatest informational-masking effect on target-speech recognition for both English speech (Freyman et al., 2004; Freyman et al., 1999; Rakerd, Aaronson, & Hartmann, 2006) and Chinese speech (X. Wu et al., 2007), and that the perceived spatial-separation-induced release of target speech from informational maskers is also highest under a two-talker masking condition, as compared to masking conditions with other numbers of masking talkers (Freyman et al., 2004; Freyman et al., 1999; X. Wu et al., 2007). Therefore, in the present study we applied a two-talker speech masker in order to simulate "cocktail-party" listening conditions, and we evaluated the unmasking effect of emotional signals on recognition of the target speech against the two-talker masking background.

In the auditory domain, emotions can be represented and expressed in various ways (Frühholz, Trost, & Kotz, 2016). Although previous studies have shown that the emotions expressed in speech prosody affect speech processing (Aue et al., 2011; Dupuis & Pichora-Fuller, 2014; Grandjean et al., 2005; Sander et al., 2005), it is still unclear whether emotionally conditioning the voice of a target speech whose prosody is emotionally neutral can also modulate recognition of the speech under "cocktail-party" listening conditions. It has been shown that the voice of a speaker is perceived as a typical attribute that is highly specific to the speaker, and that human brains show voice-selective responses in auditory cortex (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000). However, recognizing emotions from prosody involves the right frontoparietal operculum, bilateral frontal pole, and left frontal operculum (Adolphs, Damasio, & Tranel, 2002). It has also been confirmed that a speaker's voice can be effectively associated with other features (e.g., face) of this speaker through conditional learning, leading to that conditioned feature becomes an unmasking cue for improving speech recognition under "cocktail-party" listening conditions (e.g., Gao et al., 2014). Moreover, the familiarity of a target voice enhances target speech recognition under informational-masking conditions (Huang et al., 2009). Thus, it is of importance and interest to know whether emotionally conditioning a target

speaker's voice can affect recognition of the target speech under "cocktail-party" conditions.

To introduce a negative emotional valence to the target speaker's voice without altering the acoustical features of that voice, in our study we used a conditional-learning paradigm to temporally combine the target speaker's voice with an aversive sound.

In the paradigm of classical conditioning, the emotional significance of a conditioned stimulus (CS) can be obtained by pairing it with an unconditioned stimulus with negative emotional value (e.g., a loud noise, aversive picture, or electric shock; Kluge et al., 2011; Morris, Buchel, & Dolan, 2001; Weisz, Kostadinov, Dohrmann, Hartmann, & Schlee, 2007; Zou, Huang, Wu, & Li, 2007). The acquired emotion information leads to an attentional bias to conditioned threat-associated stimuli in both humans and laboratory animals (Du, Li, Wu, & Li, 2009; Du, Wu, & Li, 2010, 2011; Koster, Crombez, Van Damme, Verschuere, & De Houwer, 2005; Lei, Luo, Qu, Jia, & Li, 2014; Pischek-Simpson, Boschen, Neumann, & Waters, 2009; Z. Wu, Ding, Jia, & Li, 2016) and modulates neural activities in the auditory cortex and amygdala in both humans and laboratory animals (Armony, Quirk, & LeDoux, 1998; Du, Wang, Zhang, Wu, & Li, 2012; Morris et al., 2001; Morris, Öhman, & Dolan, 1998; for a recent review, see Grosso, Cambiaghi, Concina, Sacco, & Sacchetti, 2015).

In the present study, the emotional-learning paradigm involved pairing the target speaker's voice with loud female screaming, which has a marked negative emotional valence. The question whether the conditioning of the target voice can enhance recognition of the target speech against informational masking was examined by analyses of both speech-recognition performance and electrodermal (skin-conductance) responses. To control for the familiarity effect caused by learning, a (control) participant group received control learning by pairing the same target voice with an emotionally neutral sound. Note that this emotional-learning paradigm is able to induce negative emotional valence for a target speaker's voice while maintaining the acoustical attributes unchanged (i.e., without affecting the voice acoustical parameters).

Moreover, to compare the unmasking effects of emotional conditions with the unmasking effects of other typical unmasking cues, perceived spatial separation was used as a reference unmasking cue in this study. Perceived spatial separation between a target and a masker can induce a robust release of masking, particularly from two-talker-speech informational masking (Arbogast et al., 2002; Freyman et al., 1999; Huang et al., 2009; H. Li, Kong, Wu, & Li, 2013; L. Li et al., 2004; Schneider et al., 2007; X. Wu et al., 2005).

When binaurally presenting an auditory signal via headphones, introducing an interaural time difference (ITD) between the left and right ears can change the perceived laterality of the auditory image (e.g., Zhang, Lu, Wu, & Li, 2014). In the present study, the ITD of the target and that of the maskers were manipulated to create two types of perceived laterality relationships: (1) perceived co-location, when the target and maskers were perceived as coming from the same ear, or (2) perceived separation, when the target and maskers were perceived as coming from different ears. To our knowledge, interactions between two concurrent unmasking cues in "cocktail-party" listening conditions have not been well studied. Thus, we also designed this study to investigate the relationship between the unmasking effect of learned emotion (a non-spatial cue) and that of perceived separation (a spatial cue).

In the present study, nonsense (meaningless) sentences with three keywords were used. Previous studies had shown that pre-presented lip-reading priming cues mainly facilitate recognition of the first keyword of the target sentence (C. Wu et al., 2013), and that signals with emotional significance have priority access to selective attention and can be quickly detected in competitions among sensory inputs (Anderson, 2005; Eastwood et al., 2001; Fox, 2002; LeDoux, 1998; Öhman et al., 2001). Thus, the effect of keyword position in a target sentence was also examined.

To confirm whether a negative feeling could be reliably induced by the aversive stimulus used in this study during emotional learning, and whether an emotionally neutral feeling would be induced by the neutral stimulus during control learning, 9-point self-assessment manikin (SAM) ratings (Bradley & Lang, 1994) on each of the unconditioned stimuli (the negatively emotional one and the emotionally neutral one) were carried out by all participants at the end of the experiment. Since the participants in the emotional-learning group got more exposure to the emotionally negative sound, they might have adapted to the aversive sound. Thus, it would be of interest to know whether the two groups were different in their reports of negative feelings related to the aversive sound.

In summary, the main purpose of this study was to use an emotional-conditioning paradigm to introduce a negative emotional valence for the target speaker's voice and then examine whether the emotional conditioning of the target voice has a valid unmasking effect on speech recognition under speech-masking conditions.

Method

Participants

Thirty-two Mandarin-Chinese-speaking university students (18 females and 14 males between 18 and 27 years old, with a mean age of 21.9 years) participated in this study. They were right-handed, had normal pure-tone hearing thresholds (no more than 25 dB HL), and had bilaterally symmetric hearing (difference of no more than 15 dB between the two ears) at all tested frequencies between 0.125 and 8 kHz. All participants gave informed consent before the experiment and were paid a

modest stipend for their participation. The procedures of this study were approved by the Committee for Protecting Human and Animal Subjects at the School of the Psychological and Cognitive Sciences at Peking University.

Apparatus and materials

The participant was seated in a sound-attenuated chamber (EMI Shielded Audiometric Examination Acoustic Suite). The acoustical signals were digitized at a sampling rate of 22.05 kHz, transferred from the PCI Express Sound Blaster X-Fi Xtreme Audio software (Creative Technology Ltd, Singapore), and binaurally presented to participants using Sennheiser (Hanover, Germany) HD 265 linear headphones with a sound pressure level (SPL) of 56 dB for the target speech at each ear. Calibration of the sound level was conducted using the Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824; Larson Davis, USA). The SPL of the masker was adjusted in order to produce four signal-to-noise (masker) ratios (SNRs) of 0, – 4, – 8, and – 12 dB.

The speech stimuli were Chinese nonsense sentences, which were syntactically correct but not semantically meaningful (X. Wu et al., 2005; Yang et al., 2007). The direct English translations of the sentences were similar but not identical to the English nonsense sentences that were developed by Helfer (1997) and had been used in previous studies (e.g., Freyman et al., 1999; L. Li et al., 2004).

Each nonsense Chinese sentence has 12 syllables (also 12 characters), which included three keywords (the subject, predicate, and object) of two syllables each. For example, the English translation of a Chinese sentence was "This *polyester* will *expel* that *stomach*" (the keywords are italicized). In the nonsense sentences, no contextual support was provided for recognizing each of the keywords. The development of the Chinese nonsense sentences has been described elsewhere (X. Wu et al., 2005; Yang et al., 2007).

The target speech was spoken by a young female talker (Talker A). The speech masker was a 47-s loop of digitally combined continuous recordings of Chinese nonsense sentences spoken by two other female talkers (Talkers B and C), and the SPLs of the two masking talkers' speech sounds were the same. Each of the three talkers spoke different sentences, and the keywords of the maskers did not appear in any of the target sentences. All of the nonsense sentences were spoken in an emotionally neutral prosody.

The perceived laterality of the speech stimuli was manipulated by setting different values of the ITD. The target sentence was presented at the right ear, leading the left ear by 2 ms. Thus, due to the auditory precedence effect (Wallach, Newman, & Rosenzweig, 1949), listeners always perceived the target speech as coming from the right ear. Under the condition of perceived co-location, the masker was also presented with the right-leading ITD of 2 ms and perceived as coming from the same (right) ear as the target. Under the condition of perceived separation, the masker was presented with a left-leading ITD of 2 ms and thus was perceived as coming from the left ear. Note that shifts between the two conditions did not change either the SNR or the diffuseness/compactness of the sound images.

The intense aversive sound used for paired learning was a loud female screaming, adopted from the International Affective Digitized Sounds (IADS; Bradley & Lang, 2007) database with a peak SPL of 80 dB and a duration of 4.09 s. This screaming sound effectively caused aversive feeling according to the self-reports of the participants. After they had listened to the aversive stimulus, the 9-point SAM ratings (Bradley & Lang, 1994) from all participants confirmed the effectiveness of the aversive stimulus at inducing negative feelings (emotional valence = 1.74 ± 0.88; emotional arousal = 7.63 ± 1.37).

A 6-s sound of train whistling obtained from the IADS database with a peak SPL of 60 dB was chosen as the emotionally neutral (control) sound, which had neutral emotional valence (5.08 ± 1.42) and moderate emotional arousal (5.21 ± 1.53) across to the participants' SAM ratings.

Procedures

Half of the 32 participants were assigned to the emotional learning group, and the other half to the neutral (control) learning group, with gender, age, and hearing thresholds balanced across the groups. For the 16 participants in the emotional learning group and those in the control learning group, there were three within-subjects variables: (1) conditioning stage (before learning/control learning, after learning/control learning), (2) perceived laterality relationship (separation, co-location), and (3) SNR (0, – 4, – 8, or – 12 dB). In total there were 16 (2 × 2 × 4) conditions for each participant, and 15 trials (15 target sentences) were used in each condition. Two blocks were used for the perceived laterality relationships. The order of presentation of the co-location and separation blocks was counterbalanced across the participants in each group, and the four SNRs were arranged randomly in each block. Note that for the 16 participants in the control learning group, the experimental procedures were identical, except for that the auditory stimulus used for paired learning was an emotionally neutral sound (train whistling).

In a trial, the participant pressed a response button to start the masking speech, and 1 s later a target sentence was presented along with the masker. The masker terminated at the same time as the target sentence. After listening to the multiple-talker speech, participants were instructed to repeat the entire target sentence as best as they could. The experimenters scored whether each of the two syllables for each of the three keywords had been identified correctly. Before the formal testing, participants were trained in a session to ensure that they had fully understood the procedures and could perform the task correctly.

The target sentences used in the training session were different from those used in the formal testing.

After the initial speech-recognition testing, the participants in the emotional learning group received 21 learning trials of paired learning associating the target voice with the aversive sound. In each trial, the target talker's (Talker A) voice, speaking nonsense sentences without the presentation of the masker, was first presented binaurally, followed by the aversive sound, with a short interval that changed randomly in a range from 4 to 6 s. The nonsense sentences spoken with the target-talker voice did not appear in the speech-recognition task. After the emotional learning/control learning manipulations, participants performed the speech-recognition task again, which was the same as the one before the learning/control learning manipulation, except for the different sentences used as the target sentences.

At the end of the experiment, the participants were instructed to rate their subjective emotional feelings caused by either the aversive sound or the neutral sound, using 9-point SAM scales for both emotional valence (1, *most unpleasant*; 5, *neutral*; 9, *most pleasant*) and emotional arousal (1, *lowest*; 5, *moderate*; 9, *highest*) (Bradley & Lang, 1994). Each of the two sounds was repeated five times, and the order of the ten sound presentations was arranged randomly for each participant. Averaged emotional valence values and emotional arousal value were obtained for each individual participant.

Recordings of skin conductance responses

Skin conductance responses (SCRs) were recorded during the speech-recognition task using a Biopac MP150 (GSR100c module) with the Acknowledge software (Biopac System Inc., Santa Barbara, USA), digitized at a sampling rate of 1000 Hz.

Two Ag–AgCl disposable electrodes were attached to the palmar surface of the second and third fingers of the nondominant (left) hand. After the electrodes had been placed, participants waited 5 min and were asked to hold a deep breath for a few seconds. The experimenters observed their SCR increases to ensure that they had good contact with the recording system. The raw data were filtered offline by a 512-point digital low-pass filter with a cutting frequency at 3 Hz and then square-root-transformed to reduce data skew (Morris et al., 2001).

Results

Effects of learned emotion and perceptual separation on speech recognition

To determine the speech-recognition threshold, a logistic psychometric function,

$$y = \frac{1}{1 + e^{-\sigma(x-\mu)}}$$

was fitted to each participant's recognition accuracy measured across the four SNRs, where y is the probability of correct recognition of the keywords, x is the SNR corresponding to y, σ determines the slope of the psychometric function, and μ is the SNR corresponding to 50% correct on the psychometric function.

For each of the listening conditions, the group means that best fit the psychometric function for the learning group are plotted in Fig. 1A, along with the group-mean values for recognition accuracy. The recognition accuracy and psychometric function after emotional learning (solid lines) were higher than those before emotional learning (dash lines) under both the co-location condition and the separation condition. Also, perceived separation between the target and the masker markedly improved the recognition accuracy.

Figure 1B plots a scatter diagram presenting the speech-recognition thresholds (μ) of individual participants in the emotional learning group before and after the emotional learning manipulation. Most data points are below the diagonal line, indicating a tendency toward decreased recognition thresholds being induced by emotional learning.

Figure 2A shows statistical comparisons of the recognition thresholds (μ) both before and after emotional learning for the emotional learning group. It appears that both emotional learning and perceived separation reduced the threshold. A 2 (learning stage: before learning, after learning) × 2 (perceived laterality relationship: co-location, separation) two-way repeated measures analysis of variance (ANOVA) showed that the main effect of learning stage was significant [$F(1, 15) = 38.53$, $p < .001$, $\eta_p^2 = .72$] and the main effect of perceived laterality relationship was significant [$F(1, 15) = 264.40$, $p < .001$, $\eta_p^2 = .95$], but the interaction between two factors was not significant ($p > .05$).

Further statistical tests were conducted to examine whether the emotion-induced release of target speech from masking and the separation-induced release of target speech from masking were additive in a linear way. Figure 2B illustrates the averaged absolute values of masking release when only the emotional cue, only the spatial cue, and both cues were available. More specifically, the simple unmasking effect of emotion learning alone ("Δemotion") was the difference between the recognition threshold after emotional learning and that before emotional learning when the target and maskers were perceptually co-located (i.e., without the separation cue). The effect of separation alone ("Δseparation") was the difference between the threshold in separation minus that in co-location before emotional learning (i.e., without the emotional cue). The combined effect of both Δemotion and Δseparation was the difference between the threshold after emotional learning under the separation condition and the threshold before emotional learning under the co-location condition. Finally, the linear sum of the two simple effects was also calculated (Δemotion + Δseparation).

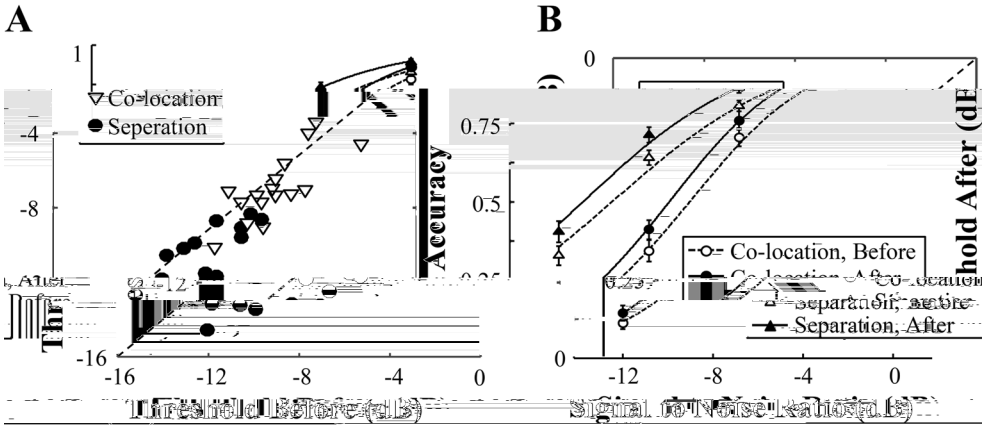

Fig. 1 (A) Group-mean recognition accuracy and the best-fitting psychometric function for each of the listening conditions in the emotional learning group. Recognition accuracy was higher after emotional learning (solid lines) than before that learning (dashed lines). Error bars represent standard errors of the means. (B) Scatter diagram of each participant's speech recognition thresholds. Dots below the diagonal line reflect recognition thresholds decreased by emotional learning for a single participant

A repeated measures one-way ANOVA showed that masking release amounts differed significantly across unmasking cues [$F(3, 45) = 58.37$, $p < .001$, $\eta_p^2 = .80$]. Post-hoc comparisons (Bonferroni-corrected) showed that the masking release of combined cues (Δemotion at separation) was significantly larger than the release caused by either cue alone (Δemotion at separation–Δemotion alone, $p < .001$; Δemotion at separation–Δseparation alone, $p < .01$). Also, the linear sum of the simple releasing effect (Δemotion + Δseparation) was larger than the release of either cue alone (as compared to Δemotion alone, $p < .001$; as compared to Δseparation alone, $p = .085$).

Moreover, the amount of release by the emotional cue alone (Δemotion ≈ 1 dB) was significantly smaller than that by the separation cue alone (Δseparation ≈ 3 dB) ($p < .001$). Finally, the combined effect of the two unmasking cues was not significantly different from the linear sum of Δemotion and Δseparation ($p = .890$), suggesting that the unmasking effects of the emotional cue and the spatial cue were linearly additive.

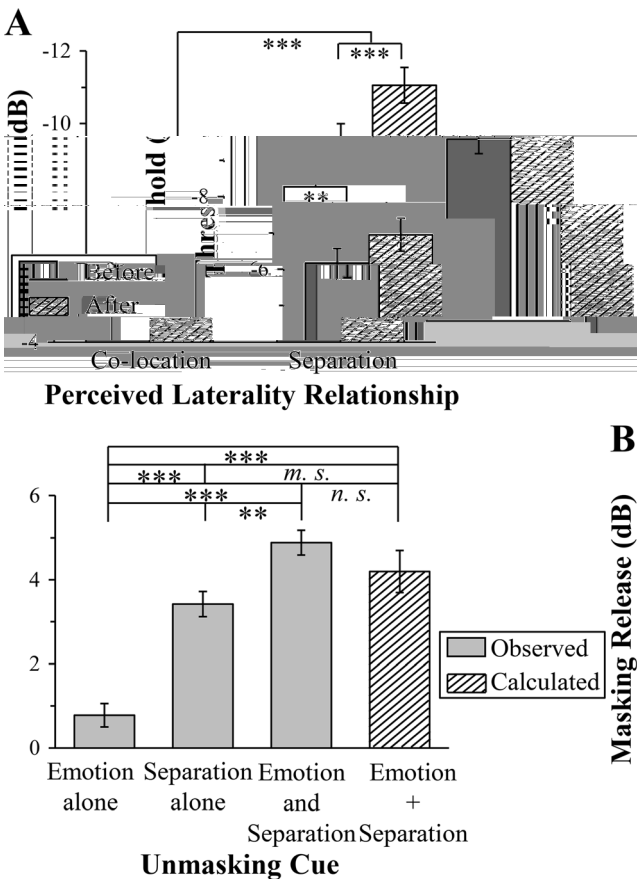

Fig. 2 (A) Speech recognition thresholds before and after emotional learning in the perceived separation or co-location condition. Both emotional and spatial cues enhanced target speech recognition. (B) Comparisons of the release of target speech from masking among different unmasking cue types. The effects of emotional and spatial cues were linearly additive. Error bars represent *SE*s. $^{**}p < .01$, $^{***}p < .001$, m.s. marginally significant, n.s. not significant

Comparisons between the emotional learning group and the control learning group

Figure 3 displays the speech-recognition performance for participants in the control learning group. The group-mean best-fitting psychometric functions are plotted in Fig. 3A, along with the values of group-mean recognition accuracy. The recognition accuracies and psychometric functions after neutral learning (solid lines) were slightly higher than those before neutral learning (dashed lines). Also, perceived separation between the target and the masker markedly improved recognition accuracy. Figure 3B plots a scatter diagram presenting the speech-recognition thresholds of each of the individual participants both before and after neutral learning in the control learning group.

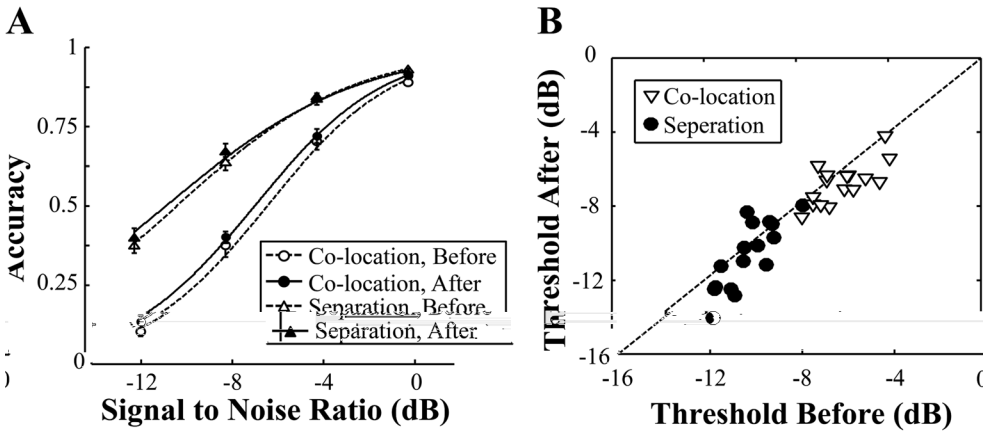

Fig. 3 (A) Group-mean recognition accuracies and the best-fitting psychometric functions for the control learning group. Recognition accuracy was slightly higher after neutral learning (solid lines) than before that learning (dashed lines). Error bars represent standard errors of the means. (B) Scatter diagram of each participant's speech-recognition threshold in the control learning group

Figure 4 shows statistical comparisons of the participants' recognition thresholds (μ) before and after neutral learning in the control learning group. Both neutral learning and perceived separation reduced the threshold. A 2 (learning stage: before control learning, after control learning) × 2 (perceived laterality relationship: co-location, separation) two-way repeated measures ANOVA showed that the main effect of learning stage was significant [$F(1, 30) = 12.47$, $p = .003$, $\eta_p^2 = .45$] and the main effect of perceived laterality relationship was significant [$F(1, 30) = 263.85$, $p < .001$, $\eta_p^2 = .95$], but the interaction between two factors was not significant ($p > .05$).

To compare the data between the emotional and control learning groups, a three-way ANOVA (Group × Learning Stage × Perceived Laterality relationship) was carried out. The results showed that the main effect of learning stage was significant [$F(1, 30) = 48.80$, $p < .001$, $\eta_p^2 = .62$], and the main effect of perceived laterality relationship was also significant [$F(1, 30) = 527.83$, $p < .001$, $\eta_p^2 = .95$]. Most importantly, the interaction between group and learning stage was significant [$F(1, 30) = 5.27$, $p = .029$, $\eta_p^2 = .15$], suggesting that the emotional-learning-induced threshold shift was significantly larger than the neutral-learning-induced one (Fig. 5A).

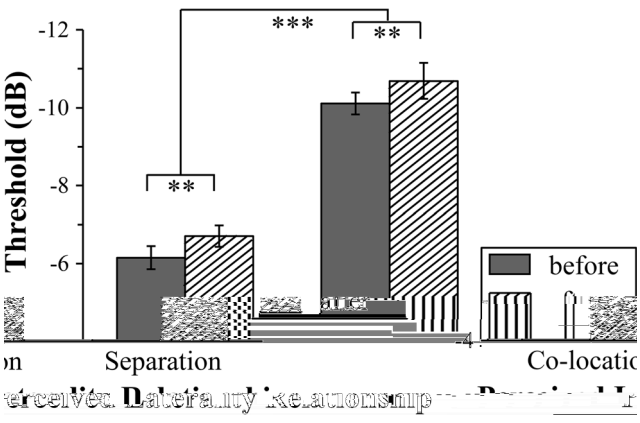

Fig. 4 Speech-recognition thresholds before and after neutral learning in the perceived separation or co-location condition. Both neutral learning and perceived separation enhanced participants' speech recognition in the control learning group. Error bars represent standard errors of the means. $^{**}p < .01$, $^{***}p < .001$

To further identify the group differences, trial-by-trial data were analyzed and the dynamics of instant total accuracy were calculated as the number of trials increased from 1 to 240. Figure 5B displays the dynamic change of accuracy between Trial 120 and Trial 240, which was normalized individually by each participant's accuracy before and after learning/control learning (instant total accuracy divided by accumulative accuracy at Trial 120). The two solid lines are the group-average changes in accuracy for the emotional learning group and the control group, respectively. Though the increasing tendencies were present in both groups, the averaged performance of the emotional learning group was apparently better than that of the control learning group, confirming the group difference in recognition thresholds.

The effect of keyword position

Since each of the target sentences had three keywords (the subject, predicate, and object), we investigate the question of whether emotional learning would affect recognition of these three keywords differently. First, we conducted a 2 (group: emotional, control) × 2 (perceived lateralized relationship: co-location, separation) × 4 (SNR: 0, – 4, – 8, – 12 dB) × 3 (keyword position: first, second, third) four-variable ANOVA for recognition accuracy before the learning/control learning manipulations. The results showed neither a significant main effect of group ($p > .05$) nor an interactions of group with the other factors (all ps > .05), confirming that there was no significant difference between the two groups before the learning/control learning manipulations.

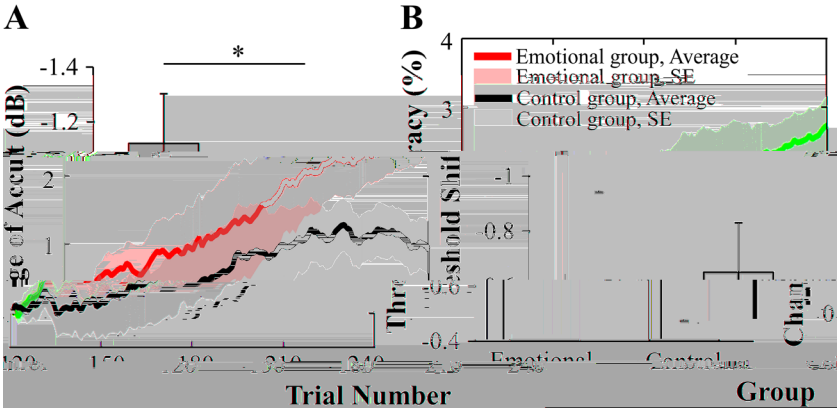

Fig. 5 Comparisons of the improvements of target-speech recognition between the emotional and control learning groups. (A) Differences in threshold shifts before and after the learning/control learning manipulations, between the emotional and control learning groups. Error bars represent standard errors of the means. $^{*}p < .05$. (B) Group differences, identified by trial-by-trial changes of cumulative accuracy

Figure 6 plots the recognition accuracy of the three keywords after the learning/control learning manipulations for the two groups. Separate three-way ANOVAs (Group × Perceived Lateralized Relationship × Keyword Position) were conducted for each of the four SNRs. The results showed that emotional learning facilitated speech recognition by improving the recognition of the first keyword, but not the second and the third ones. Specifically, when the SNR was – 4 dB, the recognition accuracy of the first keyword was significantly higher for the emotional learning group than for the control learning group [$F(1, 30) = 4.31$, $p = .047$, $\eta_p^2 = .13$]. A significant difference also occurred when the SNR was – 8 dB under the separated condition, causing a significant interaction between group and perceived spatial location [$F(1, 30) = 5.80$, $p = .022$, $\eta_p^2 = .16$] under the – 8 dB SNR. The emotional learning effect on recognizing the first keyword was present only when the task difficulty was at a moderate level, as it disappeared when the task was too easy, leading to a ceiling effect, or when the task was too demanding (accuracy below .5). The recognition accuracies for the second and third keywords were lower than the accuracy for the first keyword and were not affected by emotional learning in any of the conditions.

Skin conductance responses

The mean amplitude of SCR was calculated in a time window of 4 s following the onset of each trial during the speech recognition task. To determine the reliability of the baseline response before learning, a three-way ANOVA (Group × Perceived Lateralized Relationship × SNR) for SCR before the learning/control learning manipulations was first conducted. The results showed no significant main effects or interactions (all $ps > .05$), confirming the reliable baseline responses before the learning/control learning manipulations.

A four-way ANOVA (Group × Learning Stage × Perceived Lateralized Relationship × SNR) showed that the main effect of learning stage was significant [$F(1, 30) = 5.00$, $p = .033$, η_p^2

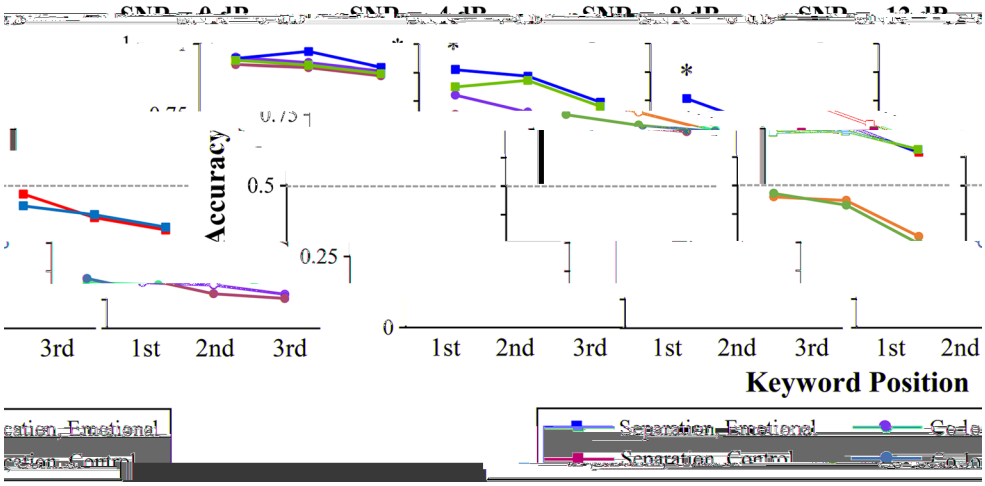

Fig. 6 Recognition accuracies for the three keywords in the emotional and control learning groups. As compared to the neutral learning in the control learning group, emotional learning improved recognition of the first keyword, but not the second and the third ones. $^{*}p < .05$

= .14], indicating that SCR was increased after learning, as compared to the baseline before learning/control learning, for both groups. The SCR increase for each participant was calculated by subtracting the SCR amplitude before learning/control learning from that after learning/control learning.

A three-way ANOVA (Group × Perceived Lateralized Relationship × SNR) for the SCR increase showed no overall interaction ($p > .05$), but only a marginally significant interaction between group and perceived lateralized relationship [$F(1, 30) = 3.17$, $p = .085$, $\eta_p^2 = .10$]. Since no significant difference in SCR increases was found between SNRs ($p > .05$), the SCR increases across the four SNRs were averaged.

As is shown in Fig. 7, for the emotional learning group the SCR increase was larger in the co-located condition than in the separated condition [$F(1, 15) = 4.77$, $p = .045$, $\eta_p^2 = .24$], indicating participants' higher physiological arousal after receiving emotional learning while recognizing target speech co-located with as compared to separated from the masker. For the control learning group, the effect of separation was not significant ($p = .628$). No significant correlation was observed between participants' SCR results and their behavioral performance at speech recognition.

SAM ratings for the emotionally negative sound and the emotional neutral sound

Independent-samples t tests showed that the 9-point self-reported emotional ratings of the negative sound and the neutral sound were significantly or marginally different between the two groups [emotional valence, $t(18.44) = 3.07$, $p < .001$; emotional arousal, $t(30) = 1.88$, $p = .07$]. When listening to the negative sound at the end of the experiment, participants who had received emotional learning reported a less negative emotional valence (2.16 ± 0.26) than did those who had received neutral learning (1.31 ± 0.09). The participants in the emotional learning group also exhibited lower emotional arousal (7.19 ± 0.36) than did those in the neutral group (8.06 ± 0.30). The differences in SAM ratings might have resulted from the longer exposure to the aversive sound in the emotional learning group during the learning session.

No significant difference was observed between the groups for the emotional valence (5.08 ± 0.25) and emotional arousal (5.21 ± 0.27) of the neutral sound.

Discussion

Learned emotion facilitates target-speech recognition against speech informational masking

In this study, recognition of the target speech in the emotional learning group was significantly improved by conditioning the target voice when the acoustical features of the speech stimuli were not changed. These results indicate that emotionally conditioning the target speaker's voice is effective for improving target-speech recognition against speech informational masking, independent of the perceived spatial separation between the target speech and masking speech.

For the control learning group, by comparing the speech

role in mediating emotional conditioning-induced unmasking of speech.

In this study, speech recognition was also tested under two conditions with different perceived spatial relationships between the target and masker: the separation and the co-location conditions. Under the co-location condition, the target and the maskers competed for listeners' spatial attention, whereas under the separation condition, listeners' spatial attention could be selectively pointed to the target speech, leading to unmasking of the target speech. The results of this study showed that the speech recognition threshold was lower when the target and the masker were perceptually separated as compared to when they were perceptually co-located. The results are consistent with previous studies in this line of investigation (Arbogast et al., 2002; Freyman et al., 1999; Huang et al., 2009; H. Li et al., 2013; L. Li et al., 2004; Schneider et al., 2007; X. Wu et al., 2005).

Interestingly, in this study no interaction was observed between the effect of perceived spatial location and the effect of emotional learning, indicating that the emotion-induced unmasking effect is independent to the separation-induced unmasking effect. By comparing the unmasking effects induced by emotional learning (about 1 dB), perceived spatial separation (about 3 dB), and the combination of the both (about 4 dB), it was found that the unmasking effect of emotional learning and that of perceived spatial separation on speech recognition were linearly additive. The present study for the first time reveals that when an emotional cue and a spatial cue are simultaneously utilized to improve the recognition of target speech against a complex auditory background, their unmasking effects are reciprocally independent and appear to be linearly additive. More studies will still be needed to extend this line of research not only to other unmasking cues, such as the visual-speech priming cue (e.g., lip-reading priming, C. Wu et al., 2013; C. Wu et al., 2017) and the voice-priming cue (Huang et al., 2010; Yang et al., 2007), but also to the related underlying brain substrates, including the amygdala, which is involved in emotional learning-induced unmasking of target signals (Du, Wu, & Li, 2011), and the superior parietal lobule (the posterior parietal cortex in rats), which is involved perceptually spatial-separation-induced unmasking of target signals (Du, Wu, & Li, 2011; Zheng et al., 2016).

In this study, emotional learning significantly improved the recognition of the first keyword in a target sentence, but not the second and the last ones. The results are in agreement with previous reports that signals with emotional significance are given the priority access to selective attention and can be quickly detected in competitions among sensory inputs (Anderson, 2005; Eastwood et al., 2001; Fox, 2002; LeDoux, 1998; Öhman et al., 2001). On the basis of the results of this study, we propose that when the target voice is conditioned with aversive sounds, the cue of learned-emotion embedded in the voice enhances the attentional capture of the target voice stream against the competing speech background, thereby improving the target-speech recognition.

Interestingly, at low SNRs when target-recognition accuracy fell below .5, the emotional-learning-induced enhancement in recognizing the first keyword disappeared. An explanation for this finding may be related to the effect of perceptual load; that is, the processing priority for a stimulus with emotional significance draws attentional resources only under low or middle, but not under high, perceptual load (Pessoa, McKenna, Gutierrez, & Ungerleider, 2002; Pessoa, Padmala, & Morland, 2005; Yates, Ashwin, & Fox, 2010). Yates et al. reported that a fear-conditioned angry face captured attention only when participants were asked to perform an easy discrimination task, but did not interfere with a difficult central task that occupied most of the participants' attentional resources. In the present study, when both the SNRs changed from 0 to –12 dB and the perceived spatial position between the target and the maskers was shifted from separation to co-location, the difficulty of the speech recognition task increased, which might have caused high perceptual load to the central executive system and inhibited the emotionally guided attention.

Skin conductance responses after emotional learning

Physiological measures revealed that the SCR was more sensitive to the perceived positions of target and masker only after participants received emotional learning. Using the SCR be-

measurement to perceived spatial relationship between the target speech and masking speech should be further studied in the future.

The results of this study also showed that the electrodermal responses did not vary across different SNR conditions, consistent with the reports by Seeman and Sims (2015) and Mackersie, MacPhee, and Heldt (2015), in which SCR was sensitive to both the task difficulty and the cognitive efforts, but not sensitive to SNRs.

Since no significant correlation was found between the

negative emotion. *Perception & Psychophysics*, *63*, 1004–1013. doi:https://doi.org/10.3758/BF03194519

Fox, E. (2002). Processing emotional facial expressions: The role of anxiety and awareness. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 52–63. doi:https://doi.org/10.3758/CABN.2.1.52

Francis, A. L., MacPherson, M. K., Chandrasekaran, B., & Alvar, A. M. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology*, *7*, 263. doi:https://doi.org/10.3389/fpsyg.2016.00263

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, *115*, 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America*, *106*, 3578–3588.

Frühholz, S., Trost, W., Kotz, S. A. (2016). The sound of emotions: Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews*, 68, 96–110.

Gao, Y., Cao, S., Qu, T., Wu, X., Li, H., Zhang, J., Li, L. (2014). Voice-associated static face image releases speech from informational masking. *PsyCh Journal*, *3*, 113–120. doi:https://doi.org/10.1002/pchj.45

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, *8*, 145–146. doi:https://doi.org/10.1038/nn1392

Grosso, A., Cambiaghi, M., Concina, G., Sacco, T., & Sacchetti, B. (2015). Auditory cortex involvement in emotional learning and memory. *Neuroscience*, *299*, 45–55.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, *40*, 432–443.

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America*, *117*, 842–849.

Huang, Y., Huang, Q., Chen, X., Wu, X., & Li, L. (2009). Transient auditory storage of acoustic details is associated with release of speech from informational masking in reverberant conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1618–1628. doi:https://doi.org/10.1037/a0015791

Huang, Y., Xu, L., Wu, X., & Li, L. (2010). The effect of voice cuing on releasing speech from informational masking disappears in older adults. *Ear and Hearing*, *31*, 579–583.

Iwashiro, N., Yahata, N., Kawamuro, Y., Kasai, K., & Yamasue, H. (2013). Aberrant interference of auditory negative words on attention in patients with schizophrenia. *PLOS ONE*, *8*, e83201.

Kluge, C., Bauer, M., Leff, A. P., Heinze, H., Dolan, R. J., & Driver, J. (2011). Plasticity of human auditory-evoked fields induced by shock conditioning and contingency reversal. *Proceedings of the National Academy of Sciences*, *108*, 12545–12550.

Koster, E. H. W., Crombez, G., Van Damme, S., Verschuere, B., & De Houwer, J. (2005). Signals for threat modulate attentional capture and holding: Fear-conditioning and extinction during the exogenous cueing task. *Cognition and Emotion*, *19*, 771–780. doi:https://doi.org/10.1080/02699930441000418

LeDoux, J. (1998). The emotional brain: The mysterious underpinnings of emotional life. New York, NY: Simon & Schuster.

Lei, M., Luo, L., Qu, T., Jia, H., & Li, L. (2014). Perceived location specificity in perceptual separation-induced but not fear conditioning-induced enhancement of prepulse inhibition in rats. *Behavioural Brain Research*, *269*, 87–94.

Li, H., Kong, L., Wu, X., & Li, L. (2013). Primitive auditory memory is correlated with spatial unmasking that is based on direct-reflection integration. *PLoS ONE*, *8*, e63106. doi:https://doi.org/10.1371/journal.pone.0063106

Li, L., Daneman, M., Qi, J. G., & Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 1077–1091.

Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, *37*, 118S.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*, 113–122.

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart-rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing*, *36*, 145–154.

Morris, J. S., Buchel, C., & Dolan, R. J. (2001). Parallel neural responses in amygdala subregions and sensory cortex during implicit fear conditioning. *NeuroImage*, *13*, 1044–1052.

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, *393*, 467–470.

Mothes-Lasch, M., Becker, M. P., Miltner, W. H., & Straube, T. (2016). Neural basis of processing threatening voices in a crowded auditory world. *Social Cognitive and Affective Neuroscience*, *11*, 821–828.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*, 85–103.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*, 466–478. doi:https://doi.org/10.1037/0096-3445.130.3.466

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences*, *99*, 11458–11463. doi:https://doi.org/10.1073/pnas.172403899

Pessoa, L., Padmala, S., & Morland, T. (2005). Fate of unattended fearful faces in the amygdala is determined by both attentional resources and cognitive modulation. *NeuroImage*, *28*, 249–255. doi:https://doi.org/10.1016/j.neuroimage.2005.05.048

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., … Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37* (Suppl), 5S-27S. doi:https://doi.org/10.1097/AUD.0000000000000312

Pischek-Simpson, L. K., Boschen, M. J., Neumann, D. L., & Waters, A. M. (2009). The development of an attentional bias for angry faces following Pavlovian fear conditioning. *Behaviour Research and Therapy*, *47*, 322–330.

Rakerd, B., Aaronson, N. L., & Hartmann, W. M. (2006). Release from speech-on-speech masking by adding a delayed masker at a different location. *Journal of the Acoustical Society of America*, *119*, 1597–1605.

Rescorla, R. A. (1973). Effect of US habituation following conditioning. *Journal of Comparative and Physiological Psychology*, *82*, 137–143.

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

Schneider, B. A., Li, L., & Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *Journal of the American Academy of Audiology*, *18*, 559–572.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research*, *58*, 1781–1792.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). A precedence effect in sound localization. *Journal of the Acoustical Society of America*, *21*, 468.

Weisz, N., Kostadinov, B., Dohrmann, K., Hartmann, T., & Schlee, W. (2007). Tracking short-term auditory cortical plasticity during classical conditioning using frequency-tagged stimuli. *Cerebral Cortex*, *17*, 1867–1876.

Wu, C., Cao, S., Wu, X., & Li, L. (2013). Temporally pre-presented lipreading cues release speech from informational masking. *Journal of the Acoustical Society of America*, *133*, L281–L285.

Wu, C., Zheng, Y., Li, J., Zhang, B., Li, R., Wu, H., . . . Li, L. (2017). Activation and functional connectivity of the left inferior temporal gyrus during visual speech priming in healthy listeners and listeners with schizophrenia. *Frontiers in Neuroscience*, *11*, 107. doi:https://doi.org/10.3389/fnins.2017.00107

Wu, X., Chen, J., Yang, Z., Huang, Q., Wang, M., & Li, L. (2007). Effect of number of masking talkers on speech-on-speech masking in Chinese. In Proceedings of Interspeech (pp. 390–393). Antwerp, Belgium.

Wu, X., Wang, C., Chen, J., Qu, H., Li, W., Wu, Y., . . . Li, L. (2005). The effect of perceived spatial separation on informational masking of Chinese speech. *Hearing Research*, *199*, 1–10. doi:https://doi.org/10.1016/j.heares.2004.03.010

Wu, Z., Ding, Y., Jia, H., & Li, L. (2016). Different effects of isolation-rearing and neonatal MK-801 treatment on attentional modulations of prepulse inhibition of startle in rats. *Psychopharmacology*, *233*, 3089–3102.

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., & Li, L. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Communication*, *49*, 892–904. doi:https://doi.org/10.1016/j.specom.2007.05.005

Yates, A., Ashwin, C., & Fox, E. (2010). Does emotion processing require attention? The effects of fear conditioning and perceptual load. *Emotion*, *10*, 822–830.

Zhang, C., Lu, L., Wu, X., & Li, L. (2014). Attentional modulation of the early cortical representation of speech signals in informational or energetic masking. *Brain and Language*, *135*, 85–95.

Zheng, Y., Wu, C., Li, J., Wu, H., She, S., Liu, S., … Li, L. (2016). Brain substrates of perceived spatial separation between speech sources under simulated reverberant listening conditions in schizophrenia. *Psychological Medicine*, *46*, 477–491. doi:https://doi.org/10.1017/S0033291715001828

Zou, D., Huang, J., Wu, X., & Li, L. (2007). Metabotropic glutamate subtype 5 receptors modulate fear-conditioning induced enhancement of prepulse inhibition in rats. *Neuropharmacology*, *52*, 476–486. doi:https://doi.org/10.1016/j.neuropharm.2006.08.016