Hearing Research 331 (2016) 119–130

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

Attentional modulation of informational masking on early cortical representations of speech signals

Changxin Zhang [a], Stephen R. Arnott [c], Cristina Rabaglia [b], Meital Avivi-Reich [b], James Qi [b], Xihong Wu [a], Liang Li [a, **], Bruce A. Schneider [b, *]

[a] Department of Psychology, Speech and Hearing Research Center, McGovern Institute for Brain Research at PKU, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing, China
[b] Department of Psychology, Human Communication Laboratory, University of Toronto Mississauga, Mississauga, Ontario, Canada
[c] Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 5 December 2013
Received in revised form
27 October 2015
Accepted 4 November 2015
Available online 10 November 2015

Keywords:
Attention
Informational masking
Energetic masking
Event-related potentials

ABSTRACT

To recognize speech in a noisy auditory scene, listeners need to perceptually segregate the target talker's voice from other competing sounds (stream segregation). A number of studies have suggested that the attentional demands placed on listeners increase as the acoustic properties and informational content of the competing sounds become more similar to that of the target voice. Hence we would expect attentional demands to be considerably greater when speech is masked by speech than when it is masked by steady-state noise. To investigate the role of attentional mechanisms in the unmasking of speech sounds, event-related potentials (ERPs) were recorded to a syllable masked by noise or competing speech under both active (the participant was asked to respond when the syllable was presented) or passive (no response was required) listening conditions. The results showed that the long-latency auditory response to a syllable (/*bi*/), presented at different signal-to-masker ratios (SMRs), was similar in both passive and active listening conditions, when the masker was a steady-state noise. In contrast, a switch from the passive listening condition to the active one, when the masker was two-talker speech, significantly enhanced the ERPs to the syllable. These results support the hypothesis that the need to engage attentional mechanisms in aid of scene analysis increases as the similarity (both acoustic and informational) between the target speech and the competing background sounds increases.

1. Introduction

Under noisy listening conditions (e.g., a cocktail-party environment; Cherry, 1953), listeners usually find it difficult to comprehend target speech and participate in conversations due to auditory masking (Miller, 1947). The mechanisms underlying auditory masking are complicated and particularly influenced by the type of masker present. Maskers can interfere with speech recognition when the peripheral neural activity elicited by a signal is overwhelmed by that elicited by a masker, leading to a degraded or noisy neural representation of the signal, making it difficult for subsequent cognitive processes to extract the signal (e.g., Freyman et al., 1999, 2001; Arbogast et al., 2002; Brungart, 2001; Brungart and Simpson, 2002; Kidd et al., 1994, 1998; Schneider et al., 2007; Li et al., 2004; Wu et al., 2005; Ezzatian et al., 2011). This type of masking effect is referred to as energetic masking.

In addition, competing sound sources can cause informational masking that interferes with the processing of the signal at levels beyond the cochlea. For example, when the masker is speech, the informational content of the masker can interfere with the processing of the target speech at both perceptual (e.g., phonemic identification) and cognitive (e.g., semantic processing) levels, making it difficult for listeners to successfully segregate the different sound sources and selectively attend to the target speech (Arbogast et al., 2002; Brungart, 2001; Brungart and Simpson, 2002; Durlach et al., 2003; Freyman et al., 1999, 2001; Kidd et al., 1994, 1998; Schneider et al., 2007; Li et al., 2004; Wu et al., 2005; Ezzatian et al., 2011).

Although a steady-state noise masker may also compete with

* Corresponding author. Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Rd. N, Mississauga, Ontario, Canada L5L 1C6.
** Corresponding author. Department of Psychology, Peking University, Beijing 100871, China.
E-mail addresses: liangli@pku.edu.cn (L. Li), bruce.schneider@utoronto.ca (B.A. Schneider).

http://dx.doi.org/10.1016/j.heares.2015.11.002
0378-5955/

the target-speech signal for the listener's attentional resources, it is likely to produce more energetic masking than informational masking since it lacks any phonetic or semantic information. However, a speech masker, in addition to producing energetic masking (due to the speech masker-elicited activities in the same or nearby regions on the basilar membrane that are processing the target speech) also will produce a considerable amount of informational masking (due to interference with the processing of the target speech at phonetic, semantic, and/or linguistic levels).[1]

Listeners can use various perceptual and/or cognitive cues to release target speech from masking, especially from irrelevant-speech-induced informational masking. These cues include perceptual familiarity with the talker's voice (Brungart, 2001; Newman and Evers, 2007; Yang et al., 2007; Huang et al., 2010), knowledge of the target talker's identity (Yonan and Sommers, 2000; Newman and Evers, 2007), knowledge of a source's location (Kidd et al., 2005; Singh et al., 2008), perceived spatial separation of target from masker (Freyman et al., 1999, 2001; Huang et al., 2008, 2009; Li et al., 2004, 2013; Wu et al., 2005), prior knowledge about part of the target-sentence content (i.e., temporally pre-presented content prime, Freyman et al., 2004; Yang et al., 2007; Wu et al., 2012), and viewing a speaker's movements of the speech articulators that are either simultaneously presented with target speech (Helfer and Freyman, 2005) or temporally pre-presented prior to target speech (Wu et al., 2013). These cues presumably are effective at unmasking the target speech because they provide information that facilitates the listener's ability to segregate and selectively attend to the target voice.

In psychoacoustic studies of speech recognition, listeners are typically asked to repeat the target sentence immediately after hearing it. Hence, it would be difficult, if not impossible, to obtain behavioral measures of speech recognition when the listener is not attending to the target speech. However, in event-related potential (ERP) recording studies of speech processing, attention can be limited and even drawn away from the acoustic stimulus to irrelevant stimuli in other modalities (Alho, 1992; Martin and Stapells, 2005; Billings et al., 2011).

The P1–N1–P2 complex, a group of components of the long-latency auditory evoked potentials can be elicited by speech stimuli (e.g., single syllables) even when a noise or a speech masker is co-presented (Martin et al., 1997, 1999, Martin and Stapells, 2005; Billings et al., 2011; Salo et al., 1995; Whiting et al., 1998; Polich et al., 1985; Muller-Gass et al., 2001). Under the latter conditions, however, the earlier aspects of this complex can become attenuated, making it difficult to identify the P1 component when the speech signal is masked (Alain et al., 2009, 2012, 2014). With respect to the N1 component, Billings et al. (2011) found that, relative to a steady-state noise masker, a four-talker speech masker with a signal-to-masker ratio (SMR) fixed at -3 dB caused a larger N1 masking effect for spoken syllables when listeners' attention was drawn away from the acoustic signals (the passive homogenous paradigm), but not when listeners paid attention to the acoustic signals (the active oddball paradigm). To further examine whether attention affects the P1–N1–P2 complex under masking conditions, Billings et al. (2011) collapsed the waveforms across the three masking conditions (continuous steady-state noise, interrupted noise, four-talker speech) and found that the N1 amplitude was significantly larger under the active paradigm than the passive paradigm, indicating a facilitating effect of attention on the ERP component. However, it is still not clear whether the attentional modulation is masker-type and/or SMR dependent.

[1] Because the target in this experiment is a single syllable, it is likely that the interference will be limited to phonetic interference.

To verify whether the unmasking effect of attentional modulation on event-related potentials (ERPs) to speech signals is masker-type dependent, this study examined the degree to which ERPs to a masked speech syllable are modulated by attention and whether the attentional modulation is different between noise- and speech-masking conditions. More specifically, ERPs to the speech syllable /*bi*/ were recorded under either a passive-listening condition (listeners attended to irrelevant video presentations) or an active-listening condition (listeners attended to the target syllable) when the masker was either noise or speech. For each of the listening condition and masker type combinations, four SMRs were used: −8, −4, 0, and 4 dB.

It has long been known that the masking effect of a speech masker depends on the number of masking voices (Carhart et al., 1975). For example, both Freyman et al. (2004) and Wu et al. (2007) have reported that the informational masking effect reaches the highest level when two-talker masking speech is used and then progressively reduces as the number of masking talkers increases. Thus, in this study, to maximize the informational masking effect under the speech-masking condition, two-talker speech was used as the speech masker.

2. Materials and methods

2.1. Participants

Twelve young adults (7 males and 5 females) with a mean age of 21 years (range = 18–24 years, SD = 2.06 years) participated in this study. They were all students recruited from the University of Toronto Mississauga who gave their written informed consent to participate in this study. All participants reported they were right handed, native-English speakers in good health. Their hearing was tested and found normal (audiometric thresholds < 20 dB HL between 250 and 8000 Hz), and balanced (interaural threshold differences in the frequency range tested did not exceed 10 dB). The participants were paid a modest stipend for their participation.

2.2. Materials and apparatus

The target signal was a naturally produced consonant-vowel syllable /*bi*/ (duration = 474 ms) obtained and modified from the standardized UCLA version of the Nonsense Syllable Test (Dubno and Schaefer, 1992), spoken by a female talker. Two types of maskers were used in the study: steady-state speech-spectrum noise and two-talker speech. The steady-state speech-spectrum noise masker was a 327-second continuous noise loop recorded from an Interacoustic AC5 audiometer (Interacoustics, Assens, Denmark). The two-talker speech masker was a set of linguistically correct but semantically meaningless sentences (e.g., “A house should dash to the bowl.” or “A frog will arrest the pit.”) spoken by two female talkers, whose waveforms were mixed with equal root-mean-square levels from the two sources (see Freyman et al., 2001; Li et al., 2004). An examination of the spectrum levels of the two types of maskers when they were presented at the same average sound pressure level (see Fig. 1) indicates that the steady-state speech-spectrum noise masker had a higher concentration of its energy in the low-frequency region than did the speech masker, with the opposite being true for the high-frequency region.

The target syllable was presented at 60 dBA. The masker level was adjusted to produce four SMRs: −8, −4, 0 and 4 dB. Calibration of these stimuli was completed by measuring the overall RMS level of 10 s of a concatenated version of each signal.

All stimuli were digitized at 20 kHz using a 16-bit Tucker Davis Technologies (TDT, Gainesville, FL) System II and custom software. The stimuli were converted to analog using the TDT system under

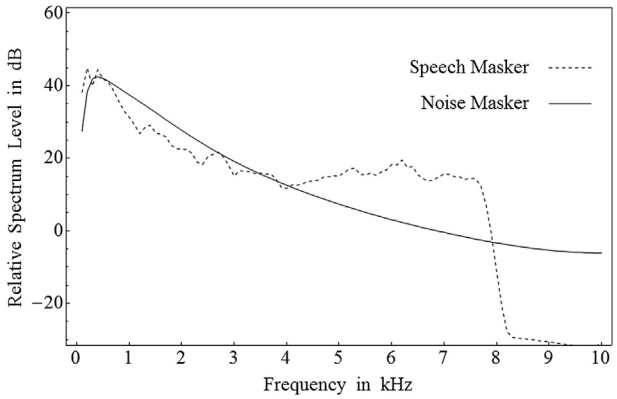

Fig. 1. Relative spectrum levels of the two-talker speech masker and the steady-state speech-spectrum noise masker, when both are equated to produce the same average sound pressure level.

the control of an Optiplex GX1 Dell computer. The stimuli were then low-pass filtered at 10 kHz, amplified by a Harmon Kardon amplifier (HK 3370), and presented to both ears via earphones. The stimuli were presented binaurally to more closely approximate everyday listening situations in which both signals and maskers stimulate both ears.

The experimental sessions were conducted in a dim sound-attenuating booth (Industrial Acoustic Company). The participants were seated 1 m from a 14-inch computer monitor placed in front of them.

2.3. Procedures

Sixteen blocks were created to encompass all possible combinations of the 2 masker types (steady-state speech-spectrum noise, two-talker speech), 2 attention conditions (passive condition, active condition), and 4 SMRs (−8, −4, 0, 4 dB). All participants were first tested in the passive conditions before experiencing the active listening conditions. This was done to avoid the possibility that previous exposure to the active conditions might predispose them to listen more "actively" when tested in the passive listening conditions. In each block, 300 trials were used during which the masker was presented continuously across trials. Half of the participants were presented with the noise masker first and then the speech masker and the other half were presented with the speech masker before the noise masker. Latin-Squares were used to balance the order of presentation of the SMRs across participants.

Under the passive conditions, participants were asked to watch a silent cartoon movie and ignore the sounds presented from the earphones during ERP recoding trials. Each trial started with a warning beep (a 500-Hz pure tone with the duration of 50 ms, 18 dB higher than the background masker). In 80% of the trials, the syllable /*bi*/ was presented 1000 ms after the warning beep; while in the other 20% of the trials, no syllable was presented. A responding beep (a 2000-Hz pure tone with the duration of 50 ms, 10 dB higher than the background masker) was presented 2000 ms after the warning beep. The next trial began randomly 2–4 s after the responding beep of the previous trial. It took about 25 min to finish one recording block under the passive condition.

Under the active condition, the stimuli and procedures were identical to those under the passive condition except that participants were asked to press one of two buttons after the responding beep to indicate whether they had heard the syllable /*bi*/ or not. Responses prior to the beep signaling the beginning of the response interval were not scored. To minimize eye movements, participants were also asked to fixate on a cross in the centre of the monitor. It took somewhat longer (30–40 min) to finish one recording block under the active condition because the next trial did not begin until a response was made.

2.4. Electrophysiology recordings

Electroencephalogram signals were recorded with a 128-channel HydroCel Geodesic Sensor Net (Electrical Geodesics, Inc, Eugene, Oregon) at a sampling rate of 1000 Hz. Electrode impedance was kept below 100 kΩ. Data were referenced online to Cz and then re-referenced offline to the common average. The waveforms were on-line amplified 500 times and band-pass filtered between 0 and 100 Hz. They were subsequently filtered offline by a 1 Hz high-pass filter and a 30 Hz low-pass filter (also see Billings et al., 2011). Ocular artifacts were removed with an eye blink threshold of 14 μV/ms. The number of accepted sweeps in the average response for participants 1–8 exceeded 200 after artifact rejection in each of the 16 conditions (2 maskers × 4 SMRs × 2 listening conditions). The number of accepted sweeps for participants 9–12 was approximately 30% higher than those for the first eight participants due to more frequent hydration of the electrode cap, leading to less noise in their recordings. For one of the twelve participants, the ERP record in the active noise condition at an SMR of 0 dB was corrupted, and could not be recovered. In this condition for this subject, the latencies and amplitudes used in the statistical analyses were interpolated between the values recorded for SMRs of −4 and +4 dB for that condition.

Recordings were divided into target-syllable epochs of 1200 ms, which included a 200 ms pre-stimulus baseline. Individual amplitudes and latencies for the N1 and P2 components were determined in the following manner. First we examined the grand mean traces over the 16 conditions (see Fig. 7). This examination suggested that the N1 peaks for each of the 12 individuals were likely to fall in a window ranging from 95 to 175 ms. With respect to P2, an examination of the active condition for the speech masking suggested that a wider range was involved. Hence, we defined a window ranging from 180 to 350 ms. Within each of the windows the location of the relevant peak was determined using NetStation software for each individual in each of the 16 conditions. Finally, the average waveform of each subject at each of the 16 conditions was visually examined to ensure that the location of the peak identified by the NetStation software as being within the relevant window corresponded to the location determined through visual inspection. In eleven cases, visual inspection indicated that the relevant peak was located just outside the respective window. In an additional two cases visual inspection indicated that the location of a P2 peak, of lesser magnitude than the one identified by the software, may have occurred earlier in the observation window. Because statistical analyses conducted with and without the substitution of the visually-identified peak data did not alter the significance of any of the main effects or interactions, the data presented here are those identified by the software.

To verify that our algorithm was correctly locating the N1 and P2 peaks, we overlayed the peaks identified by the algorithm on the plots of the ERP waveforms. Fig. 2 shows such an overlay when the SMR was equal to 4 dB. An examination of this plot shows that, when the masker was noise, the peaks identified by the algorithm were located where a visual inspection of the ERP waveforms suggested that they should be located. However, when the masker was speech and listening was passive, there were several instances where it was difficult to visually identify the presence of N1 and P2 peaks, because such peaks, if present, were too close to the noise floor. Hence, as noted above, we decided to use the peaks identified by the software to avoid biasing on the part of the person visually inspecting the data.

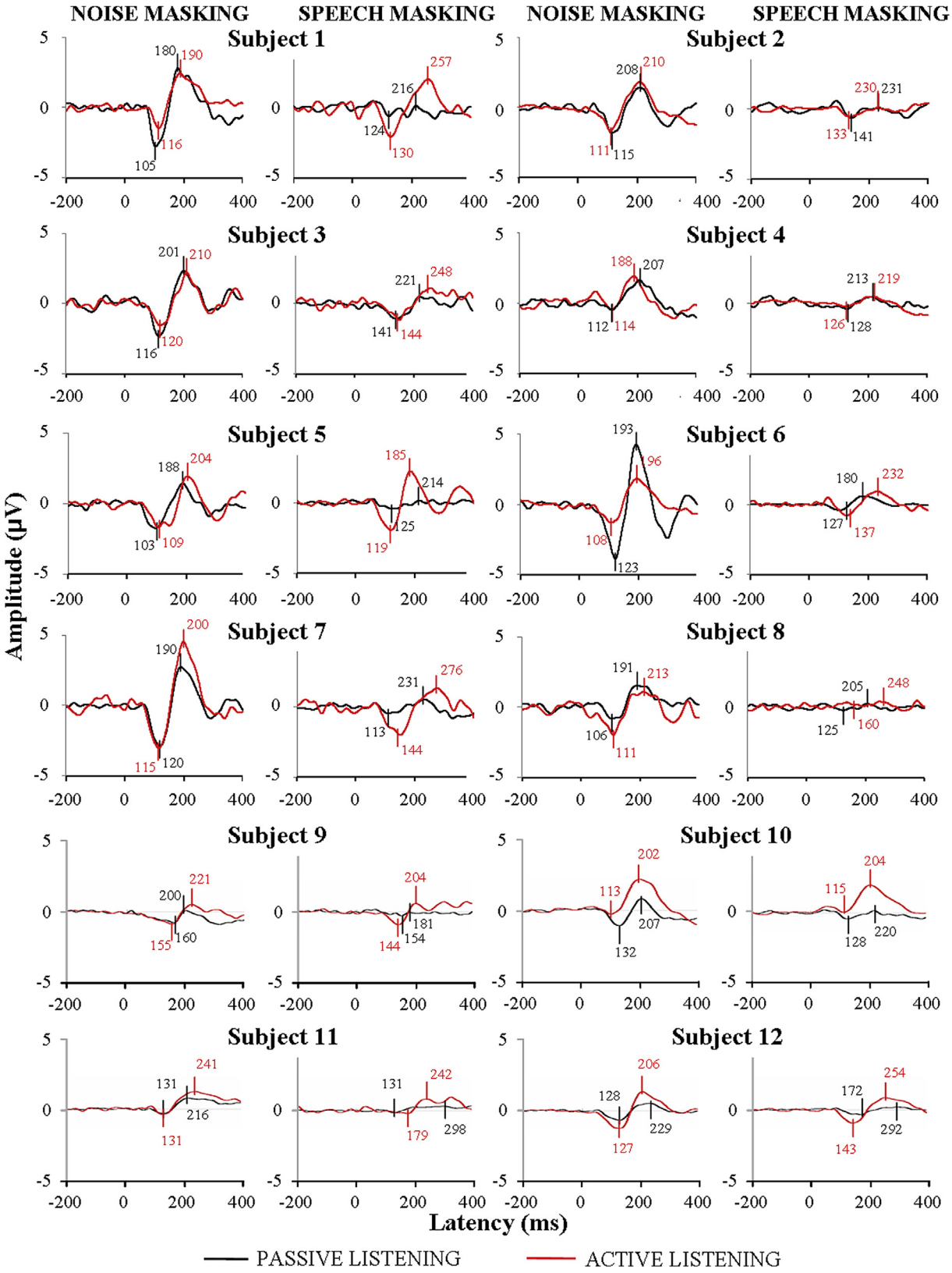

Fig. 2. Mean ERP waveforms evoked by the target syllable /*bi*/ recorded from electrode site Cz for each participant under both noise-masking and speech masking conditions when the listening condition was either passive (black trace) or active (red trace). The locations of the N1 and P2 peaks identified by the algorithm, along with their latency values, are indicated by vertical lines.

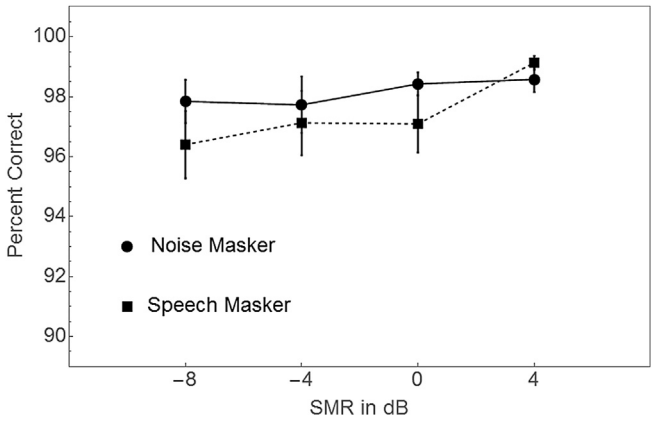

Fig. 3. Percentage of correct detection responses (averaged across participants) in the active listening condition as a function of SMR when the masker was noise (filled circles) or speech (filled squares). Standard error bars are shown.

Note that any uncertainty with respect to the location of the peaks will not affect any observed differences in their amplitudes between the passive versus active listening conditions because a peak buried in the noise floor will have a smaller amplitude than one that is clearly distinguishably above the noise floor, as are the peaks in the speech-masked active-listening condition for all but one of the subjects (subject 8) in Fig. 2. Hence, for all participants, the peak amplitudes of N1 and P2 evoked by the target syllable /*bi*/ under each of the 16 conditions were analyzed and statistically compared.

A glass window in the sound attenuating booth allowed the experimenter to monitor the participant's state of alertness at all times. In addition, the experimenter entered the sound booth between blocks to ensure that the electrodes were sufficiently hydrated.

3. Results

Fig. 3 plots the percentage of correct responses in the active listening condition as a function of SMR when the speech syllable was masked by noise (filled circles) or by speech (filled squares). This figure indicates that the syllable was well above detection threshold at all SMRs, and that the average detection accuracy at SMR4 was essentially at asymptote, and equivalent for the two types of maskers (99% in both cases).

Fig. 4 (top two panels) shows the ERP waveforms, averaged over participants, at selected electrode sites when the syllable /bi/ was masked by noise. The lower two panels of Fig. 4 present comparable plots when the masker was two-talker speech. As has been reported in other studies (Alain et al., 2009, 2012, 2014), the P1 component of the P1–N1–P2 complex, in some instances, is severely attenuated, especially when the masker is two-talker speech. Hence, our focus in this paper is restricted to the N1 and P2 components of the response. In general, when the masker was noise, ERP waveforms appear to be approximately the same under passive and active listening conditions and to increase systematically with SMR. However, when the masker was two-talker speech, the N1 and P2 components of the waveform appear to be more prominent and well-defined under active as opposed to passive listening conditions. Because of the prominence of the N1 and P2 components at Cz, and in the absence of a significant laterality effect when we compared left sites (C1/C3/C5) to the right sites (C2/C4/C6), subsequent analyses and statistical tests are based on

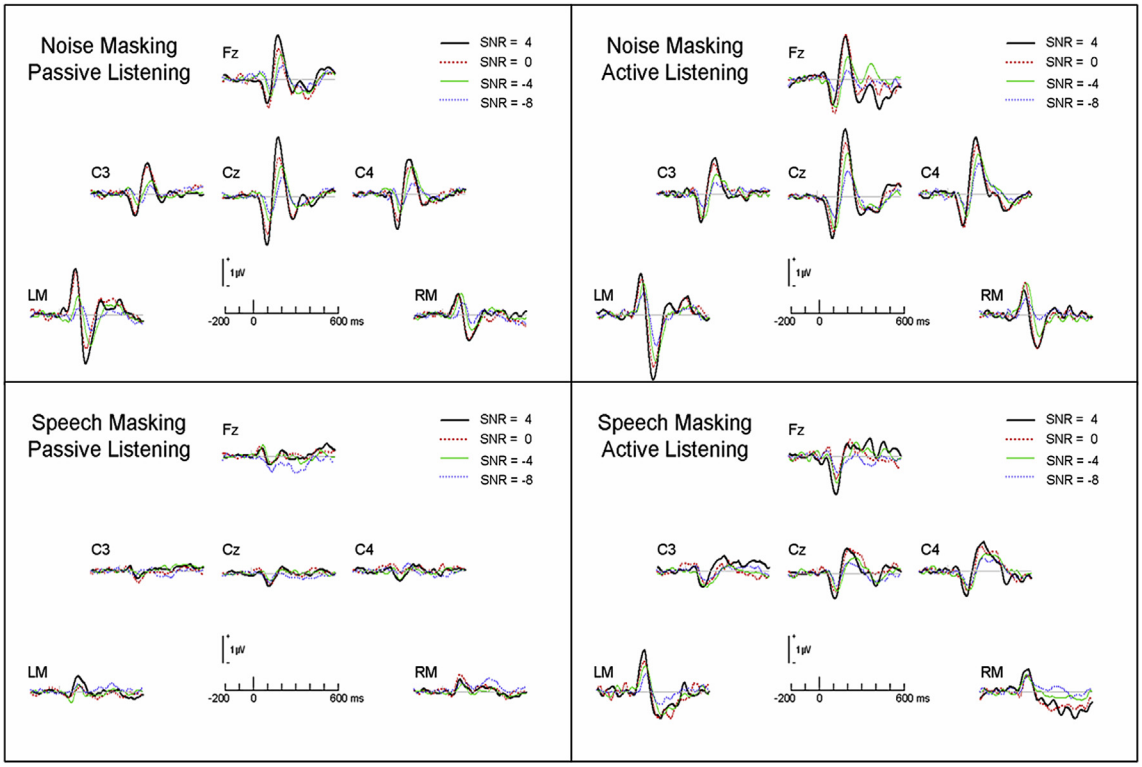

Fig. 4. Grand mean ERP waveforms evoked by the target syllable /*bi*/ recorded from six electrode sites (Fz, C3, Cz, C4, LM, and RM) under four conditions: Left upper panel, noise masking, passive listening; Right upper panel, noise masking, active listening; Left lower panel, speech masking, passive listening; Right lower panel, speech masking, active listening. The four signal-to-masker ratios (SMRs, −8, −4, 0, 4 dB) are represented with four different colors.

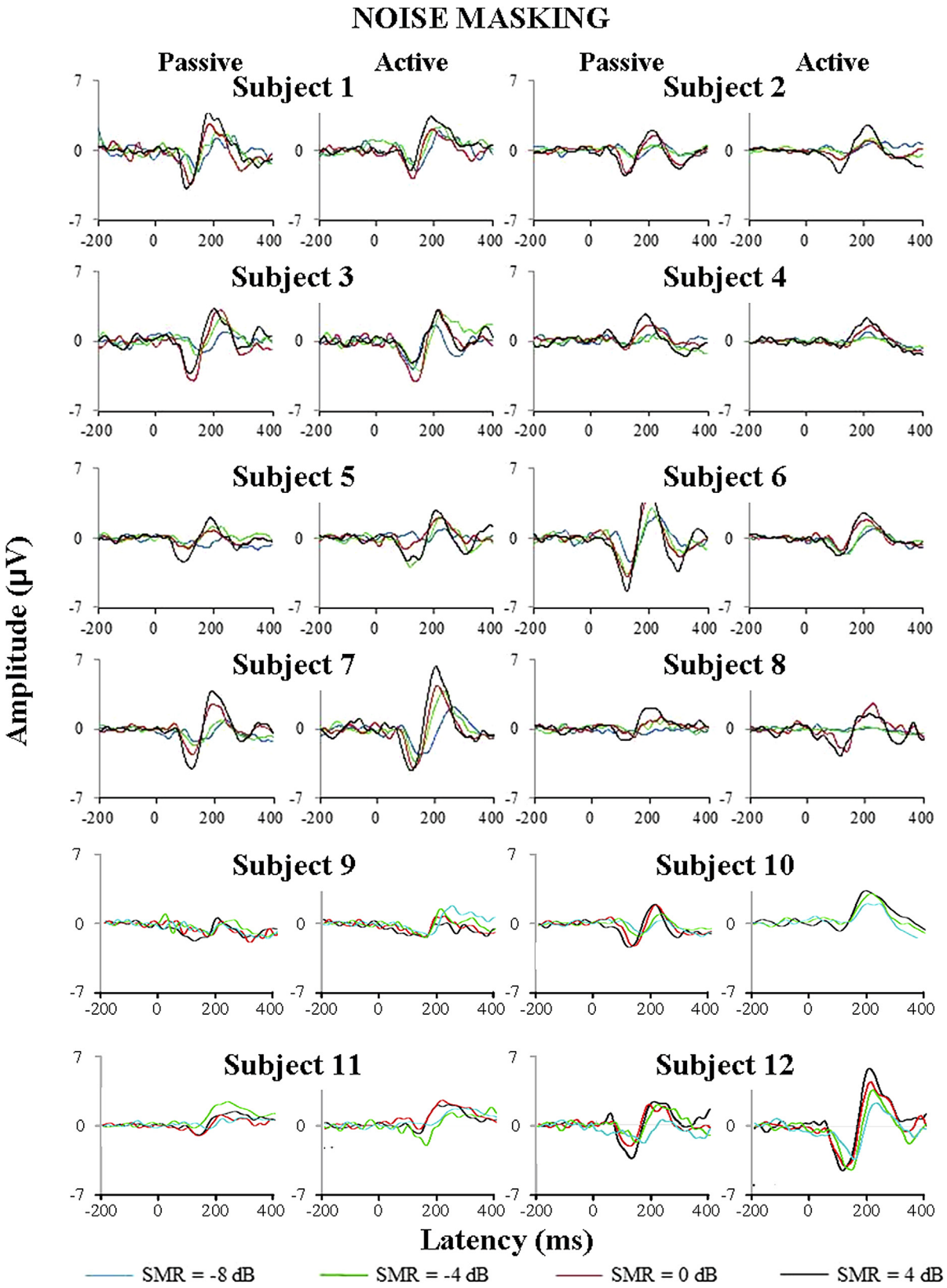

Fig. 5. Mean ERP waveforms evoked by the target syllable /*bi*/ recorded from electrode site Cz for each participant under the noise-masking condition, when the listening condition was either passive (left panel for each participant) or active (right panel for each participant). The four signal-to-masker ratios (SMRs, −8, −4, 0, 4 dB) are represented with four different colors. Note that the N1−P2 amplitudes were quite similar between passive- and active-listening conditions for all the participants except Participant 6, for whom the N1−P2 amplitudes were smaller under the active-listening condition than under the passive-listening condition, and Participants 7, 10, and 12 whose N1−P2 waveforms were larger under active than passive conditions. Generally, for all the participants, a shift from the passive condition to the active condition did not improve the ERPs evoked by the target syllable.

waveforms recorded at Cz.

Figs. 5 and 6 display the ERP waveforms evoked by the target syllable /bi/ recorded from electrode site Cz for each of the 12 subjects under the noise-masking (Fig. 5) and speech-masking (Fig. 6) conditions. When the masker type was noise (Fig. 5), the N1–P2 amplitudes were quite similar between passive- and active-listening conditions for all the subjects except Participant 6, for whom the N1–P2 amplitudes were smaller under the active-listening condition than under the passive-listening condition and Participants 7, 10, and 12 whose N1–P2 waveforms were somewhat larger under active than passive conditions. It also appears that the amplitudes of the two peaks tend to increase with increases in SMR, and their latencies to decrease with increases in SMR.

When the masker type was speech (Fig. 6), the N1–P2 amplitudes tend to increase when the listening condition was shifted from passive to active in all but two of the participants (Subjects 2 and 4).

The average ERP waveforms elicited by the syllable */bi/* at site Cz, are plotted in Fig. 7 for each of the 16 experimental conditions. As shown in Fig. 7, the mean amplitudes of the N1–P2 complex appear to be both larger and more sensitive to SMR changes under the noise-masking condition than under the speech-masking condition. When the masker is noise, a change from passive to active listening appears to have very little effect on these average waveforms. However, when the masker is speech, a change from passive to active listening appears to have a much stronger effect on the N1 and P2 components of the waveform.

3.1. Amplitudes of the N1 component

To indicate how listening condition (active versus passive) and SMR affect N1 amplitude, the top panels of Fig. 8 plot the average values of N1 amplitudes as a function of SMR for noise masking (left upper panel) and speech masking (right upper panel) under both active and passive listening conditions. When the masker is noise, the amplitude of the N1 component appears to increase linearly with SMR at approximately the same rate when listening is passive as it does when listening is active. When the masker is two-talker speech, the amplitude of the N1 component when listening is active appears to be greater than when listening is passive and may grow at a faster rate as SMR increases.

A 2 (attention condition) by 4 (SMR) two-way repeated-measures ANOVA on N1 amplitude was conducted for the noise-masking condition and the speech-masking condition, separately. Under the noise-masking condition, the main effect of SMR was found significant [$F(3,33) = 14.838$, $p < .001$], but neither the effect of attention condition [$F(1,11) < 1$] nor the two-way interaction [$F(3,33) = 2.163$, $p = .111$] was found significant. Under the speech-masking condition, there were significant main effects of attention condition [$F(1,11) = 10.434$, $p = .008$], and SMR [$F(3,33) = 6.094$, $p = .002$], but the two-way interaction [$F(3,33) = 1.431$, $p = .251$] was not significant.

3.2. Amplitudes of the P2 component

The mean values of P2 amplitudes across participants are displayed in the lower panels of Fig. 8. When the masker was noise, a 2 (attention condition) by 4 (SMR) two-way repeated-measures ANOVA showed a significant main effect of SMR [$F(3,33) = 20.011$, $p < .001$], but no statistically significant main effect of attention condition [$F(1,11) = 1.938$, $p = .191$] or of the interaction between attention and SMR [$F(3,33) < 1$] was found. When the masker was speech, a 2 (attention condition) by 4 (SMR) two-way repeated-measures ANOVA showed a significant main effect of attention condition [$F(1,11) = 20.063$, $p = .001$], a significant main effect of SMR [$F(3,33) = 3.601$, $p = .024$], but the two-way interaction [$F(3,33) = 1.018$, $p = .397$] was not found to be statistically significant.

3.3. Latencies of the N1 and P2 component

The mean values of N1 latencies across participants when the masker was noise are displayed in right top panel of Fig. 9. A 2 (attention condition) by 4 (SMR) repeated-measures ANOVA on N1 latency in noise showed that the main effect of SMR was significant [$F(3,33) = 19.736$, $p < .001$], but not that of the attention condition [$F(1,11) = 2.502$, $p = .142$]. However, the two-way interaction [$F(3,33) = 3.253$, $p = .034$] was found to be statistically significant. An examination of this figure indicates that the interaction is due to the rather large difference in N1 latencies in the lowest SMR condition.

The N1 latencies in the two-talker masker condition (left-hand side of the top panel of Fig. 9) suggest that these latencies do not vary across the attention condition, and do not appear to decrease substantially with increasing SMR. A 2 (attention condition) by 4 (SMR) repeated-measures ANOVA on N1 latency when speech was a masker did not find any significant effects due to the attention condition [$F(1,11) < 1$], SMR [$F(3,33) = 1.866$, $p = .154$], nor any interaction between the main two factors [$F(3,33) < 1$].

The bottom two panels show how P2 latency varies with SMR and Attention when the masker was Noise (left panel), and Speech. A 2 (attention condition) by 4 (SMR) repeated-measures ANOVA on P2 latencies for Noise maskers (lower left-hand panel) found a significant effect of SMR [$F(3,33) = 21.901$, $p < .001$], but no effect of Attention [$F(1,11) < 1$], nor any interaction between the two [$F(3,33) = 1.632$, $p = .201$]. The equivalent analysis for the right-hand panel (Speech Masker) did not find any significant effects of Attention [$F(1,11) = 1.522$, $p = .243$], SMR [$F(3,33) = 2.450$, $p = .081$], nor any interaction between the two factors [$F(3,33) = 1.066$, $p = .377$].

4. Discussion

4.1. Effects of masker type

In this study, regardless of whether the listening condition was passive or active, the amplitudes of the N1 and P2 components of ERPs evoked by the syllable /bi/ were much smaller under the speech-masking condition than the noise-masking condition at all SMR levels despite the fact that the syllable was readily detectable in both kinds of maskers. As mentioned in the Introduction a number of behavioral studies suggest that stream segregation is more difficult to achieve when the masking background is acoustically similar to the speech signal. The amplitude envelope of a steady-state noise is relatively flat. Hence, the frequency-dependent amplitude fluctuations that are produced when a speech sound is superimposed on this steady-state background are quite likely to elicit recognizable transient responses in the auditory pathway. However, when there are many frequency-dependent amplitude fluctuations in the background (as there would be when the background is two-talker speech), any transient response due to the speech syllable would be one of many that are continually elicited by the variable nature of the competing speech (e.g., Billings et al., 2011; Kozou et al., 2005; Skoe and Kraus, 2010). As a result we would expect the N1–P2 components of the P1–N1–P2 complex to be less prominent when the masker is speech as opposed to noise, irrespective of attentional state. Indeed, as Fig. 7 clearly shows, despite the degree of inter-participant variability present in Fig. 6, the average N1–P2 components are

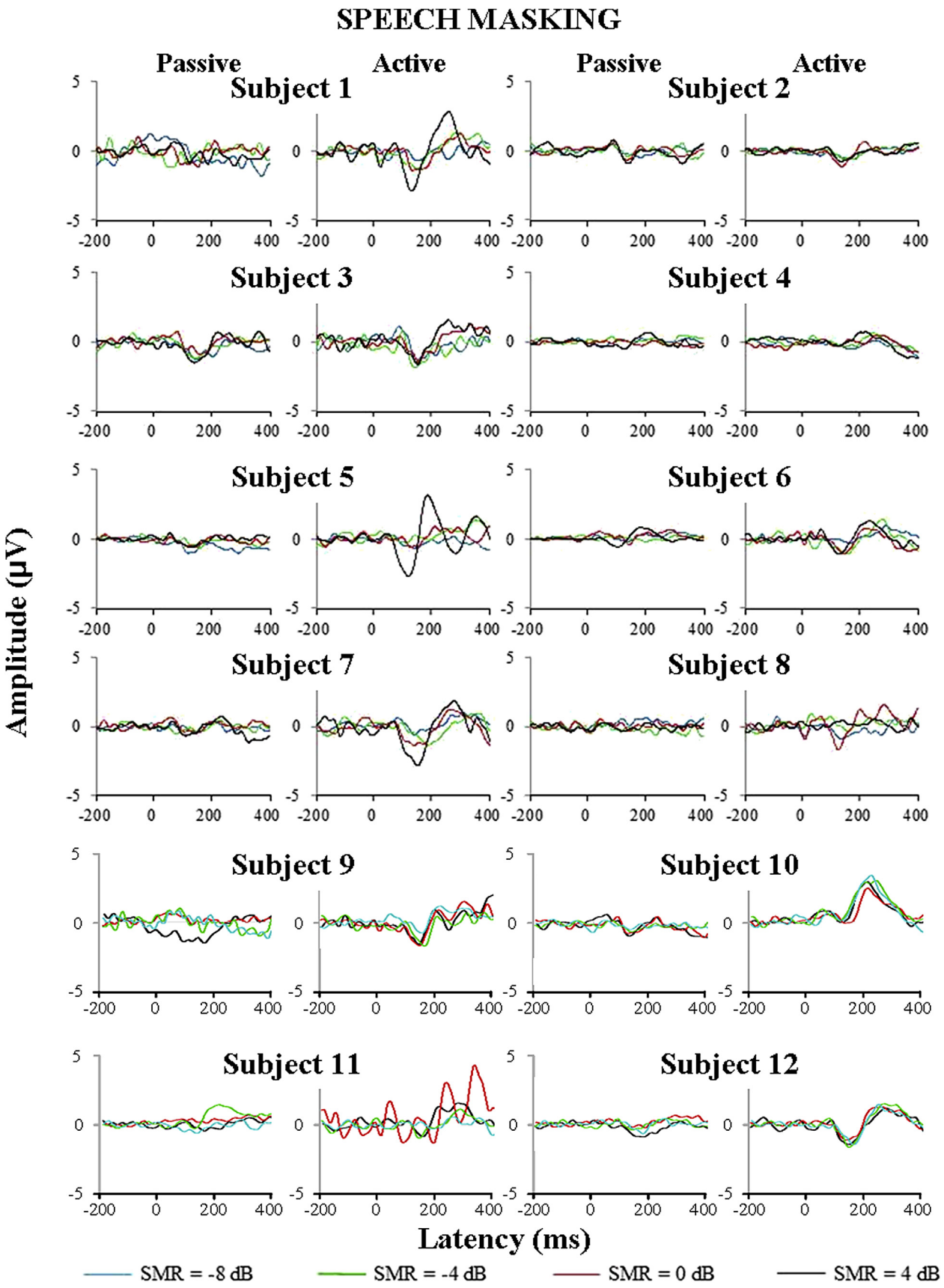

Fig. 6. Mean ERP waveforms evoked by the target syllable /bi/ recorded from electrode site Cz for each participant under the speech-masking condition, when the listening condition was either passive (left panel for each participant) or active (right panel for each participant). The four SMRs are represented by four different colors. Note that for all the participants, the N1–P2 amplitudes increased to some extent when the listening condition was shifted from passive to active.

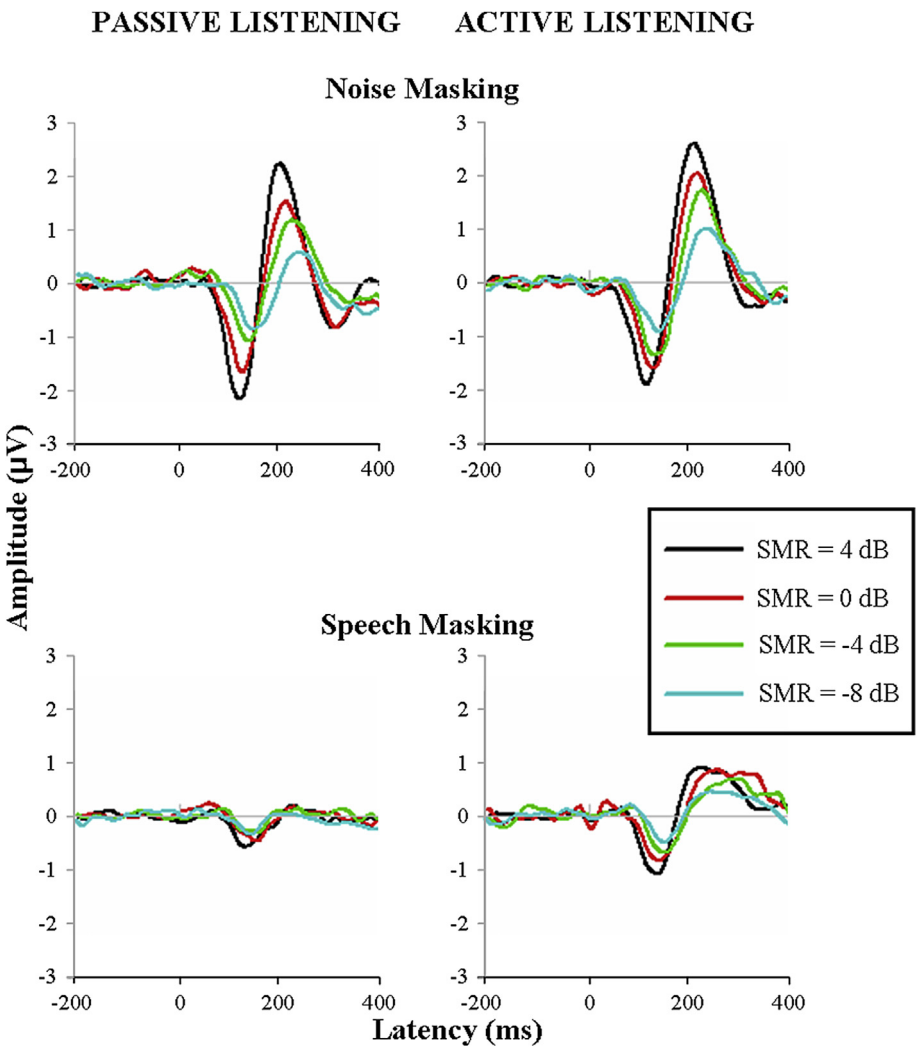

Fig. 7. Grand mean ERPs recorded from electrode site Cz to the target syllable /*bi*/ , when the masker was steady-state speech-spectrum noise (upper panels) or two-talker speech (lower panels). The left panels show ERPs under the passive-listening condition, and the right panels represent ERPs under the active-listening condition. The four SMRs are represented by four different colors. Note that the amplitudes of the N1–P2 complex appear to be larger and more sensitive to the change in SMR under the noise-masking condition than those under the speech-masking condition, but were more vulnerable to the change in attention condition under the speech-masking condition than under the noise-masking condition.

attenuated more by a speech masker than by a noise masker in both active and passive listening conditions. However, there is clear evidence that the N1–P2 components are more prominent in the active than in the passive condition when the masker is speech. Since the stimuli and the masker did not change from the active to the passive conditions, the fact that the N1–P2 complex is more prominent in the active listening condition than in the passive listening condition suggests that top-down attentional processes are sharpening the cortical response to the stimulus when participants are required to actively attend to the stimulus.

Other electro-physiological studies have also found evidence that selective attention to a speech target enhances cortical responses to speech targets being masked by speech. Mesgarani and Chang (2012) presented a target sentence masked by a competing sentence to epilepsy patients implanted with electrode arrays in the posterior temporal lobe (as part of their workup for surgery). The simultaneously presented sentences were modeled after those found in the coordinate response measure corpus (Bolia et al., 2000). For example, the two sentences might be “Ready Baron, go to blue two now,” and “Ready Tiger, go to red one now,” with the target sentence being identified prior to the simultaneous presentation of the two sentences by specifying its ‘call sign’ (Baron or Tiger). Mesgarani & Chang were able to show that neural activity recorded from epilepsy patients in this region was highly correlated with the spectral–temporal features of the sentence designated as the target, not only when the target sentence was presented alone but also when it was being masked by a competing sentence of the same type. In addition, Golumbic et al. (2013) have also shown that attention enhances the cortical representation of target speech being masked by competing speech in epilepsy patients. Hence these two studies, along with the present results support the hypothesis that selective attention can enhance cortical responses to a speech target being masked by speech.

The present study supports the hypothesis that the listener, in order to process the target syllable in a background of speech, has to engage attentional resources to segregate the target voice from the background. In contrast, when the background is steady-state noise, the transient response initiated by the speech syllable may be able to gain cortical access without having to engage top-down attentional processes to isolate it from the background. The greater

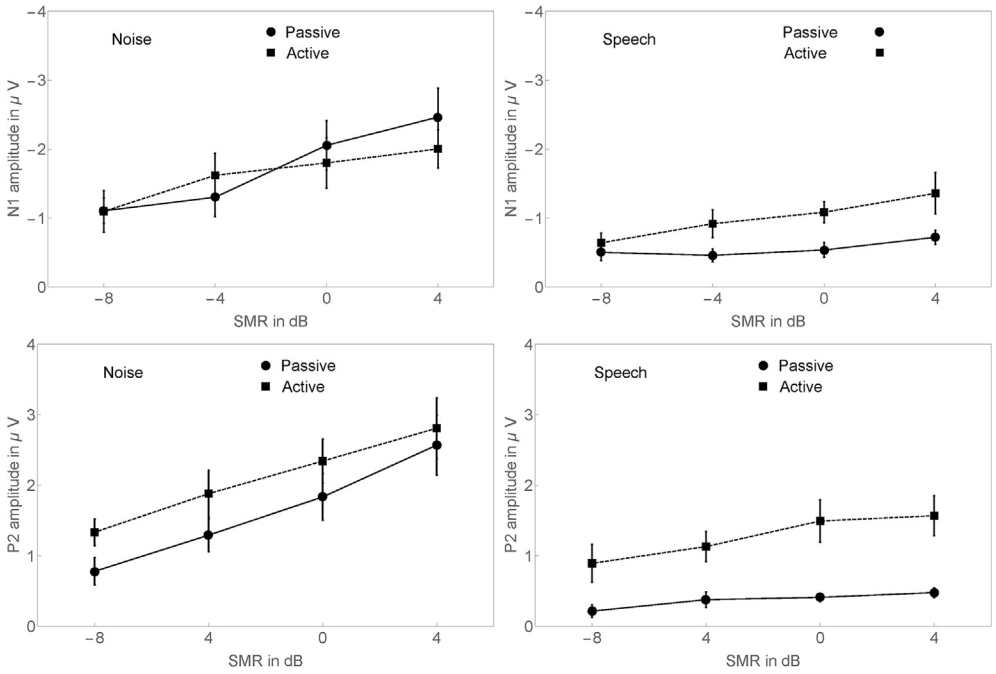

Fig. 8. The average of the individual values for N1 and P2 amplitudes recorded from the electrode site Cz under each of the 16 conditions. In all conditions both N1 and P2 amplitudes increased with SMR. Both N1 and P2 amplitudes were larger when the masker was noise than when it was speech. When the masker was noise the difference between active and passive conditions was statistically significant different for P2 but not for N1. When the masker was speech, both N1 and P2 amplitudes were significantly larger under active as opposed to passive listening conditions. Standard error bars are shown.

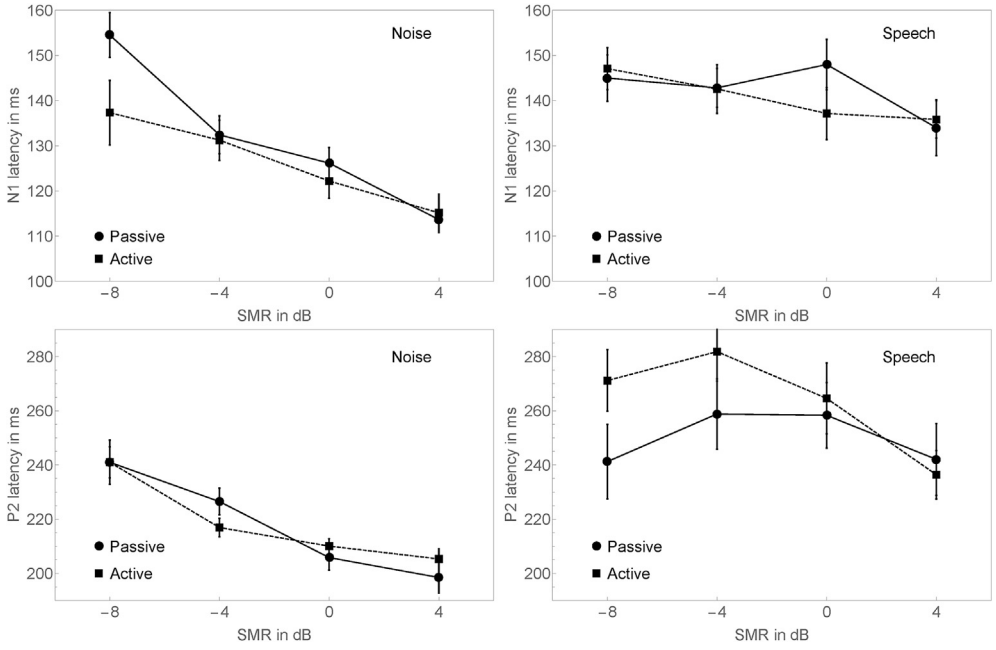

Fig. 9. The average of the individual N1 and P2 latencies recorded from the electrode site Cz under each of the 16 conditions. When the masker was Speech, N1 and P2 latencies did not differ significantly with respect to the attention condition or vary significantly with SMR. When the masker was noise, N1 and P2 latencies decreased as a function of SMR. Standard error bars are shown.

demand on attentional resources when listening to speech masked by speech than when listening to speech masked by noise is likely to increase the cognitive load placed on the listener, thereby straining the cognitive resources available for higher-order processing of the speech signal.

4.2. Effects of SMR on ERP amplitudes

The present study was the first to investigate how attention affects the interaction between the SMR and the masker type (noise, speech) on speech-evoked ERPs. The results showed that the SMR modulation of N1 and P2 amplitudes is masker-type

dependent. As the SMR was increased, the N1 and P2 amplitudes became significantly larger when the masker was noise, and did not differ between active and passive listening. However, when the masker was speech, ERP amplitudes were larger under active than passive listening. This is consistent with the view that there is a greater need to engage attentional processes when the background is speech than when it is steady-state noise.

4.3. Effects of listening condition on ERP amplitudes

As mentioned in the Introduction, ERP recordings make it possible to examine how attention affects auditory processing of target signals when a masker background is present. The ERP study of Tervaniemi et al. (2009) has shown that musicians displayed larger mismatched negativity and N2b to speech sounds than did non-musicians under the attentive-listening condition but not the passive-listening condition, indicating certain enhanced top-down strategies for processing fine structure obtained from musical training (also see Warren, 1999). In the present study, one of the most important results is that shifting the listener's attention from irrelevant visual stimuli to the target stimulus significantly released both the N1 and P2 components of ERPs to the target from speech masking but not from noise masking.

4.4. Effects of listening condition and SMR on ERP latencies

In this study, under either the passive- or active-listening condition, when the masker was noise, both the N1 and P2 latencies decreased as the SMR increased. However, there was some indication of an interaction between SMR and the listening condition with respect to N1 insofar as the latency in the passive condition at the lowest SMR (−8 dB) was considerably longer than in the active condition (see Fig. 9). This would be consistent with the notion that a switch from passive to active listening has an effect on neural processing in a noise background even when the masker is steady-state noise, when listening becomes difficult (lower SMR).

The failure to find any significant effect of either attention or SMR on N1 and P2 latencies when the masker was speech most likely reflects the difficulty in specifying the location of these two peaks when the background sound is variable, as it is when the masker is two-talker speech. The fact that attention had a much more prominent effect on N1 and P2 amplitudes when the masker was speech than when it was noise, is consistent with the hypothesis that the need to engage top-down attentional processes is increased as the informational content of the masker is increased (two-talker speech versus steady-state noise).

5. Summary

(1) Under either the active-listening condition or the passive-listening condition, the two-talker-speech masker induced a much larger masking effect than the steady-state-noise masker on both the N1 and the P2 components of the ERPs to the syllable /*bi*/ , suggesting that the need for top-down attentional processing of the speech signal is increased as the masking background becomes more informationally complex.

(2) A shift from the passive listening condition to the active one affects the magnitude of the ERPs to the target syllable when the masker is speech, again indicating that there is a greater need for cortical processing when the auditory background is informationally complex.

Acknowledgments

This work was supported by the National High Technology Research and Development Program of China (863 program: 2015AA016306), the National Natural Science Foundation of China (31170985), the CJN13J004 Grant, and grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 995), the Canadian Institutes of Health Research (MOP 1539, and TEA-12497).

References

Alain, C., Quan, J., McDonald, K.L., Van Roon, P., 2009. Noise-induced increase in human auditory evoked neuromagnetic fields. Eur. J. Neurosci. 30, 132–142.
Alain, C., McDonald, K.L., Van Roon, P., 2012. Effects of age and background noise on processing a mistuned harmonic in an otherwise periodic complex sound. Hear. Res. 283, 126–135.
Alain, C., Roye, A., Salloum, C., 2014. Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex. Front. Syst. Neurosci. 8, 8. http://dx.doi.org/10.3389/fnsys.2004.00008.
Alho, K., 1992. Selective attention in auditory processing as reflected by event-related brain potentials. Psychophysiology 29, 247–263.
Arbogast, T.L., Mason, C.R., Kidd, G., 2002. The effect of spatial separation on informational and energetic masking of speech. J. Acoust. Soc. Am. 112, 2086–2098.
Billings, C.J., Bennett, K.O., Molis, M.R., Leek, M.R., 2011. Cortical encoding of signals in noise: effects of stimulus type and recording paradigm. Ear Hear. 32, 53–60.
Bolia, R.S., Nelson, W.T., Ericson, M.A., Simpson, B.D., 2000. A speech corpus for multitalker communications research. J. Acoust. Soc. Am. 107, 1065–1066.
Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109, 1101–1109.
Brungart, D.S., Simpson, B.D., 2002. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J. Acoust. Soc. Am. 112, 664–676.
Carhart, R., Johnson, C., Goodman, J., 1975. Perceptual masking of spondees by combinations of talkers. J. Acoust. Soc. Am. 58, 35.
Cherry, E.C., 1953. Some experiments on the recognition of speech with one and two ears. J. Acoust. Soc. Am. 25, 975–979.
Durlach, N.I., Mason, C.R., Shinn-Cunningham, B.G., Arbogast, T.L., Colburn, H.S., Kidd, G., 2003. Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. J. Acoust. Soc. Am. 114, 368–379.
Dubno, J.R., Schaefer, A.B., 1992. Comparison of frequency-selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners. J. Acoust. Soc. Am. 91, 2110–2121.
Ezzatian, P., Li, L., Pichora-Fuller, K., Schneider, B.A., 2011. The effect of priming on release from informational masking is equivalent for younger and older adults. Ear Hear. 32, 84–96.
Freyman, R.L., Helfer, K.S., McCall, D.D., Clifton, R.K., 1999. The role of perceived spatial separation in the unmasking of speech. J. Acoust. Soc. Am. 106, 3578–3588.
Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2001. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Am. 109, 2112–2122.
Freyman, R.L., Balakrishnan, U., Helfer, K.S., 2004. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. Acoust. Soc. Am. J. 115, 2246–2256.
Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., et al., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron 77 (5), 980–991.
Helfer, K.S., Freyman, R.L., 2005. The role of visual speech cues in reducing energetic and informational masking. J. Acoust. Soc. Am. 117, 842–849.
Huang, Y., Huang, Q., Chen, X., Qu, T.-S., Wu, X.-H., Li, L., 2008. Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults. Hear. Res. 244, 51–65.
Huang, Y., Huang, Q., Chen, X., Wu, X.-H., Li, L., 2009. Transient auditory storage of acoustic details is associated with release of speech from informational masking in reverberant conditions. J. Exp. Psychol. Hum. Percept. Perform. 35, 1618–1628.
Huang, Y., Xu, L.-J., Wu, X.-H., Li, L., 2010. The effect of voice cuing on releasing speech from informational masking disappears in older adults. Ear Hear. 31, 579–583.
Kidd, G., Mason, C.R., Deliwala, P.S., Woods, W.S., Colburn, H.S., 1994. Reducing informational masking by sound segregation. J. Acoust. Soc. Am. 95, 3475–3480.
Kidd, G., Mason, C.R., Rohtla, T.L., Deliwala, P.S., 1998. Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. J. Acoust. Soc. Am. 104, 422–431.
Kidd, G., Arbogast, T.L., Mason, C.R., Gallun, F.J., 2005. The advantage of knowing where to listen. J. Acoust. Soc. Am. 118, 3804–3815.
Kozou, H., Kujali, T., Shtyrov, Y., Toppila, E., Starck, J., Arku, P., Näätänen, R., 2005.

The effect of different noise types on the speech and non-speech mismatch negativity. Hear. Res. 199, 31–39.
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? J. Exp. Psychol. Hum. Percept. Perform. 30, 1077–1091.
Li, H.-H., Kong, L.-Z., Wu, X.-H., Li, L., 2013. Primitive auditory memory is correlated with spatial unmasking that is based on direct-reflection integration. PLoS One 8 (4), e63106.
Martin, B.A., Sigal, A., Kurtzberg, D., Stapells, D.R., 1997. The effects of decreased audibility produced by high-pass noise masking on cortical event-related potentials to speech sounds/ba/and/da/. J. Acoust. Soc. Am. 101, 1585–1599.
Martin, B.A., Kurtzberg, D., Stapells, D.R., 1999. The effects of decreased audibility produced by high-pass noise masking on N1 and the mismatch negativity to speech sounds/ba/and/da/. J. Speech, Lang. Hear. Res. 42, 271–286.
Martin, B.A., Stapells, D.R., 2005. Effects of low-pass noise masking on auditory event-related potentials to speech. Ear Hear. 26 (2), 195–213.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236.
Miller, G.A., 1947. The masking of speech. Psychol. Bull. 44, 105–129.
Muller-Gass, A., Marcoux, A., Logan, J., Campbell, K.B., 2001. The intensity of masking noise affects the mismatch negativity to speech sounds in human participants. Neurosci. Lett. 299, 197–200.
Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. J. Phon. 35, 85–103.
Polich, J., Howard, L., Starr, A., 1985. Stimulus frequency and masking as determinants of P300 latency in event-related potentials from auditory stimuli. Biol. Psychol. 21, 309–318.
Salo, S.K., Lang, A.H., Salmivalli, A.J., 1995. Effect of contralateral white noise masking on the mismatch negativity. Scand. Audiol. 24, 165–173.
Skoe, E., Kraus, N., 2010. Auditory brainstem response to complex sounds: a tutorial. Ear Hear. 31 (3), 302–324.
Schneider, B.A., Li, L., Daneman, M., 2007. How competing speech interferes with speech comprehension in everyday listening situations. J. Am. Acad. Audiol. 18, 559–572.
Singh, G., Pichora-Fuller, M.K., Schneider, B.A., 2008. The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. J. Acoust. Soc. Am. 124, 1294–1305.
Tervaniemi, M., Kruck, S., De Baene, W., Schröger, E., Alter, K., Friederici, A.D., 2009. Top-down modulation of auditory processing: effects of sound context, musical expertise and attentional focus. Eur. J. Neurosci. 30, 1636–1642.
Warren, J.D., 1999. Variations on the musical brain. J. R. Soc. Med. 92, 571.
Whiting, K.A., Martin, B.A., Stapells, D.R., 1998. The effects of broadband noise masking on cortical event-related potentials to speech sounds/ba/and/da/. Ear Hear. 19, 218–231.
Wu, C., Cao, S.-Y., Zhou Wu, X.-H., Li, L., 2013. Temporally pre-presented lipreading cues release speech from informational masking. J. Acoust. Soc. Am. 133, EL281–EL285.
Wu, M.-H., Li, H.-H., Gao, Y.-Y., Lei, M., Teng, X.-B., Wu, X.-H., Li, L., 2012. Adding irrelevant information to the content prime reduces the prime-induced unmasking effect on speech recognition. Hear. Res. 283, 136–143.
Wu, X., Wang, C., Chen, J., Qu, H., Li, W., Wu, Y., Schneider, B.A., Li, L., 2005. The effect of perceived spatial separation on informational masking of Chinese speech. Hear. Res. 199, 1–10.
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. INTERSPEECH 390–393.
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 Commun. 49, 892–904.
Yonan, C.A., Sommers, M.S., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol. Aging 15, 88–99.