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