Hearing Research 316 (2014) 73–81

Contents lists available at ScienceDirect

Hearing Research

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

Research paper

Differential modulation of auditory responses to attended and unattended speech in different listening conditions

Ying-Yee Kong [a, b, *], Ala Mullangi [b], Nai Ding [c]

[a] Department of Speech Language Pathology & Audiology, Northeastern University, Boston, MA 02115, United States
[b] Bioengineering Program, Northeastern University, Boston, MA 02115, United States
[c] Department of Psychology, New York University, NY 10012, United States

ARTICLE INFO

Article history:
Received 4 May 2014
Received in revised form
24 July 2014
Accepted 31 July 2014
Available online 11 August 2014

ABSTRACT

This study investigates how top-down attention modulates neural tracking of the speech envelope in different listening conditions. In the quiet conditions, a single speech stream was presented and the subjects paid attention to the speech stream (active listening) or watched a silent movie instead (passive listening). In the competing speaker (CS) conditions, two speakers of opposite genders were presented diotically. Ongoing electroencephalographic (EEG) responses were measured in each condition and cross-correlated with the speech envelope of each speaker at different time lags. In quiet, active and passive listening resulted in similar neural responses to the speech envelope. In the CS conditions, however, the shape of the cross-correlation function was remarkably different between the attended and unattended speech. The cross-correlation with the attended speech showed stronger N1 and P2 responses but a weaker P1 response compared to the cross-correlation with the unattended speech. Furthermore, the N1 response to the attended speech in the CS condition was enhanced and delayed compared with the active listening condition in quiet, while the P2 response to the unattended speaker in the CS condition was attenuated compared with the passive listening in quiet. Taken together, these results demonstrate that top-down attention differentially modulates envelope-tracking neural activity at different time lags and suggest that top-down attention can both enhance the neural responses to the attended sound stream and suppress the responses to the unattended sound stream.

1. Introduction

Top-down attention plays an important role in auditory perception in complex listening environments. Using an event-related design, previous electroencephalography (EEG) and magnetoencephalography (MEG) studies showed greater brain responses to attended sounds relative to the responses to unattended sounds (e.g., Hillyard et al., 1973; Näätänen, 1992; Alain and Woods, 1994; Teder-Sälejärvi et al., 1999; Snyder et al., 2006). The attentional effect can occur at various processing stages. For example, the effects of attentional modulation appear as early as ~100 msec in some experiments (e.g., Hillyard et al., 1973; Hansen and Hillyard, 1988; Alho et al., 1994; Melara et al., 2002; Choi et al., 2014) but later in other experiments (>150 msec; e.g, Picton and Hillyard, 1974; Neelon et al., 2006; Snyder et al., 2006; Ross et al., 2010). Stimulus properties and the subject's task are likely to determine the latency of top-down attentional modulations. In addition to the amplification of brain responses to the attended signals, responses to the unattended signals are sometimes attenuated, suggesting active suppression mechanisms for irrelevant stimuli (e.g., Rif et al., 1991; Alain et al., 1993; Alho et al., 1994; Alain and Woods, 1994; Bidet-Caulet et al., 2010). The effect of suppression is usually found in a later processing stage about 200 msec post stimulus onset.

Recent works have extended neurophysiology studies on top-down auditory attention from using relatively simple sounds, e.g. tones, to using more ecologically realistic stimuli, such as continuous speech (see a recent review by Ding and Simon, 2014). When a single speech stream is presented in a quiet listening environment, the neural responses from the auditory cortex phase lock to the temporal envelope of the speech signal (e.g., Luo and Poeppel, 2007; Aiken and Picton, 2008; Howard and Poeppel, 2010; Lalor

Abbreviations: ANOVA, analysis of covariance; CI, confidence interval; CS, competing speaker; DSS, denoising source separation; EEG, electroencephalography; Hz, hertz; kHz, kilohertz; MEG, magnetoencephalography; msec, millisecond; Q, quiet; RMS, root-mean-square; TMR, target-to-masker ratio

* Corresponding author. Department of Speech Language Pathology & Audiology, 226 Forsyth Building, 360 Huntington Ave., Northeastern University, Boston, MA 02115, United States. Tel.: +1 617 373 3704; fax: +1 617 373 2239.
E-mail address: yykong@neu.edu (Y.-Y. Kong).

http://dx.doi.org/10.1016/j.heares.2014.07.009
0378-5955/

and Foxe, 2010; Pasely et al., 2012). When two speech streams are presented simultaneously, neural activity shows stronger phase locking to the temporal envelope of the attended speech stream, compared with the phase locking to the unattended speech (Kerlin et al., 2010; Ding and Simon, 2012a, 2012b; Mesgarani and Chang, 2012; Horton et al., 2013; O'Sullivan et al., 2014).

Although most previous studies have shown attentional modulation of speech tracking neural activity (see, however, Zion Golumbic et al., 2013), the latency of the attentional modulation effects has been controversial. While some studies reported an early attentional gain effect at a time lag around 100 msec (Ding and Simon, 2012a, 2012b), others reported a longer-latency attentional effect near 200-msec time lag (Power et al., 2012; O'Sullivan et al., 2014). Furthermore, more complicated patterns were observed by Horton et al. (2013), who reported that the EEG responses were correlated with the attended and unattended speech streams with opposite signs at time lags between 150 and 450 msec. Based on the neurophysiological findings that the phase of slow neural oscillations modulates the excitability of neurons (Lakatos et al., 2008, 2013; Schroeder and Lakatos, 2009), Horton et al. (2013) suggested that the opposite polarities of the cross-correlations reflect enhancement of the attended speech and suppression of the unattended speech.

The goal of the present study is to investigate whether top-down attention differentially modulates neural tracking of the speech envelope when the target speech stream is competed with different types of sensory interference. Specifically, when the properties of the sensory interference varied, we tested if the effect of top-down attention may change from modulating the response gain (e.g., Ding and Simon, 2012b) to modulating the response timing (Horton et al., 2013; O'Sullivan et al., 2014), or to not significantly modulating cortical activity at all (Zion Golumbic et al., 2013).

First, we investigated whether cortical responses may be differentially modulated by attention when competing information was presented via the same or different sensory modalities. In the cross-modality condition, a narrated story was presented in a quiet listening environment, and the subjects were instructed to either listen to the story or watch a silent movie instead. In the within-modality condition, the subjects heard a mixture of two simultaneous speakers, one male and one female, and had to selectively attend to one of them based on the instruction. Second, we compared the neural responses to an attended or unattended speech stream when that speech stream was presented together with a speech stream of the same sound intensity, a speech stream of a lower sound intensity, or in a quiet listening environment. Using these conditions, we probed how irrelevant information sources are filtered out by top-down attention when competitors are presented from different sensory modalities, and how the filtering process may depend on the amount of interference within the auditory modality.

2. Methods

2.1. Subjects

Eight normal-hearing, right-handed (Oldfield, 1971), adult native speakers of American English between the ages of 21 and 36 years participated in the study. This study was conducted according to the protocols approved by the Institutional Review Board of Northeastern University. Written informed consent was obtained prior to the experiment.

2.2. Test stimuli and procedures

Auditory stimuli were continuous speech extracted from two chapters in a public domain children's book, "A Child's History of England" by Charles Dickens (http://librivox.org/a-childs-history-of-england-by-charles-dickens/), narrated by one male and one female speaker. The first chapter was 22, read by a male speaker, and the second was chapter 35, read by a female speaker. The sampling rate of the recordings was 22.05 kHz. All silent intervals longer than 300 msec were shortened to 300 msec to maintain continuous flow of the speech streams. The extracted passages were divided into sections with durations of approximately 1 min each. The actual length of the segment varied slightly to include a complete sentence. All of the 1-min speech segments were normalized to have equal root-mean-square (RMS) amplitude. In addition, for the competing speaker conditions described below, speech mixtures were constructed by mixing two speakers digitally with one speaker beginning 1 s after the other speaker, and both speakers ending the same time. The RMS level of one speaker was fixed in the mixture, while the other speaker (the speaker with a delayed start) was either the same or 6 dB weaker, resulting in two target-to-masker ratio (TMR) conditions. All stimuli were presented diotically using insert earphones. All experiments were conducted in a double-walled sound-isolated booth.

There were two main experiments – speech comprehension in quiet and speech comprehension in a competing background. Prior to the main experiments, each subject was presented with 150 repetitions of a short (100-msec with 10-msec on- and off-ramps) 1000-Hz tone pip to elicit the auditory N1 response. After confirming that N1 response was present for the subject, (s)he was tested with clean speech (quiet [Q] condition) and with speech mixtures (competing speech [CS] condition). The clean speech in Q conditions or the attended speech in the CS conditions was presented at 65 dB A. Each subject completed the Q conditions before the CS conditions. For the Q listening conditions (see Fig. 1A), subjects were asked to either pay attention to the presented speech stimuli (active listening) while fixating their eyes on a crosshair ("+") on a computer screen in front of them, or not attend to the speech sounds but pay attention to a silent movie on a computer screen in front of them (passive listening). The silent movie was extracted from the animated film "Snow White." Four randomly chosen 1-min speech segments (two from each speaker) were presented to each subject for each quiet listening condition. Each of the 1-min speech segments was presented 10 times, resulting in a total of four blocks of testing per listening condition. All subjects were tested with the active listening condition first, followed by the passive listening condition. The active and passive conditions were tested in two different 2-h test sessions.

For the CS conditions (see Fig. 1B), each subject was presented with two randomly chosen 1-min speech mixtures for each TMR condition. The experiment was divided into four blocks per TMR, such that each of the two speech mixtures was used for two blocks. Each trial began with a written text cue (the word "male" or "female") on a computer screen to indicate which speaker the subject should pay attention to in the mixture. The cue lasted for 1 s and was the replaced by a crosshair, where subjects maintained visual fixation for the duration of the trial. Subjects were asked to focus on the male speaker in block 1 and the female speaker in block 2 for the first and second speech mixture, respectively. In block 3 and block 4, subjects switched their attention to the other speaker for the same speech mixtures as in block 1 and block 2, respectively. For each mixture, the unattended speaker started 1-sec after the attended speaker to help the subject listen to the correct speaker as cued. Similar to the Q conditions, each of the 1-min speech mixtures were presented 10 times per block. Blocks 1 and 2 for each of the two TMR conditions were tested in one 2-h test session. Blocks 3 and 4 were tested in a different 2-h session. The TMR of 6 dB (i.e., the level of the attended speech was 6 dB higher than that of the