NeuroImage 156 (2017) 435–444

Contents lists available at ScienceDirect

NeuroImage

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

Noise-robust cortical tracking of attended speech in real-world acoustic scenes

Søren Asp Fuglsang[a,*], Torsten Dau[a], Jens Hjortkjær[a,b,*]

[a] Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, 2800 Kgs. Lyngby, Denmark
[b] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Kettegaard Allé 30, 2650 Hvidovre, Denmark

ARTICLE INFO

Keywords:
Auditory attention
Speech
Cortical entrainment
EEG
Decoding
Acoustic simulations
Delta rhythms
Theta rhythms

ABSTRACT

Selectively attending to one speaker in a multi-speaker scenario is thought to synchronize low-frequency cortical activity to the attended speech signal. In recent studies, reconstruction of speech from single-trial electroencephalogram (EEG) data has been used to decode which talker a listener is attending to in a two-talker situation. It is currently unclear how this generalizes to more complex sound environments. Behaviorally, speech perception is robust to the acoustic distortions that listeners typically encounter in everyday life, but it is unknown whether this is mirrored by a noise-robust neural tracking of attended speech. Here we used advanced acoustic simulations to recreate real-world acoustic scenes in the laboratory. In virtual acoustic realities with varying amounts of reverberation and number of interfering talkers, listeners selectively attended to the speech stream of a particular talker. Across the different listening environments, we found that the attended talker could be accurately decoded from single-trial EEG data irrespective of the different distortions in the acoustic input. For highly reverberant environments, speech envelopes reconstructed from neural responses to the distorted stimuli resembled the original clean signal more than the distorted input. With reverberant speech, we observed a late cortical response to the attended speech stream that encoded temporal modulations in the speech signal without its reverberant distortion. Single-trial attention decoding accuracies based on 40–50 s long blocks of data from 64 scalp electrodes were equally high (80–90% correct) in all considered listening environments and remained statistically significant using down to 10 scalp electrodes and short (< 30-s) unaveraged EEG segments. In contrast to the robust decoding of the attended talker we found that decoding of the unattended talker deteriorated with the acoustic distortions. These results suggest that cortical activity tracks an attended speech signal in a way that is invariant to acoustic distortions encountered in real-life sound environments. Noise-robust attention decoding additionally suggests a potential utility of stimulus reconstruction techniques in attention-controlled brain-computer interfaces.

Introduction

Speech communication is remarkably robust to the signal distortions encountered in everyday acoustic environments. Successful speech comprehension in noisy situations relies both on the ability of the auditory system to segregate simultaneous sound sources, but also on the listeners' ability to direct attentional focus to a potentially degraded sound stream while suppressing irrelevant information. Numerous electrophysiological studies in humans have reported a synchronization between the slow (< 20 Hz) temporal modulations inherent in speech signals and low-frequency cortical activity in the delta (1–4 Hz) and theta (4–8 Hz) frequency ranges (Ahissar et al., 2001; see Ding and Simon, 2014 for a review). In scenarios with more than one talker, selective attention has been shown to enhance the cortical tracking of the attended speech and to suppress synchronization of the ignored speech (Ding and Simon, 2012a, 2012b; Horton et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012; Power et al., 2012; Zion Golumbic et al., 2013). A number of recent studies have employed 'stimulus reconstruction' techniques (Bialek et al., 1991; Rieke et al., 1995; Mesgarani et al., 2009) to reconstruct the envelopes of competing speech signals from the EEG response. It has been shown that an enhanced neural reconstruction of the attended speech signal can be used to decode which talker a listener is attending to with less than one minute of unaveraged EEG data (Mirkovic et al.,

* Corresponding authors at: Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, 2800 Kgs. Lyngby, Denmark.
E-mail addresses: soerenf@elektro.dtu.dk (S.A. Fuglsang), jhjort@elektro.dtu.dk (J. Hjortkjær).

http://dx.doi.org/10.1016/j.neuroimage.2017.04.026
Received 10 December 2016; Accepted 10 April 2017
Available online 13 April 2017

2015; O'Sullivan et al., 2014). However, successful decoding of attention from single-trial EEG has only been demonstrated in acoustically controlled environments with two competing talkers. It is currently unknown how well these results generalize to real-life acoustic scenarios where the speech signals to be decoded are distorted, e.g. by reverberation or background noise.

In order to extract speech sources from a complex scene, the brain must represent the relevant signal in a way that is robust to noise. Converging evidence from animal electrophysiology suggests that noise-invariant representations of sounds emerge at later stages in the auditory pathway by neuronal adaptation to stimulus statistics (Mesgarani et al., 2014; Moore et al., 2013; Rabinowitz et al., 2013). Using stimulus reconstruction with population responses in ferrets, Mesgarani et al. (2014) demonstrated that speech stimuli with added reverberation or stationary noise reconstructed from auditory cortex resembled the original clean signal more than the distorted signal. Temporal coding of amplitude modulations in the auditory midbrain may even be enhanced in real-world reverberation compared to anechoic conditions (Slama and Delgutte, 2015). In cortex, the entrainment of low-frequency activity to speech envelope fluctuations has been shown to be robust to stationary noise, even when the intensity of the background noise is greater than the speech signal (Ding and Simon, 2013). However, in situations where the 'background' sound stream is itself a potential auditory object of interest (e.g., another speech signal), perception cannot rely solely on bottom-up mechanisms. In this case, top-down attention plays a critical role in ignoring irrelevant information. Yet, it remains unclear how attention may contribute to the formation of noise-invariant representations in human cortex.

In real-world listening environments, speech signals are inevitably distorted, e.g. by sound reflections and reverberation. Unlike background sounds that can form a separate sound stream to be ignored, such 'transmission' distortions degrade the attended speech signal itself. Previous studies have mainly considered how acoustic degradations that reduce speech intelligibility affect envelope entrainment. Noise-vocoded speech with a temporal envelope that resembles that of the original signal but degrades speech intelligibility has been found to reduce cortical speech tracking responses in the theta range (Peelle et al., 2013; Ding et al., 2014) and to diminish the differential cortical response between the attended and the unattended talker (Rimmele et al., 2015; Kong et al., 2015). However, such artificial signal manipulations also affect the statistics of natural speech stimuli and degrade the acoustic cues that listeners typically use to maintain stable speech recognition. In real-world reverberant rooms, normal-hearing listeners can maintain robust speech recognition even when the acoustic envelope of a signal has been substantially distorted (Darwin and Hukin, 2000; Ruggles and Shinn-Cunningham, 2010). Robust discrimination of sound sources has been proposed to rely on the statistical regularities that are found in real-world reverberation (Traer and McDermott, 2016). It is currently unknown whether robust speech perception, despite the envelope distortions imposed by real-world rooms, is mirrored by a cortical entrainment mechanism that is robust to these distortions.

In the current study, we investigated this question using speech stimuli embedded in real-world acoustic scenes. Using acoustic simulations, we created virtual auditory scenes with multiple talkers and different reverberant decay properties. In a two-tiered approach for reproduction of these sound scenes, we first reproduced the scenes over earphones in an electromagnetically shielded environment to control for noise in the EEG measurements. Next, we reproduced the sound scenes using a multi-loudspeaker virtual reality facility providing accurate free-field reconstructions of real-world rooms. This provided listeners with the full range of acoustic cues typically encountered in everyday listening situations. Since reverberation and background talkers can severely disrupt the envelope of an attended speech signal, this approach allowed us to examine the cortical tracking of distorted envelopes with intact speech recognition. We measured ongoing scalp EEG while the subjects selectively attended to a particular talker embedded in these different scenes. Using stimulus reconstruction to derive envelopes of attended speech streams from single-trial EEG, we investigated (i) whether the clean envelope of an attended speech stream could be reconstructed from a distorted input when speech perception is robust, and (ii) whether this might facilitate the decoding of the attended talker in real-world acoustic scenes. We hypothesized that attention promotes noise-robust neural representations of attended sound streams such that attended speech envelopes reconstructed from cortical EEG resemble the clean signals to a similar or even higher degree than the distorted input. Such a cortical robustness would allow attention decoding accuracies in real-world acoustic scenes that are similar to those observed with undistorted signals.

Material and methods

Participants

Twenty-nine subjects (13 females, 25 right-handed), aged from 19 to 30 years, participated in the experiment. Three subjects were excluded from the analysis because of missing data from several trials. All participants were students with self-reported normal hearing and no history of neurological disorders. The subjects received financial compensation for their participation in the experiment. The experimental procedure was approved by the Science Ethics Committee for the Capital Region of Denmark, and written informed consents were obtained from all participants before the experiment in accordance with the Declaration of Helsinki.

Stimuli and virtual room simulations

Two hours of speech material were recorded from a male and a female professional story teller narrating fictional stories. The speech material was recorded in an anechoic chamber at the Technical University of Denmark (DTU) and sampled at a frequency of 48 kHz. The naturally spoken stories were subsequently segmented into consecutive 50-second long segments that were each accompanied by multiple-choice questions.

Virtual auditory environments (VAEs) were simulated using the room acoustic modeling software Odeon (version 13.02). To create auditory scenes representative of everyday listening environments, we simulated the acoustics of a mildly reverberant room and a highly reverberant room. A model of a square classroom at DTU (9×7×3 m^3) was used to represent a mildly reverberant environment and a model of the Hagia Irene church (~39.000 m^3) represented a highly reverberant listening scenario. Binaural impulse responses were derived for each of the VAEs at two source-receiver positions, with source-receiver distances of 2.4 m and target sources positioned at ±60° along the azimuth with 0° elevation angles (see Fig. 1C). The impulse responses of the simulated rooms and the 50-s long excerpts of the speech material were used to create virtual auditory scenes with the two speakers of different gender talking at the same time from different positions. The two concurrent speech streams were normalized to have the same root-mean-square (RMS) value and were presented at a sound pressure level (SPL) of 65 dB. For the mildly reverberant room, a scenario with additional 6 talkers (3 male, 3 female) positioned uniformly along the azimuth direction 2.4 m from the listener was simulated in addition to the two target talkers positioned at ±60° azimuth. The multi-talker babble of the six additional speakers was presented at a level of 55 dB SPL. The average decay time constant of the mildly- and highly reverberant rooms were T_{30}=0.9 s and T_{30}=4 s, respectively. The clarity, defined as the ratio of the direct 80-ms sound energy to the remaining energy, ranged between $C_{80,63\ Hz}$=5.7 dB and $C_{80,4\ kHz}$=7.4 dB for the mildly reverberant room and between $C_{80,63\ Hz}$=6.7 dB and $C_{80,4\ kHz}$=9.7 dB for the highly reverberant room.