NeuroImage 88 (2014) 41–46

Contents lists available at ScienceDirect

NeuroImage

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

Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure

Nai Ding [a,b], Monita Chatterjee [c], Jonathan Z. Simon [a,d,e,*]

[a] Department of Electrical and Computer Engineering, University of Maryland, College Park, College Park, MD 20742, USA
[b] Department of Psychology, New York University, New York, NY 10003, USA
[c] Boys Town National Research Hospital, Omaha, NE 68131, USA
[d] Department of Biology, University of Maryland, College Park, College Park, MD 20742, USA
[e] Institute for Systems Research, University of Maryland, College Park, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Accepted 27 October 2013
Available online 2 November 2013

Keywords:
Envelope entrainment
Auditory cortex
Auditory scene analysis
MEG

ABSTRACT

Speech recognition is robust to background noise. One underlying neural mechanism is that the auditory system segregates speech from the listening background and encodes it reliably. Such robust internal representation has been demonstrated in auditory cortex by neural activity entrained to the temporal envelope of speech. A paradox, however, then arises, as the spectro-temporal fine structure rather than the temporal envelope is known to be the major cue to segregate target speech from background noise. Does the reliable cortical entrainment in fact reflect a robust internal "synthesis" of the attended speech stream rather than direct tracking of the acoustic envelope? Here, we test this hypothesis by degrading the spectro-temporal fine structure while preserving the temporal envelope using vocoders. Magnetoencephalography (MEG) recordings reveal that cortical entrainment to vocoded speech is severely degraded by background noise, in contrast to the robust entrainment to natural speech. Furthermore, cortical entrainment in the delta-band (1–4 Hz) predicts the speech recognition score at the level of individual listeners. These results demonstrate that reliable cortical entrainment to speech relies on the spectro-temporal fine structure, and suggest that cortical entrainment to the speech envelope is not merely a representation of the speech envelope but a coherent representation of multiscale spectro-temporal features that are synchronized to the syllabic and phrasal rhythms of speech.

Introduction

Normal hearing listeners exhibit a surprising ability to understand speech in noisy acoustic environments, even in the absence of visual cues. A number of studies have suggested that the target speech and the listening background are separated in auditory cortex (Ding and Simon, 2012a; Zion Golumbic et al., 2013; Horton et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012; Power et al., 2012). In particular, when a listener attends to a speech stream, auditory cortical activity is reliably entrained to the temporal envelope of that stream, regardless of the listening background. This reliable neural representation of the speech envelope, i.e. slow temporal modulations below 16 Hz, is a key candidate mechanism underlying the reliable recognition of speech, since the temporal envelopes carry important cues for speech recognition (Shannon et al., 1995). It remains mysterious, however, how such reliable cortical entrainment to the speech envelope is achieved, since envelope is not an effective cue for segregation of speech from noise (Friesen et al., 2001).

Moreover, even the nature of cortical entrainment to the speech envelope is heavily debated, especially about whether it encodes the temporal envelope per se or instead other speech features that are correlated with the speech envelope (Obleser et al., 2012; Peelle et al., 2013). Many speech features, including pitch and spatial cues, are temporally coherent and correlated with the temporal envelope (Shamma et al., 2011). Therefore it has been proposed that the envelope entrainment in fact reflects a collective neural representation of multiple speech features that are synchronized to the syllabic and phrasal rhythm of speech (Ding and Simon, 2012a). Because of the collective nature of this representation, it has been suggested as a representation of speech as a whole auditory object.

If envelope entrainment indeed reflects an object-level, collective representation of speech features, reliable envelope entrainment in complex auditory scenes is likely to involve an analysis-by-synthesis process (Poeppel et al., 2008; Shamma et al., 2011; Shinn-Cunningham, 2008): In such a process, multiple features of a complex auditory scene are extracted subcortically in the analysis phase and then, based on speech segregation cues such as pitch, features belonging to the same speech stream are grouped into an auditory object in the synthesis phase. In

* Corresponding author at: Department of Biology, University of Maryland, College Park, College Park, MD 20742, USA. Fax: +1 301 314 9281.
E-mail addresses: gahding@gmail.com (N. Ding), monita.chatterjee@boystown.org (M. Chatterjee), jzsimon@umd.edu (J.Z. Simon).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.10.054

contrast, if envelope entrainment involves only direct neural processing of the envelope, its robustness to noise may arise from more basic processes such as contrast gain control (Ding and Simon, 2013; Rabinowitz et al., 2011).

In this study, we investigate whether noise-robust cortical entrainment to the speech envelope involves merely envelope processing or instead reflects an analysis-by-synthesis process that includes the processing of spectro-temporal fine structure and reflects envelope properties of the re-synthesized auditory object. Here, the spectro-temporal fine structure refers to the acoustic information not included in the broadband envelope of speech (<16 Hz), including, for example, the acoustic cues responsible for the pitch and formant structure of speech. We degrade the spectro-temporal fine structure of speech or speech–noise mixtures using noise vocoders and investigate whether vocoded stimuli are cortically represented differently from natural speech using MEG. If cortical entrainment only depends on the temporal envelope, it will not be affected by degradation of the spectro-temporal fine structure, even in a noisy listening environment. In contrast, if reliable cortical entrainment to speech requires an analysis-by-synthesis process that relies on the spectro-temporal fine structure, it should be severely degraded for vocoded speech.

Materials & methods

Subjects

Twelve normal hearing, right-handed (Oldfield, 1971) young adults (6 females), all between 19 and 32 years old (23 years old on average) participated in the experiment. Subjects were paid, and the experimental procedures were approved by the University of Maryland institutional review board. Written informed consent form was obtained before the experiment.

Stimuli

The stimuli were selected from of a narration of the story *Alice's Adventures in Wonderland* (Chapter One, http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-4/). The sound recording was low-pass filtered below 4 kHz and divided into twelve 50-second duration segments, after long speaker pauses (>300 ms) were shortened to 300 ms. All sound stimuli were presented binaurally (diotically). Six types of stimuli were created (2 noise levels × 3 vocoding conditions).

Background noise

Half of the speech segments (N = 6) were presented in a quiet listening environment (no noise added in), while the other half were mixed with spectrally matched stationary noise generated using a 12th-order linear predictive model estimated from the speech recording. The intensity ratio between speech and noise was fixed at 3 dB, measured by RMS.

Noise vocoding

Each stimulus is either noise vocoded (through a 4-channel or 8-channel vocoder) or unprocessed. The noise vocoder filters the stimulus, either speech in quiet or speech in noise, into 4 or 8 frequency channels between 123 and 3951 Hz using a 4th order Butterworth filter. All frequency channels are evenly distributed in the Cam scale (Glasberg and Moore, 1990; Qin and Oxenham, 2003). In each frequency band, the envelope of the stimulus, either speech or a speech–noise mixture, is extracted by taking the absolute value of the Hilbert Transform, low-pass filtering below 160 Hz using a 4th order Butterworth filter, and then half-wave rectifying the filtered signal. The extracted envelope is used to modulate white noise filtered into the same frequency band from which the envelope was derived. The envelope-modulated-noises are then summed over frequency bands to create the noise-vocoded stimulus. The RMS intensity of the noise-vocoded stimulus is adjusted to match that of the unprocessed stimulus.

Stimulus characterization

The auditory spectrogram of the stimulus was calculated using a sub-cortical auditory model (Yang et al., 1992) and expressed in a logarithmic amplitude scale. The frequency by time auditory spectrogram has 128 logarithmically spaced frequency channels and a 10-ms resolution in time. The broadband temporal envelope of the stimulus was extracted by summing the auditory spectrogram over frequency.

Procedure

The stimuli were presented in two orders, each to half of the subjects. In either order, the story continued naturally between stimuli and was repeated twice after the first presentation (3 trials in total). In the progressive order, the first two speech segments were natural speech presented in quiet, followed by 8-band vocoded speech in quiet and then 4-band vocoded speech in quiet. Then, natural speech in noise, 8-band vocoded speech in noise, and 4-band vocoded speech in noise were presented sequentially. To control for the effect of presentation order, we also created a random order condition, in which each acoustic manipulation (e.g. vocoding or background noise) was assigned randomly to a segment for each subject. The two presentation orders did not result in any difference in speech intelligibility or neural synchronization spectrum and were therefore not distinguished in the following analysis.

The subjects were asked to listen to the story and keep their eyes closed. Questions about the story were asked after each 50-second duration stimulus to ensure subjects' attention. The subjects were also asked to rate the percent of words they understood after the first presentation of each stimulus (on a scale of 0% (not intelligible) to 100% (fully intelligible)). The grand averaged subjectively rated intelligibility is highly correlated with the grand averaged percent of questions correctly answered ($R = 0.96$). Before the experiment, the subjects listened to 100 repetitions of a 500-Hz tone and the responses were used to extract the M100 response, a salient MEG response localized to auditory cortex (Lütkenhöner and Steinsträter, 1998).

The magnetic field generated by cortical activity was recorded using a 157-channel whole-head MEG system (KIT, Kanazawa, Japan). The signal was sampled at 1 kHz and was filtered by a 200-Hz lowpass filter and a notch filter at 60 Hz online. Environmental noise was further removed using TS-PCA (de Cheveigné and Simon, 2007). The whole-head MEG recording was used for analysis unless otherwise specified. When the two hemispheres were analyzed separately, hemisphere-specific responses were extracted using 55 sensors located above each hemisphere. More details of the recording procedure are as described in Ding and Simon (2012a).

Inter-trial correlation analysis

The phase locking of a neural response was evaluated by the inter-trial correlation of the neural response in narrow frequency bands (2-Hz wide) (Ding and Simon, 2013; Zion Golumbic et al., 2013). The inter-trial correlation is the Pearson correlation coefficient between two trials of the neural responses to the same stimulus (averaged over all possible combinations of two trials). It measures the reliability of the neural response when the same stimulus repeats, and reflects the strength of phase-locked neural activity. The major phase-locked component of the MEG response was extracted using a blind source separation method, Denoising Source Separation (DSS) (de Cheveigné and Simon, 2008). The first DSS component was used for this analysis.