# Auditory Object Based Cortical Representation during Speech Segregation

Nai Ding[1], Jonathan Z. Simon[1,2]

[1]Department of Electrical & Computer Engineering, [2]Department of Biology
University of Maryland College Park

Computational Sensorimotor Systems Lab

## Cocktail Party Problem & Auditory Objects

How is a complex auditory scene consisting of multiple auditory objects/ streams represented in human auditory cortex? What is the neural correlate of the segregation of the auditory scene into auditory objects or streams?

Is each auditory object represented by a distinct neural code? If so, when and where does this auditory scene analysis process occur? Is it robust against bottom-up acoustic saliency (e.g. the loudness of each object) and acoustic/perceptual similarity between the auditory objects?

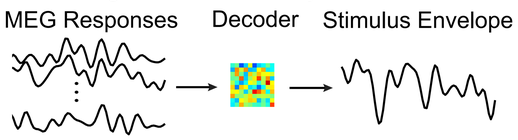
We address these questions by recording from human subjects who selectively listen to simultaneous natural auditory narrations, using the noninvasive physiological method of magnetoencephalography (MEG).

## Methods

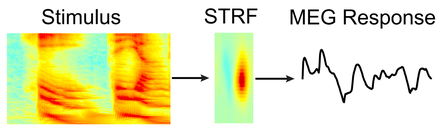
**Stimulus & Procedures:** Two speakers were mixed into a single acoustic channel presented diotically. Listeners were instructed to attend to one speaker and answer comprehension questions after every 1 minute section (2 sections per condition). The listeners switched attention to the other speaker when the same stimulus repeated. All was repeated 3 times, resulting in 3 trials in each attentional condition.
In the ***main experiment*** *(N=11)*, the two speakers were of different gender and were mixed with equal intensity. In the *varying* ***target-to-masker ratio (TMR) experiment*** *(N=6)*, the same two speakers were used, but the intensity of one speaker was varied. In the ***same gender experiment*** *(N=3)*, both speakers were female, and the listeners were given a training section.

**MEG:** 157-channel, whole-head MEG. 1 kHz sampling rate, downsampled to 40 Hz. The neural source of MEG activity is localized using a equivalent current dipole model, one per hemisphere.

**Neural Reconstruction:** We reconstructed the envelope of speech using a linear decoder that integrates MEG activity over time and sensors.




**STRF:** The spectro-temporal response function (STRF) models the neural response evoked by a unit power increase in the stimulus (by frequency).




STRFs were estimated by boosting. Since two speakers were presented, the full model is: Response = $\text{Speaker}_1 * \text{STRF}_1 + \text{Speaker}_2 * \text{STRF}_2$.

## Conclusions:

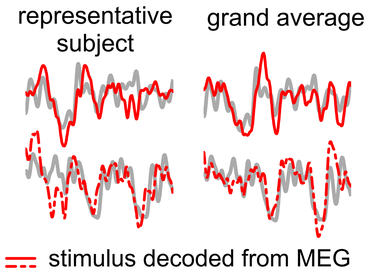
**A *complex auditory scene* containing two simultaneous speakers mixed into a single acoustic channel is behaviorally, and neurally *parsed into auditory objects* in auditory cortex.**

**Attention routes the neural representation of auditory object, i.e. speaker, into distinct spatial-temporal neural networks. The attended object is more strongly represented in posterior association auditory cortex at a latency of ~100 ms.**

**The separate spatial-temporal representations of each speaker can be decoded non-invasively using MEG: We can tell who are you listening to!**

## Spatial-temporally Distinct Representations of Attended and Unattended Speakers *(Decoding Individual Speakers)*

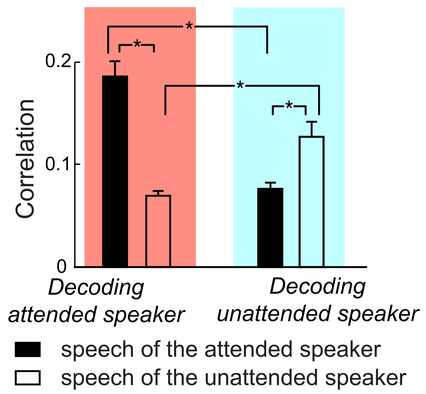
### Neural representation of the attended speaker




The temporal envelope of the attended speaker can be reconstructed from the cortical response to the speech mixture.

Different envelopes (in the upper and lower panels) are decoded from neural responses to the ***same stimulus***, depending on whether the listener ***attends to one or the other speaker*** in the speech mixture.

### Decoding the neural representation of each speaker, either attended or unattended




The grand averaged correlation between the reconstructed envelope and actual envelope of the stimulus is shown in the left.

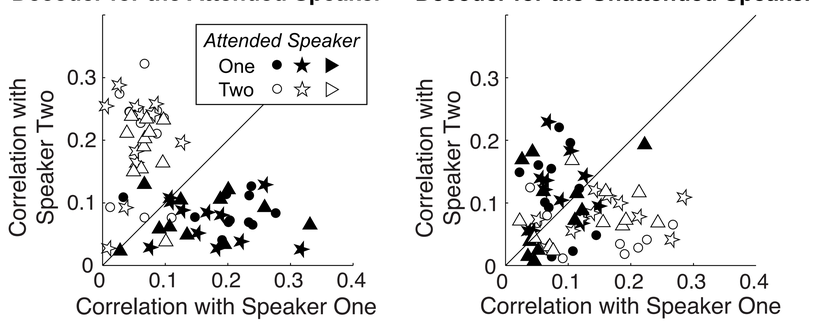
Two decoders (spatial-temporal weighting matrices) are designed to reconstruct the attended and unattended speaker respectively.

These two decoders integrate spatial-temporal neural activity differentially, and extract the attended and unattended speakers respectively, when applied to the same neural recording.

Therefore, the two speakers are represented by distinct spatial-temporal patterns of neural activity, each of which can be noninvasively extracted.

### Decoding at a single subject and single trial level

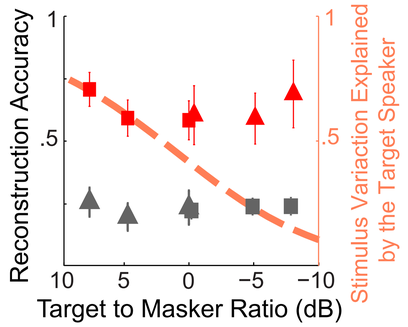
**Decoder for the Attended Speaker** **Decoder for the Unattended Speaker**




**The two speakers are represented by distinct spatial-temporal patterns of cortical activity during active listening. From these separate spatial-temporal patterns, the temporal envelopes of the two speakers can be separately reconstructed neurally.**

**Acknowledgement:** work supported by NIH R01 DC-005660
**Reference:** Ding & Simon, J. Neurophys. 2012

## Speaker Intensity Invariant Representation




Dashes: $R^2$ between stimulus envelope and the target speaker envelope.

To test how robust the speaker-specific representation is, the intensity ratio between the two speakers is varied between -8 dB and 8 dB.

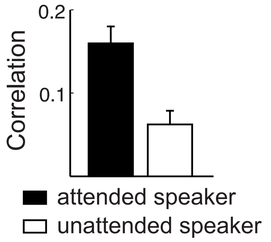
The envelope of the attended speaker can be reliably decoded at all test TMR, and the decoding performance is TMR independent.

Subjectively rated intelligibility decreases with TMR, though not the percent of questions correctly answered (~70%).

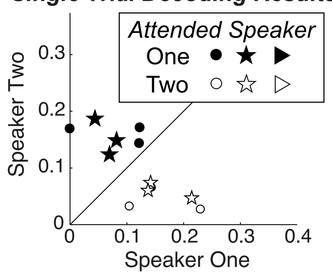
The TMR-independent responses suggest that a speaker-specific neural adaptation to sound intensity compensates variations in speaker intensity.

## Segregation of Speakers of the Same Gender

***Decoding the Attended Speaker***




***Single Trial Decoding Results***



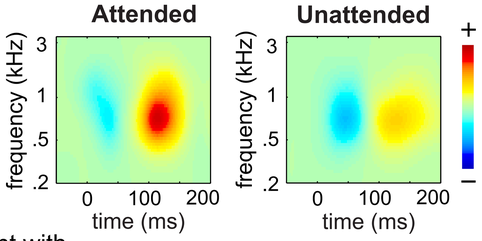

The attended and unattended speakers are differentially represented even when the two speakers are of the same gender.

(74% questions correctly answered)

**The speaker-specific cortical representations are robustly formed as long as the two speakers are perceptually separable.**

## Spatial-temporally Distinct Representations of Attended and Unattended Speakers *(Forward Model)*

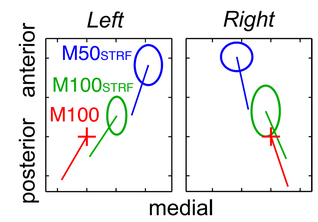
The $\text{M100}_{\text{STRF}}$ is significantly modulated by attention while the shorter latency response component $\text{M50}_{\text{STRF}}$ is not.




The neural source of the $\text{M100}_{\text{STRF}}$ is roughly consistent with that of the M100 evoked by a tone pip, which is commonly localized to planum temporale (PT).

The neural source of the $\text{M50}_{\text{STRF}}$ is more anterior than the neural sources of the $\text{M100}_{\text{STRF}}$ and M100, and therefore is probably more consistent with core auditory cortex.

**Neural Source Locations**




**Cortical representations are transformed from feature-based to object-based up the cortical hierarchy: from shorter latency activity in core auditory cortex to longer latency activity in posterior association auditory cortex.**