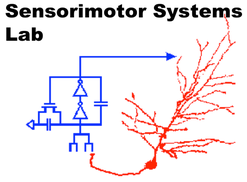
Neural representations of Background speakers at the Cocktail Party

Krishna Puvvada[1], Jonathan Z Simon[1,2,3]

[1]Dept of Electrical & Computer Engineering, [2]Dept of Biology, [3]Institute for Systems Research

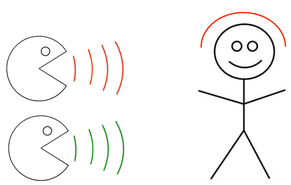
University of Maryland College Park

Computational Sensorimotor Systems Lab

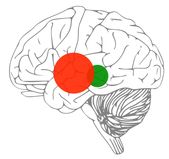


How are unattended background speakers represented neurally?

Stimulus world

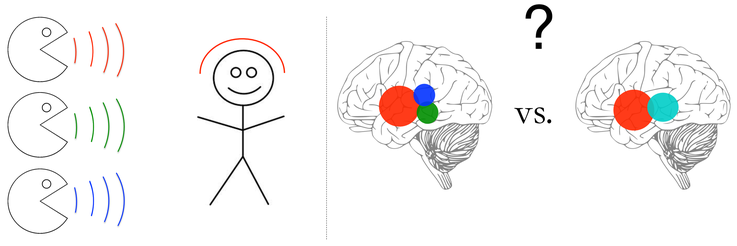


Neural world



With two auditory sources to choose from, attended (foreground) and unattended (background) have distinct neural representations (Ding & Simon, 2012).

But when there are more than two sources to choose from, are the background sources' representations *distinct* or *merged*?



Does selective attention work by selecting pre-formed auditory streams out of a complex auditory scene? Or by identifying only the attended (foreground) stream? Or a different mechanism?

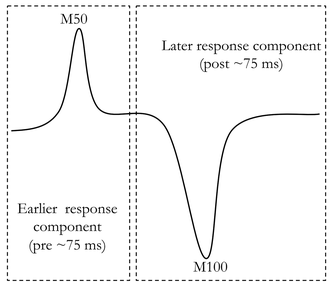
Experimental Design

- 3 co-located speakers simultaneously narrating separate stories.
- 220 s stimulus duration, 3 attention conditions x 3 repetitions
- $N = 7$ Subjects
- In each trial the subject counts the number of times a given keyword is heard in the attended story.
- At the end of 3 trials, the subject reports a summary of story.
- Magnetoencephalography (MEG) recordings, 157 channels.
- 1 kHz sampling , Time-shifted PCA based de-noising.
- Spatial filtering used to reduce 157 channels to fewer, more reliable virtual channels.

Analysis & Results

Stimulus reconstruction

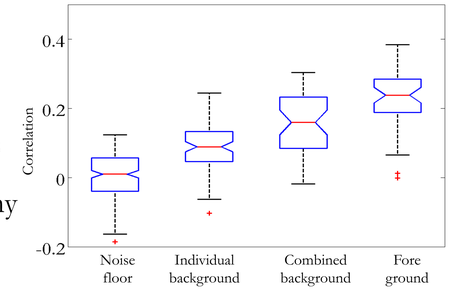
- Temporal envelope of stimulus is reconstructed from cortical responses using optimum linear filters.
- Correlation between reconstructed and actual envelope is used as metric as how faithfully the foreground or background is represented in cortical responses.
- Reconstruction is based on integrating neural responses over a temporal window.
- Different latency ranges in the temporal integration window represent different neural areas, with different processing specializations and representations.



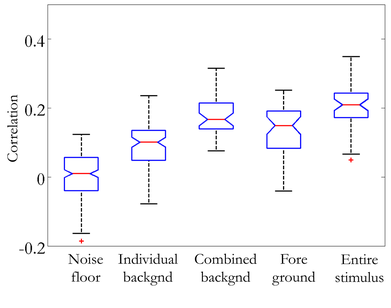
Stereotypical neural response

Stimulus reconstruction using only late responses

- *Foreground* is most accurately reconstructed of all.
- *Combined background* is more accurately reconstructed than any *individual background*.



Stimulus reconstruction using only early responses

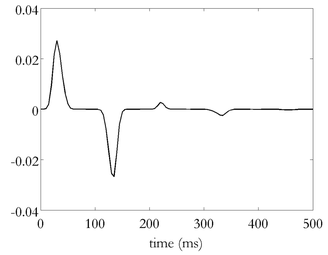


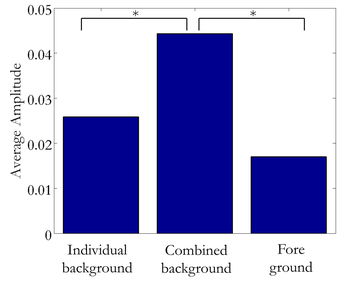
- *Physical stimulus* is most accurately reconstructed of all.
- *Combined background* is more accurately reconstructed than *any individual background*.
- *Combined background* is more accurately reconstructed than *foreground*.

Temporal Response function (TRF) analysis

- Neural response is modeled as linear (filtered) function of temporal envelope of speech stimulus.
- Optimum filters are derived using Boosting algorithm with 10-fold cross validation.

A representative TRF showing M50-like and M100-like responses for speech stimuli.



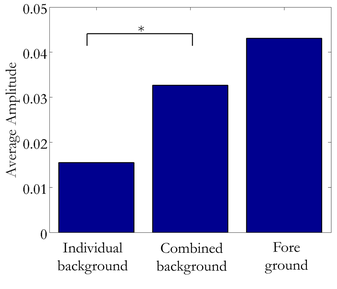


Average amplitudes for M50-like response

- Higher amplitudes observed for combined background than individual background or foreground.

Average amplitudes for M100-like response

- Higher amplitudes observed for combined background than any individual background.
- Highest amplitudes observed for Foreground but not significant.



Discussion

In a complex auditory scene with more than two sources:

- Longer latency cortical areas represent the auditory scene as a single foreground source but a merged background (everything other than foreground), not as having distinct neural representations for each source.
- Earlier latency cortical areas are consistent with representing the entire acoustic scene, without distinct neural representations for each source.

References: Ding N. and J. Z. Simon, (2012) *The Emergence of Neural Encoding of Auditory Objects While Listening to Competing Speakers*, PNAS, 109(29), 11854-11859.

Acknowledgements: Funding from NIH R01 DC 008342