Robust Cortical Encoding of Speech in Human Auditory Cortex

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

[1]Department of Electrical & Computer Engineering, [2]Department of Biology

University of Maryland College Park

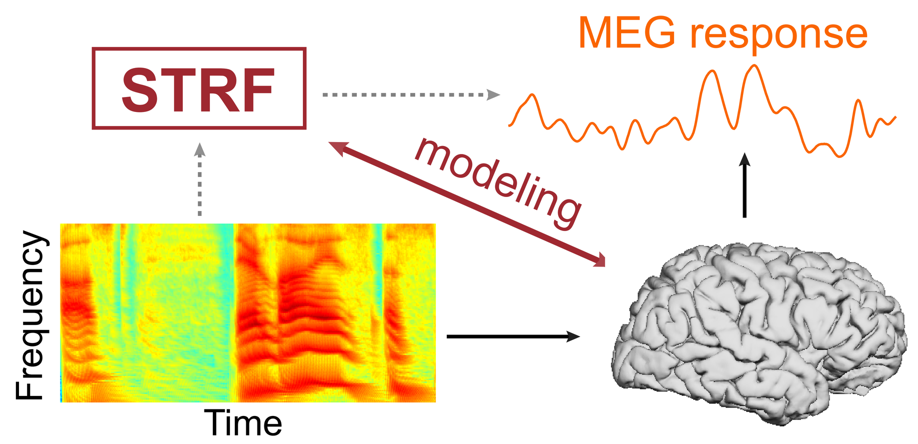
Computational Sensorimotor Systems Lab

Introduction

✓ **How are the spectro-temporal features of speech encoded in auditory cortex in realistic listening environments?**

✓ **How is the neural coding scheme modulated by top-down attention and the bottom-up saliency of speech?**

These two questions are addressed by recording the *magnetoencephalography* (MEG) response from human subjects actively listening to spoken narratives. MEG is an non-invasive neural recording tool, with millisecond level time resolution.



The MEG response to natural spoken narratives is analyzed using the *Spectro-Temporal Response Function* (STRF), a counterpart of the spectro-temporal receptive field in single unit neurophysiology.

Experimental Procedures

➢ **Stimulus & Procedure**

Monaural Cue based (Diotic) Speech Segregation

- Two 2 minute long spoken narratives, one read by a male and the other by a female, were mixed and played simultaneously to both ears.
- The stimulus was played 6 times. The subjects focused on one speaker at a time and switched focus after every repetition.
- After every minute, the subjects were asked a comprehension question. 70% of the questions were correctly answered.
- In a separate session, each spoken narrative was played individually 4 times.
- 11 subjects participated in the experiment.

Binaural Cue based (Dichotic) Speech Segregation

- Two 2 minute long spoken narratives, read by the same male speaker, were played simultaneously to each ear.
- The stimulus was played 6 times. The subjects focused on one ear at a time and switched focus after every repetition.
- After every minute, the subjects were asked a comprehension question. 90% of the questions were correctly answered.
- In a separate session, each spoken narrative was played monaurally 4 times.
- 10 subjects participated in the experiment.

➢ **MEG Recording and Processing**

- 157 channel whole-head MEG system, sampled at 1 kHz, with a 60 Hz notch filter.
- The STRF is estimated using boosting with 10-fold cross validation, based on a sub-cortical spectro-temporal representation of speech.
- The neural source is localized by a equivalent current dipole model, containing one dipole in each hemisphere.

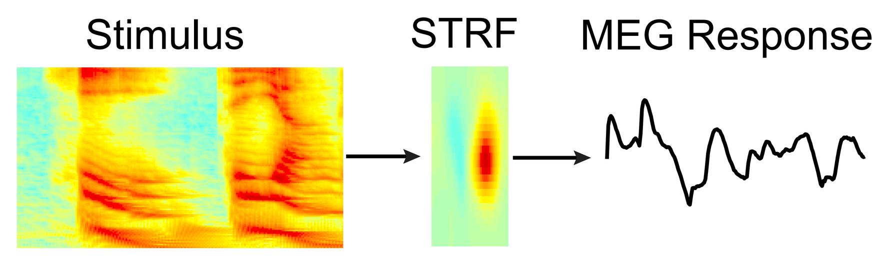
Summary

✦ The large-scale synchronized neural activity in human auditory cortex is phase locked to the slow temporal modulations of speech and can be modeled by the STRF.

✦ Speech segregation occurs in auditory cortex within 150 ms and each speech stream is processed differently under the modulation of top-down attention.

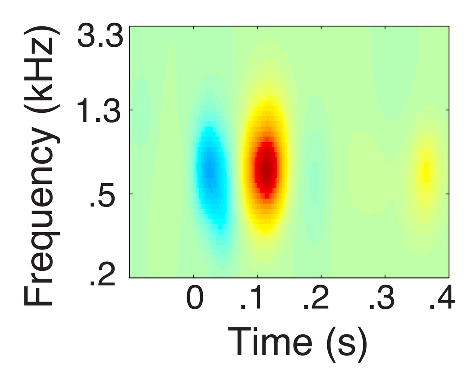
✦ Auditory processing of sounds from one ear is delayed and suppressed by sounds from the other ear.

Spectro-Temporal Response Function

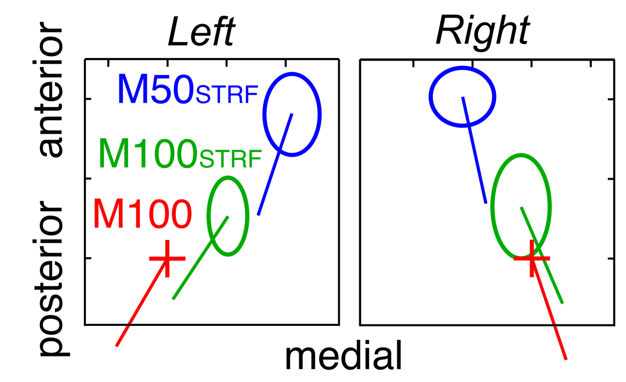


The neural encoding of spectro-temporal features of continuous speech in MEG is modeled by the STRF.

STRF for speech presented binaurally



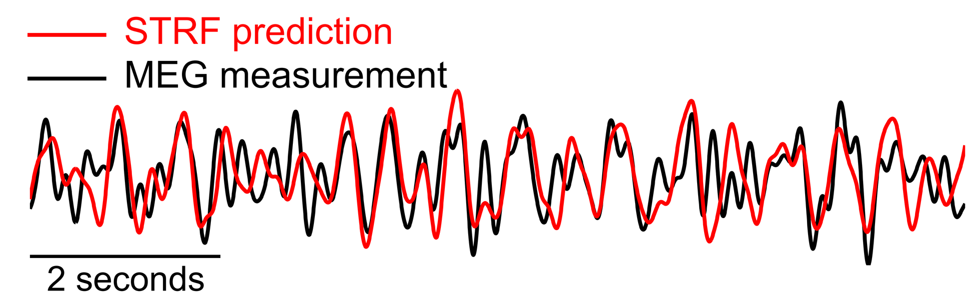
Neural Source Localization



The STRF have two major peaks, named $M50_{STRF}$ and $M100_{STRF}$ based on their response latency. The two peaks have opposite polarity and a dipole analysis shows they are generated from different neural sources.

Response Prediction

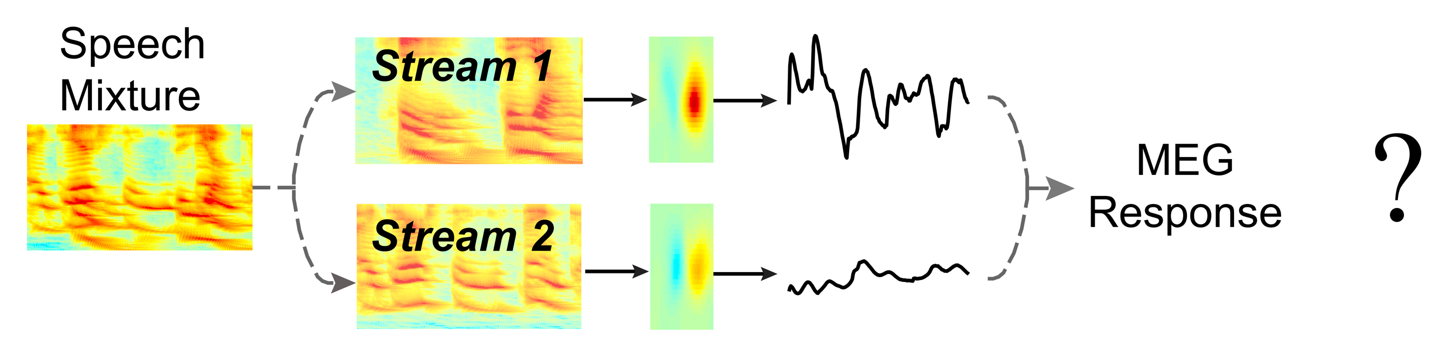
As a functional model of cortical auditory processing, the STRF can *predict* the cortical response based on stimulus.



The spatially coherent activity in human auditory cortex is synchronized to the slow modulations of speech.

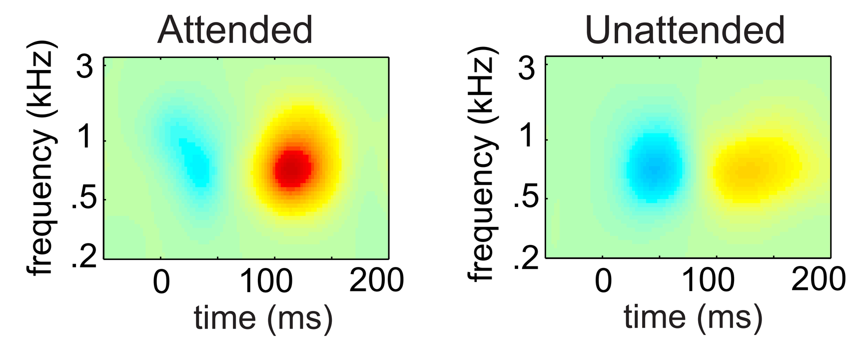
Monaural Speech Segregation

Hypothesis: A complex auditory scene (e.g. a mixture of speech) is decomposed into auditory streams, each being encoded differently under the control of attention.



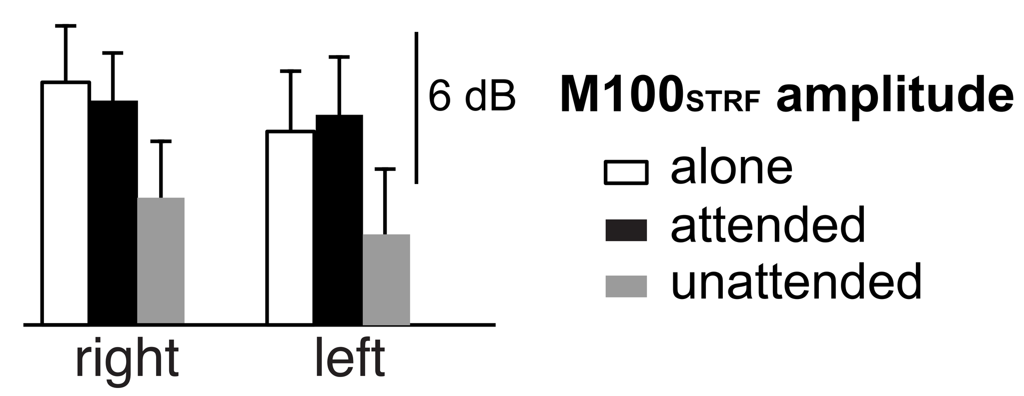
Alternative hypothesis: A complex auditory scene is not decomposed into auditory streams or the selective processing of each stream is not reflected in the large-scale stimulus-synchronized activity.

Results: The STRF derived for each speaker is strongly modulated by attention.



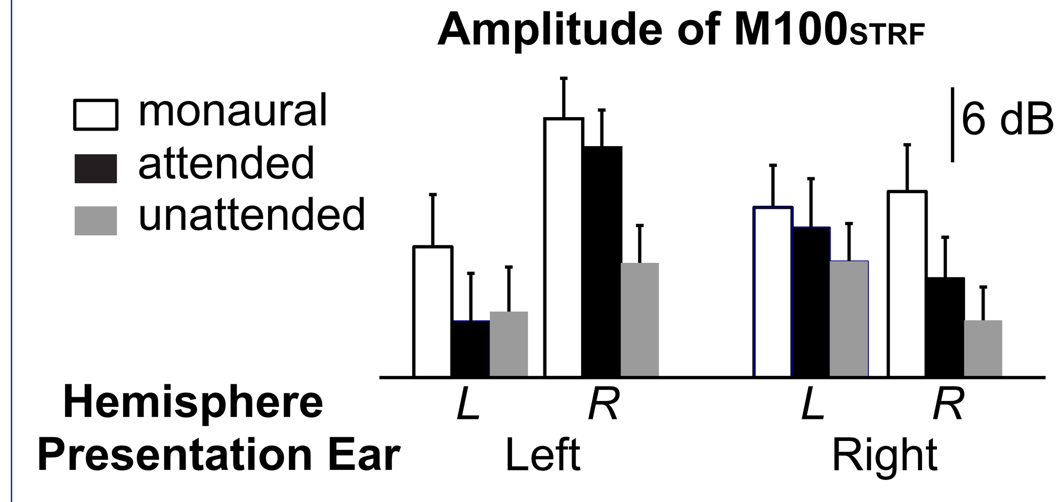
The attentional modulation of the neural representation of individual speakers indicates the physically mixed speech signals are neurally segregated.

The $M100_{STRF}$ component is enhanced when a speaker is attended to, in both left and right hemispheres.

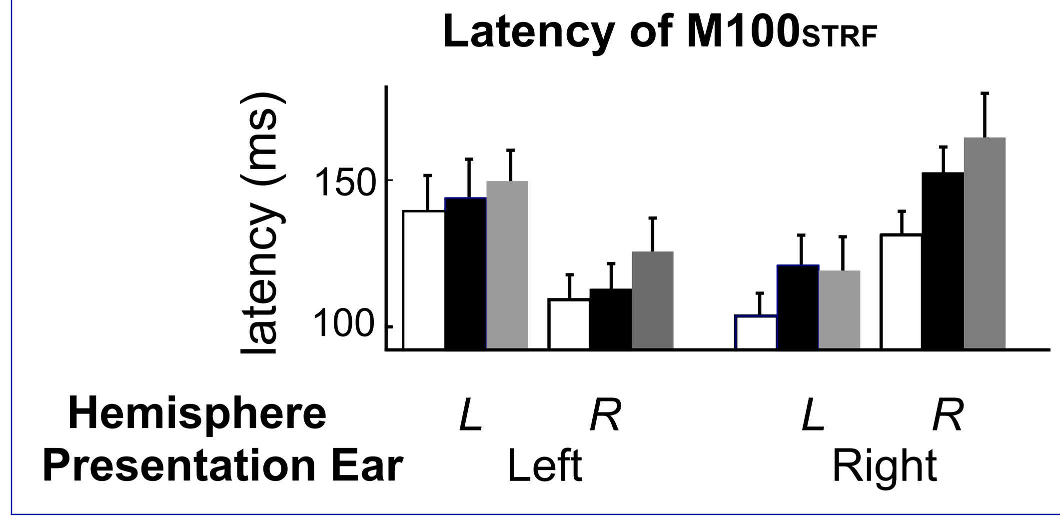


Speech segregation occurs pre-lingually, in the auditory system. The response to the attended speech dominates cortical activity.

Dichotic Speech Segregation

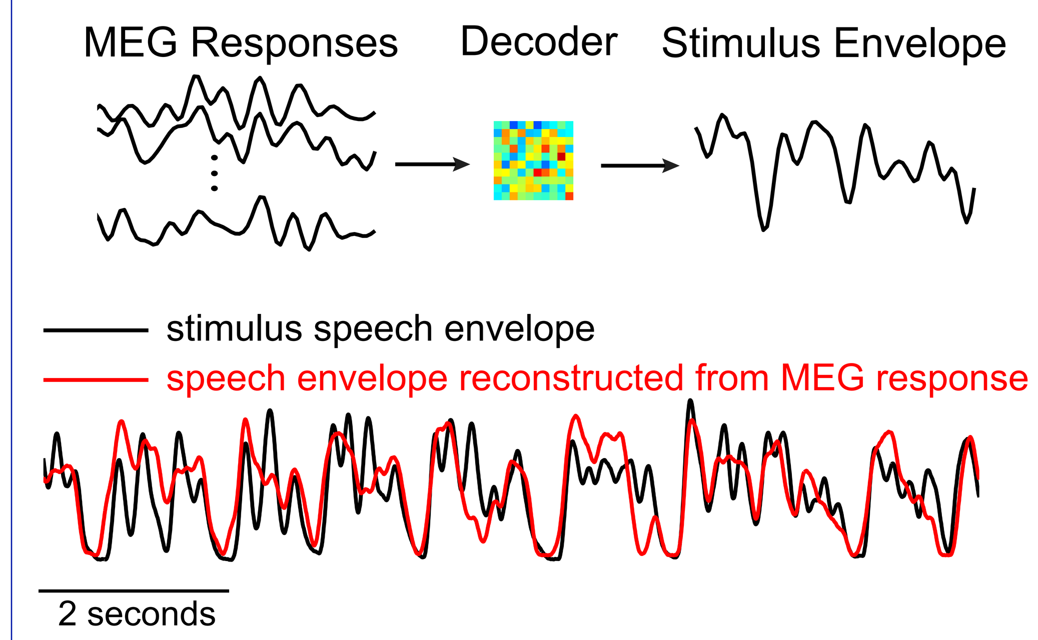


The cortical response is (1) stronger in the right hemisphere, especially for monaural stimuli, (2) suppressed by sounds in the other ear, (3) enhanced when the ear is attended to, especially in the contralateral hemisphere.



The response is also (1) slower in the contralateral hemisphere, (2) delayed by sounds in the other ear. (3) slower for unattended sounds.

Decoding Speech & Attentional Focus

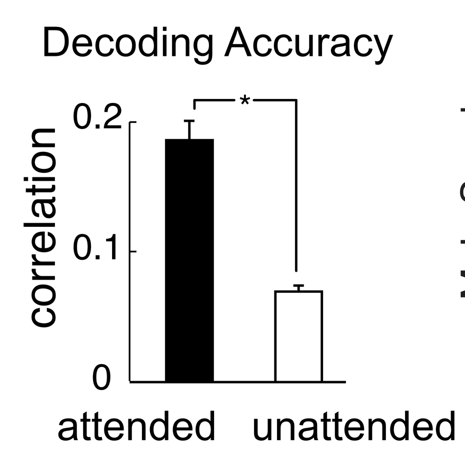


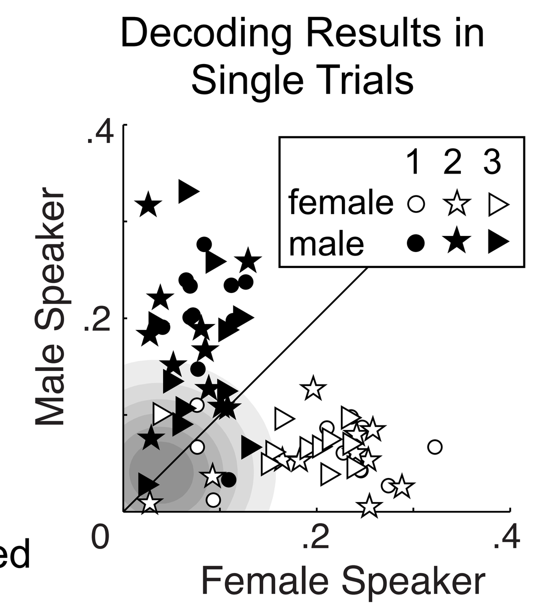
Stimulus Reconstruction

The inverse of the STRF can *decode* speech information from the MEG response. Since the STRF has broad frequency tuning, the decoding analysis focuses on the temporal envelope of the stimulus.

Decoding Attentional Focus

During speech segregation, a decoder is designed to reconstruct the envelope of the attended speech. The success of this decoder allows us to decode the attentional focus of the listener.





The attentional focus of a listener can be decoded by reconstructing the envelope of the attended speech.

Reference

S.V. David, N. Mesgarani & S.A. Shamma, Network: *Comput. Neural Syst.* 18 (2007)

N. Ding & J.Z. Simon, *J. Neurophysiol.* (in press)

Acknowledgement

Supported has been provided by R01-DC-008342. We thank Stephen David, David Poeppel, Mary Howard, Monita Chatterjee and Shihab Shamma for discussions, and Max Ehrmann for technical support.