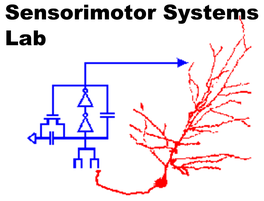
Cortical Neural Coding of Speech in Simple and Complex Auditory Scenes

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

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

University of Maryland College Park

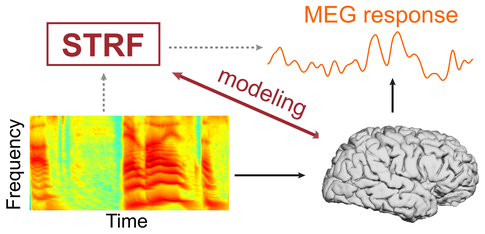
Computational Sensorimotor Systems Lab



Introduction

1. **What is the neural code of spectro-temporal features of continuous speech in human auditory cortex?**
2. **How does the brain segregate and maintain attention on a speech target in the presence of a concurrent speech masker?**

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

Speech mixture

• Two 2 minute long spoken narratives, from *the Legend of Sleepy Hollow*, by Washington Irving, were played simultaneously, one in each ear **(dichotic stimuli)**.
• 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 subject were asked a question about the comprehension of the story attended to. 90% of the questions were correctly answered.

Monaural speech

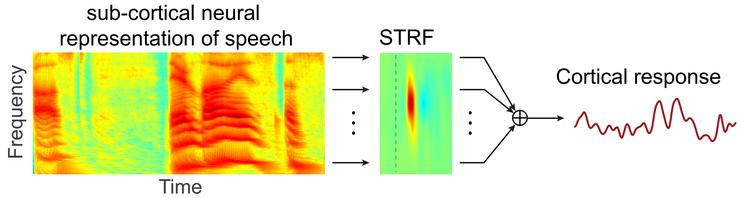
• In a separate session, each spoken narrative was played monaurally 4 times.

➢ MEG Recording and Processing

• 157 channel whole-head MEG system, sampled at 1 kHz, with a 60 Hz notch filter.
• Denoising Source Separation (DSS) assisted equivalent current dipole fitting localizes the neural source bilaterally to the *superior temporal gyrus*.
• The moment of the equivalent current dipole in each auditory cortex is reconstructed using the generalized least squares method, as a function of time.

➢ Spectro-Temporal Response Function (STRF)

• The STRF is estimated using boosting with 10-fold cross validation, based on a sub-cortical spectro-temporal representation of speech.



$$r(t) = \sum_f \sum_\tau STRF(f,\tau) S(f, t-\tau) + \varepsilon(t)$$

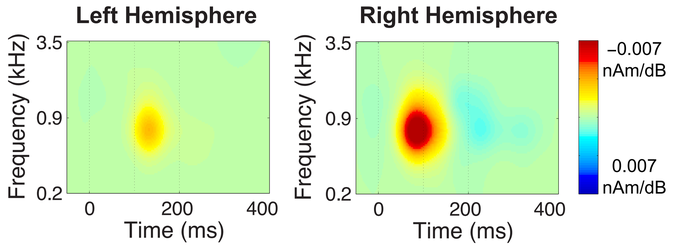
Reference

A. de Cheveigné & J.Z. Simon, *J. Neurosci. Methods* (2008)
S.V. David, N. Mesgarani & S.A. Shamma, Network: *Comput. Neural Syst.* **18** (2007)
X. Yang, K. Wang & S.A. Shamma, *IEEE Trans. Info. Theory* (1992)

Acknowledgements

Supported has been provided by R01-DC-008342.We thank Stephen V. David, David Poeppel, Mary F. Howard, Jonathan B. Fritz and Shihab A. Shamma for valuable discussions and Max Ehrmann for excellent technical support.

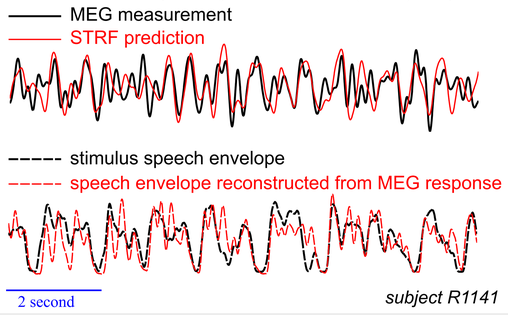
Spectro-temporal response function for speech



The STRF for the speech signal **monaurally** presented to the left ear. (grand average)

The STRF indicates the neural response is following the spectro-temporal modulations of speech at latency of about 100 ms, where the contralateral response is enhanced.

The peak at about 100 ms is called the **M100-like response**.The polarity of this peak is the same as the polarity of the M100 (N1m) response to tone pip.

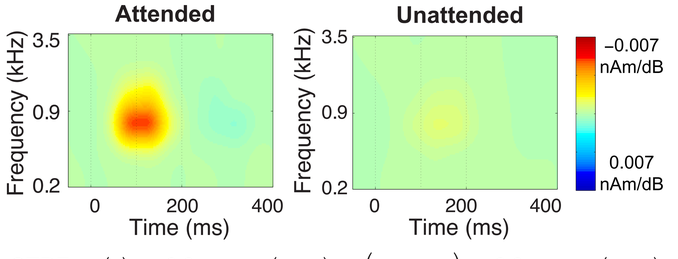


As a functional model of cortical auditory processing, the STRF can *predict* the cortical neural response based on the spectro-temporal representation of speech.

The inverse of the STRF can be used to *decode* the speech envelope information from the MEG response.

The neural activity in human auditory cortex precisely encodes spectro-temporal modulations of speech.

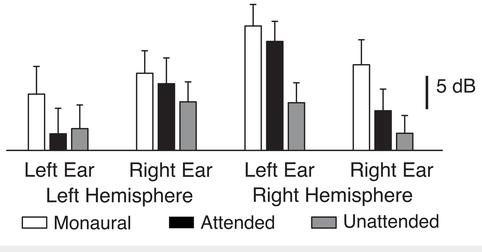
Spatial cue based speech segregation



Right hemisphere STRF to speech from the left ear, when listening to **dichotic speech mixture** (grand average).

STRF model: $$r(t) = \iint_{f\,\tau} STRF_A(f,\tau) S_A(f, t-\tau) + \iint_{f\,\tau} STRF_U(f,\tau) S_U(f, t-\tau) + \varepsilon(t)$$

Amplitude of the M100-like Response

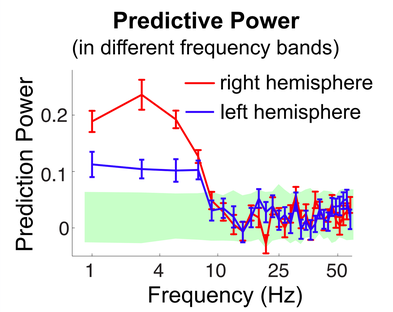


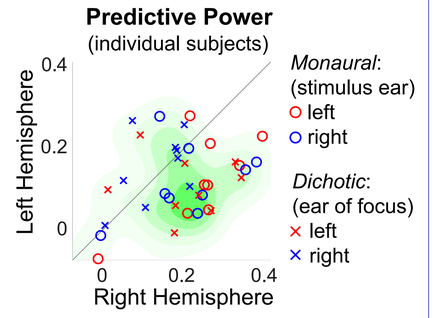
• substantial attentional gain control
• stronger response in the right hemisphere.
• The response is stronger for monaural stimuli than dichotic stimuli.
• Each hemisphere is more responsive to the contralateral stimulus.

Purely spatial cue based speech segregation occurs in human auditory cortex, within 100 ms, reflected by the different attentional gains of the target and masker speech.

Properties of MEG STRF

The predictive power is the correlation between STRF model prediction and real MEG measurement.





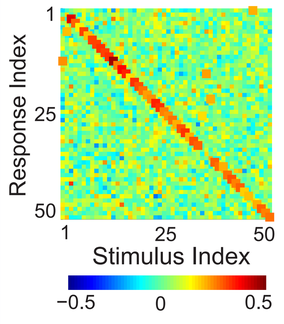
Slow temporal modulations (<10 Hz) are precisely encoded in the MEG response, especially in the right hemisphere.

Decoding the slow temporal modulations of speech from MEG

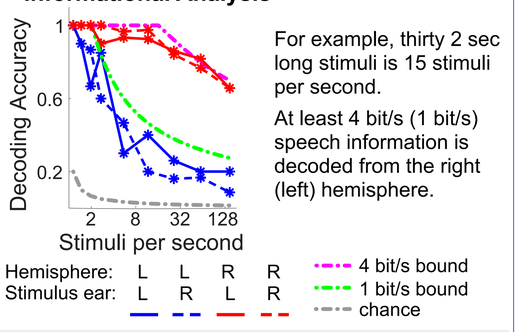
Decoding fifty stimuli (1.2 second long each)

The decoding performance is characterized using Fano's inequality: $H(P_e) + P_e \log(N-1) > \log(N) - I(s,r)$

Correlation



Informational Analysis



For example, thirty 2 sec long stimuli is 15 stimuli per second.

At least 4 bit/s (1 bit/s) speech information is decoded from the right (left) hemisphere.

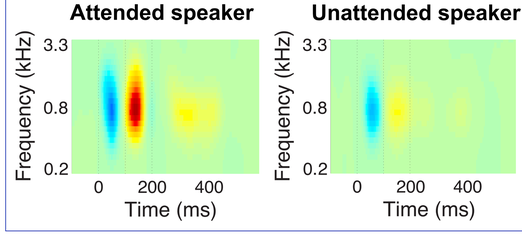
At least 4 bit/s information about the temporal structure of speech can be decoded from MEG responses.

Summary & Future work

• A faithful and robust neural coding of the temporal features of speech is observed using noninvasive neural recording methods. This neural code is strongly modulated by attention.

• A paradigm is provided to study the neural processing of continuous speech in various natural auditory scenes.

To examine whether speech segregation as a general phenomenon occurs in human auditory cortex, we investigate the neural processing underlying ***speaker feature cue based speech segregation***.



Stimulus: a diotic speech mixture consisting of a male and a female voice

Preliminary results: The strong attentional gain effect suggests **collocated target and masker speech are represented separately in auditory cortex**.