Real-Time Tracking of the Selective Auditory Attention from M/EEG via Bayesian Filtering

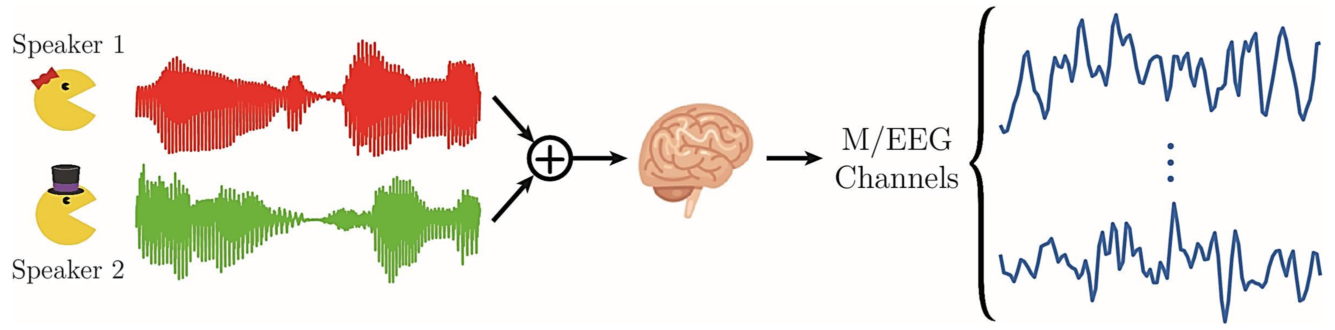
Sina Miran[1], Sahar Akram[2], Jonathan Z. Simon[1], Tao Zhang[2], Behtash Babadi[1]

[1] University of Maryland College Park (UMD) [2] Starkey Hearing Technologies

Problem Overview

Cocktail Party Effect: the ability to identify and track a target speaker amid a cacophony of acoustic interference [1]

→ humans can rapidly switch attention across multiple speakers



Simplified Computational Problem: In a *dual-speaker* environment, can we decode the attentional state in *real-time* from the *clean speech signals* of the two speakers and the multi-channel *magnetoencephalography (MEG) or electroencephalography (EEG) measurements* of the listener's brain?

attention decoding in *real-time* from *non-invasive neuroimaging data*

→ applications in Brain-Computer Interface (BCI) systems and smart hearing aids

Existing methods:

linear decoding models → linearly map M/EEG data to stimulus

linear encoding models → linearly map stimulus to a neural response from M/EEG

Examples:

- **reverse-correlation or stimulus reconstruction in decoding models (EEG) [2]:**
  1. train a decoder on the attended speech using training data
  2. use the *attended* decoder on the EEG data to reconstruct a stimulus
  3. speech signal which has the highest correlation with the reconstructed stimulus considered as the *attended* speech
- **important stimulus time lags in encoding models (MEG) [3][4]:**
  1. estimate the encoding coefficients for each speaker, i.e., Temporal Response Function (TRF), in a test trial
  2. the attended speaker has a larger M100 (the peak close to 100ms delay)

Shortcomings for Real-Time Attention Decoding:

- temporal resolution ~ tens of seconds (too slow given the dynamics of auditory processing)
- operation in the batch-mode regime (requiring the entire data from one or multiple trials at once for processing)
- need *large training datasets* to estimate the *attended* encoder/decoder reliably for use in test trials (not available in real-time applications)

References

[1] Cherry, E. Colin. "Some experiments on the recognition of speech, with one and with two ears." *The Journal of the acoustical society of America* 25.5 (1953): 975-979.

[2] O'sullivan, James A., et al. "Attentional selection in a cocktail party environment can be decoded from single-trial EEG." *Cerebral Cortex* 25.7 (2014): 1697-1706.

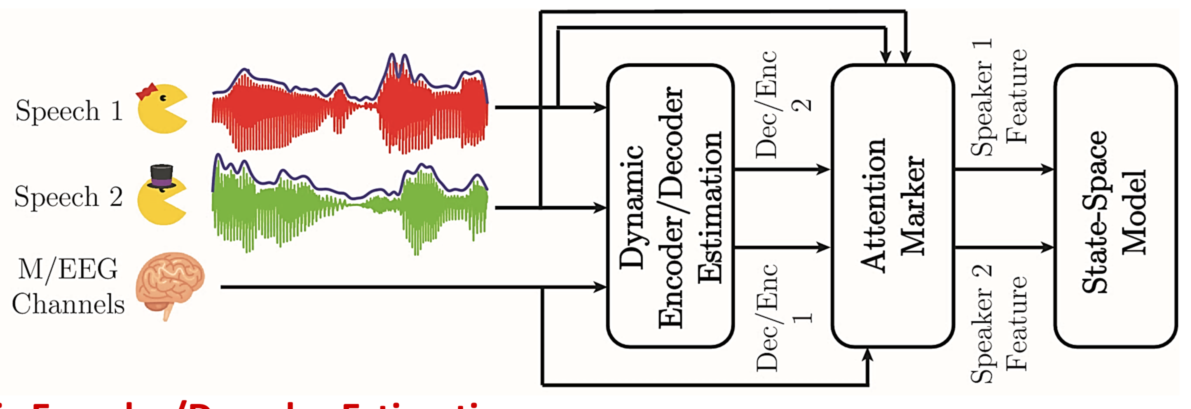
[3] Ding, Nai, and Jonathan Z. Simon. "Emergence of neural encoding of auditory objects while listening to competing speakers." *Proceedings of the National Academy of Sciences* 109.29 (2012): 11854-11859.

[4] Akram, Sahar, Jonathan Z. Simon, and Behtash Babadi. "Dynamic Estimation of the Auditory Temporal Response Function From MEG in Competing-Speaker Environments." *IEEE Transactions on Biomedical Engineering* 64.8 (2017): 1896-1905.

[5] Akram, Sahar, et al. "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling." *NeuroImage* 124 (2016): 906-917.

Proposed Framework

our proposed framework for *real-time* attention decoding includes three modules:



Dynamic Encoder/Decoder Estimation:

- consider K consecutive non-overlapping windows of length W samples
- update the encoder/decoder estimates $\hat{\boldsymbol{\theta}}_k$ for *each speaker* in every window:

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}}{\arg\min} \underbrace{\sum_{j=1}^{k} \lambda^{k-j}}_{\text{forgetting factor}} \big\| \boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\theta} \big\|_2^2 + \underbrace{\gamma \|\boldsymbol{\theta}\|_1}_{\ell_1 \text{ regularization penalty}}, \qquad k = 1,2,\dots,K$$

$\boldsymbol{y}_j$: speech envelopes (dec.) / neural response (enc.)

$\boldsymbol{X}_j$: M/EEG covariates (dec.) / envelope covariates (enc.)

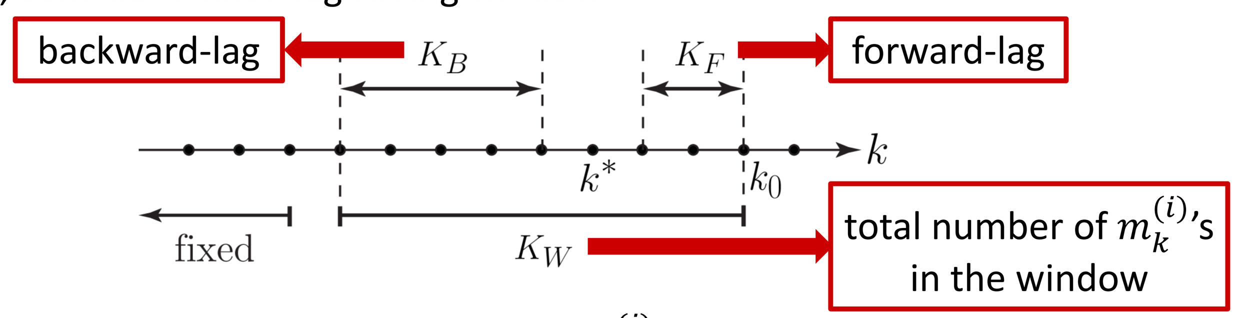
- γ chosen by cross-validation, λ chosen considering the inherent dynamics of data
- **estimation alg.:** Forward-Backward Splitting (FBS) with *real-time* implementation

Attention Marker:

- compute a feature for each speaker from the set of measurements and estimated encoder/decoder coefficients in every window k → $m_k^{(i)}$ for $i = 1,2$
- **potential examples:**
  - reverse-correlation in decoding models: $m_k^{(i)} = \left|\mathrm{corr}\big(\boldsymbol{y}_k^{(i)}, \boldsymbol{X}_j \hat{\boldsymbol{\theta}}_k^{(i)}\big)\right|$
  - M100 peak magnitude in MEG encoding models: $\left|\hat{\boldsymbol{\theta}}_k^{(i)}\right|$ near the 100ms delay

Dynamic State-Space Model:

- at $k = k_0$, consider a fixed-lag sliding window:



- **dynamic state-space model:** defined on the $m_k^{(i)}$'s in the sliding window
  → for an interpretable, probabilistic, and robust measure of attentional state

State-Space Model

$$\begin{cases} p_k = P(n_k = 1) = \dfrac{1}{1 + \exp(-z_k)} \\ z_k = z_{k-1} + w_k \\ w_k \sim N(0, \eta_k) \end{cases}$$

Observation Equations

$$\begin{cases} m_k^{(i)} \big| n_k = i \sim \mathrm{LogNormal}\big(\rho^{(a)}, \mu^{(a)}\big) \\ m_k^{(i)} \big| n_k \neq i \sim \mathrm{LogNormal}\big(\rho^{(u)}, \mu^{(u)}\big) \end{cases}$$

- **model parameters:** $z_{1:K_W}, \eta_{1:K_W}, \rho^{(a)}, \rho^{(u)}, \mu^{(a)}, \mu^{(u)}$
- **goal at** $k = k_0$: estimate $p_{k^*} = \mathrm{logistic}(z_{k^*})$ where $k^* = k_0 - K_F$
- **inference algorithm:** apply the EM algorithm in the sliding window [5]
- quality of the chosen feature in attention marker ∝ separation between the fitted attended and unattended LogNormal distributions

EEG Analysis (Decoding Model)

Experiment Specifications:

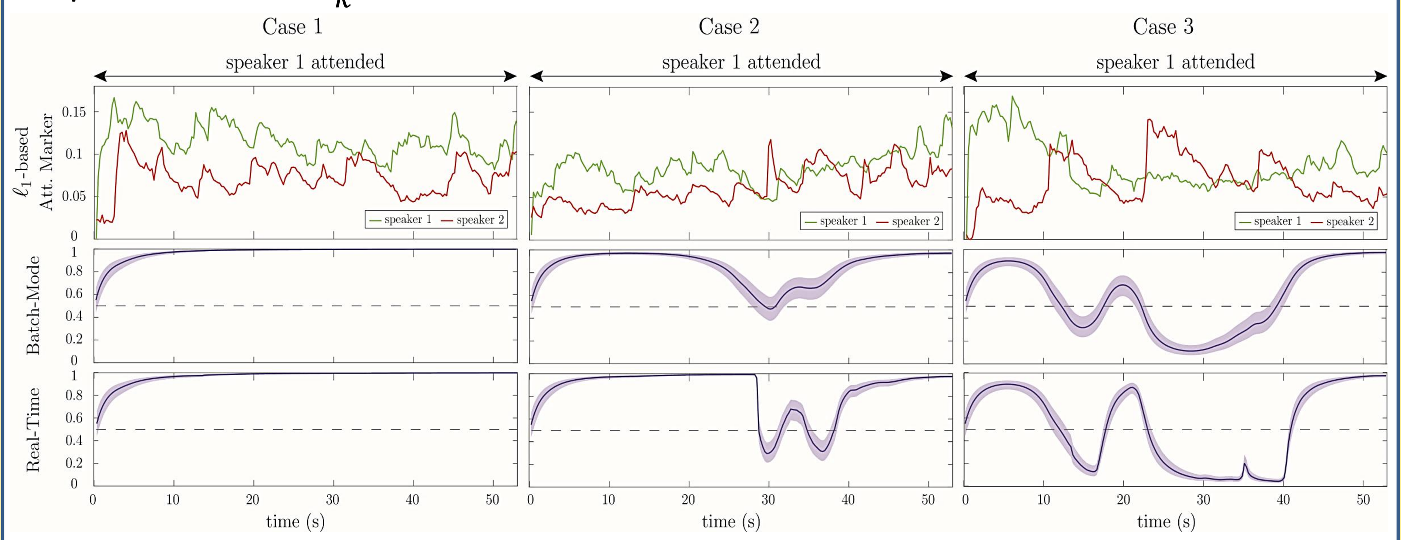
- 3 subjects, *instructed* constant attention on speaker 1, two speakers
- 64-channel EEG recording, 24 trials each 60s, downsampled to $f_s = 64Hz$

Attention Decoding Framework:

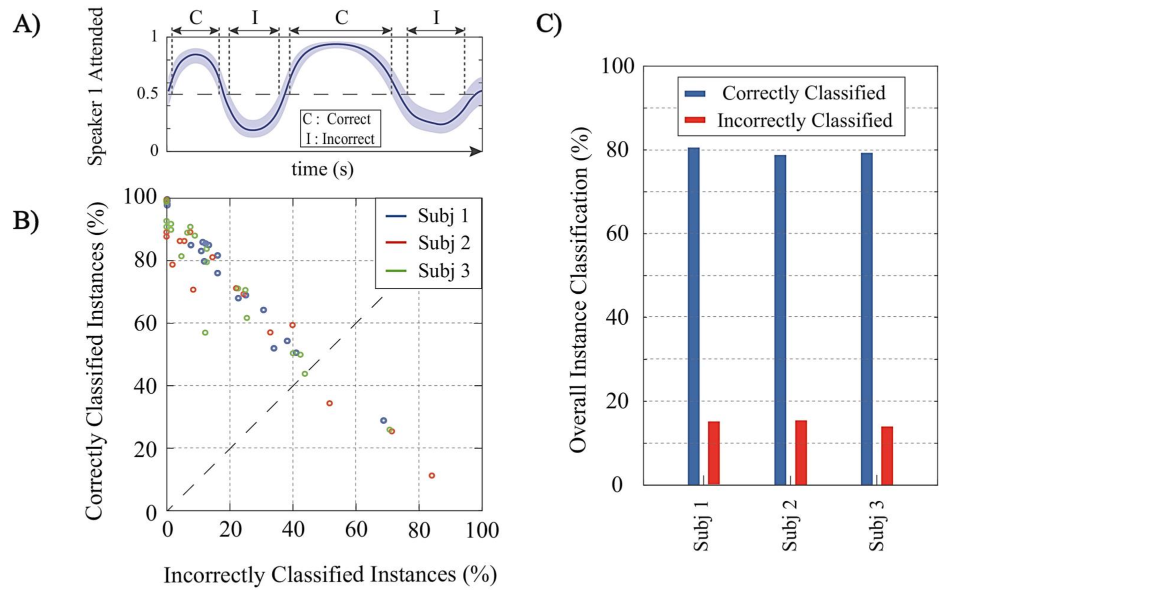
- **decoder estimation parameters:** $W = 0.25 f_s$, considered EEG delays up to 0.25s, $\gamma = 0.4$, $\lambda = 0.975$ (*effective* data length of $\frac{W}{(1-\lambda) f_s} = 10s$)
- **attention marker:** ℓ_1 norm of the decoder, i.e., $m_k^{(i)} = \left\|\hat{\boldsymbol{\theta}}_k^{(i)}\right\|_1$
  → **rationale:** detects significant decoder peaks
- **fixed-lag sliding window parameters:** $K_W = 15 f_s$, $K_F = 1.5 f_s$
- **total attention decoding delay:** 1.5s + 0.25s = 1.75s

Example Trial Outputs:

- separating power of the attention marker decreasing from case 1 to 3
- second row shows inferred p_k's in our real-time framework
- third row shows inferred p_k's in the batch-mode case, where the state-space processes all $m_k^{(i)}$'s at once

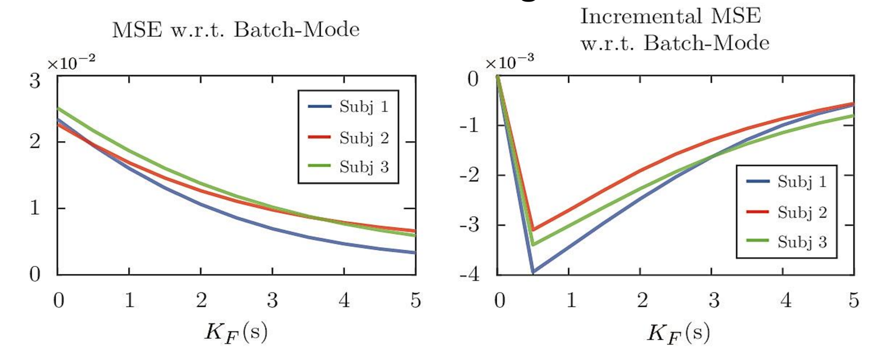


average classification accuracy in a trial for each subject:



MSE of inferred p_k's w.r.t. the batch-mode as a function of K_F:

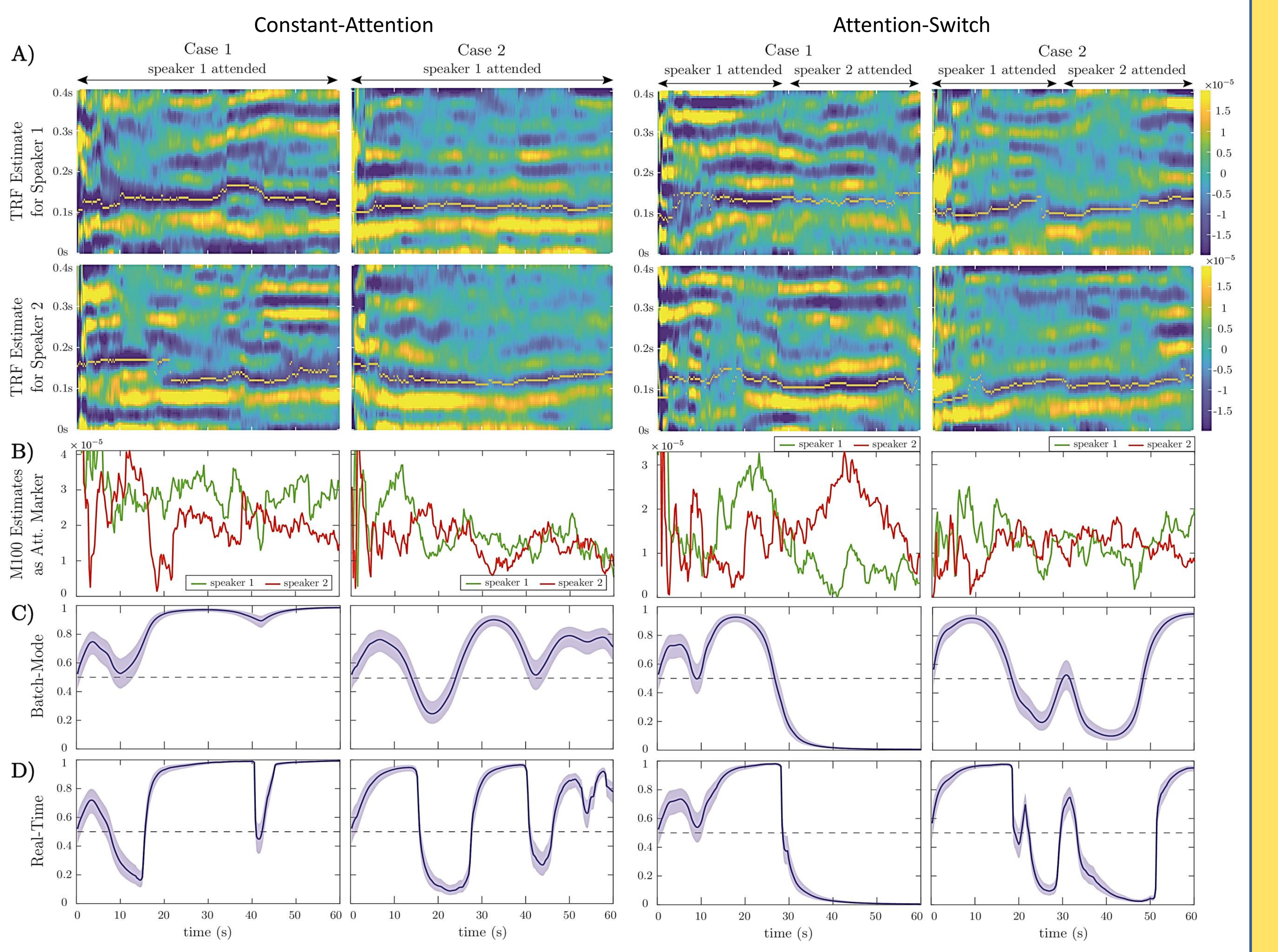
→ **rationale:** batch-mode can serve as the ground truth for our framework



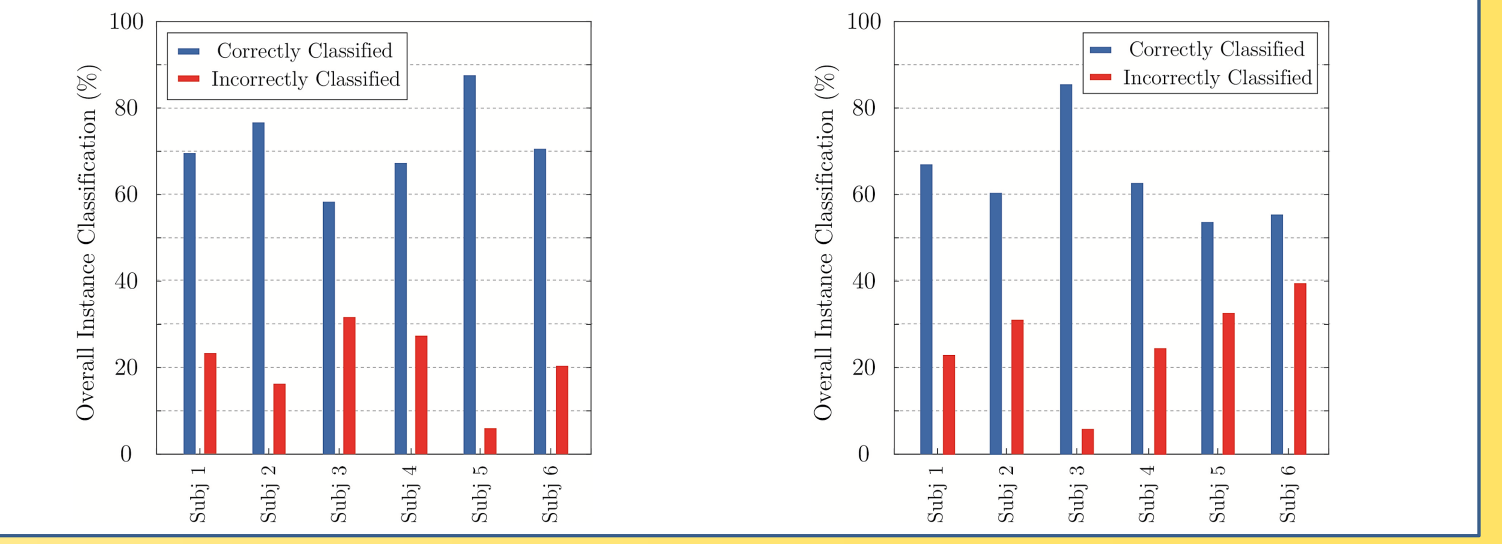
MEG Analysis (Encoding Model)

- 6 subjects, dual-speaker setting, constant-attention and attention-switch experiments
- estimation parameters similar to the EEG analysis
- **attention marker:** real-time M100 magnitude estimates in the TRFs

example TRF estimation results and state-space outputs:



average classification accuracy in a trial for each subject:



Summary

- a new framework for real-time attention decoding in competing speaker environments:
  - real-time estimation of encoding or decoding coefficients
  - computing a feature from the estimates and recorded data
  - apply a state-space model on the features for a statistically interpretable and robust measure of the attentional state
- high temporal resolution and no need for large training datasets, unlike existing methods
- serves as a step towards attention decoding for emerging real-time applications

Supported in part by the National Science Foundation Award No. 1552946 and a Collaborative Research Grant from Starkey Hearing Technologies

Poster PDF available at: http://ter.ps/simonpubs , Email: smiran@umd.edu