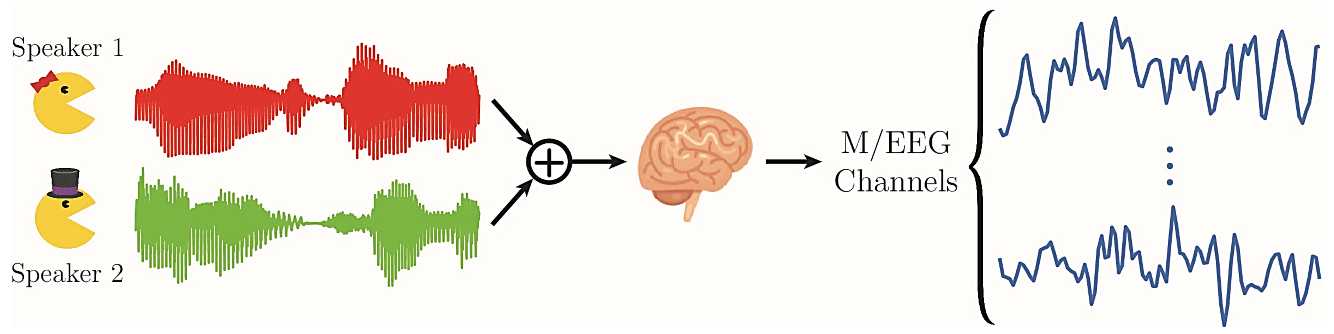
Real-Time Tracking of the Selective Auditory Attention from M/EEG via Bayesian Filtering

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



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?

Applications: brain-computer Interface (BCI) systems and smart hearing aids

decoding models ➡ linearly map M/EEG data to stimulus
encoding models ➡ linearly map stimulus to a neural response from M/EEG

Existing Methods:

- **reverse-correlation or stimulus reconstruction in decoding models (EEG) [2]:** train a decoder on the *attended* speech using training data; apply the trained decoder on recorded EEG to reconstruct a stimulus; speech that best matches the reconstruction is classified as the attended speech
- **important stimulus time lags in encoding models (MEG) [3][4]:** estimate the encoding coefficients for each speaker, i.e., Temporal Response Function (TRF); the speaker with a larger M100 (the TRF peak close to 100ms delay) is classified as the attended speaker

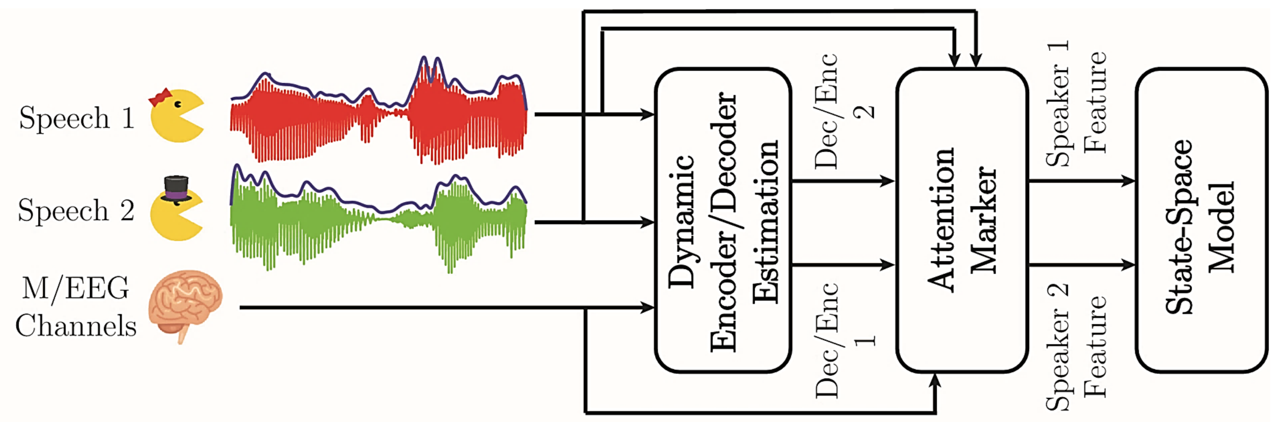
Shortcomings for Real-Time Attention Decoding:

- attention decoding accuracy drops significantly at high temporal resolutions, e.g. 1s (unreliable performance in real-time settings)
- need *large training datasets* to pre-estimate the *attended* encoder/decoder coefficients reliably (may not be accessible 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



Dynamic Encoder/Decoder Estimation:

- consider K consecutive non-overlapping windows of length W samples
- update the enc./dec. estimates $\widehat{\boldsymbol{\theta}}_k^{(i)}$ for *each speaker* in *every window*:

$$\widehat{\boldsymbol{\theta}}_k^{(i)} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{k} \lambda^{k-j} \left\|\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\theta}\right\|_2^2 + \gamma\|\boldsymbol{\theta}\|_1, \quad k = 1,2,\ldots,K, \quad i = 1,2$$

forgetting factor | speech envelopes (dec.) / neural response (enc.) | M/EEG covariates (dec.) / envelope covariates (enc.) | ℓ_1 regularization penalty

Attention Marker:

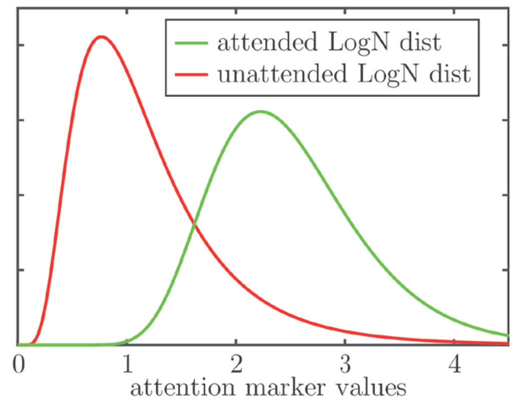
- an *attention-modulated* feature for *each speaker* in *every window*
  - $m_k^{(i)} = f\left(\boldsymbol{y}_k, \boldsymbol{X}_j, \widehat{\boldsymbol{\theta}}_k^{(i)}\right)$
- **potential examples:**
  - reverse-correlation in dec. models: $m_k^{(i)} = \left|\mathrm{corr}\left(\boldsymbol{y}_k^{(i)}, \boldsymbol{X}_j\widehat{\boldsymbol{\theta}}_k^{(i)}\right)\right|$
  - M100 peak magnitude in MEG enc. models: $\left|\widehat{\boldsymbol{\theta}}_k^{(i)}\right|$ near the 100ms delay

Dynamic State-Space Model:

- $n_k = 1$ (resp. 2) if speaker 1 (resp. 2) is attended at window k
- fit the following model on $m_k^{(i)}$'s in a *fixed-lag sliding window* fashion for real-time attention decoding

Observation Equations

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



State-Space Model

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

- **model parameters:** z_k's, η_k's, $\rho^{(a)}$, $\rho^{(u)}$, $\mu^{(a)}$, $\mu^{(u)}$
- **goal at window** $k = k_0$: estimate $p_{k^*} = \mathrm{logistic}(z_{k^*})$ where $k^* = k_0 - K_F$ through nested EMs [4] as a *dynamic*, *probabilistic*, and *robust* measure of the attentional state

EEG Analysis (Decoding Model)

Experiment Specifications:

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

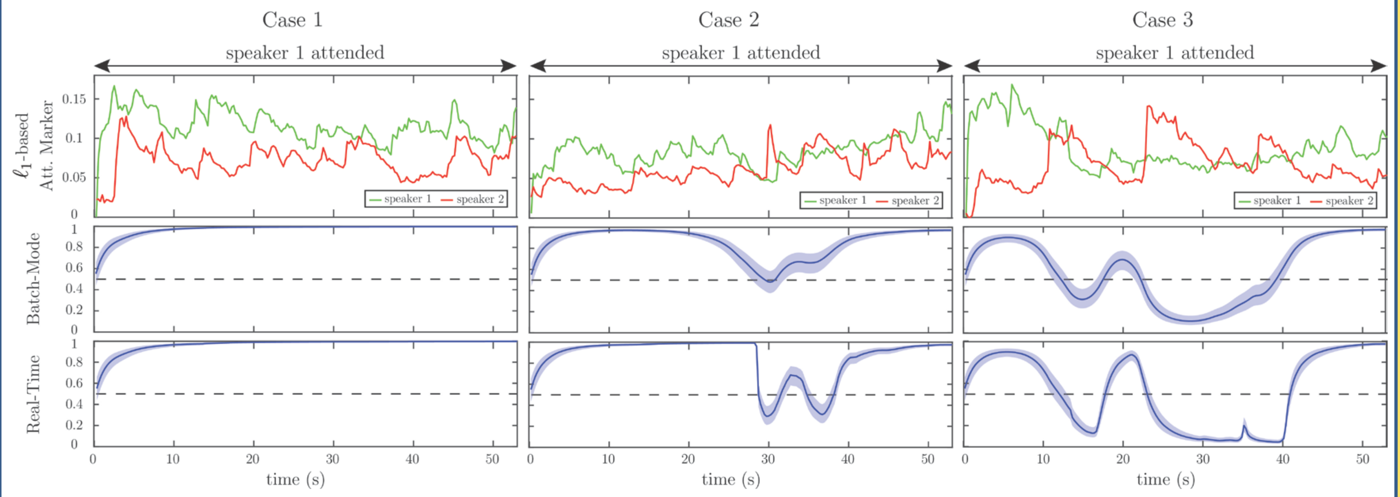
Attention Decoding Framework:

- **decoder estimation parameters:** $W = 0.25f_s$, decoder lag of 0.25s
- **attention marker:** ℓ_1 norm of the decoder, i.e., $m_k^{(i)} = \left\|\widehat{\boldsymbol{\theta}}_k^{(i)}\right\|_1$
  - **rationale:** detects significant peaks in dec. coefficients
- **total built-in attention decoding delay:** 1.5s + 0.25s = 1.75s

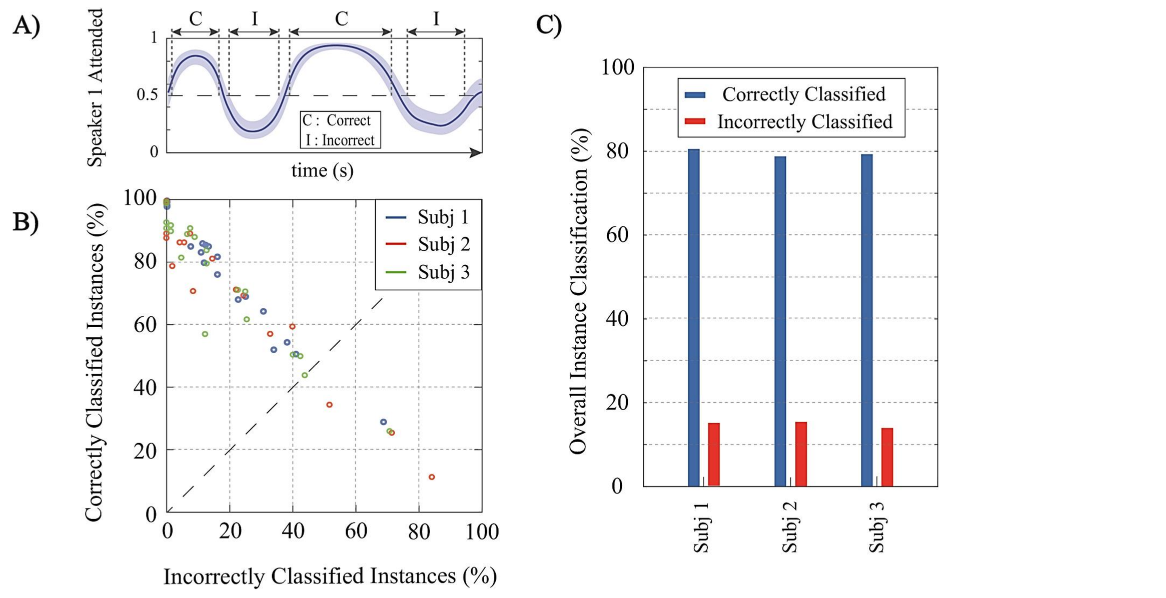
forward-lag in state-space model application | decoder lag

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



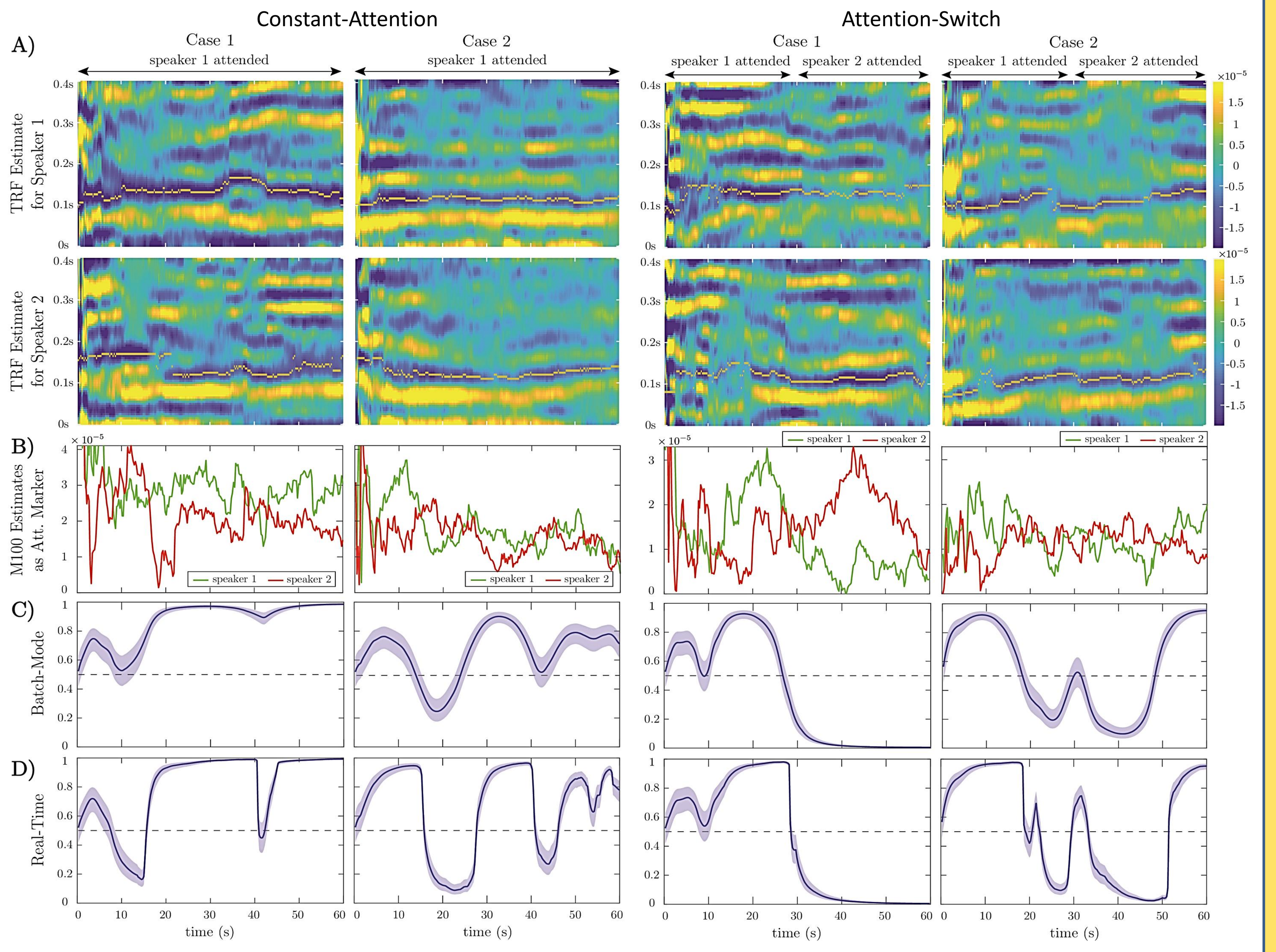
average classification accuracy in a trial for each subject:



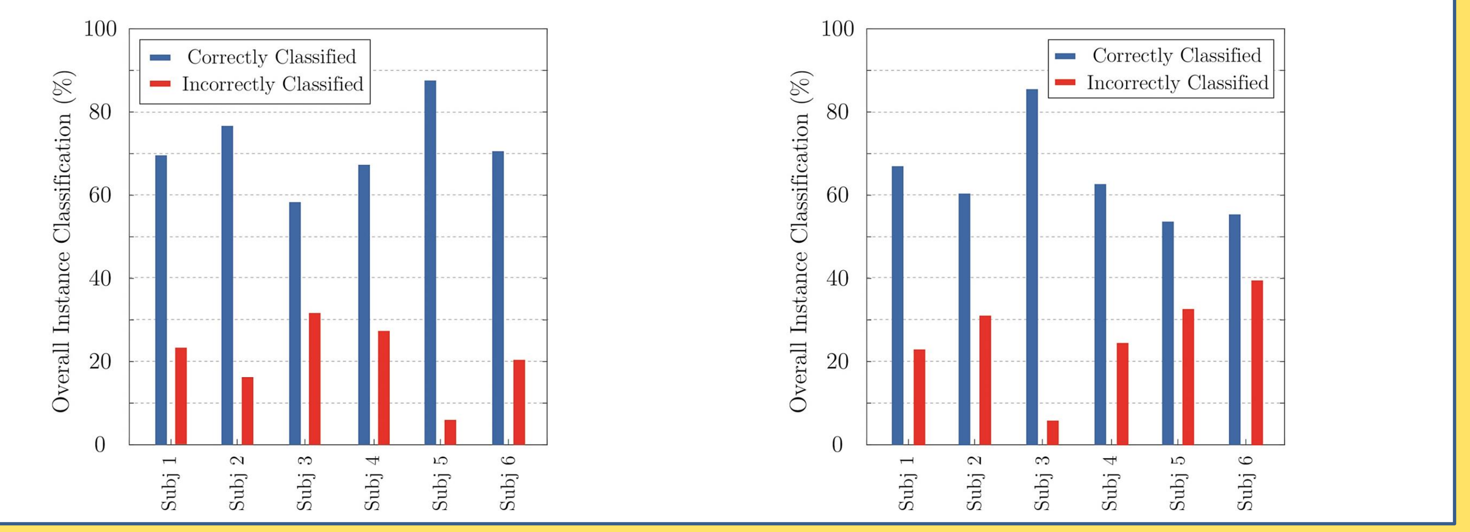
MEG Analysis (Encoding Model)

- 6 subjects, two speakers, constant-attention (6 trials) and attention-switch (3 trials) experiments
- **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 settings
  - *real-time* estimation of encoding or decoding coefficients
  - computing an att.-modulated 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
- can operate at high temporal resolution with no need for large training datasets, unlike existing methods

Supported in part by the National Science Foundation Awards No. 1552946 and 1734892, and a Collaborative Research Grant from Starkey Hearing Technologies

Poster PDF available at: http://ter.ps/simonpubs , Email: {jzsimon,behtash}@umd.edu