INFORMATION-THEORETIC LIMITS ON THE PERFORMANCE OF AUDITORY ATTENTION DECODERS

Ruwanthi Abeysekara[1,2], Christopher J. Smalt[4], I. M. Dushyanthi Karunathilake[1,2], Jonathan Z. Simon[1,2,3], and Behtash Babadi[1,2]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, [2]Institute for Systems Research, University of Maryland, College Park, MD, [3]Department of Biology, University of Maryland, College Park, MD,
[4]Human Health & Performance Systems Group, MIT Lincoln Laboratory, Lexington, MA.

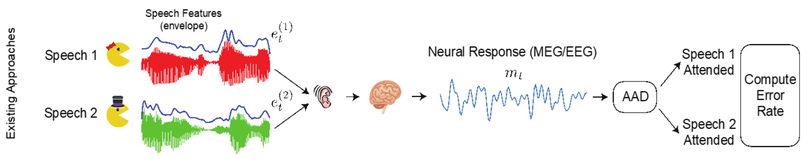
INTRODUCTION

Goal

To provide an information-theoretic framework for AAD, offering a fundamental understanding of its capabilities and limitations.

What is AAD?

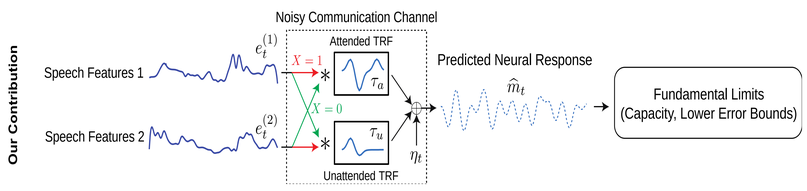
The goal of AAD is to decode the attentional state of a listener to a target speech stream in a multi-speaker environment, and thereby use it as a feedback to the hearing aid device to enhance the attended speech.



Why a theoretical bound?

Establishing the capabilities and limitations of AAD algorithms is important in practice, especially for manufacturers of hearing devices. Is it possible to achieve a near 100% decoding accuracy in practice?

Key contribution : AAD as a noisy communication channel



CHALLENGES

1. Estimating the covariance presents a challenge due to the high dimensionality of the problem and limited data, which is addressed through shrinkage covariance estimation.
2. Mutual information does not admit a closed-form expression. This is addressed through Monte Carlo sampling.
3. Monte Carlo sampling requires drawing samples from high-dimensional distributions over a large number of iterations. This is mitigated by utilizing a reparameterization technique.

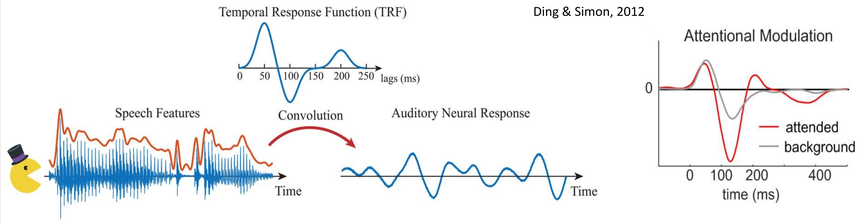
CONCLUSIONS

- Quantified AAD limits using **channel capacity** and **Fano's lower bound**.
- Established **algorithm-agnostic error bound** for AAD performance evaluation.
- Emphasized limited decision window impact on AAD accuracy in **real-time scenarios.**
- Future research extends to complex auditory environments, more speech features, multi-channel EEG/MEG data, and theoretical bounds used to benchmark a broader range of AAD algorithms.

PARAMETER ESTIMATION

Encoding model based on TRF

- The neural response observed by MEG is a linear convolution of a kernel, known as TRF, with the envelope of the speech.



- With two equally-loud speakers, the neural response, $m_t = \underbrace{\tau_a * e_t^{(a)} + \tau_u * e_t^{(u)}}_{\mu} + \underbrace{\eta_t}_{\sim\mathcal{N}(0,\Sigma)}$

where $e_t^{(a)}, e_t^{(u)}, \tau_a, \tau_u$, and η_t denote attended speech, unattended speech, attended TRF, unattended TRF, and noise term representing un-modeled neural processes and measurement noise, respectively.

Estimating TRF model parameters

- TRFs are computed using ridge regression or boosting. Then, the two predicted MEG responses would be:

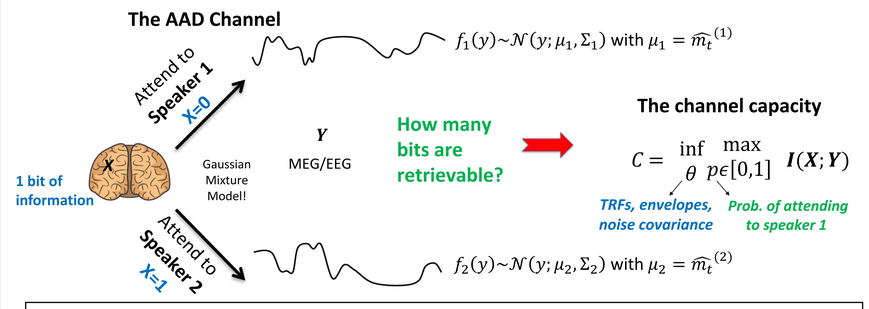
Attended to speaker 1: $\widehat{m_t}^{(1)} = \tau_a * e_t^{(1)} + \tau_u * e_t^{(2)}$

Unttended to speaker 2: $\widehat{m_t}^{(2)} = \tau_a * e_t^{(2)} + \tau_u * e_t^{(1)}$

- The noise covariance for each model is estimated via covariance shrinkage.
- In conventional AAD, the Pearson correlation between the two predictions and the observed signal determine the attended speaker, and the AAD's error rate is evaluated empirically.

CAPACITY/ ERROR RATE CHARACTERIZATION

Information theoretic framework



$$I(X;Y) = -pE_{f_1}\{\log_2[pf_1(y) + (1-p)f_2(y)]\} - (1-p)E_{f_2}\{\log_2[pf_1(y) + (1-p)f_2(y)]\} - \frac{p}{2}\log_2(2\pi e)^T|\Sigma_1| - \frac{(1-p)}{2}\log_2(2\pi e)^T|\Sigma_2|$$

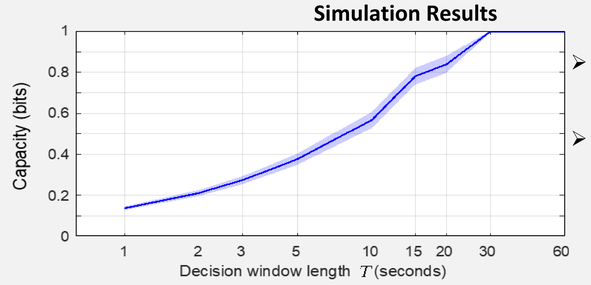
Lower Error Bound

- Utilizing Fano's inequality, the lower bound for the probability of error is $\boxed{P_e \geq H_b^{-1}(H(X|Y))}$

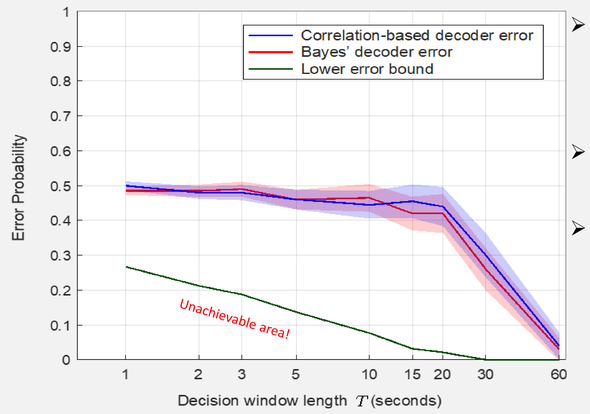
where, $P_e = P[\hat{X} \neq X]$, $H(\cdot)$ is the binary entropy function, and $H_b^{-1}(\cdot)$ is the inverse binary entropy function over the domain $[0,1]$.

RESULTS

Simulation Results

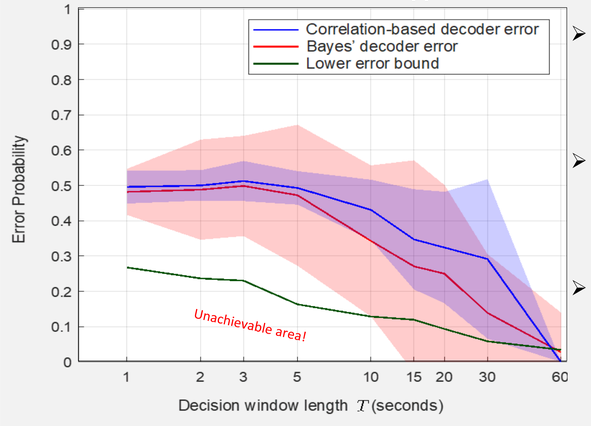


- Capacity reaches ~**1 bit** at ~**30 s**!
- Implication: 30 s of neural response information for reliable attention recovery.



- 100 trials, 60 s long experiment, attended TRF with deeper M100 peak.
- Error bound reaches ~0% at ~**30 s**.
- Decoders used: Correlation-based and Bayes' with Gaussian assumption; Both decoders perform better than chance level at >20 s.

Real Data Application



- 10 subjects, 6 trials, 60 s long experiment; TRFs estimated via boosting (500 ms long).
- The two decoders perform better than chance level (single-trial decoding) at > 20 s.
- Some room for improvement, but not too much!

REFERENCES

[1] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," Nature, vol. 485, no. 7397, pp. 233–236, 2012.

[2] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," IEEE Signal Processing Magazine, vol. 38, no. 4, pp. 89–102, 2021.

[3] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, "Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid," Neural Networks, vol. 140, pp. 136–147, 2021.