

Assessing the contribution of visual speech features to audiovisual speech perception in noise

Aaron R. Nidiffer¹ Aisling O'Sullivan² and Edmund C. Lalor¹

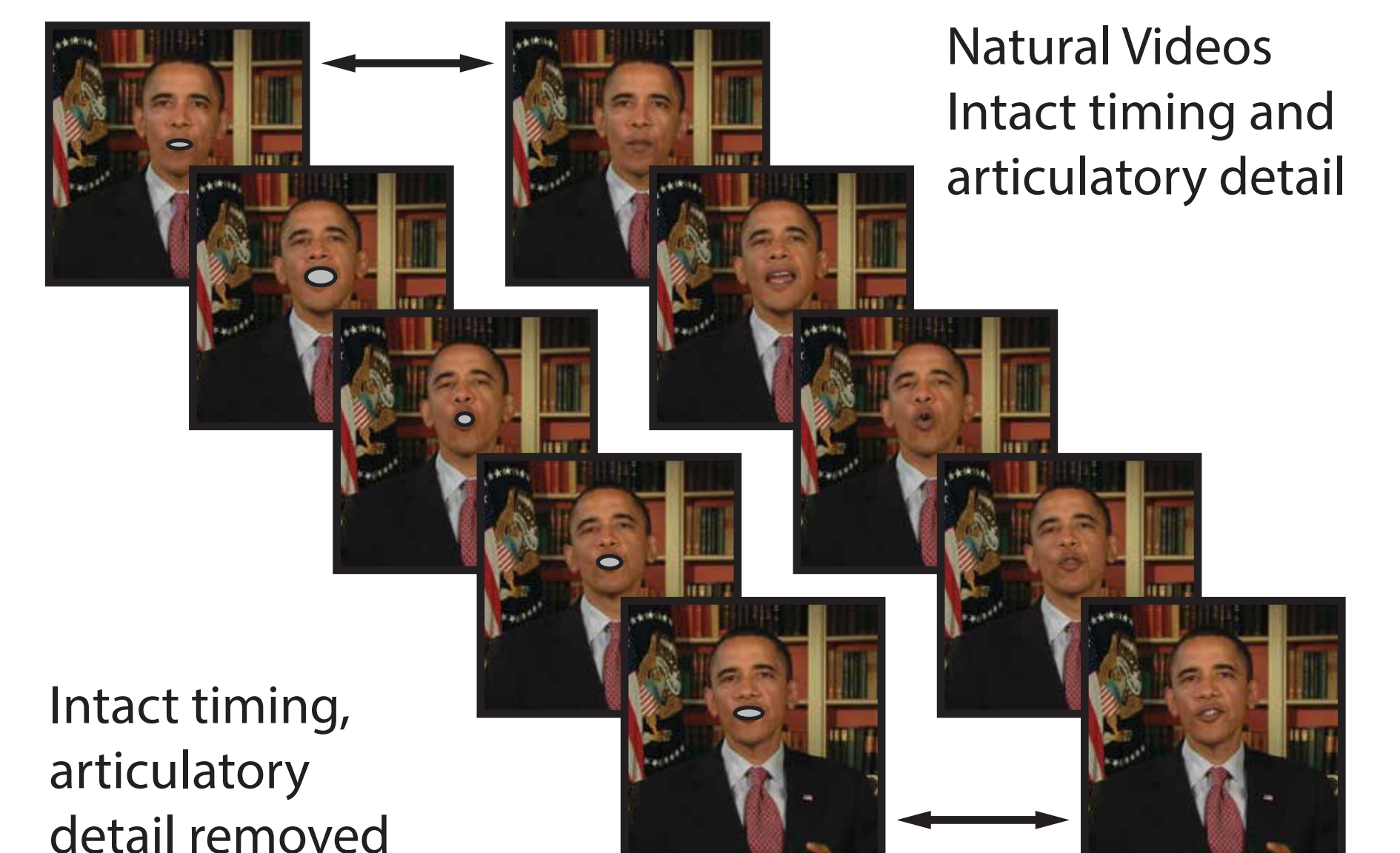
¹Depts. Biomedical Engineering and Neuroscience, University of Rochester, Rochester, NY, USA

²Trinity Centre for Biomedical Engineering, Trinity College Dublin, Dublin, Ireland

Introduction

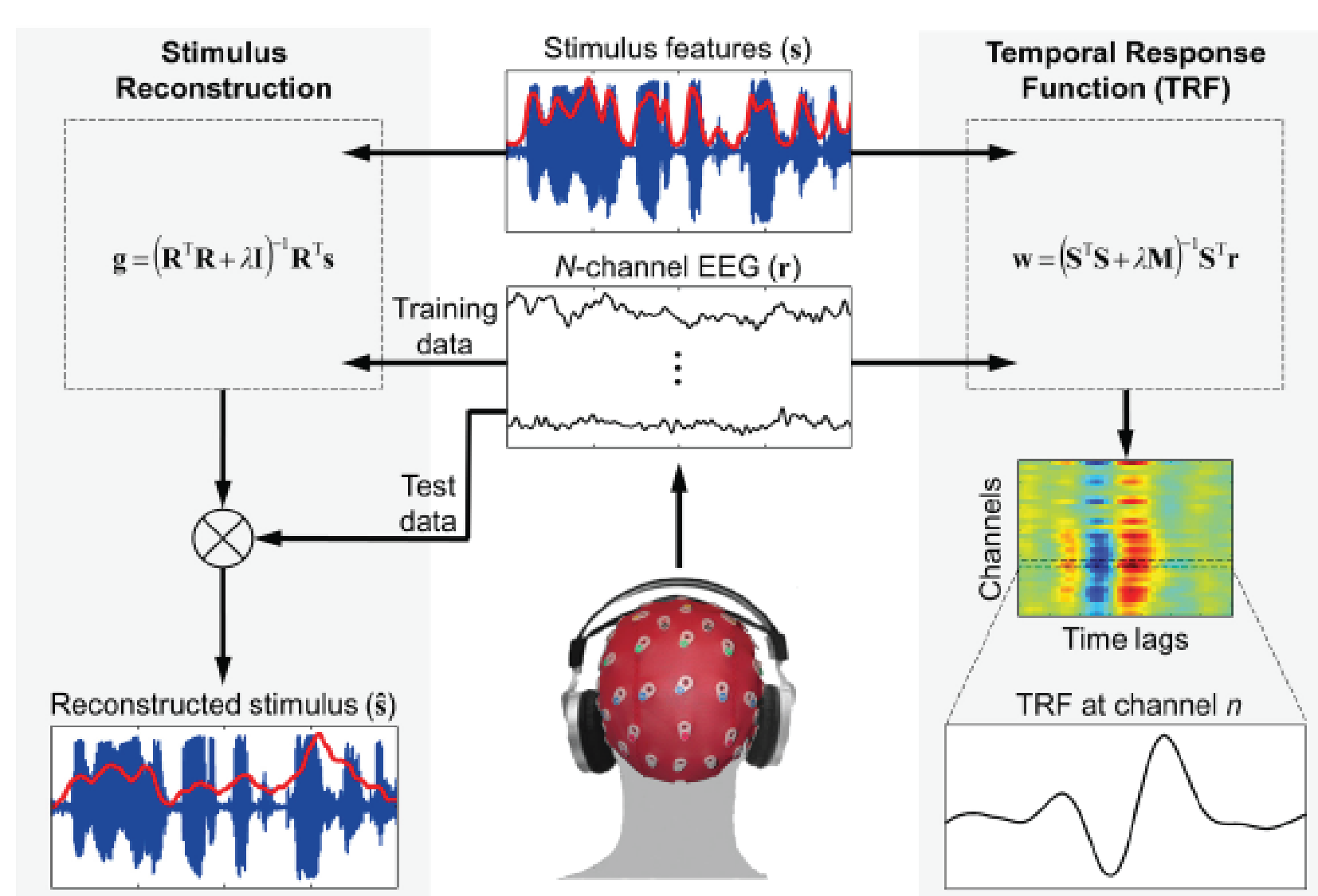
- Seeing the face of a speaker improves their intelligibility - particularly when noise obscures the speech signal.
- Listeners predominantly direct their gaze toward a speaker's lips which convey general dynamic information that is correlated with the acoustic envelope and detailed articulatory shapes which convey complementary linguistic information.
- Neuroimaging work has also found an enhancement of lip processing regions in visual cortex when the acoustics are missing.

Together, this suggests that the lips are an important feature of visual speech which the brain exploits to assist speech processing. Yet it remains unclear whether the information that confers the improved intelligibility of noisy audiovisual speech is derived from the correlated lip dynamics or the complementary lip shape. Here we present an experiment where we have modulated the amount of facial information available to listeners as they listen to audiovisual speech in noise (-9 dB).



Methods - Linear Modeling of EEG Responses

Temporal Response Function Analysis

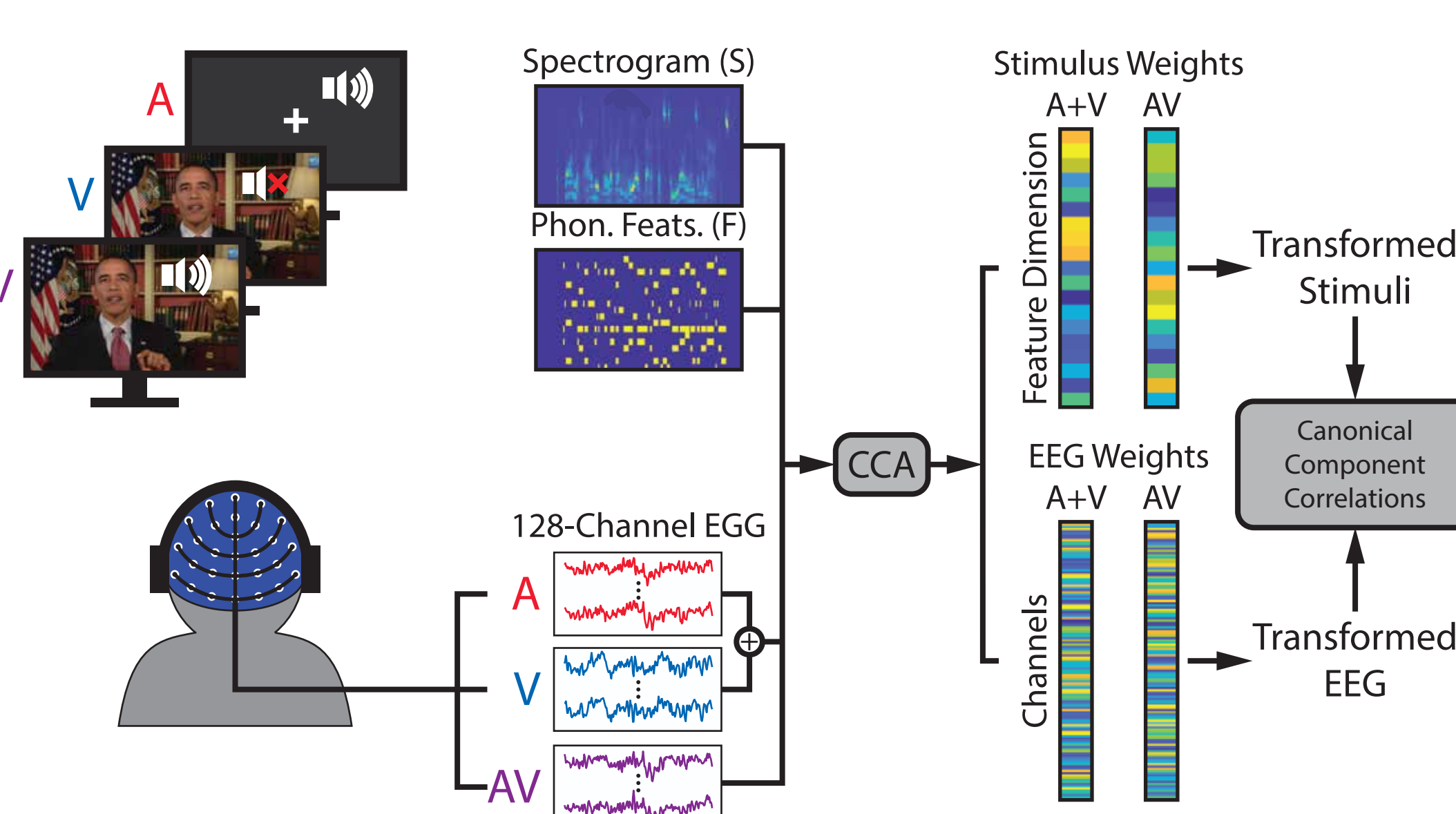


We recorded high-density EEG and eye position data while participants listened to minute-long trials of natural, connected speech presented acoustically, visually, and audiovisually.

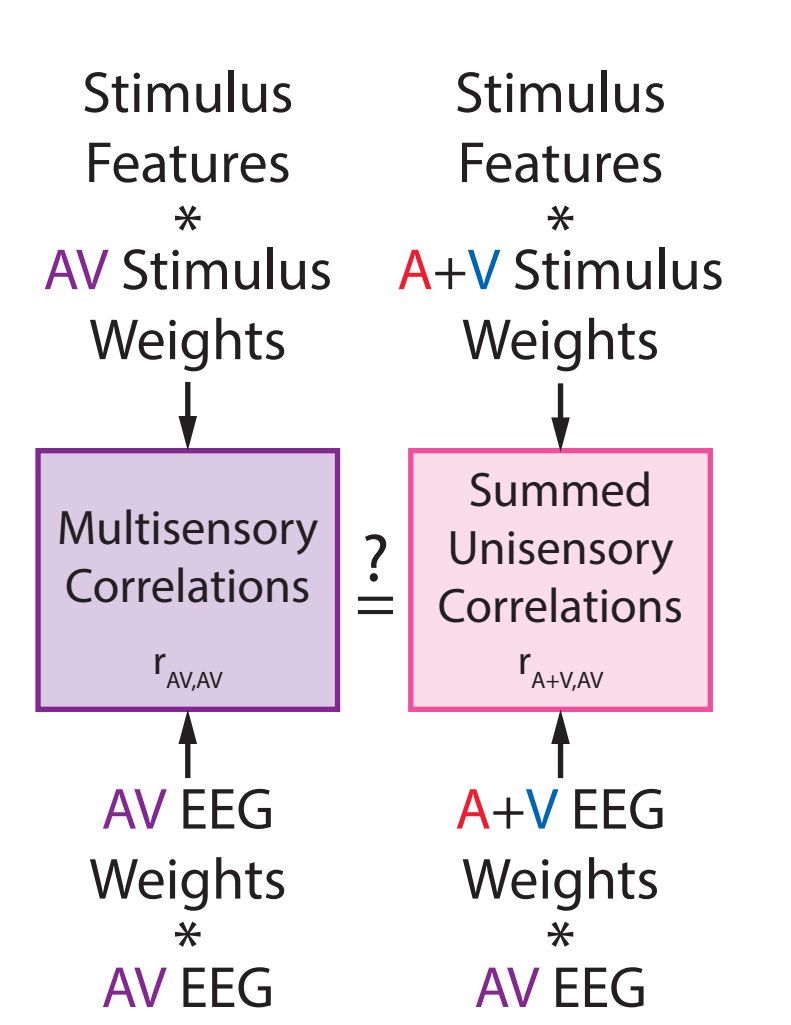
Participants were asked to detect target words which were given at the beginning of each trial and also to rate the intelligibility of the speech they heard/saw.

Using linear modeling, we extracted Temporal Response Functions (TRFs) or Canonical Components (CCs) and used those to predict unseen data.

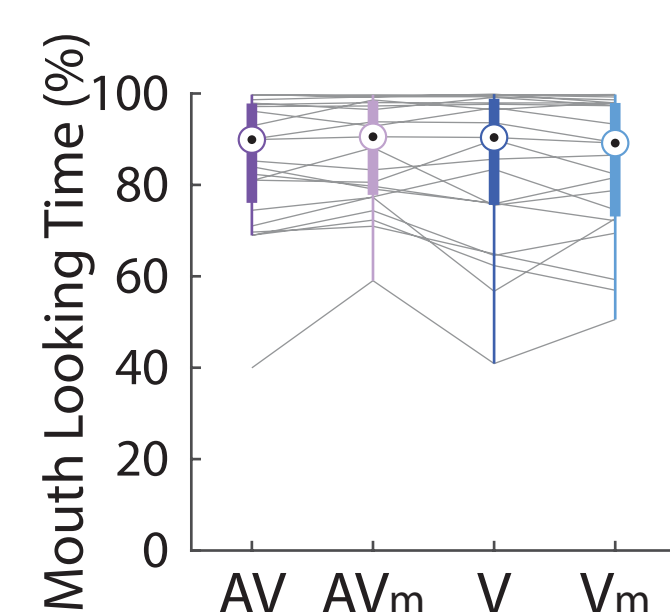
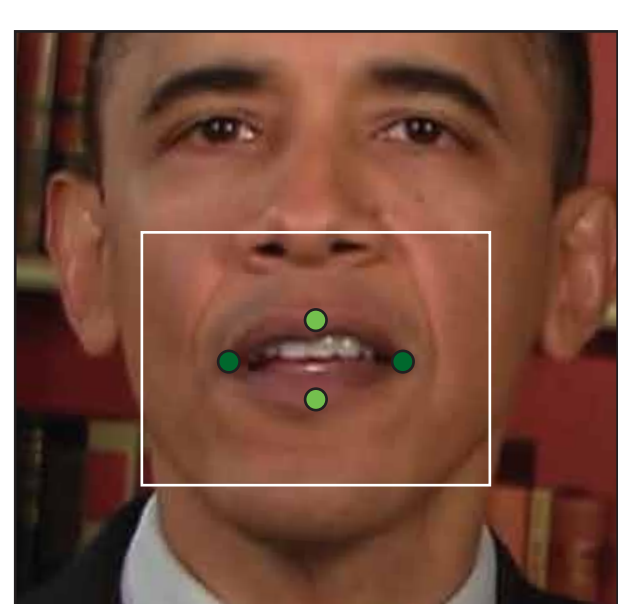
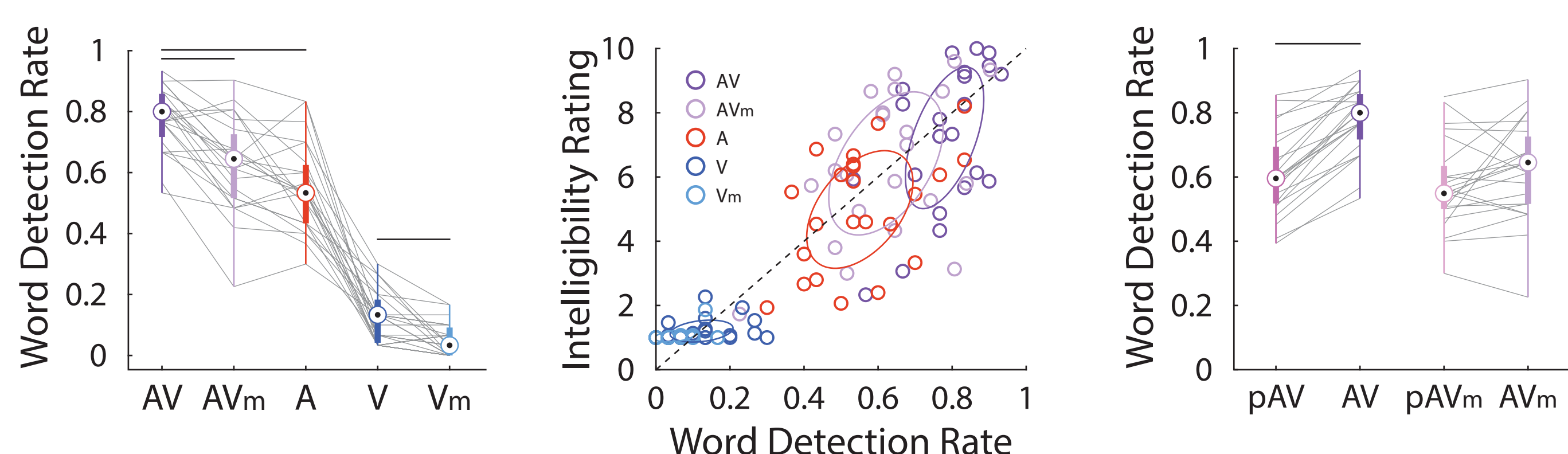
Canonical Component Analysis



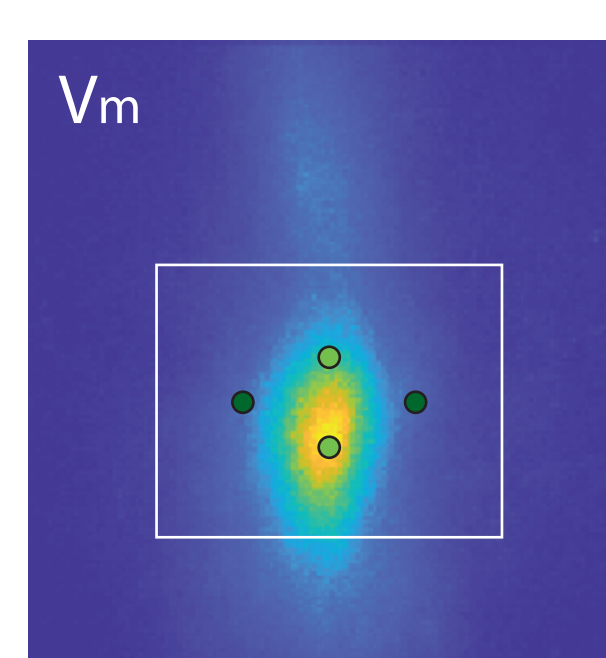
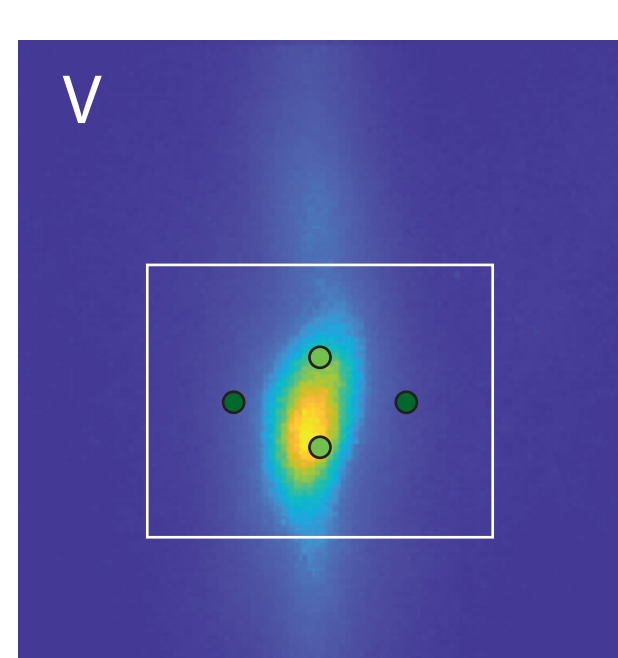
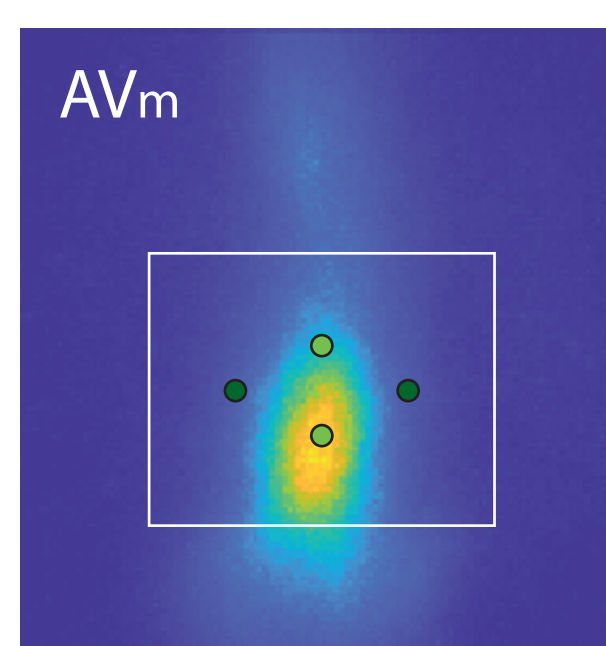
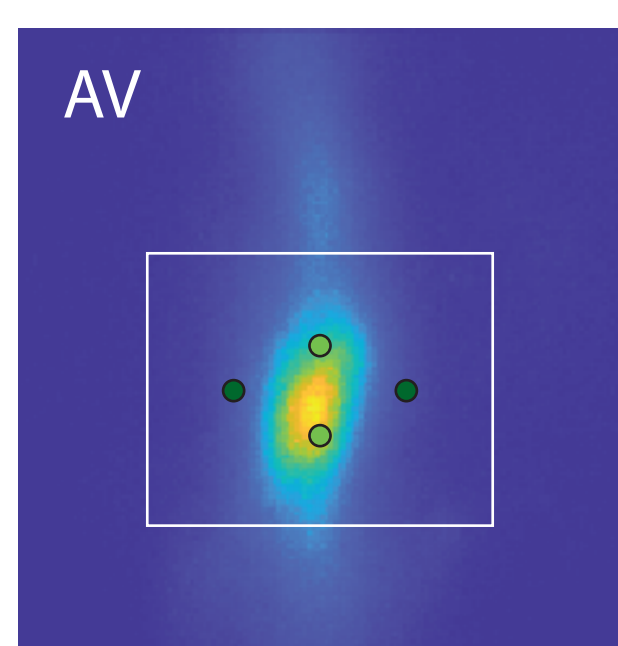
Multisensory Effects



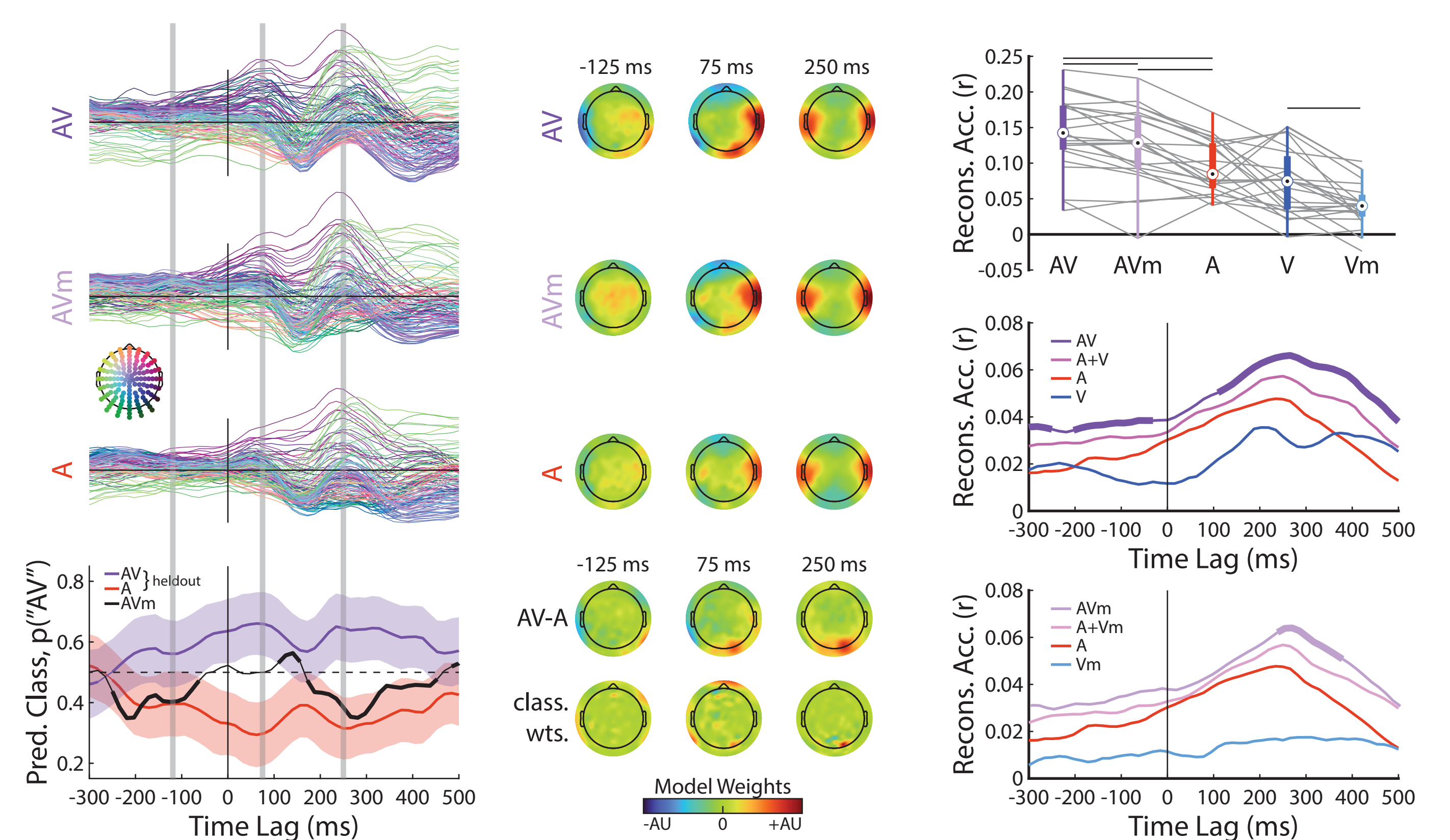
Clear visual speech enhances speech comprehension and mouth viewing tendencies.



- Unsurprisingly, participants were better at detecting target words when the detailed articulatory information was visible, suggesting the importance of visual linguistic information speech comprehension.
- Generally, detection and subjective intelligibility scaled proportionally, with the exception of the visual conditions. Participants underrated their subjective perception relative to their objective performance.
- Participant spent a majority of their time viewing the mouth ROI in the visual and audiovisual conditions. Even when there was little articulatory detail, participants persisted with appropriate gaze.
- Importantly, objective behavior (word detection) didn't scale with the amount of time viewing the lips. Interestingly, participants found the speech clearer the more they looked at the mouth (or vice-versa), regardless of articulatory detail.



Visual articulatory detail improves speech tracking, likely due to linguistic availability



- Cortical tracking of the speech envelope was enhanced above auditory tracking for both audiovisual conditions. Orthogonally, envelope tracking was enhanced when detailed articulatory information was present.
- Clear visual speech produce multisensory benefits at broad time lags compared to the masked condition which produced benefits at restricted lags.
- Using CCA we found evidence of multisensory enhancement at the linguistic level of representation of our speech stimulus. There was no difference unique acoustic responses between visual conditions, but phonetic feature tracking was stronger when participants could extract linguistic information from articulatory details.

