Published in final edited form as:

Hear Res. 2023 June ; 433: 108767. doi:10.1016/j.heares.2023.108767.

Incorporating models of subcortical processing improves the ability to predict EEG responses to natural speech

Elsa Lindboom[1], **Aaron Nidiffer**[1,2], **Laurel H. Carney**[1,2,3], **Edmund C. Lalor**[1,2]

[1]Department of Biomedical Engineering, University of Rochester, Rochester, NY

[2]Department of Neuroscience and Del Monte Institute for Neuroscience, University of Rochester, Rochester, NY

[3]Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY

Abstract

The goal of describing how the human brain responds to complex acoustic stimuli has driven auditory neuroscience research for decades. Often, a systems-based approach has been taken, in which neurophysiological responses are modeled based on features of the presented stimulus. This includes a wealth of work modeling electroencephalogram (EEG) responses to complex acoustic stimuli such as speech. Examples of the acoustic features used in such modeling include the amplitude envelope and spectrogram of speech. These models implicitly assume a direct mapping from stimulus representation to cortical activity. However, in reality, the representation of sound is transformed as it passes through early stages of the auditory pathway, such that inputs to the cortex are fundamentally different from the raw audio signal that was presented. Thus, it could be valuable to account for the transformations taking place in lower-order auditory areas, such as the auditory nerve, cochlear nucleus, and inferior colliculus (IC) when predicting cortical responses to complex sounds. Specifically, because IC responses are more similar to cortical inputs than acoustic features derived directly from the audio signal, we hypothesized that linear mappings (temporal response functions; TRFs) fit to the outputs of an IC model would better predict EEG responses to speech stimuli. To this end, we modeled responses to the acoustic stimuli as they passed through the auditory nerve, cochlear nucleus, and inferior colliculus before fitting a TRF to the output of the modeled IC responses. Results showed that using model-IC responses in traditional systems analyses resulted in better predictions of EEG activity than using the envelope or spectrogram of a speech stimulus. Further, it was revealed that model-IC derived TRFs predict different aspects of the EEG than acoustic-feature TRFs, and combining both types of TRF models provides a more accurate prediction of the EEG response.

Introduction

Decades of research have sought to understand how the human brain processes the many sounds we encounter in everyday life. For example, since the 1930s researchers have used the electroencephalogram (EEG) to derive event-related potentials (ERPs) by averaging responses immediately following repeated presentations of brief, isolated stimuli (Davis,

Correspondence: elalor@ur.rochester.edu.

1939; Handy 2005, Sur and Sinha, 2009). However, most sensory information available to human listeners is continuous, non-repeated, and occurs within a noisy environment, forcing listeners to discern important on-going signals from surrounding irrelevant information with only a single presentation. This includes that all-important of human signals – speech. To better approximate normal listening conditions – with a particular emphasis on speech – research turned towards using longer, more natural stimuli to elicit the EEG (Connolly et al., 1994; Näätänen, 1997). Initially, these studies used relatively short segments of speech to produce ERPs, which still provided only a limited view on how the brain parses and processes continuous segments of acoustically, lexically, and semantically rich speech.

In recent years, researchers have increasingly emphasized the use of continuous, natural speech in their experiments (Hamilton and Huth, 2020). One fruitful approach to analyzing the resulting neural data involves modeling those data based on the speech stimuli that elicited them (Brodbeck and Simon, 2020). This approach, which is known as system identification, treats the brain as something of a 'black box' and seeks to develop quantitative mappings between various speech features and the resulting neurophysiological responses. In particular, electroencephalography (EEG) has often been the recording modality of choice given its noninvasive nature, ease of use, and high temporal resolution (Gevins et al., 1995; Regan, 1989; Murakami and Okada, 2006; Buzsaki et al., 2012; Lopes da Silva and Niedermeyer, 2005). One particularly popular and tractable analysis involves treating the brain as a linear time-invariant (LTI) system and obtaining a so-called temporal response function (TRF) via regularized linear regression (Crosse et al., 2016, 2021). This framework allows researchers to study how the brain processes speech at different hierarchical levels by modeling the relationship between EEG and both acoustic (e.g., acoustic envelope, spectrogram) and linguistic (e.g., phonetic features, semantic surprisal) features (Di Liberto et al., 2015, Broderick et al., 2018; Brodbeck et al., 2018, 2022; Gillis et al., 2021). For example, a model involving spectrogram and phonetic features out-performs either constituent model, indicating that each feature contributes to unique aspects of the EEG signal (Di Liberto et al., 2015).

One limitation of the TRF approach – as noted by Drennan and Lalor (2019) – is that it makes a strong assumption about linearity and time invariance. In effect, this assumes that EEG responses to a particular speech feature always have the same timing and morphology; because the speech feature changes in intensity, the EEG response will scale, but will not change in terms of its timing or shape. However, this assumption is incorrect; it has long been known that EEG responses to auditory stimuli vary in both amplitude and latency with the intensity of the sound (Beagley and Knight, 1967). Drennan and Lalor (2019) proposed to relax this assumption in the context of modeling EEG responses based on the speech envelope. Specifically, they allowed the TRF to vary in morphology for different envelope intensities by binning the speech envelope based in amplitude deriving a multivariate TRF (mTRF). The resulting amplitude-binned (AB) envelope mTRF produced significant improvements in the ability to predict EEG responses to novel stimuli (Drennan and Lalor, 2019).

While this approach produced significant improvements, it was based on a simple, somewhat arbitrary manipulation of the stimulus (amplitude binning). A more principled approach not yet considered would be to formally incorporate the substantial processing of sound input that occurs along subcortical pathways before reaching cortex and contributing to scalp-recorded EEG. For example, it is well known that neurons in the inferior colliculus (IC) are tuned to sound frequency and amplitude-modulation rate (Krishna and Semple, 2000; Nelson and Carney, 2007). Thus, IC neurons represent a population of cells that can integrate responses to sound features (e.g., spectrogram, amplitude envelope) extracted at lower levels of the auditory system (e.g., Carney et al., 2015). Given that such a transformed representation of the speech input is what cortex actually receives (rather than the stimulus itself), incorporating a model of such subcortical processing might lead to improved predictions of cortical EEG.

In the present study, we hypothesized that accounting for subcortical processing of speech sounds would improve predictions of EEG responses to natural speech. To test this, we modeled IC responses to speech sounds using the phenomenological same-frequency, inhibitory-excitatory (SFIE) model based on Nelson and Carney (2004; Fig. 1). This model transforms a sound input into simulated responses at the levels of the auditory nerve (AN), ventral cochlear nucleus, and inferior colliculus that have been validated against neurophysiological recordings across a series of studies (Zilany et al., 2009, 2014; Nelson and Carney, 2004; Carney et al., 2015; Carney and McDonough, 2019). We fitted TRF models to these simulated IC responses, broadband speech envelopes, AB envelopes, and spectrograms and measured how well each TRF could predict EEG responses recorded while participants listened to speech. Overall, the mTRF fit to IC responses produced more robust EEG prediction than the speech envelope or spectrogram. The IC response and AB-envelope mTRFs performed at comparable levels. However, analysis of the correlations between predicted EEGs from these two models revealed that they predicted different aspects of the EEG. Thus, combining the IC-response and AB-envelope mTRFs further improved the EEG predictions.

Methods

EEG Data and Stimuli

Stimuli and corresponding EEG responses from 19 subjects were obtained from two previous studies (DiLiberto et al., 2015; Broderick et al., 2018). In those experiments, subjects were presented with 20 three-minute-long segments of audio from an audiobook read by a male American English speaker (*The Old Man and the Sea* by Ernest Hemingway), using Sennheiser HD650 headphones. Each segment was ~155 s in duration and segments were presented in sequential order. During stimulus presentation, 128 scalp channels (+ 2 mastoid channels) of EEG data were recorded from each participant with a sampling rate of 512 Hz using the BioSemi ActiveTwo system. The recordings were digitally filtered between 1 and 15 Hz with a 2nd order, zero-phase (non-causal) Butterworth filter. The EEG signal was then referenced to the average of the two mastoid channels and down-sampled to 128 Hz to decrease computation time during further analyses.

Speech Representations

As mentioned, our goal was to attempt to account for early stage acoustic encoding when modeling EEG responses to natural speech. To that end, we wanted to compare the EEG predictions that included modeled IC responses to those based on acoustic representations computed directly from the speech stimulus. In particular, we derived TRFs based on four distinct representations of the speech stimulus. These representations were presented as either single or multivariate feature vectors. Before feature extraction, all audio samples were lowpass filtered with a Chebyshev Type 2 filter having a cutoff frequency of 20 kHz. The broadband-envelope representation of the audio was calculated as

$$env = |x_a(t)|, \quad x_a(t) = x(t) + j\hat{x}(t), \tag{1}$$

where $x_a(t)$ is the analytical representation of the signal, taken as the sum of the original speech, $x(t)$, and its Hilbert transform $\hat{x}(t)$ (Fig. 2A).

The spectrogram of the speech was obtained by filtering the speech into 20 log-spaced frequency bands ranging from 200 to 8-kHz (Di Liberto et al., 2015). The amplitude envelope of each frequency band was calculated using Eq. 1 (Fig. 2C). All audio-feature signals were down sampled to 128 Hz to match the sampling rate of the EEG data. To create the AB-envelope feature vector, the SPL (sound pressure level) envelope (Fig. 2B) was binned into 10 8-dB level ranges using the `histcounts` function in MATLAB (as outlined in Drennan and Lalor, 2019). This binning resulted in 10-variable feature vectors at each time point (Fig. 2D).

Neural responses to the acoustic stimuli were modeled in two steps. First, an AN model (Zilany et al., 2014) was used to simulate the responses of 20 AN fibers with characteristic frequencies (CF, the frequency that elicits a response at the lowest sound pressure level, SPL) that were matched to the spectrogram frequency bands. Then, the SFIE model (Carney and McDonough, 2019) was used to simulate responses of two types of IC neurons: band-enhanced (BE) and band-suppressed (BS) neurons, as described by modulation transfer functions, average rates as a function of modulation frequency in response to sinusoidally amplitude-modulated sounds (Kim et al., 2020). BE IC neurons are excited by amplitude-modulated stimuli with modulation frequencies near the peak of the modulation transfer function (MTF), whereas BS IC neurons are suppressed by stimuli that are modulated near a trough frequency in the MTF. The IC models had peak or trough modulation frequencies in the MTFs that were set to 100 Hz, which is near the center of the distribution for MTFs recorded in the IC (Kim et al., 2020). The IC-model feature vector consisted of responses from 20 BE and 20 BS neurons, with CFs ranging from 200 Hz to 8 kHz to match the frequencies used for the spectrogram analysis (Fig. 2E, F). BE and BS responses were concatenated, resulting in a 40-variable feature vector at each time point. The responses of the neural models were also down sampled to 128-Hz for correlation analysis with the EEG signals. Code for the AN and IC models is available at https://urhear.urmc.rochester.edu.

TRF calculation and EEG prediction

The TRF is a linear transformation from a stimulus feature vector, $S(t)$, to the neural response vector, $R(t)$, i.e.,

$$R(t) = TRF * S(t), \tag{2}$$

where * represents the convolution operator (Crosse et al., 2016). The TRF for each feature is calculated over a series of time lags between the stimulus and the response, producing a set of temporal TRF weights for each EEG channel. To estimate the TRF we used ridge regression (see Crosse et al., 2016 for details on the TRF; code available at https://github.com/mickcrosse/mTRF-Toolbox). In brief, this involves solving for the TRF using the following equation:

$$TRF = \left(S^T S + \lambda I\right)^{-1} S^T r,$$

where S is the lagged time series of the stimulus property, $s(t)$, and is defined as follows:

$$\mathbf{S} = \begin{bmatrix} s(1-\tau_{\min}) & s(-\tau_{\min}) & \cdots & s(1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & s(1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & s(1) \\ s(T) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & s(T) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & s(T) & s(T-1) & \cdots & s(T-\tau_{\max}) \end{bmatrix}.,$$

where the values $\tau_{\min}$ and $\tau_{\max}$ represent the minimum and maximum time lags (in samples), respectively. In S, each time lag is arranged column-wise. The center column, beginning with s(1), represents zero lag with columns to the right representing positive lag and left-side columns representing negative lags. Non-zero lags are padded with zeros to ensure causality (Mesgarani, et al., 2009). The window over which the TRF is calculated is defined as $\tau_{\text{window}} = \tau_{\max} - \tau_{\min}$ and the dimensions of S are thus $T \times \tau_{\text{window}}$ (where T is the total length of the stimulus/data used for fitting). To include the constant term (y-intercept) in the regression model, a column of ones is concatenated to the left of S. The neural response data is organized into a matrix r with the N EEG channels arranged column-wise (i.e., a $T \times N$ matrix). The resulting TRF, w, is a $\tau_{\text{window}} \times N$ matrix with each column representing the univariate mapping from s to the neural response at each channel. Λ is a regularization parameter that controls for overfitting (see below for how this was determined). TRFs were estimated separately for each stimulus feature representation. For initial inspection (Fig. 3), $\tau_{\min}$ and $\tau_{\max}$ were set to −500 to 500 ms, respectively. Subsequently, the window of analysis was narrowed to 0 to 275 ms for prediction analyses.

Nested cross-validation was completed to first select the optimal ridge parameter that would prevent overfitting (a detailed description of this process is provided in Crosse et al., 2016) and to assess the EEG's sensitivity to different speech features or combinations of features. To summarize, data were split into training (19 trials) and testing (1 trial) sets. On the training set, 'leave-one out' cross-validation was used in which TRF models were fit to all but one trial for each ridge-parameter value ($\lambda = 10^{-6}$,10^{-4}, 10^{-2}…10^{6}). TRFs were averaged and used to predict responses from the remaining trial. This process was repeated until all trials had been left out and all ridge parameters within the predetermined range were exhausted. The best ridge parameter was chosen based on the correlation value between predicted and measured EEG data and then used to fit a new TRF on all training trials. This TRF was used to predict the EEG responses to the remaining test trial. Training and testing steps were repeated until all trials had been used as the test trial. TRFs were optimized and tested separately for each speech feature model and participant.

Assessment of Model Performance

The Pearson's correlation coefficient, *r*, between predicted and measured EEGs was computed as the dependent measure used to assess how well each feature or groups of features were represented in the EEG. As such, *r* was used to compare EEG prediction accuracy between the different TRF models using a forward selection approach. Briefly, the logic here is that if adding a feature to an existing model improves EEG prediction accuracy, then that EEG is encoding that feature independent the other features. A repeated measures ANOVA test was used to compare distributions of prediction correlation values across the different TRF models. Post-hoc comparisons between TRF models were done using Bonferroni corrected paired t-tests.

Results

High-density (128 channels) EEG responses were collected from 19 subjects as they listened to excerpts from an audiobook (*The Old Man and the Sea* by Ernest Hemingway) containing narrative speech from an American male speaker. We used these responses and several features derived from the speech heard by the participants to fit TRF models and predict unseen EEG. In particular, we were interested to explore whether simulated IC responses driven by the stimulus could better predict the EEG compared to previously used features derived directly from the stimulus. Pearson's correlation coefficients, *r*, between the predicted and measured EEGs were used to assess how well each TRF model predicted the EEG.

Twelve electrode channels over the frontocentral region of the scalp were used for analysis (Fig. 4, blue dots). Analysis channels were chosen based on examining the distribution of prediction correlations across the scalp for all four TRF models. The distributions were not significantly different ($p>0.05$, ANOVA) allowing for a single set of 12 electrodes with the highest prediction correlations to be chosen that did not bias the results towards any of the models (consistent with the approach in DiLiberto et al., 2015).

Comparing Performance of Feature-Specific TRFs

The grand means of prediction correlation values were compared across the different TRF models (Fig. 5A). We first performed a one-way repeated-measures ANOVA with factor TRF model which revealed a significant main effect ($F_{(4,72)}=30.87$, $p=5.8\times10^{-15}$), indicating that some models were better at predicting EEG than others. The IC-response mTRF outperformed the envelope and spectrogram mTRFs (IC vs Envelope: $T_{(18)} = 4.63$, $p = 3.6\times10^{-7}$; IC vs. Spectrogram: $T_{(18)} = 4.63$, $p = 4.5\times10^{-7}$). The IC-response mTRF performance was not significantly different from the AB-envelope-derived TRF ($T_{(18)}$ = 0.062, p>0.5).

Because both the AB envelope and IC response predicted the EEG with comparable performance, we explored whether the two speech representations predicted different aspects of the EEG signal. To test this hypothesis, we analyzed the correlation between the predictions derived from each of these models and compared that to the correlation between envelope- and spectrogram-based prediction. For comparison, the EEG predicted from the envelope-derived and spectrogram-derived TRFs were strongly correlated (r=0.79), suggesting that these TRFs predict similar features of the EEG signal. In contrast, there was a significantly weaker correlation between the AB-envelope and IC predicted EEGs (r = 0.59; $T_{(18)}$ = 6.59, p = 3.4×10^{-6}), suggesting that these two speech representations were capturing more complementary aspects of the EEG compared to envelope and spectrogram model. Given this evidence, we fit a joint model (AB+IC) and hypothesized that it would predict both aspects of EEG, thus improving predictions of the overall EEG signal. As expected, the combined AB+IC mTRF predicted the EEG significantly better than its constituent features (Fig. 5A; AB+IC vs AB: $T_{(18)}$ = 4.82, p = 0.0064; AB+IC vs. IC: $T_{(18)}$ = 6.77, p = 0.0076).

Analyzing Inter-subject Variability

To better assess model performance given inter-subject variability, the recorded *r*-values for each TRF-model were plotted individually and compared across subjects (Fig. 5B). Although the results show variability across subjects, the AB+IC mTRF model produced higher correlation coefficients than all other TRF models across 18 of 19 subjects. Further, the envelope- and spectrogram-TRF models consistently performed the poorest of the five models.

Discussion

Auditory stimuli such as speech undergo substantial processing as they ascend from the cochlea to the cortex. Despite awareness of such transformations and despite the ability to extract subcortical responses to continuous speech (Forte et al., 2017; Maddox and Lee, 2018; Polonenko and Maddox, 2021) to our knowledge subcortical processing above the level of cochlear filters (Kulasingham et al., 2020; Gillis et al., 2021; Weineck et al., 2022) has not been incorporated into EEG analyses of continuous speech. Here, we have shown that including subcortical processing, in the form of auditory-midbrain model responses, allows for the derivation of TRFs that predict EEG responses with higher accuracy than previous TRFs based on acoustic features derived directly from the speech stimuli.

More specifically, in this work, the SFIE midbrain model was used to produce model IC responses to speech which, in turn, were used to derive TRFs for predicting EEG responses. The Pearson’s correlation coefficient between predicted and measured EEG was used to evaluate if incorporating the IC responses into the mTRF pipeline could produce better EEG predictions than those based on the envelope, spectrogram, or AB-envelope derived from the speech. Such acoustic features have previously been reported as successful methods for predicting EEG (Lalor et al., 2009, Ding and Simon, 2012, DiLiberto et al., 2015;

Drennan and Lalor, 2019). However, those approaches generally ignore the substantial amount of processing that occurs before the input signal reaches the cortex and influences the EEG signal. Thus, incorporating a model of the IC into the pipeline could lead to better predictions of the EEG. Similar hierarchical, predictive models have been implemented, for example, in the nonhuman primate visual neuroscience literature (Mineault et al., 2012). As expected, the IC-response TRF outperformed the envelope and spectrogram TRF models; but, it did not predict the EEG better than the AB envelope-derived TRFs. However, the IC-response and AB-envelope-based predictions were not as highly correlated with each other as envelope and spectrogram predictions and combining IC responses and the AB envelope features into one mTRF model provided the best of the tested mappings to EEG responses. This suggests that the IC responses and the AB envelope are capturing unique information from the EEG signal and supports our original hypothesis. While the IC model is capturing nonlinearities of the subcortical auditory system (e.g., saturating transduction, compressive amplification, neural adaptation), we think the AB envelope could be capturing cortical nonlinearities such as non-monotonic rate-level functions (Schreiner et al., 1992).

IC neurons lend themselves well to TRF analyses of speech encoding, as most cells are rate-tuned to both audio frequency and amplitude-modulation (AM) frequency. The display of spectral tuning is often characterized by a strong sensitivity to a certain frequency, or best frequency (BF), while low-frequency AM tuning is often characterized by a best modulation frequency (BMF; Krishna and Semple, 2000; Joris et al., 2004; Nelson and Carney, 2007). Further, a majority of IC BMFs fall within the range of voice pitch (Langner, 1992) making them suitable for analyzing speech stimuli (Delgutte et al., 1998; Carney et al., 2015). In the current study, we have selected a single BMF near the f0 of our speaker. While it is possible that the selection of a single modulation frequency might negatively impact our ability to encode the time-varying f0 of the speaker, the neural modulation filters are quite broad (Q~1) and should be sensitive to the range of frequency fluctuations in the speech. It seems likely that incorporating a bank of population responses spanning multiple BMFs (and likewise increasing number of simulated CFs within each bank) would improve EEG predictions, but we worry about our data being underpowered or overfit. One could conceivably test the specificity of the responses by fitting IC model parameters to EEG data recorded during speech from two speakers with very different f0s.

BMFs are best represented by the peaks (or troughs) in MTFs, which depict average discharge rate as a function of AM frequency and can be classified as either band-enhanced, exhibiting an increased discharged rate at BMF, or band-suppressed, exhibiting a decreased firing rate at BMF. The SFIE model simulates responses from both cell types, providing a robust population response to our stimulus that is intended to represent a simplified midbrain encoding of complex sounds. Given this evidence, it is not surprising that the IC-response mTRF was able to represent the EEG significantly better than previous models, which typically only incorporate acoustic features of the speech stimuli (albeit sometimes passed through a very simple gammatone filter model of the cochlea). In particular, it is likely that the addition of AM tuning properties, typically present at the level of the midbrain provides, has provided much of the improvement in modeling the EEG.

The analysis presented in the current work focused specifically on EEG responses to speech. This was a natural choice given the importance of speech in everyday life, as well as the wealth of previous research aimed at modeling EEG responses to natural, continuous speech (Lalor and Foxe, 2010; Ding and Simon, 2014; Myers et al., 2019; Brodbeck and Simon, 2020). And while the inclusion of the IC-model within the framework has improved our EEG modeling, we think it is possible that the use of speech may have produced more modest benefits than would have derived for other types of audio stimuli. This is because envelope/spectrogram modeling has already been shown to work quite well for speech in particular – given the large amplitude modulation depths seen in natural speech and the importance of envelopes for speech intelligibility in general (Shannon et al., 1995; Smith et al., 2002). As such, it might be the case that the framework we have introduced here – including an IC model in an EEG modeling pipeline – would produce larger benefits in the context of other stimuli. In particular, greater improvements in EEG prediction accuracy might derive for signals with a more heterogeneous pattern of amplitude modulations across frequencies, such as music. Indeed, EEG tracking of the envelope of music has often been shown to be much weaker than for speech (Zuk et al., 2021). Future work will apply the framework presented here to modeling EEG responses to music.

Another interesting possible use of the framework presented here could be for the refinement of auditory subcortical models themselves. Although models of the auditory periphery have been used successfully to predict human speech perception (Heinz, 2010; Moncada-Torres et al., 2017; Bruce, 2017; Zaar and Carney, 2022), such models are often fit using data recorded from non-human mammals (Carney, 1993; Zhang et al., 2001; Zilany and Bruce, 2006, 2007; Zilany et al., 2009, 2014). While these models should work well given evolutionary homologies in the midbrain (Webster, 1992; Grothe et al., 2004; Woolley and Portfors, 2013), it is also true that speech is a particularly special signal for humans. As such, the processing of speech by humans involves predictions (Kutas and Hillyard, 1980, 1984; Leonard et al., 2016; Zoefel, 2018; Broderick et al., 2018) and attention (Cherry, 1953; McDermott, 2009; Mesgarani and Chang, 2012; Golumbic et al., 2013; O’Sullivan et al., 2015) effects that are unlikely to be present in non-human animals. One could imagine constraining and refining parameters of a subcortical model based on EEG prediction accuracy to have those subcortical models better capture human-specific subcortical auditory processing. Of course, validation of such models would be extremely important given the relatively low SNR of EEG and the risk of overfitting, and could perhaps be carried out using intracranial recordings in human neurosurgical patients.

Two additional issues are worth considerating, both of which relate to filtering EEG signals. First, it is important to note that filters affect the interpretation of TRF components (and TRF component latencies in particular). This is because the TRFs fit to filtered EEG contain a convolution of the filter response and the impulse response of the neural system (de Cheveigne and Nelken 2019; see Fig. 3a inset showing the impulse and step response for filters used in this study). As such, interpreting TRF components at different latencies as reflecting different stage of processing – as has been done for ERPs/ERFs (e.g., Salmelin, 2007) – is not straightforward. An alternative approach for interpreting TRF models, and one that we and others have used regularly, is to model EEG responses based on different explicitly defined acoustic and linguistic speech representations (e.g., Di Liberto et al, 2015;

Brodbeck et al., 2018). As stated in the introduction, our goal in the present study was to improve the modeling of the acoustic features of speech in particular. Second, filtering high-frequency and difficult-to-explain features from the EEG signal increases the signal-to-noise ratio (SNR) of the TRF at the expense of a fuller explanatory model. Indeed, it seems that activity outside the frequency range tested in the current study can be useful to decode stimulus features in some individuals (Synigal et al, 2020), but generally has low signal strength on the scalp. As such, varying how one filters one's data can produce differences in prediction accuracy simply because one is essentially varying the amount of signal one is trying to predict. In our study, we don't think this is a major concern, as we always compared the performance of models based on different features in predicting EEG data that had been filtered in the same way. Other approaches have been developed in an effort to contend with the complex effects of filtering on modeling neural data. We were mostly interested in the interpretability of our models, so we elected to use linear TRFs. However future work might explore other methods such as mTRF stimulus reconstruction (Crosse et al 2016), canonical components analysis (CCA; de Cheveigne et al., 2018), or back-to-back regression (King et al 2020).

In sum, we have shown that incorporating a well-established model of IC neuronal activity can improve models of EEG responses to natural speech. Given the relatively low SNR of EEG, any improvements in the ability to model that EEG could have important benefits in electrophysiological research on speech and language processing. Future work will aim to extend the framework to other auditory stimuli.

Acknowledgements

This work was supported by NIH-DC010813 and the Del Monte Institute for Neuroscience at the University of Rochester.

References

Beagley HA, & Knight JJ (1967). Changes in auditory evoked response with intensity. The Journal of Laryngology & Otology, 81(8), 861–873. [PubMed: 6036752]

Brodbeck C, Hong LE, & Simon JZ (2018). Rapid transformation from auditory to linguistic representations of continuous speech. Current Biology, 28(24), 3976–3983. [PubMed: 30503620]

Brodbeck C, Bhattasali S, Heredia AAC, Resnik P, Simon JZ, & Lau E (2022). Parallel processing in speech perception with local and global representations of linguistic context. Elife, 11, e72056. [PubMed: 35060904]

Brodbeck C, & Simon JZ (2020). Continuous speech processing. Current Opinion in Physiology, 18, 25–31. [PubMed: 33225119]

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, & Lalor EC (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Current Biology, 28(5), 803–809. [PubMed: 29478856]

Bruce IC (2017). Physiologically based predictors of speech intelligibility. Acoustics Today, 13(1), 28–35.

Buzsáki G, Anastassiou CA, & Koch C (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nature reviews neuroscience, 13(6), 407–420. [PubMed: 22595786]

Carney LH (1993). A model for the responses of low-frequency auditory-nerve fibers in cat. The Journal of the Acoustical Society of America, 93(1), 401–417. [PubMed: 8423257]

Carney LH, Li T, & McDonough JM (2015). Speech coding in the brain: representation of vowel formants by midbrain neurons tuned to sound fluctuations. Eneuro, 2(4).

Carney LH, & McDonough JM (2019). Nonlinear auditory models yield new insights into representations of vowels. Attention, Perception, & Psychophysics, 81(4), 1034–1046.

Cherry EC (1953). Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America, 25(5), 975–979.

Connolly JF, & Phillips NA (1994). Event-Related Potential Components Reflect Phonological and Semantic Processing of the Terminal Word of Spoken Sentences. Journal of Cognitive Neuroscience, 6(3), 256–266. [PubMed: 23964975]

Crosse MJ, Di Liberto GM, Bednar A, & Lalor EC (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience, 10.

Crosse MJ, Zuk NJ, Di Liberto GM, Nidiffer AR, Molholm S, & Lalor EC (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research. Frontiers in Neuroscience, 15.

Davis PA (1939). Effects of acoustic stimuli on the waking human brain. Journal of neurophysiology, 2(6), 494–499.

de Cheveigné A, & Nelken I (2019). Filters: When, Why, and How (Not) to Use Them. Neuron, 102(2), 280–293. 10.1016/j.neuron.2019.02.039 [PubMed: 30998899]

de Cheveigné A, Wong DD, Di Liberto GM, Hjortkjær J, Slaney M, & Lalor E (2018). Decoding the auditory brain with canonical component analysis. NeuroImage, 172, 206–216. [PubMed: 29378317]

Delgutte B, Hammond BM, & Cariani PA (1998). Neural coding of the temporal envelope of speech: relation to modulation transfer functions. Psychophysical and physiological advances in hearing, 595–603.

Di Liberto Giovanni M., O'Sullivan James A., & Lalor Edmund C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology, 25(19), 2457–2465. [PubMed: 26412129]

Ding N, & Simon JZ (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophysiology, 107(1), 78–89. 10.1152/jn.00297.2011 [PubMed: 21975452]

Ding N, & Simon JZ (2014). Cortical entrainment to continuous speech: functional roles and interpretations. Frontiers in human neuroscience, 8, 311. [PubMed: 24904354]

Drennan DP, & Lalor EC (2019). Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity. Eneuro, 6(3), ENEURO.0082–19.2019.

Forte AE, Etard O, & Reichenbach T (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. elife, 6, e27203. [PubMed: 28992445]

Gevins A, Leong H, Smith ME, Le J, & Du R (1995). Mapping cognitive brain function with modern high-resolution electroencephalography. Trends in neurosciences, 18(10), 429–436. [PubMed: 8545904]

Gillis M, Vanthornhout J, Simon JZ, Francart T, & Brodbeck C (2021). Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. Journal of Neuroscience, 41(50), 10316–10329. [PubMed: 34732519]

Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, … & Schroeder CE (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron, 77(5), 980–991. [PubMed: 23473326]

Grothe B, Carr CE, Casseday JH, Fritzsch B, & Köppl C (2004). The evolution of central pathways and their neural processing patterns. In: Evolution of the vertebrate auditory system, Editors: Manley GA, Fay RR, Popper AN; Springer Handbook of Auditory Research (SHAR, Vol. 22), pp. 289–359.

Hamilton LS, & Huth AG (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. Language, cognition and neuroscience, 35(5), 573–582. [PubMed: 32656294]

Handy TC (Ed.). (2005). Event-related potentials: A methods handbook. MIT press.

Heinz MG (2010). Computational modeling of sensorineural hearing loss. In Computational models of the auditory system (pp. 177–202). Springer, Boston, MA.

Joris PX, Schreiner CE, & Rees A (2004). Neural Processing of Amplitude-Modulated Sounds. Physiological Reviews, 84(2), 541–577. [PubMed: 15044682]

Kim DO, Carney L, & Kuwada S (2020). Amplitude modulation transfer functions reveal opposing populations within both the inferior colliculus and medial geniculate body. Journal of Neurophysiology, 124(4), 1198–1215. [PubMed: 32902353]

Krishna BS, & Semple MN (2000). Auditory Temporal Processing: Responses to Sinusoidally Amplitude-Modulated Tones in the Inferior Colliculus. Journal of Neurophysiology, 84(1), 255–273. [PubMed: 10899201]

Kulasingham JP, Brodbeck C, Presacco A, Kuchinsky SE, Anderson S, & Simon JZ (2020). High gamma cortical processing of continuous speech in younger and older listeners. NeuroImage, 222, 117291. [PubMed: 32835821]

Kutas M, & Hillyard SA (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207(4427), 203–205. [PubMed: 7350657]

Kutas M, & Hillyard SA (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature, 307(5947), 161–163. [PubMed: 6690995]

Lalor EC, Power AJ, Reilly RB, & Foxe JJ (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. Journal of Neurophysiology, 102(1), 349–359. [PubMed: 19439675]

Lalor EC, & Foxe JJ (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European journal of neuroscience, 31(1), 189–193. [PubMed: 20092565]

Langner G (1992). Periodicity coding in the auditory system. Hearing Research, 60(2), 115–142. [PubMed: 1639723]

Leonard MK, Baud MO, Sjerps MJ, & Chang EF (2016). Perceptual restoration of masked speech in human cortex. Nature communications, 7(1), 1–9.

Niedermeyer E, & da Silva FL (Eds.). (2005). Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins.

Maddox RK, & Lee AK (2018). Auditory brainstem responses to continuous natural speech in human listeners. Eneuro, 5(1).

McDermott JH (2009). The cocktail party problem. Current Biology, 19(22), R1024–R1027. [PubMed: 19948136]

Mesgarani N, & Chang EF (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233–236. [PubMed: 22522927]

Mesgarani N, David SV, Fritz JB, & Shamma SA (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. Journal of neurophysiology, 102(6), 3329–3339. [PubMed: 19759321]

Mineault PJ, Khawaja FA, Butts DA, & Pack CC (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. Proceedings of the National Academy of Sciences, 109(16), E972–E980.

Moncada-Torres A, van Wieringen A, Bruce IC, Wouters J, & Francart T (2017). Predicting phoneme and word recognition in noise using a computational model of the auditory periphery. The Journal of the Acoustical Society of America, 141(1), 300–312. [PubMed: 28147586]

Myers BR, Lense MD, & Gordon RL (2019). Pushing the envelope: Developments in neural entrainment to speech and the biological underpinnings of prosody perception. Brain Sciences, 9(3), 70. [PubMed: 30909454]

Murakami S, & Okada Y (2006). Contributions of principal neocortical neurons to magnetoencephalography and electroencephalography signals. The Journal of physiology, 575(3), 925–936. [PubMed: 16613883]

Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, & Alho K (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature, 385(6615), 432–434. [PubMed: 9009189]

Nelson PC, & Carney LH (2004). A phenomenological model of peripheral and central neural responses to amplitude-modulated tones. The Journal of the Acoustical Society of America, 116(4), 2173–2186. [PubMed: 15532650]

Nelson PC, & Carney LH (2007). Neural rate and timing cues for detection and discrimination of amplitude-modulated tones in the awake rabbit inferior colliculus. Journal of neurophysiology, 97(1), 522–539. [PubMed: 17079342]

O'Sullivan Power, Mesgarani Rajaram, Foxe Shinn-Cunningham, Slaney Shama, Lalor 2015 Cerebral Cortex

Polonenko MJ, & Maddox RK (2021). Exposing distinct subcortical components of the auditory brainstem response evoked by continuous naturalistic speech. Elife, 10, e62329. [PubMed: 33594974]

Regan D (1989). Human brain electrophysiology. Evoked potentials and evoked magnetic fields in science and medicine. Elsevier.

Schreiner CE, Mendelson JR, & Sutter ML (1992). Functional topography of cat primary auditory cortex: representation of tone intensity. Experimental brain research, 92, 105–122. [PubMed: 1486946]

Shannon RV, Zeng FG, Kamath V, Wygonski J, & Ekelid M (1995). Speech recognition with primarily temporal cues. Science, 270(5234), 303–304. [PubMed: 7569981]

Smith ZM, Delgutte B, & Oxenham AJ (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature, 416(6876), 87–90. [PubMed: 11882898]

Sur S, & Sinha V (2009). Event-related potential: An overview. Industrial Psychiatry Journal, 18(1), 70. [PubMed: 21234168]

Synigal SR, Teoh ES, & Lalor EC (2020). Including measures of high gamma power can improve the decoding of natural speech from EEG. Frontiers in human neuroscience, 14, 130. [PubMed: 32410969]

Webster DB (1992). An overview of mammalian auditory pathways with an emphasis on humans. In: The mammalian auditory pathway: Neuroanatomy, Editors: Webster DB, Fay RR, Springer, pp.1–22.

Weineck K, Wen OX, & Henry MJ (2022). Neural synchronization is strongest to the spectral flux of slow music and depends on familiarity and beat salience. ELife, 11, e75515. [PubMed: 36094165]

Woolley SM, & Portfors CV (2013). Conserved mechanisms of vocalization coding in mammalian and songbird auditory midbrain. Hearing research, 305, 45–56. [PubMed: 23726970]

Zaar J, & Carney LH (2022). Predicting speech intelligibility in hearing-impaired listeners using a physiologically inspired auditory model. Hearing Research, 108553.

Zhang X, Heinz MG, Bruce IC, & Carney LH (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. The Journal of the Acoustical Society of America, 109(2), 648–670. [PubMed: 11248971]

Zilany MS, & Bruce IC (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. The Journal of the Acoustical Society of America, 120(3), 1446–1466. [PubMed: 17004468]

Zilany MS, & Bruce IC (2007). Representation of the vowel/ε/in normal and impaired auditory nerve fibers: model predictions of responses in cats. The Journal of the Acoustical Society of America, 122(1), 402–417. [PubMed: 17614499]

Zilany MS, Bruce IC, & Carney LH (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. The Journal of the Acoustical Society of America, 135(1), 283–286. [PubMed: 24437768]

Zilany MS, Bruce IC, Nelson PC, & Carney LH (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. The Journal of the Acoustical Society of America, 126(5), 2390–2412. [PubMed: 19894822]

Zoefel B (2018). Speech entrainment: Rhythmic predictions carried by neural oscillations. Current Biology, 28(18), R1102–R1104. [PubMed: 30253150]

Zuk NJ, Murphy JW, Reilly RB, & Lalor EC (2021). Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. PLoS computational biology, 17(9), e1009358. [PubMed: 34534211]

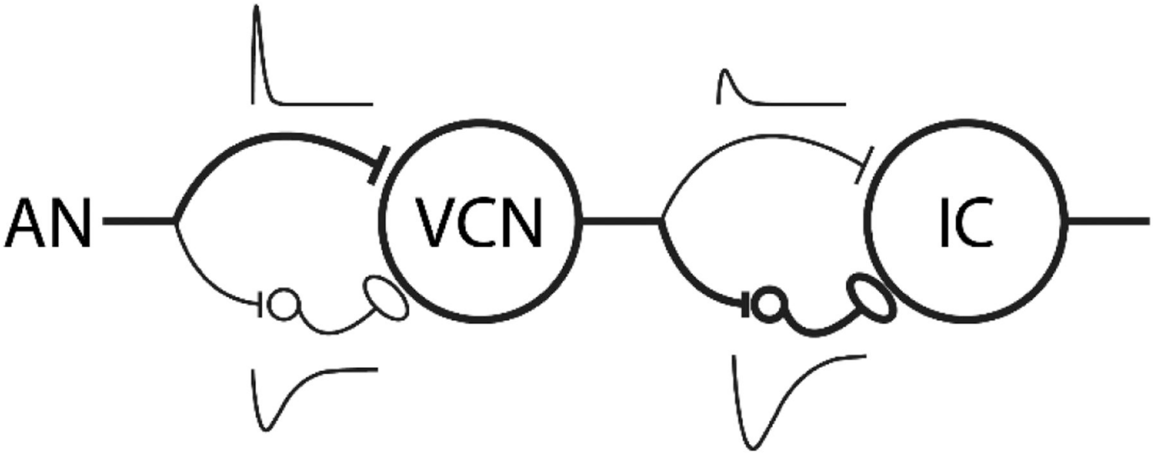

Figure 1:
Schematic diagram of the same-frequency inhibition and excitation (SFIE) model. A single model AN fiber provides the postsynaptic cell with both excitatory and inhibitory input, via an inhibitory interneuron. The thickness of the lines corresponds to the relative strength of the inhibition and excitation at each level. Alpha functions representing the assumed membrane and synaptic properties are also shown above or below corresponding synapses. Adapted from Nelson and Carney (2004).

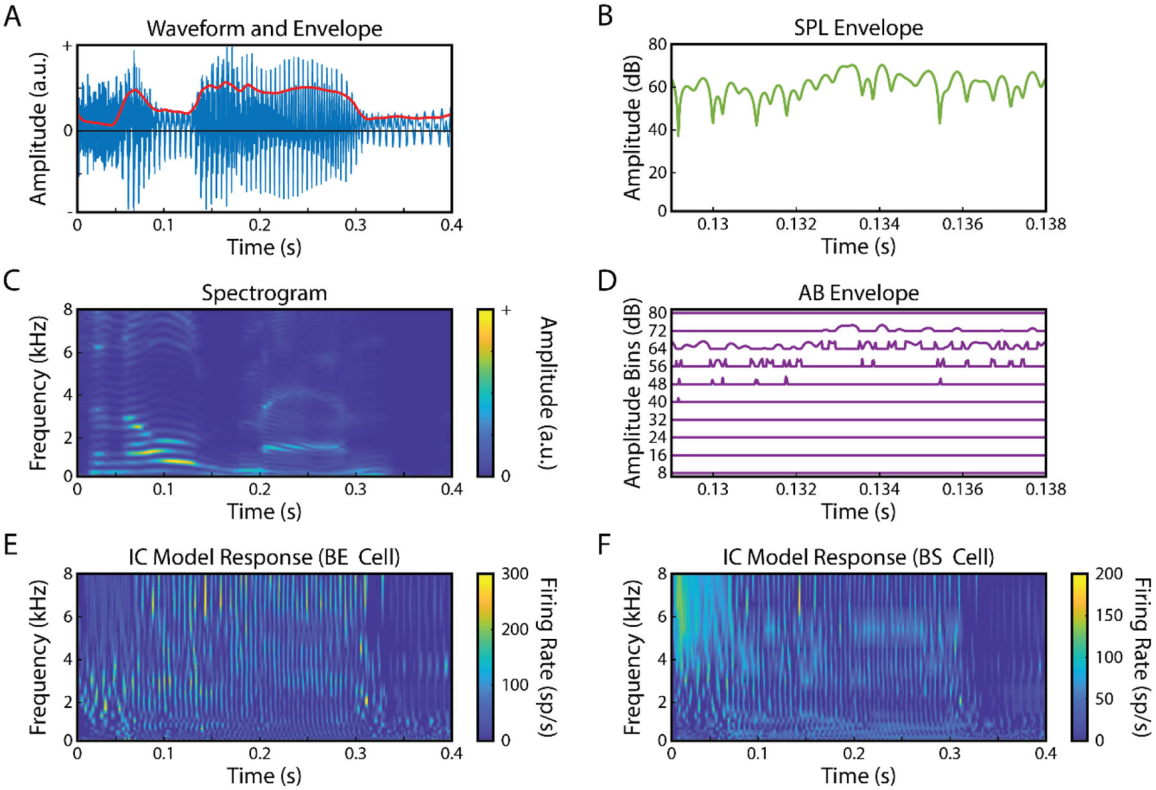

Figure 2: Representations of speech stimulus.

Four different representations of the speech stimuli were used to derived TRFs and predict EEG responses: **A)** broadband envelope, **B)** SPL envelope used for amplitude binning, **C)** spectrogram, **D)** amplitude-binned envelope and **E, F)** IC BE and BS model responses.

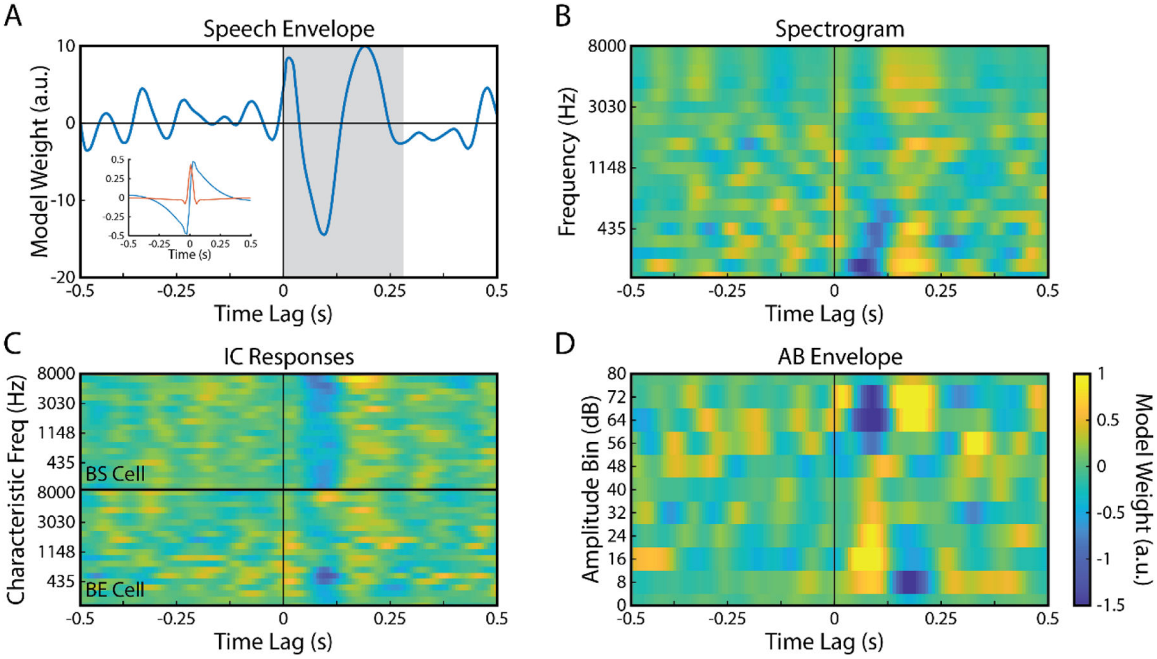

Figure 3: TRF model weights plotted over wide range of time-lags.

mTRF plotted for the **A)** Envelope (inset: filter impulse and step response), **B)** Spectrogram, **C)** IC responses provided by SFIE model; top half shows IC BS responses; bottom half shows IC BE responses, and **D)** AB envelope representations of the speech stimulus. Time lags ranged from −500 to 500-ms and mTRFs were averaged across the 12 electrodes of interest (see Fig. 4). Shaded region in **A)** indicates the 275-ms analysis window used.

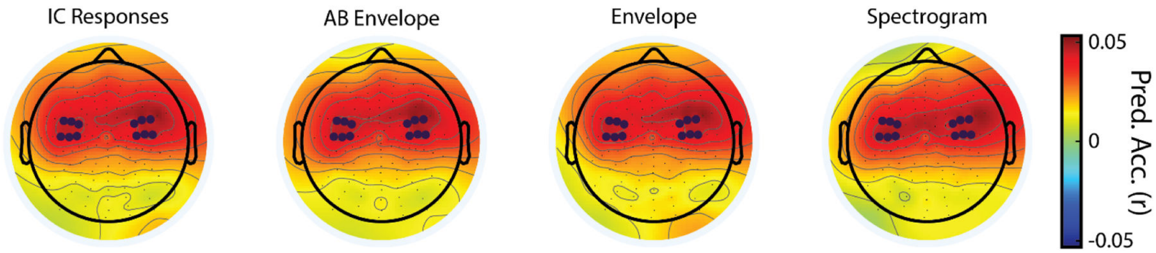

Figure 4: Topographical representation of prediction correlations between predicted and actual EEGs.

Topographical distribution of prediction correlation values plotted onto a schematic diagram of the scalp. The electrodes used in analysis are emphasized in dark blue. The colormap indicates the Pearson correlation between the predicted and recorded (unaveraged) EEG.

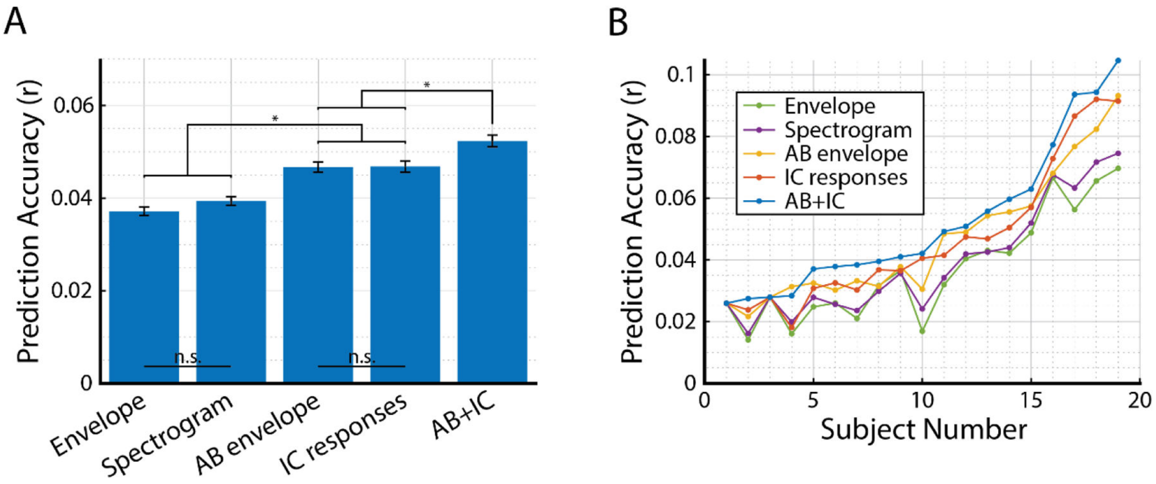

Figure 5: Comparison of IC-model derived mTRF to acoustic feature derived TRFs.

A) The grand mean prediction correlation values for each type of TRF (mean ±SEM). The IC-model and AB-envelope TRF models were significantly higher than the envelope and spectrogram models. There was no significant difference between envelope and spectrogram prediction accuracy or between IC-model and AB-envelope prediction accuracy (p>0.05). The combined AB+IC TRF outperformed all other models. **B)** Correlation values were plotted for each subject individually. Data was sorted according to the prediction correlation values for the AB+IC model. Despite some variability across subjects, it is clear that the AB+IC mTRF model performs the best and the envelope TRF has the poorest performance across subjects.