The effects of data quantity on performance of temporal response function analyses of natural speech processing

Juraj Mesik[1*] and Magdalena Wojtczak[1]

[1] Department of Psychology, University of Minnesota, Minneapolis, MN, USA

*** Correspondence:**
Juraj Mesik
mesik002@umn.edu

Abstract

In recent years, temporal response function (TRF) analyses of non-invasive recordings of neural activity evoked by continuous naturalistic stimuli have become increasingly popular for characterizing response properties within the auditory hierarchy. However, despite this rise in TRF usage, relatively few educational resources for these tools exist. Here we use a dual-talker continuous speech paradigm to demonstrate how a key parameter of experimental design, the quantity of acquired data, influences TRF analyses fit to either individual data (subject-specific analyses), or group data (generic analyses). We show that although model performance monotonically increases with data quantity, the amount of data required to achieve significant prediction accuracies can vary substantially based on whether the fitted model contains densely (e.g., acoustic envelope) or sparsely (e.g., lexical surprisal) spaced features, especially when the goal of the analyses is to capture the aspect of neural responses that co-vary with the amplitude of the modelled features. Moreover, we demonstrate that generic models can exhibit high performance on small amounts of test data (4-8 min), as long as they are trained on a sufficiently large data set. As such, they may be particularly useful for clinical and multi-task study designs. Finally, we show that the regularization procedure used in fitting TRF models can interact with the quantity of data used to fit the models, with larger training quantities resulting in systematically larger TRF amplitudes. Together, demonstrations in this work should aid the learning process of new users of TRF analyses, and in combination with other tools, such as piloting and power analyses, may serve as a detailed reference for choosing acquisition duration in future studies.

Keywords: electroencephalography; temporal response function (TRF); forward modeling; generic model; continuous speech processing; cortical tracking of speech; envelope processing; semantic processing

1. Introduction

Characterizing how acoustic and linguistic features are encoded in human cortex is a key goal of auditory cognitive neurosciences and neurolinguistics. In recent years, substantial progress has been made in utilizing noninvasive electroencephalographic (EEG) and magnetoencephalographic (MEG) responses to continuous speech in order to uncover a diverse set of neural signatures associated with different aspects of language processing. These range from responses to relatively low-level features associated with the speech envelope (e.g., Ding and Simon, 2012; Power et al., 2012; Kong et al., 2014), mid-level features implicated in phonemic processing (e.g., Di Liberto et al., 2015, 2019; Brodbeck et al., 2018), and higher-order linguistic features associated with semantic and syntactic processing (e.g., Broderick et al., 2018; Weissbart et al., 2019; Donhauser and Baillet, 2020; Mesik et al., 2021).

An important catalyst for this work has been the popularization of regularized linear regression methods for mapping the relationship between features in the stimulus space and the brain response (Lalor and Foxe, 2010; Crosse et al., 2016). This relationship can be characterized in both the *forward* direction, mapping from the stimulus to the brain response, and the *backward* direction, specifying how to reconstruct stimulus features from patterns of brain activity. The forward modeling approach models the continuous M/EEG data as a convolution between a set of to-be-estimated feature-specific impulse responses, known as the *temporal response functions* (TRFs), with the known time courses of the corresponding features (e.g., acoustic envelope, lexical surprisal, etc). These analyses contrast with the more traditional event-related potential (ERP) method, which relies on averaging of hundreds of identical trials in order to estimate the stereotypical neural response for a given stimulus (e.g., Luck, 2005; Woodman, 2010). Unlike ERP methods, which rely on repetition, the TRF approach allows for analyzing brain responses to naturalistic time-varying stimuli, including continuous speech and music.

While the popularity of TRF methods has increased substantially, there remains a relative lack of literature exploring how these methods perform in the context of EEG and MEG analyses under various constraints, such as the type of features utilized (e.g., temporally sparse vs. dense) or the quantity of the data to which the models are applied. This information resource gap has more recently received some attention (Sassenhagen, 2019; Crosse et al., 2021), but published work has provided a broader overview of issues in TRF research (e.g., effects of correlated variables, missing features, preprocessing choices, etc.) without a more thorough examination of any one issue. As such, there is a continued need for further methodological resources to allow researchers interested in adopting TRF techniques to optimize the experimental design for their efficient use.

One of the most fundamental decisions in study design is the choice of how much data to collect (i.e., number of subjects and acquisition duration per subject). This decision has broad consequences affecting the cost of the study, complexity of applicable models, data quality (due to fatigue / comfort level changes over the course of long experimental sessions), and others. With respect to TRF modeling, understanding how analysis outcomes are influenced by data quantity is important for making decisions about duration of data acquisition. Specifically, at the low end of the spectrum (small amounts of data), TRF models may be unable to isolate neural responses of interest due to poor data signal-to-noise ratio (SNR) and/or limited sampling of the stimulus feature space used in the analysis. At the upper end of the spectrum (large amounts of data), model performance may saturate, making additional data wasteful both in research costs and subject discomfort. Understanding these tradeoffs is particularly important for studies of special populations, such as the elderly or clinical patients, who may not tolerate longer experimental sessions.

To date, majority of work exploring effects of data quantity on analyses of speech-evoked EEG data have focused on attention decoding, especially with *backward* TRF models (e.g., Mirkovic et al., 2015; O'Sullivan et al., 2015; Fuglsang et al., 2017; Wong et al., 2018). However, because the driving force behind the interest in attention decoding is innovation in hearing aid technologies, much of this work has focused on the performance of trained models on decoding attention using limited amount of data. In other words, majority of this work has explored the effects of data quantity at the level of model *evaluation*, rather than on the model *training* itself (but see Mirkovic et al., 2015). In the context of forward modeling, the effect of *training* data quantity on model performance has only received a limited amount of attention. Di Liberto and Lalor (2017) explored the impact of data acquisition duration on the performance of TRF models of phonemic processing, to assess whether small amounts of data can reliably support detection of phonemic responses in individuals. They showed that although models trained on data from individual participants required 30+ min to detect these signals, *generic* models derived from data from multiple participants could detect phonemic signals with as little as 10 min of data per participant. More recently, in their overview of TRF methods, Crosse et al. (2021) used simulations to demonstrate the impact of noise and data quantity on a single-feature TRF prediction accuracies and the fidelity of the derived TRFs. However, beyond these works, a more thorough exploration of TRF forward model performance in the context of more realistic listening scenarios, with a more diverse set of modelled speech features has not been performed thus far.

The goal of the present work was to provide a detailed, practical demonstration of how data quantity and model feature sparsity affect the outcome of TRF analyses in the context of measured, noninvasive EEG responses to a dual-talker continuous speech paradigm (Mesik et al., 2021). In a series of analyses, we repeatedly fit TRF models to progressively larger segments of the data, estimating both individual subject models, as well as generic models based on data pooled across multiple subjects. For each analysis, we demonstrate how data quantity influences TRF estimates, overall model prediction accuracy, and prediction accuracy attributable to the natural variation in feature amplitudes. Additionally, we show the effect of the interplay between data quantity and regularization on the amplitudes of resulting TRF estimates. Given the unique nature of each auditory study (design, analyses, etc.), we caution readers against taking our work as a sole prescription for how much data should be collected in future TRF studies. Instead, we believe our work should serve as a detailed demonstration of general patterns of TRF model performance as a function of data quantity, and a reference that should be carefully used in conjunction with other tools and sources of information (e.g., piloting and power analyses) for informing study design.

2. Materials and Methods

EEG data used in the present manuscript was acquired and previously analyzed to investigate age effects in cortical tracking of word-level features in competing speech (Mesik et al., 2021). Extensive description of the details associated with the experiment and the data are openly accessible in the original manuscript. For brevity, we highlight key aspects of this data below.

2.1 Participants

Data from 41 adult participants (18-70 years old, mean ± SD age: 41.7±14.3 years; 15 male, 26 female) was used in the present study. The broad age range was utilized to assess age effects on speech-driven EEG responses in the original study. Consequently, a subset of participants (n = 18) had mild-to-moderate hearing loss (HL), largely concentrated in the high-frequency region (≥ 4 kHz), which was compensated-for via amplification. All participants provided a written informed consent and received either course credit or monetary compensation for their participation. The procedures were approved by the Institutional Review Board of the University of Minnesota.

2.2 Stimuli

Stimuli were four public-domain short-story audiobooks (*Summer Snow Storm* by Adam Chase; *Mr. Tilly's Seance* by Edward F. Benson; *A Pail of Air* by Fritz Leiber; *Home Is Where You Left It* by Adam Chase; source: LibriVox.org) read by two male speakers (2 stories per speaker). Each story had a duration of approximately 25 minutes. Stories were pre-processed to truncate silent gaps exceeding 500 ms to 500 ms, and the levels in each one-minute segment were root-mean-square normalized and scaled to 65 dB SPL. In participants with HL, the audio was then amplified to improve audibility at frequencies affected by HL (see Mesik et al., 2021 for details of amplification). Stimuli were presented using ER1 Insert Earphones (Etymotic Research, Elk Grove Village, IL).

2.3 Experimental procedure

Participants completed two experimental runs in which they listened to pairs of simultaneously presented audiobooks narrated by different male talkers. The stories were presented at equal levels (i.e., 0 dB SNR) and were spatially co-located (i.e., diotic presentation of same stimuli to the two ears). One story was designated as the target story and participants were instructed to attend to the target talker for the duration of the experimental run, while ignoring the other talker. Runs were divided into 1-minute blocks, each of which was followed by a series of four multiple-choice comprehension questions about the target story, along with several questions about the subjects' state of attentiveness and story intelligibility. This behavioral data was not analyzed in the context of the present manuscript. In the second experimental run, participants listened to a new pair of stories spoken by the same two talkers, with the to-be-attended and to-be-ignored talker designations switched, to eliminate talker-specific effects in the analysis results. The order of the story pairs as well as the to-be-attended talker designations were counter-balanced across participants. All experimental procedures were implemented via the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) in MATLAB (Mathworks, Natick, MA, United States; version R2019a).

2.4 EEG procedure

Data were acquired using a non-invasive 64-electrode BioSemi ActiveTwo system (BioSemi B.V., Amsterdam, The Netherlands), sampled at 4096 Hz. Electrodes were placed according to the international 10-20 system. Additional external electrodes were used to obtain activity at mastoid sites, as well as a vertical electro-oculogram. Data from these external electrodes were not analyzed in the present study.

2.5 EEG preprocessing

Here we include a brief overview of pre-processing steps applied to the data. For more detailed description of pre-processing, see Mesik et al. (2021). Unless otherwise stated, pre-processing steps were implemented using the EEGLAB toolbox (Delorme and Makeig, 2004; version 14.1.2b) for MATLAB. To reduce computational load, raw data were downsampled to 256 Hz and band-pass filtered using *pop_eegfiltnew* function between 1 and 80 Hz using a zero-phase Hamming windowed sinc FIR filter (846th order, 1 Hz transition band width). Next, the data were pre-processed using the PREP pipeline (Bigdely-Shamlo et al., 2015), in order to reduce the impact of noisy channels on the referencing process. This procedure involved three stages: 1) power line noise removal via multi-taper regression, 2) iterative referencing procedure to detect noisy channels based on abnormally high signal amplitude, abnormally low correlations with neighboring channels, poor predictability of channel data based on surrounding channels, and excessive amount of high-frequency noise, and 3) spherical-interpolation of the noisy channels detected in stage 2. For this procedure, we used the default parameters outlined in Bigdely-Shamlo et al. (2015). Following up to 4 iterations of stages 2-3 (or until no further noisy channels were identified), the cleaned estimate of global mean activation was used to reference the dataset.

Subsequently, we epoched all 1-minute blocks and applied independent component analysis (ICA; Jutten and Herault, 1991; Comon, 1994) to remove components of data corresponding to muscle artifacts and other sources of noise. ICA decomposes EEG signal into a set of statistically independent components that reflect various underlying contributors to the channel data (e.g., eye blinks, different aspects of cognitive processing, etc.), allowing for removal of components driven largely by nuisance factors such as muscle activity.

The data were then band-pass filtered between 1 and 8 Hz using a Chebyshev type 2 filter (80 dB attenuation below 0.5 Hz and above 9 Hz, with 1 dB band-pass ripple), applied with the *filtfilt* function in MATLAB. This was done given the existing evidence that cortical speech processing mechanisms track speech predominantly via low-frequency dynamics in the 1-8 Hz range (e.g., Ding and Simon, 2012; Zion Golumbic et al., 2013). Finally, the data were transformed into z-scores to account for variability in overall response amplitudes due to inter-subject variability in nuisance factors such as skull thickness. Data from the first block of each run were excluded from analysis due to a small subset of participants accidentally confusing the attended and ignored speakers in the initial block (this became apparent during behavioral task following the first block, which pertained to the to-be-attended story).

2.6 TRF analyses

The time courses of speech-evoked responses, or TRFs, were extracted from the pre-processed EEG data using regularized linear regression, implemented via the mTRF Toolbox (Crosse et al., 2016, version 2.3). Briefly, TRFs are estimated by regressing a set of *n* time-lagged copies of the time series of a given speech feature (e.g., acoustic envelope) against the EEG time course at *m* channels. This results in *m* separate TRFs, each representing how the response at a given electrode site is affected at the *n* time lags relative to that feature's presentation times. This procedure can be simultaneously applied to multiple features (see section 2.6.3 for details of features used in our analyses), enabling the decomposition of EEG signals into contributions from different stages of speech processing (e.g., acoustic vs. semantic processing).

To avoid overfitting, the procedure was implemented using leave-one-trial-out cross-validation in which all-but-one (training) trials or subjects (see sections 2.6.1 and 2.6.2 for details) were used to estimate the TRFs, and the remaining held-out (test) trial/subject was used to evaluate the prediction accuracy of the estimated model parameters. In the fitting stage of each cross-validation loop, the data were first used to select the optimal regularization parameter λ. This was done via a separate leave-one-out cross-validation loop utilizing only the training trials. In each of these cross-validation folds, the model was fit using a range of different λ parameter values and evaluated by predicting the data from the left-out trial. The prediction accuracies for each λ value were then averaged across all cross-validation folds and electrodes. The λ corresponding to the highest average prediction accuracy was used in the final fit to the entire training set. The resulting fit was then evaluated on the held-out "test" trial or subject to determine the ability of the TRF model to predict EEG responses to speech. The Pearson's correlation between the predicted and the actual data represents the prediction accuracy. This procedure was repeated *j*-times, where *j* denotes the number of trials or subjects, each time holding out data from a different trial/subject, resulting in *j* sets of TRF estimates and prediction accuracies.

In addition to overall model prediction accuracies, for each feature with time-varying amplitudes, we further estimated the degree to which these variations were tracked by the EEG responses. Briefly, this was done by comparing the full model's prediction accuracy to "null model" prediction accuracies, computed with the same model estimate but using test trial regressors in which a given feature's values were shuffled, while maintaining their timing (Broderick et al., 2021; Mesik et al., 2021). As such, the true and null model regressors had identical dimensionality and only differed in the veracity of one feature (with different null models having different feature shuffled). The difference between their prediction accuracies, therefore, reflected the aspect of the overall prediction accuracy that systematically varied with the feature shuffled within the null regressor. Because the shuffling process introduces variability into null model's performance, we computed 10 null model prediction accuracies within each cross-validation fold of the main TRF fitting procedure for each feature and subtracted the average of these from the full model performances. In the remainder of this manuscript, these differentials are referred to as feature-specific model contributions.

The main goal of the present work was to explore how estimates of cortical responses to a range of speech features are influenced by the quantity of EEG data included in the analysis. To explore this question, we iteratively applied identical sets of analyses to progressively larger amounts of the pre-processed EEG data. This was done via two distinct analysis approaches: 1) subject-specific analyses, where data from each participant was fit independently, and 2) generic analyses, where data from multiple subjects were fit jointly. Below we provide details of each analysis approach.

2.6.1 Subject-specific analyses

The subject-specific analyses involved repeated TRF estimation using data from individual subjects with progressively larger amounts of their data. Cross-validated fitting procedure was repeated with 11 distinct data quantities (3, 4, 6, 8, 10, 14, 18, 24, 30, 36, and 42 min of data), with data in each analysis selected in chronological order to reflect real constraints of data collection. While this may potentially bias analysis outcomes via temporally systematic phenomena such as fatigue or adaptation, such order effects are a natural aspect of most experiments and thus reflect realistic data acquisition scenarios. Group-level analyses were performed using each subject's average TRFs and prediction accuracies across all cross-validation folds (see section 2.6.4).

2.6.2 Generic subject analyses

In generic subject analyses, we repeatedly estimated TRFs using data pooled across progressively larger number of subjects. Cross-validated model fitting was again repeated with 11 distinct numbers of subjects (3, 4, 6, 8, 10, 14, 18, 24, 30, 36, and 41 subjects) at two distinct data acquisition durations per subject (4 and 8 min), yielding 22 unique analysis outputs per model. The data per subject was constrained both because of memory limitations for analyses involving larger numbers of subjects, but also to explore the extent to which small amount of data per subject can support accurate TRF estimation and robust prediction accuracies.

A notable limitation of generic analyses, as implemented here, is that all of the data is utilized within a single cross-validation procedure, resulting in a single set of average prediction accuracy and TRF estimates. While individual subject prediction accuracies from cross-validation can be used for group level statistics, these statistics are highly susceptible to noise from outlier data when utilizing small numbers of subjects. As such, to more accurately assess the central tendencies of generic analysis performance as a function of subject count, we utilized a resampling approach to obtain distributions of model performances for each sample size. Briefly, for each analysis, we randomly resampled, with replacement, participant data 20 times, and for each resampling instance we fitted a given model and computed mean TRFs and prediction accuracies across the cross-validation folds. To aid statistical evaluation of the analyses (see section 2.7) we corrected the prediction accuracies using estimates of the noise floor, i.e. range of prediction accuracies that may be expected by chance, by mismatching regressor-data pairings and computing their corresponding prediction accuracies. This was done at the level of each cross-validation fold (i.e., held-out subject), where we computed prediction accuracies for each 1-min data segment for all mismatched regressor-data pairings. Each subject's true prediction accuracy was then noise-floor corrected by subtracting the average mismatched prediction accuracy. Results of the 20 resampling analyses were then interpreted as a distribution of mean TRFs and prediction accuracies expected for a given sample size. Note that to allow for more reliable statistical inference, bootstrap-based analyses such as these are typically performed ≥1000 times in order to more precisely estimate the degree of overlap between the distributions of parameter estimates for different conditions. However, because the goal of this work was to demonstrate general behavior of TRF analyses rather than to draw strong statistical conclusions about our data, and the computational load of running ≥1000 resamplings for each of 44 analyses (2 models x 11 subject counts x 2 data quantities/subject) would have been very high, we chose to perform the more modest bootstrap procedure with 20 iterations.

2.6.3 Modelled speech features

To explore effects of feature choice on subject-specific and generic model performance, each modeling approach was evaluated using two distinct models of *attended* speech processing that emphasized features with different levels of sparseness. In the denser "envelope" model we included log-transformed acoustic envelope, along with word-onset regressor intended to capture responses to acoustic onsets. The sparser "surprisal and audibility" model included lexical surprisal for each word, audibility of each word against the to-be-ignored speaker, and the word-onset regressor. The latter was shared between the two models to help account for onset-driven neural activity, which is known to have particularly large amplitudes. Thus, the models differed in that one included a dense feature containing a nearly continuous time-varying signal, whereas the other model contained two much sparser features containing just one value per word. Because the envelope responses have previously been shown to occur at latencies < ~400 ms (e.g., Power et al., 2012; Kong et al., 2014; Fiedler et al., 2019), we modelled these responses using time intervals between -100 to 500 ms relative to feature onsets. Since the responses to the features in the sparser model tend be longer (e.g., Weissbart et al., 2019; Mesik et al., 2021), we modelled them using time intervals between -100 to 800 ms.

Across the two models, the four features were estimated as follows. The word onset regressors contained unit-amplitude impulses time-aligned to the onset of each word. The low frequency acoustic envelopes were extracted by half-wave rectifying the speech stimuli and lowpass filtering this representation below 8 Hz. This representation was then log-transformed to more closely approximate encoding of sound level in the human auditory system. Lexical surprisal regressors were the inverse of each word's probability given the multi-sentence preceding context, as estimated using the GPT-2 artificial neural network (Radford et al., 2019). Word audibility regressors were obtained by computing the ratio of each word's root-mean-square (RMS) of its acoustic waveform and the RMS of background speaker's acoustic waveform. As with envelopes, these ratios were log-transformed to more closely mimic sound level encoding in the human auditory system. Features in both surprisal and audibility regressors utilized the same timing as word-onset regressors, as prior work has shown that onset timing results in reasonable characterization of responses to higher-order speech features (e.g., Broderick et al., 2018; Weissbart et al., 2019; Mesik et al., 2021). Finally, non-zero regressor values for all features were RMS scaled to have an RMS value of 1. This was done to make TRFs for different features more similar in amplitudes in order to optimize regularization performance. With the exception of the acoustic envelope, a more detailed description of the derivation of these features can be found in Mesik et al. (2021).

2.6.4 Regions of interest

Due to the relatively low spatial resolution of EEG data and for simplicity of analysis result presentation, we chose to limit the spatial dimensionality of the results to two regions of interest (ROIs), the frontal and parietal ROIs. These ROIs were chosen based on the peak locations of activation for the two models. Both ROIs contained 13 electrodes, which symmetrically surrounded electrode Fz in the frontal ROI (i.e., AF3, AFz, AF4, F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, and FC4), and electrode Pz in the parietal ROI (i.e., CP3, CP1, CPz, CP2, CP4, P3, P1, Pz, P2, P4, PO3, POz, and PO4). ROI-specific TRFs and prediction accuracies were computed by averaging TRF analysis results from these sets of electrodes. All statistical analyses were performed on these ROI-averaged results.

2.7 Statistical analysis

The primary goal of this work was to describe general patterns of TRF model behavior as a function of data quantity. For subject-specific analyses, to test whether a given quantity of data was sufficient for the derived TRF to yield prediction accuracies that were significantly greater than zero, we utilized either t-tests or Wilcoxon signed-rank test, based on the outcome of the Anderson-Darling test of normality. These tests were conducted at the group level, using all 41 individual prediction accuracies. Note that statistics were not corrected for multiple comparisons, as we treated the analysis of each data quantity as a quasi-independent experiment, emulating the scenario where only that amount of data was acquired. Because analyses on different data quantities were not independent due to utilizing partially overlapping data, we abstained from direct pairwise comparisons of model prediction accuracies, and instead focused on more general description of model performance patterns (e.g., trajectory of mean model performance and between-subject variance) as a function of data quantity.

We additionally used the subject-specific fits to explore the relationship between the size of participant pool and the data quantity per subject required to achieve statistically reliable detection of cortical tracking of attended speech for different significance levels. To do this, we utilized subject pool sizes ranging from 2 to 41 subjects, and for each pool size, we resampled with replacement the prediction accuracies from analyses of each training data quantity (i.e., minutes of data per participant) 10,000 times. Within these samples, we conducted t-tests on results derived from each of the 11 data quantities per subject, searching for the minimum data quantity for which at least 80% of the 10,000 analyses exceeded the t-score thresholds corresponding to p < 0.05, p < 0.01, and p < 0.001. For these analyses we used parametric statistics, although non-parametric tests resulted in similar patterns of results.

For generic analyses, we estimated prediction accuracy noise floors as described in section 2.6.2. Within each cross-validation fold of each resampling analysis, we computed the difference between the true prediction accuracy and the "mismatched" prediction accuracy estimated using mismatched regressor-data pairings for the held-out participant. Thus, the distribution of these corrected prediction accuracies across resampling analyses reflects the proportion of times in which the true regressor-data pairings enabled more accurate data predictions than mismatched-pairings. Given that our analyses included 20 resamplings, only analyses where all data points exceeded the 0-point were deemed to be significant (i.e., since 1/20 corresponds to 5%). Note that noise floor correction in generic analyses was motivated by the possibility that very small positive prediction accuracies could occur by chance, making it difficult to determine the proportion of analyses with reliably elevated prediction accuracies. This correction was particularly important given our use of only 20 resampling analyses. In principle, this approach could also be used in subject-specific analyses. However, we abstained from the use of noise floors in subject-specific analyses because of their greater statistical power (i.e., more data points), our use of cross-validation, and the observation that in generic analyses, noise floor prediction accuracies were concentrated around zero.

To assess whether subject-specific and generic analyses resulted in estimation of morphologically similar TRFs, Pearson's correlations were used. For these analyses, we focused on models that utilized the greatest amounts of minutes (subject-specific analyses) and subjects (generic analyses).

3. Results

3.1 Subject-specific analyses

In subject-specific analyses, each participant's data was individually fit using each of the two models and evaluated on held-out data from the same participant. Group-level pattern of overall model prediction accuracies as a function of data quantity is depicted in Fig 1 for the denser (Fig 1A) and sparser (Fig 1B) models. In general, increases in training data quantity resulted in monotonic increases in performance for both models, along with reductions in inter-subject variability in prediction accuracies. With 41 participants used in these analyses, both models reached high degree of statistical significance with as little as 5 minutes of data per participant. However, at low training data quantities (e.g., < 10 min), a subset of participants exhibited prediction accuracies (i.e., correlations between the predicted and actual EEG data) that did not exceed 0. At larger data quantities (e.g. > 20 min), prediction accuracy distributions were elevated and largely did not span the value of zero. Finally, while prediction accuracies for the denser model exhibited signs of saturation beyond ~17 min of training data, saturation was less apparent for the sparser model.

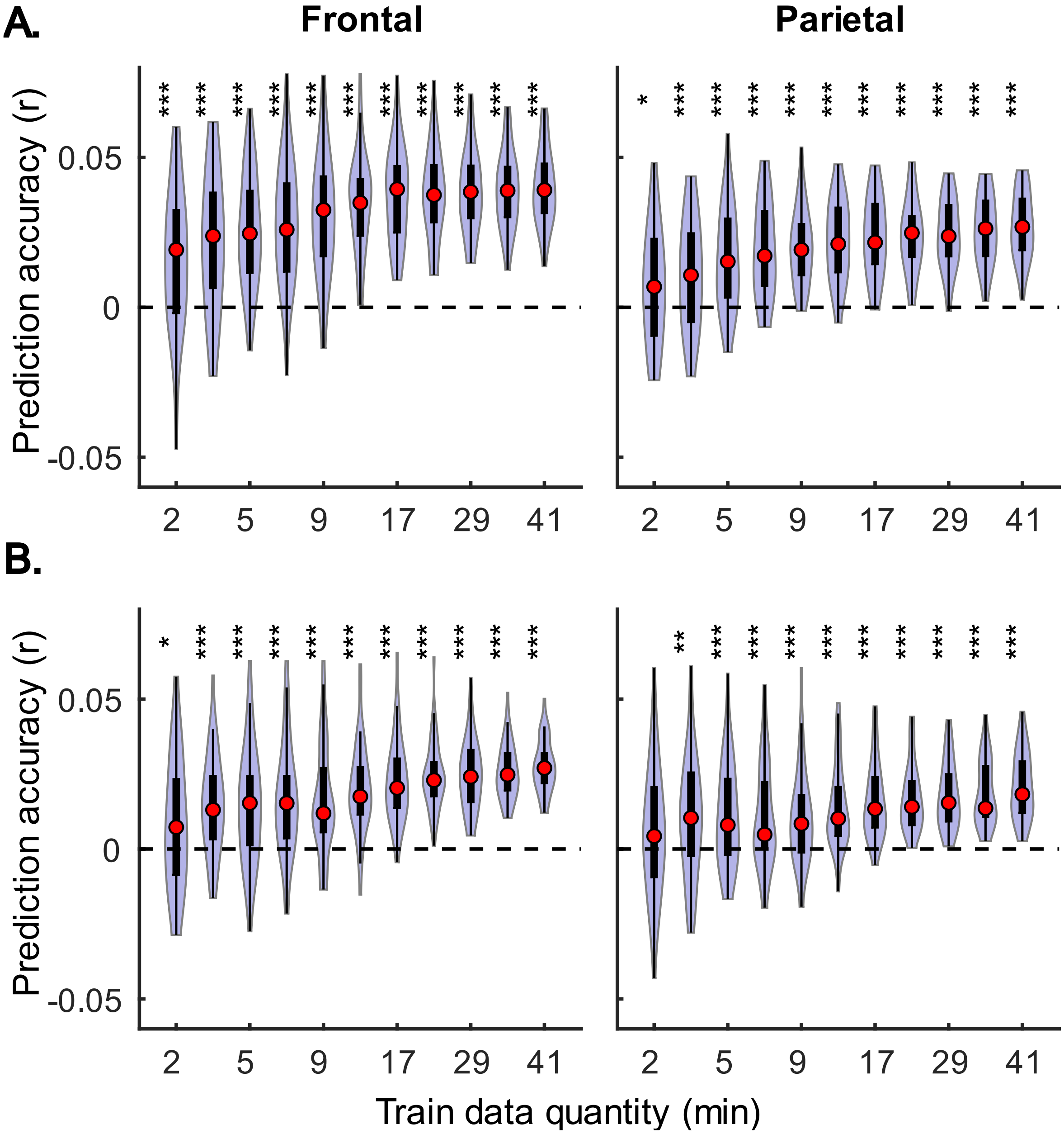

Figure 1. Overall subject-specific model prediction accuracies for the denser (A; onset and envelope features) and sparser (B; onset, surprisal, and audibility features) models displayed as a function of training data quantity (data acquisition duration). The violin plots depict the distribution of these values for all 41 participants, with a box plot depiction (interquartile range) shown within each violin. Red circles denote median values of the distributions. The uncorrected significance levels of each statical comparison against zero are displayed at the top of each plot: * p < 0.05, ** p < 0.01, ** p < 0.001

Although overall prediction accuracy is a key metric in evaluating TRF model performance, it does not directly reflect the extent to which the model captures systematic variation in neural responses as a function of the amplitude of the modelled features. As such, much of the prediction accuracy may be driven by "simple" evoked responses that are time-locked to the modelled features. To assess the extent to which the trained models captured systematically varying aspects of feature-driven responses, we compared the true model prediction accuracies to those estimated using regressors in which a given feature's values were permuted, while maintaining their timings (see section 2.6; Broderick et al., 2021; Mesik et al., 2021). Model fit contributions from systematically varying features (i.e., excluding onsets) as a function of training data quantity are depicted in Fig 2 for the denser (Fig 2A) and sparser (Fig 2B) models. Like overall prediction accuracies, these analyses exhibit similar general patterns of increasing model fit contributions and decreasing across-subject variability as the data quantity used in model fitting increases. However, this analysis revealed a key difference between denser and sparser models. While the denser envelope feature showed highly significant group-level model fit contributions even at low data quantities (< 5 min), the sparser surprisal and audibility features generally exhibited much smaller contributions to prediction accuracy and required considerably more data (> ~17 min) to allow reliable detection of these contributions.

In addition to prediction accuracies, a key output of TRF models are the TRFs themselves: the impulse responses to each of the modelled features. Fig 3 shows TRFs for the denser model and Fig 4 shows TRFs for the sparse model. Mirroring the high prediction accuracies and model fit contributions, the denser model TRFs (Fig 3) showed a high degree of morphological similarity across the different data amounts, albeit with a systematic increase in amplitudes seen with increasing data quantity. In line with its lower model fit contributions, the sparser model TRFs (Fig 4) generally exhibited greater noisiness at low data quantities, with TRFs reaching stable appearance once substantial data quantity (> 17 min) was used in model fitting. While there was a trend for lower TRF amplitudes with increasing data quantity, this pattern is opposite to that seen for the denser model and generally appears less systematic than that seen in Fig 3. Notably, onset TRFs in Figs 3 and 4 exhibit highly dissimilar temporal morphologies, with the TRF from the denser model showing a prominent parietal negativity around 400 ms that is not seen in the sparser TRF. This discrepancy is caused by the fact that in the latter model, higher order features related to the N400 response are captured by the surprisal and audibility features, leaving little variance in this response to be captured by the onset regressor. This demonstrates that the onset regressor captures response components time-locked to word onsets that reflect both lower- and higher-order neural processes contributing to the data variance.

The systematic increase in TRF amplitudes seen in Fig 3 could, in principle, reflect a genuine neural phenomenon related to the chronological inclusion of data into models trained on more data. For example, it could reflect improved neural entrainment to the target audiobook over the course of the study session. However, examination of the regularization parameter (Fig 5A), which controls the penalty assigned to large TRF values during the fitting procedure, indicates a systematic decrease in this parameter with increasing data quantity. In other words, models fitted to greater data quantity penalized large TRF amplitudes less, likely resulting in the systematic increase in their amplitudes seen in Fig 3. This interpretation was supported by a supplemental analysis (not shown), in which we fit the denser TRF model using a 6-min moving window to assess whether the TRF amplitude changes over the course of the study session. This analysis revealed virtually identical TRF amplitudes in all analysis windows, lending support to the effect in Fig 3 being entirely technical in nature. Regularization parameter of the sparser model (Fig 5B) showed a similar systematic decrease with increasing data quantity without the corresponding increase in TRF amplitudes (Fig 4). This may reflect the overall greater noisiness in these TRFs at low data quantities, as well as greater similarity in general TRF amplitudes across the three features (in contrast to the large amplitude discrepancy between onset and envelope TRFs seen in Fig 3).

Although utilizing data from all participants provides the most accurate estimate of the average model performance for a particular training data quantity, it is less informative about the statistical performance of our TRF models with more limited samples. To provide a more complete description of how subject-specific models perform under more limited sample sizes, we performed a resampling analysis using subsets of participant from our 41-subject pool. More specifically, for each participant pool ranging from 2-41 participants, we resampled (with replacement) the subject-specific result pool 10,000-times at each of the data quantities and determined how many minutes of data per participants were required to reach significance at three commonly used significance thresholds ($p < 0.05$, 0.01, and 0.001). The results of these analyses are shown in Fig 6 for overall prediction accuracy, and Fig 7 for feature-specific model fit contributions. These results show the expected downward sloping pattern whereby smaller participant pools require greater amount of data per participant. Additionally, these results mirror those in Figs 1-2 in that the sparser model generally requires more data per participant, and capturing significant feature-specific model contributions requires both more participants and more data per participant. Finally, while patterns of minimum data per participant required to reach significance from Figs 1-2 (i.e., when $n = 41$) are consistent with those in Figs 6-7 (rightmost data points in each plot), the exact minutes per participants don't match between the two in some cases, since the true participant sample in Figs 1-2 corresponded to just one data point within the larger bootstrap distribution used for Figs 6-7.

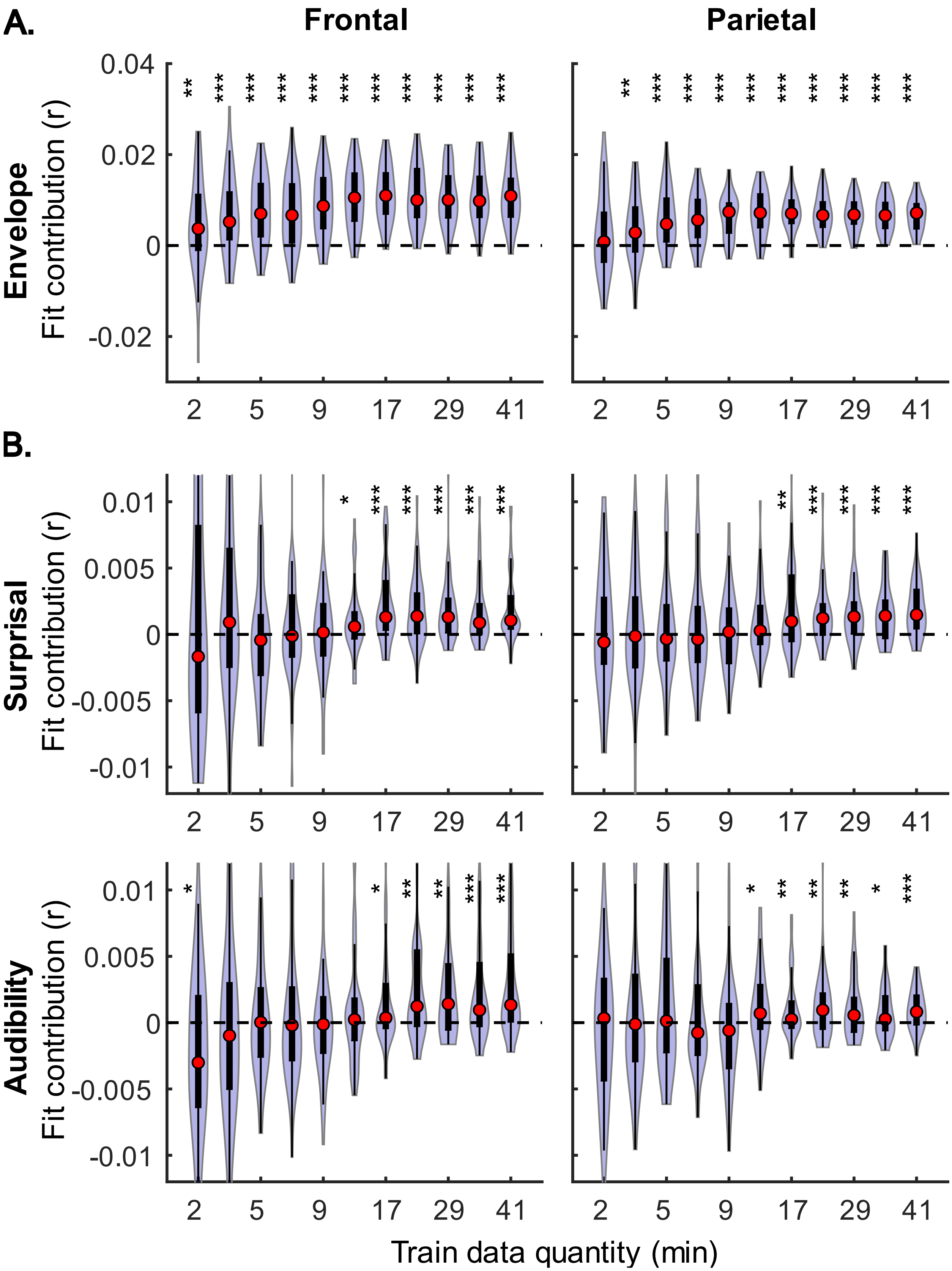
A.
Frontal
Parietal
Envelope
Fit contribution (r)
0.04
0.02
0
-0.02
** *** *** *** *** *** *** *** *** *** ***
** *** *** *** *** *** *** *** *** ***
2 5 9 17 29 41
B.
Surprisal
Fit contribution (r)
0.01
0.005
0
-0.005
-0.01
* *** *** *** *** ***
** *** *** *** ***
Audibility
Fit contribution (r)
* * ** ** *** ***
* ** ** ** * ***
Train data quantity (min)

Figure 2. Model fit contributions as a function of training data quantity for features in denser (A, top row) and sparser (B, middle and bottom rows) models. Model fit contributions reflect the degree to which the true sequence of feature values allows for more accurate prediction of EEG data compared to an arbitrary ordering of that feature. Note that tails of some of the violin plots in B are truncated to facilitate visualization of the central portion of the distributions. Asterisks denote significance levels as in Fig 1.

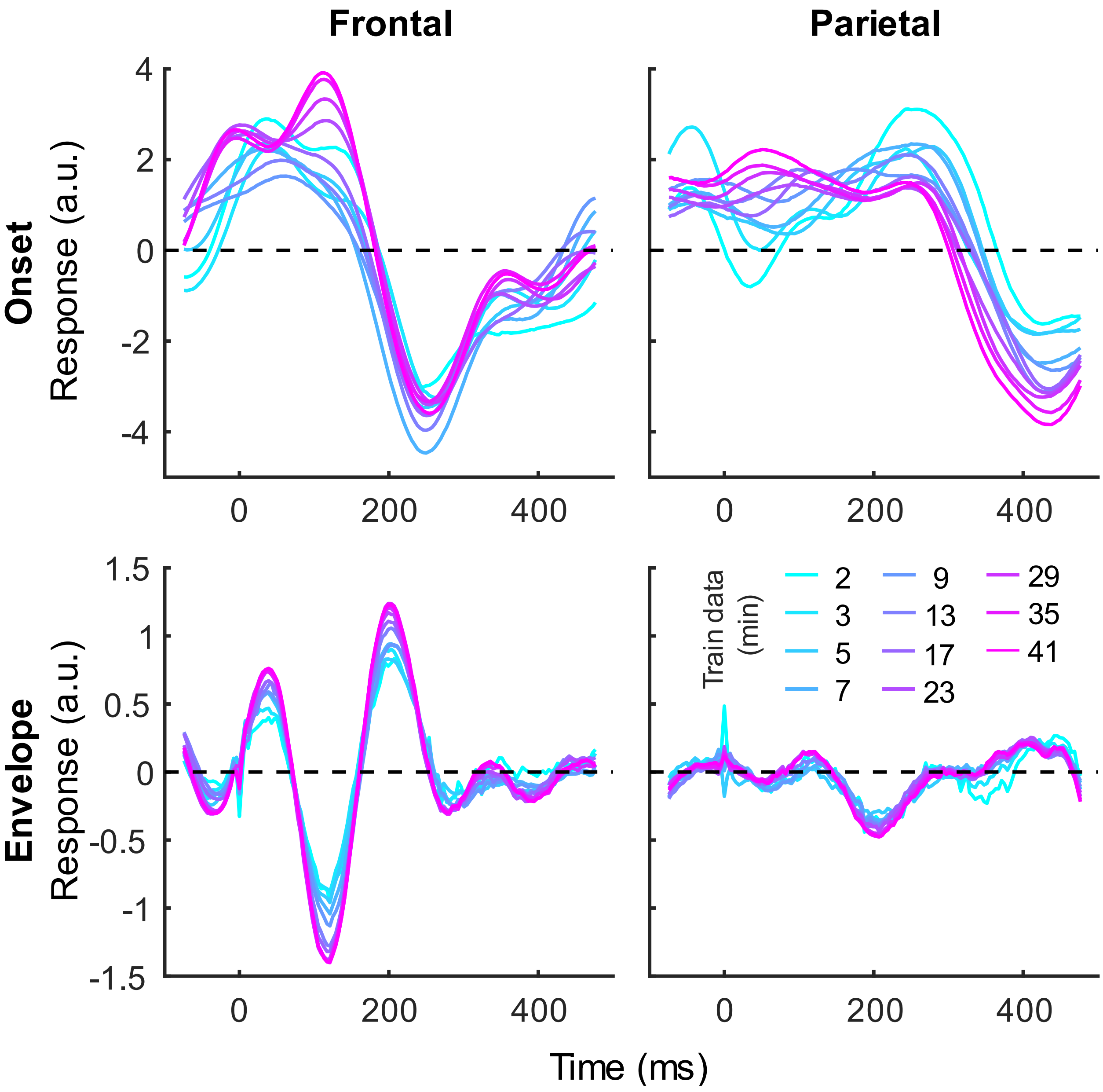

Figure 3. Group-averaged TRF time courses for onset (top row) and envelope (bottom row) features from the denser model in frontal (left column) and parietal (right column) ROIs. TRFs estimated using different amounts of data are depicted in different colors (see legend above

the lower right plot). The sharp deflections in envelope response around t = 0 ms reflect mild leakage of electrical artifact from earphones into the EEG signal.

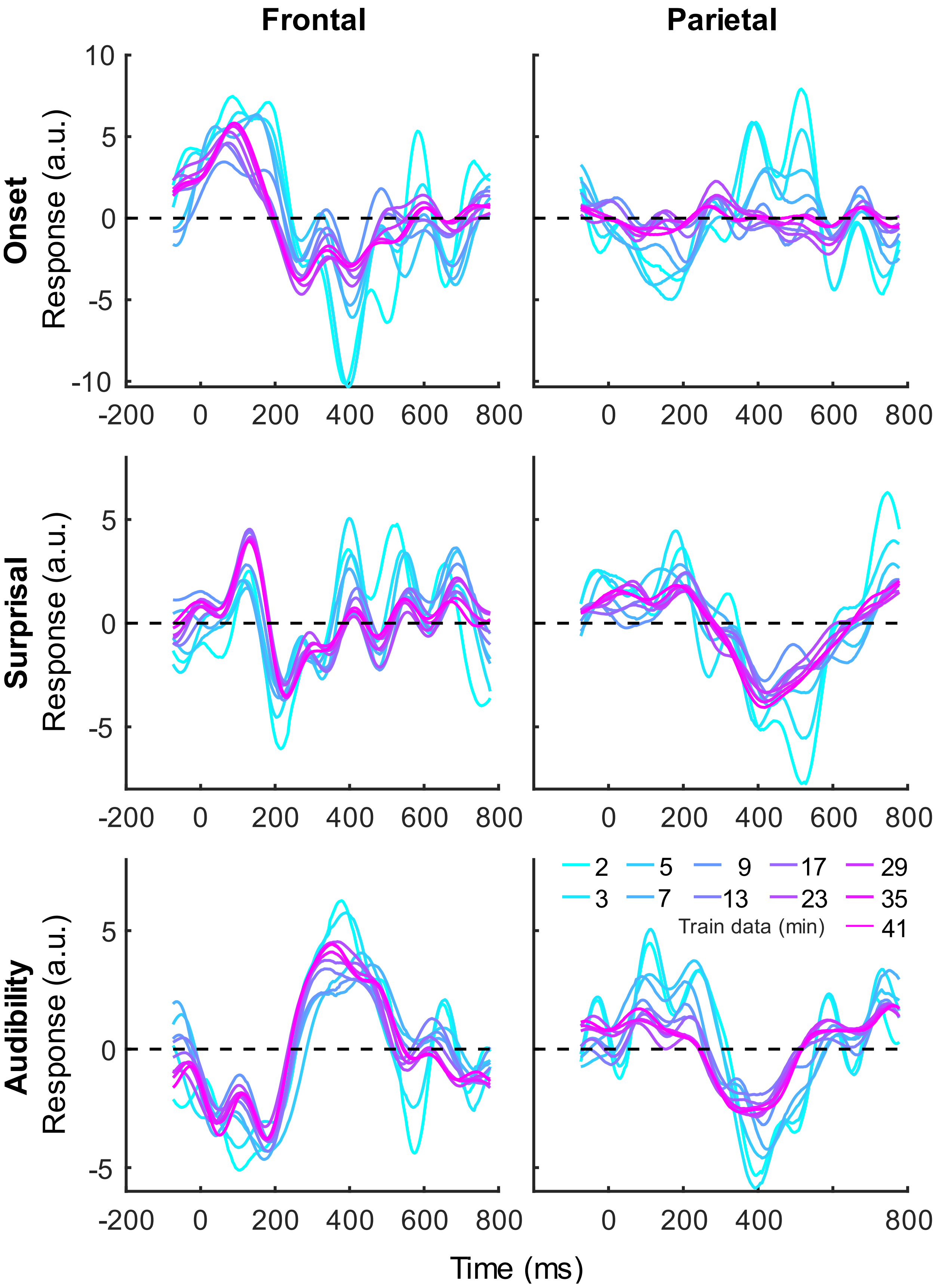
Frontal
Parietal
Onset
Response (a.u.)
10
5
0
-5
-10
-200
0
200
400
600
800
Surprisal
Response (a.u.)
Audibility
Response (a.u.)
2
3
5
7
9
13
17
23
29
35
41
Train data (min)
Time (ms)

Figure 4. Group-averaged TRF time courses for onset (top row), surprisal (middle row), and audibility (bottom row) features from the sparser model in frontal (left column) and parietal (right column) ROIs, as a function of training data quantity (see legend in bottom right). Note that these TRFs contain larger latencies than those in Fig 3 due to these features engaging higher-order processing, reflected in key TRF features such as the N400 response seen in parietal ROI.

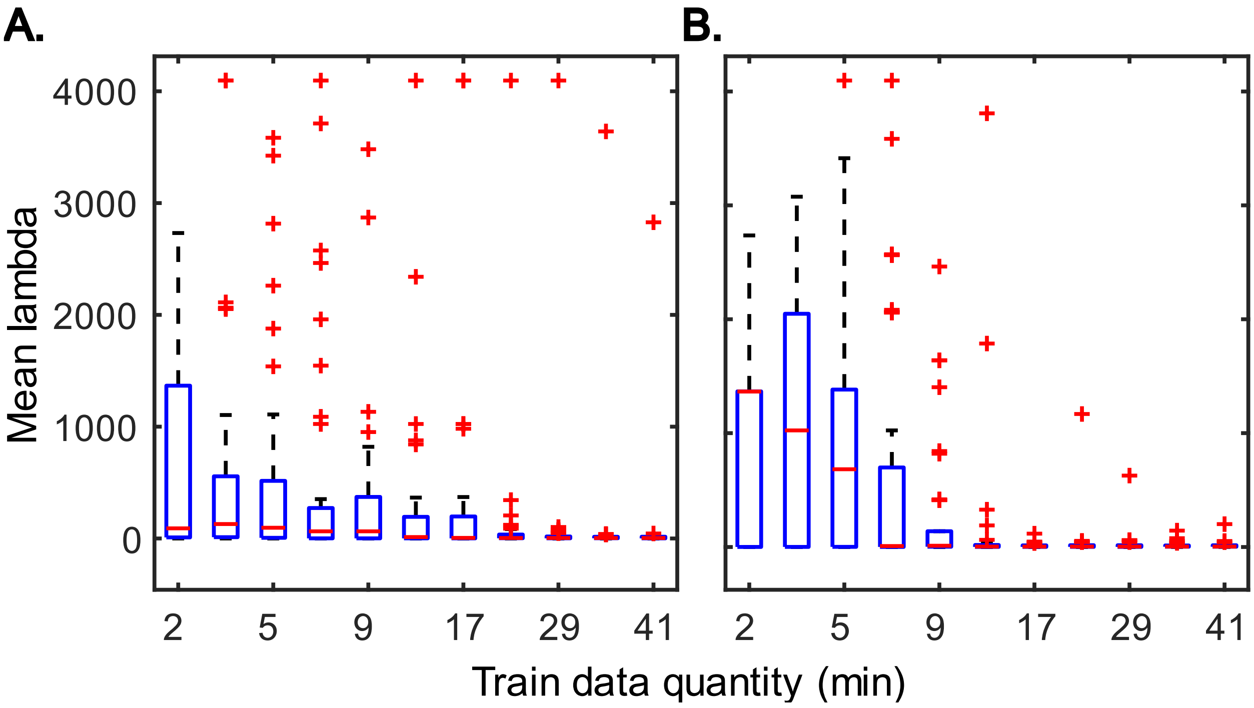

Figure 5. Distributions of subject-specific regularization parameters, lambda, from the denser (A) and sparser (B) models, as a function of training data quantity. Red horizontal lines within each box represent the median value while the lower and upper bounds of the blue boxes depict the 25th and 75th percentiles of the distributions. To aid the visualization of where bulk of the data points were, unusually large parameters were marked as outliers with red “+” symbols.

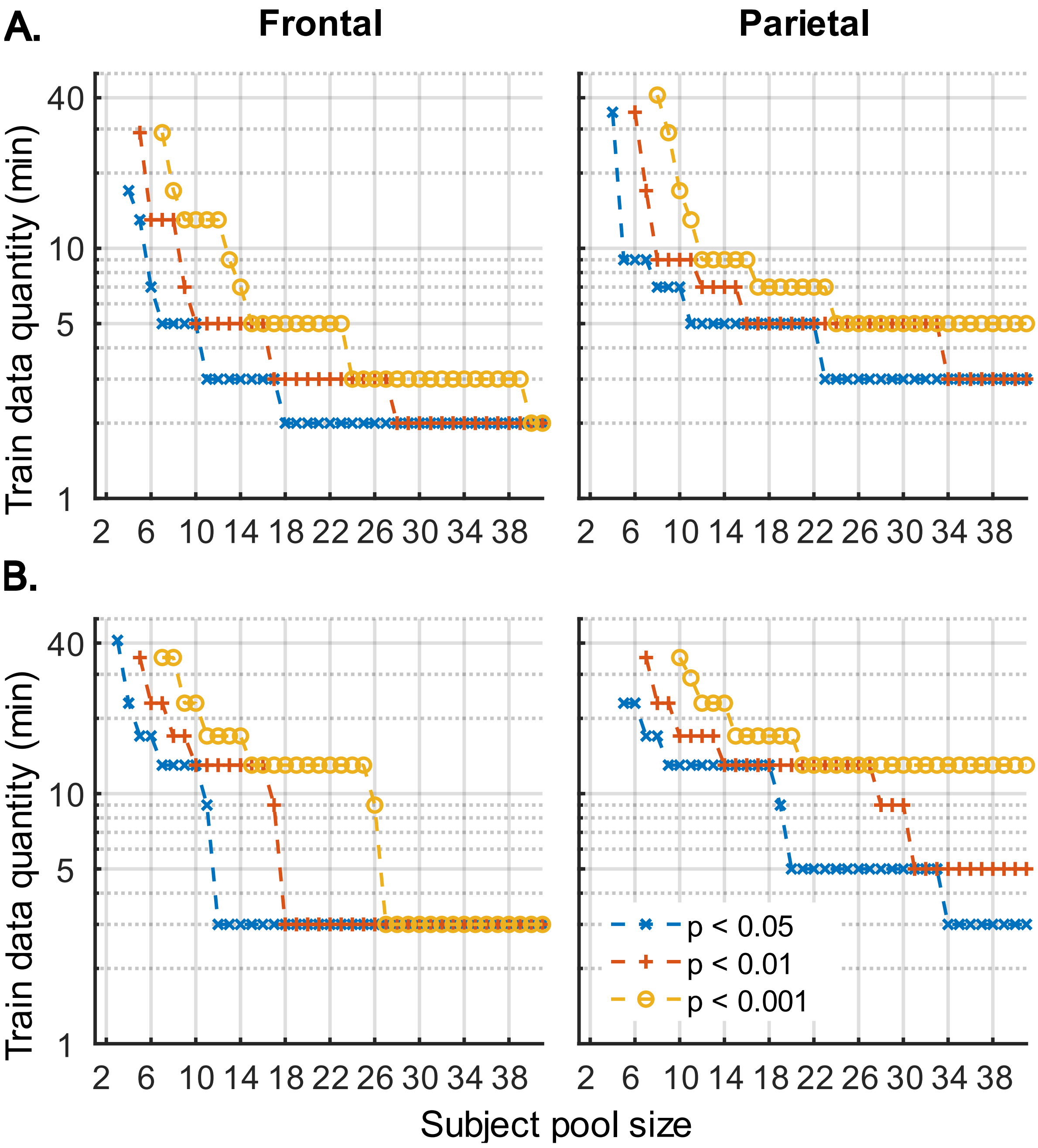

Figure 6. Amount of training data per participant required to reach significant overall prediction accuracy as a function of participant sample size for the denser (A) and sparser (B) TRF models in frontal (left column) and parietal (right column) ROIs. Different line colors represent different significance levels (see legend in lower right). For visualization purposes, minutes of data are shown on a log axis. The discrete steps along the y-axis stem from the fact that subject-specific analyses were run using 11 discrete data quantities per subject. Therefore, values on the plots represent approximate threshold data quantities needed for each sample size.

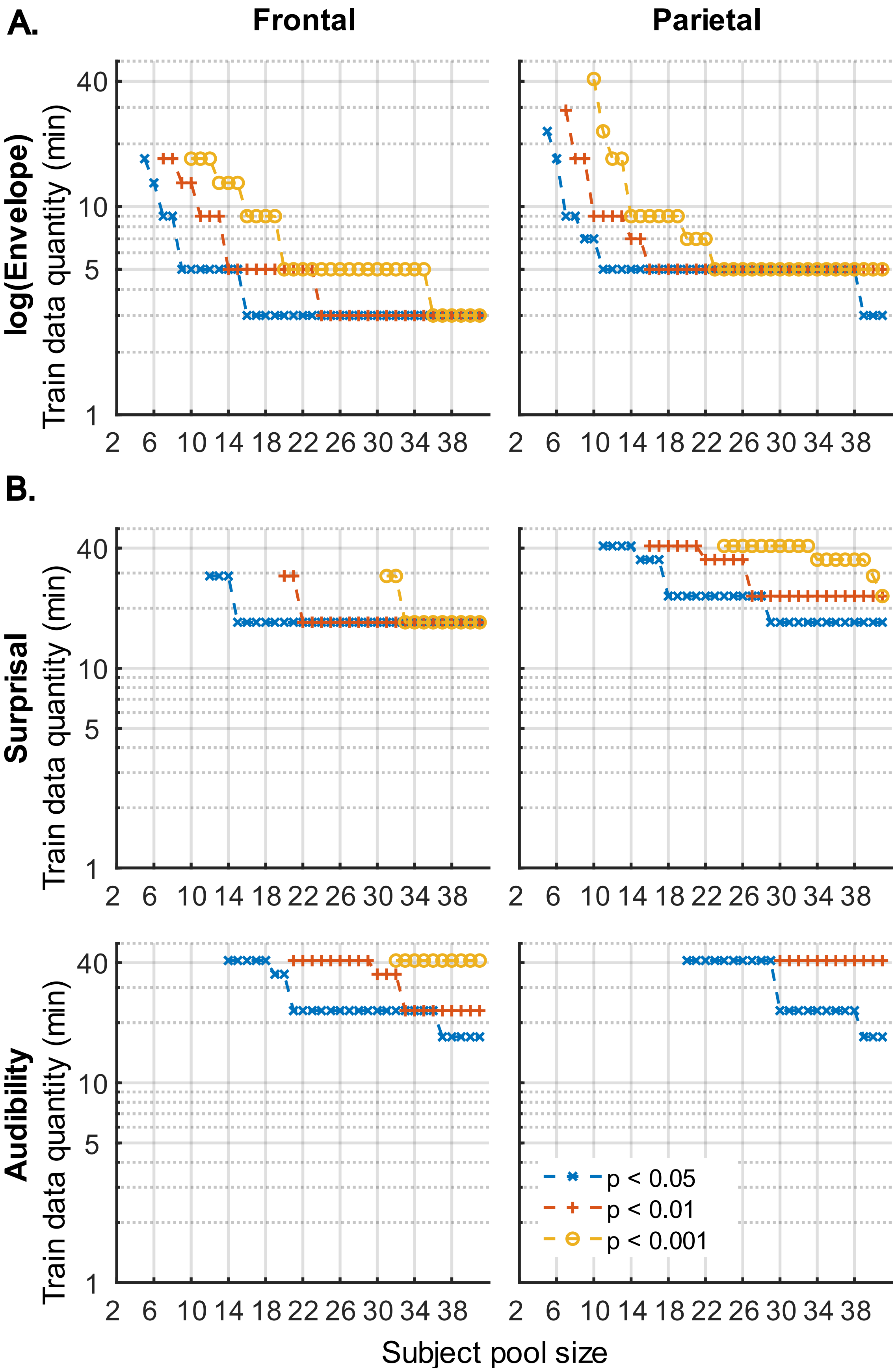
A.
Frontal
Parietal
log(Envelope)
Train data quantity (min)
40
10
5
1
2 6 10 14 18 22 26 30 34 38
B.
Surprisal
Audibility
Subject pool size
p < 0.05
p < 0.01
p < 0.001

Figure 7. Amount of training data per participant required to reach significant feature-specific prediction accuracies as a function of participant sample size for the denser (A) and sparser (B) TRF models in frontal (left column) and parietal (right column) ROIs. Different rows of plots represent different features. Line colors represent different significance levels (see legend in lower right).

3.2 Generic analyses

In the second set of analyses, we explored the effects of subject count on performance of *generic* analyses using the same two models (with denser and sparser features; section 2.6.3) used for subject-specific analyses (section 3.1). In contrast to subject-specific analyses, generic TRF models are simultaneously fit to data from multiple participants and evaluated on their ability to predict data of held-out participants. Due to the higher computational load of fitting models to multiple subjects simultaneously, these models were tested on 4 and 8 minutes of data per participant. Figures 8-12 depict results of generic analyses analogous to those shown in Figures 1-5 for subject-specific analyses. Note, that the two sets of results are not directly comparable, as the latter results depict distributions over and averages of the central tendency of 20 resampled generic analyses, rather than distributions of individual subject results. The lower variability in these analyses is therefore not directly indicative of generic analyses performing better than subject-specific analyses. The resampling approach used here was important due to inherent noisiness and strong influence of outlier data in analyzing small subject counts (see section 2.6.2).

Consistent with subject-specific analyses, the generic model prediction accuracies (Fig 8) as well as model fit contributions (Fig 9) both exhibited monotonically increasing performance and decreasing variability as the number of participants used to train the model increased. Note that because these plots depict distributions of average performances across 20 resampling analyses, only subject counts for which the entire distribution is elevated above zero can be deemed as reliably achieving non-zero performance. As such, depending on the ROI, about 7-9 participants were needed to achieve elevated overall model prediction accuracy, while ≥ 12 and ≥ 17 participants were needed to observe elevated feature-specific model contributions for the denser and sparser models, respectively. Interestingly, comparisons of 4- and 8-min of data per subject used in model fitting (left and right pairs of columns in each figure) made only a modest difference in performance, with the models trained on more data per participant producing tighter performance distributions, and in the case of the denser model, reaching significance with fewer participants. Finally, it is noteworthy that the feature-specific model contributions for the sparse model (Fig 9B) trained on 8-min of data/participant showed a trend of a decreased peak prediction accuracy compared to model trained on 4-min of data. The cause of this is unclear and may warrant further exploration on different data sets in the future.

TRFs derived from denser generic models, averaged across the 20 resampling analyses, are depicted in Fig 10. The TRFs are highly stereotypical across different subject pool sizes and the two per-subject data quantities. In line with subject-specific analyses, there is some evidence of increasing onset TRF amplitudes as the overall amount of data used to fit the model increases (i.e., with increasing number of subjects; see TRFs depicted by different colors in each subplot). On the other hand, the envelope TRFs exhibited no change in amplitude akin to that seen in subject-specific analyses (Fig 3). Although this discrepancy is puzzling, the general patterns of increasing onset TRF amplitudes was again accompanied by a systematic decrease in the regularization parameters (Fig 12A). Notably, because most generic models were overall trained on far more data than any of the subject-specific models (for n > 10 and n > 5, in models using 4-min and 8-min of data per subject, respectively), the upper bound of regularization parameter values seen in generic analyses was substantially lower than those seen in subject-specific analyses. These differences likely account for the differences in the patterns of TRF amplitude increases between the two sets of analyses.

TRFs from the sparser model (Fig 11) also exhibited a high degree of similarity for models trained on different amounts of data, although they generally show slightly higher degree of noisiness across different sample sizes. Mirroring the slight decrease in feature-specific model contributions of models trained on 8-min vs 4-min of data/subject (Fig 9B), there is an analogous decrease in TRF amplitudes (Fig 11, left vs right pairs of columns). At the same time, we also observed a pattern of slight increases in TRF amplitudes for models trained on more participants (i.e., different plot colors in each subplot of Fig 11), mirroring the decreasing amplitudes of regularization parameters in these analyses (Fig 12B). We speculate that this apparently paradoxical discrepancy between effects of more data per participant vs more participants could reflect tradeoff between phenomena driven by cognitive (e.g., waning attention or increased adaptation) vs technical (i.e., decreasing regularization parameter with more data) factors.

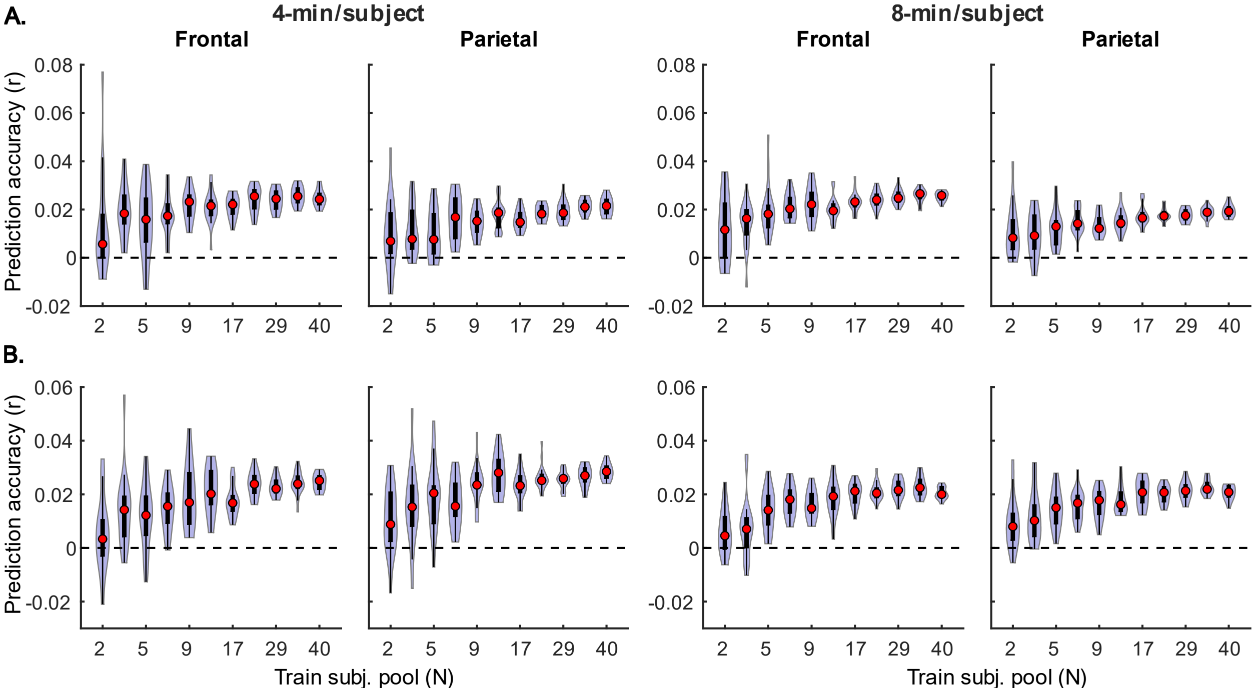

Figure 8. Noise floor-corrected generic model prediction accuracy distributions as a function of training subject pool size, for denser (A) and sparser (B) models with 4-min (left pair of columns) and 8-min (right pair of columns) of training data per subject. The distributions depicted by violin plots represent collection of mean prediction accuracies across 20 resampling analyses (instead of across-subject variability seen in Figs 1-2). As such, only subject counts with distributions with no overlap with zero (i.e. above the dashed horizontal lines) can be considered to reliably yield > 0 mean prediction accuracy.

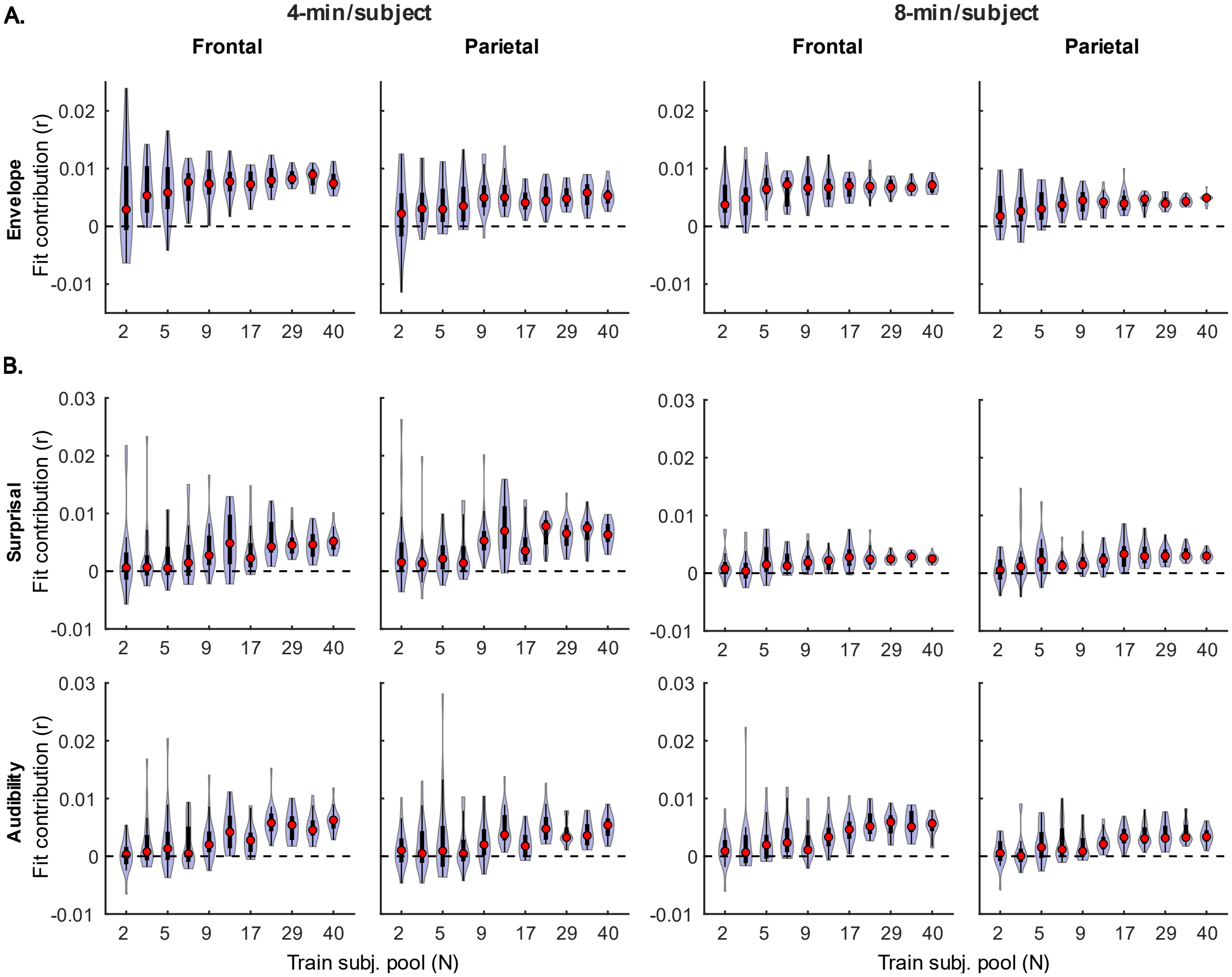

Figure 9. Feature-specific generic model contributions as a function of training subject pool size, for denser (A) and sparser (B) models with 4-min (left pair of columns) and 8-min (right pair of columns) of training data per subject.

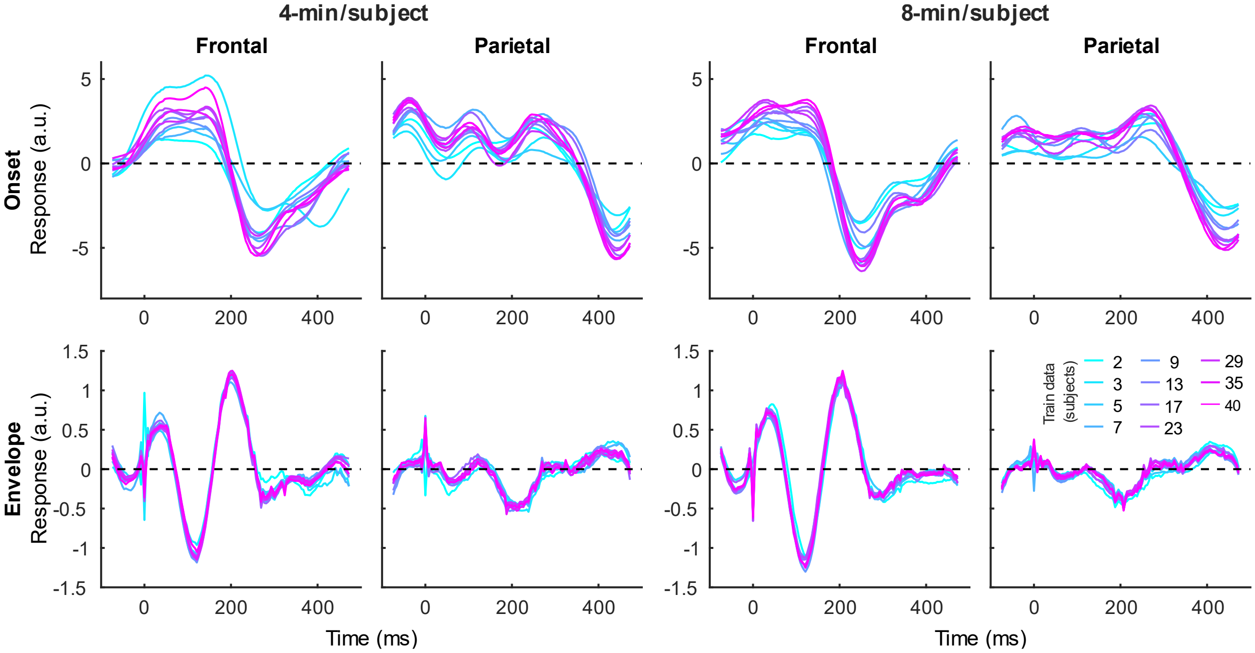

Figure 10. Average generic model TRF time courses for onset (top row) and envelope (bottom row) features from the denser model trained with 4-min (left pair of columns) and 8-min (right pair of columns) of data per subject. TRFs estimated using different amounts of data are depicted in different colors (see legend above the lower right plot). Note that the sharp deflections in envelope response around t = 0 ms reflect mild leakage of electrical artifact from earphones into the EEG signal.

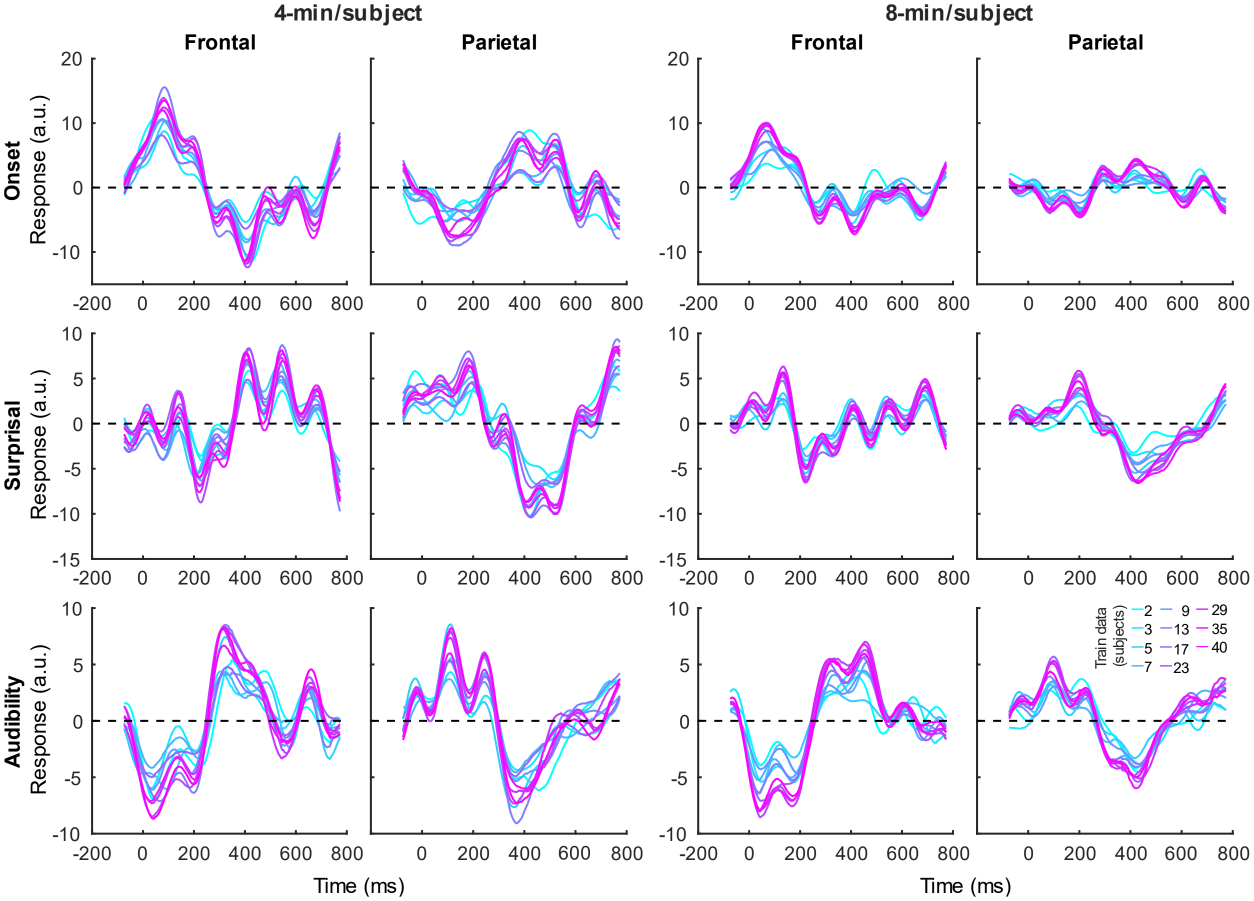

Figure 11. Average generic model TRF time courses for onset (top row), surprisal (middle row), and audibility (bottom row) features from the sparser model trained with 4-min (left pair of columns) and 8-min (right pair of columns) of data per subject. TRFs estimated using different amounts of data are depicted in different colors (see legend above the lower right plot).

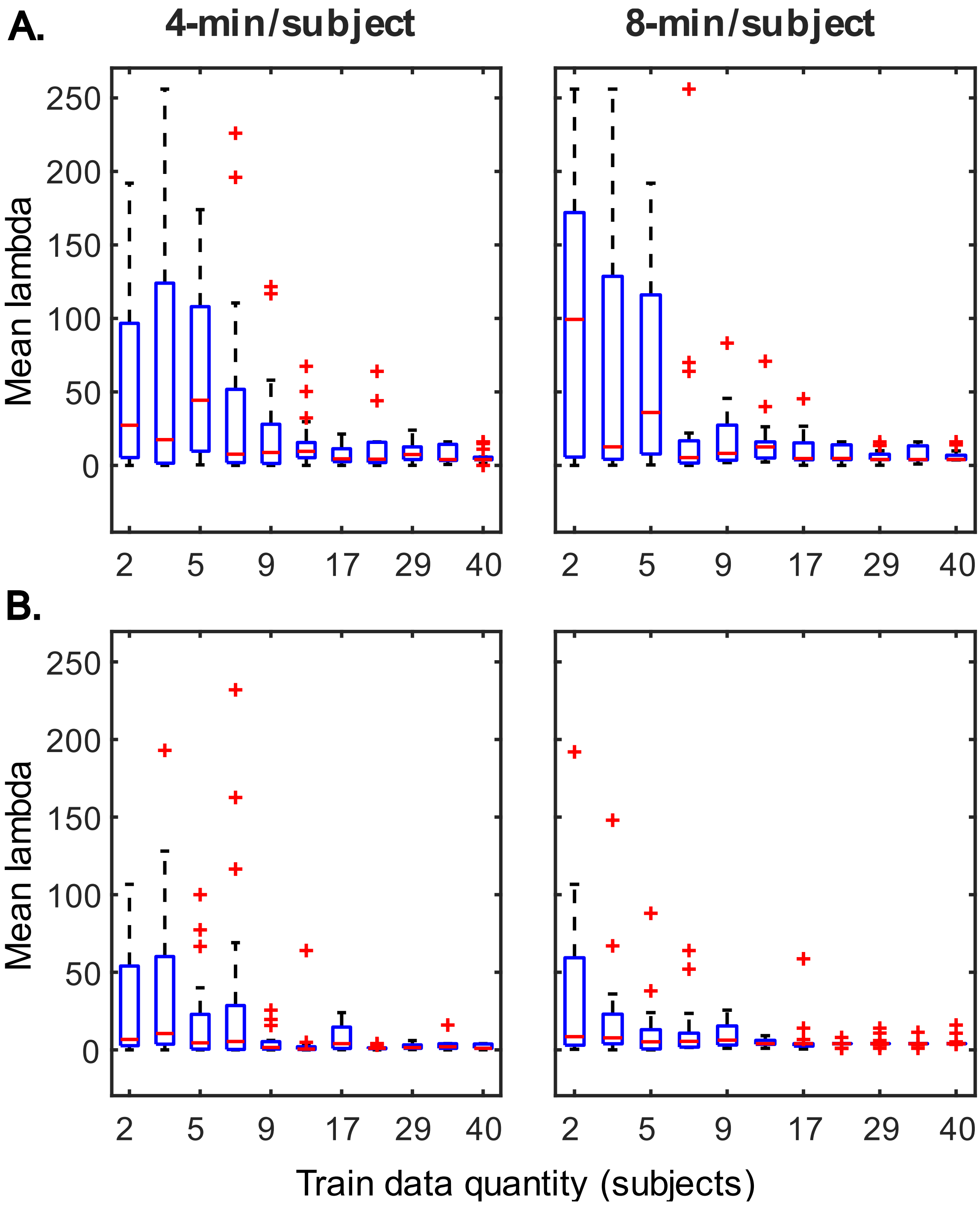

Figure 12. Distributions of generic model regularization parameters, lambda, from the denser (A) and sparser (B) models, as a function of training data quantity. Details of visualization are as in Fig 5.

3.3 Similarity between subject-specific and generic TRFs

While subject-specific and generic analyses produced qualitatively similar TRFs (Figs 3 and 4 vs Figs 10 and 11), we sought to assess this similarity quantitatively using a correlation analysis

(Fig 13). To this end, we focused on TRFs derived from the largest quantity of data in each analysis (including generic TRFs based on both 4-min and 8-min of data per subject), as these TRFs were deemed to most accurately capture the underlying neural responses to each of the features. For each feature, we computed the Pearson's correlation between TRFs derived in subject-specific and generic analyses. This analysis confirmed that in vast majority of cases, the two analyses produce highly similar TRFs, deviating below r = 0.85 only for sparser features in ROIs where the TRF amplitudes are generally quite low (e.g., surprisal response in the frontal ROI). These results demonstrate that at least for features modelled here, subject-specific and generic analyses enable extraction of highly similar neural responses.

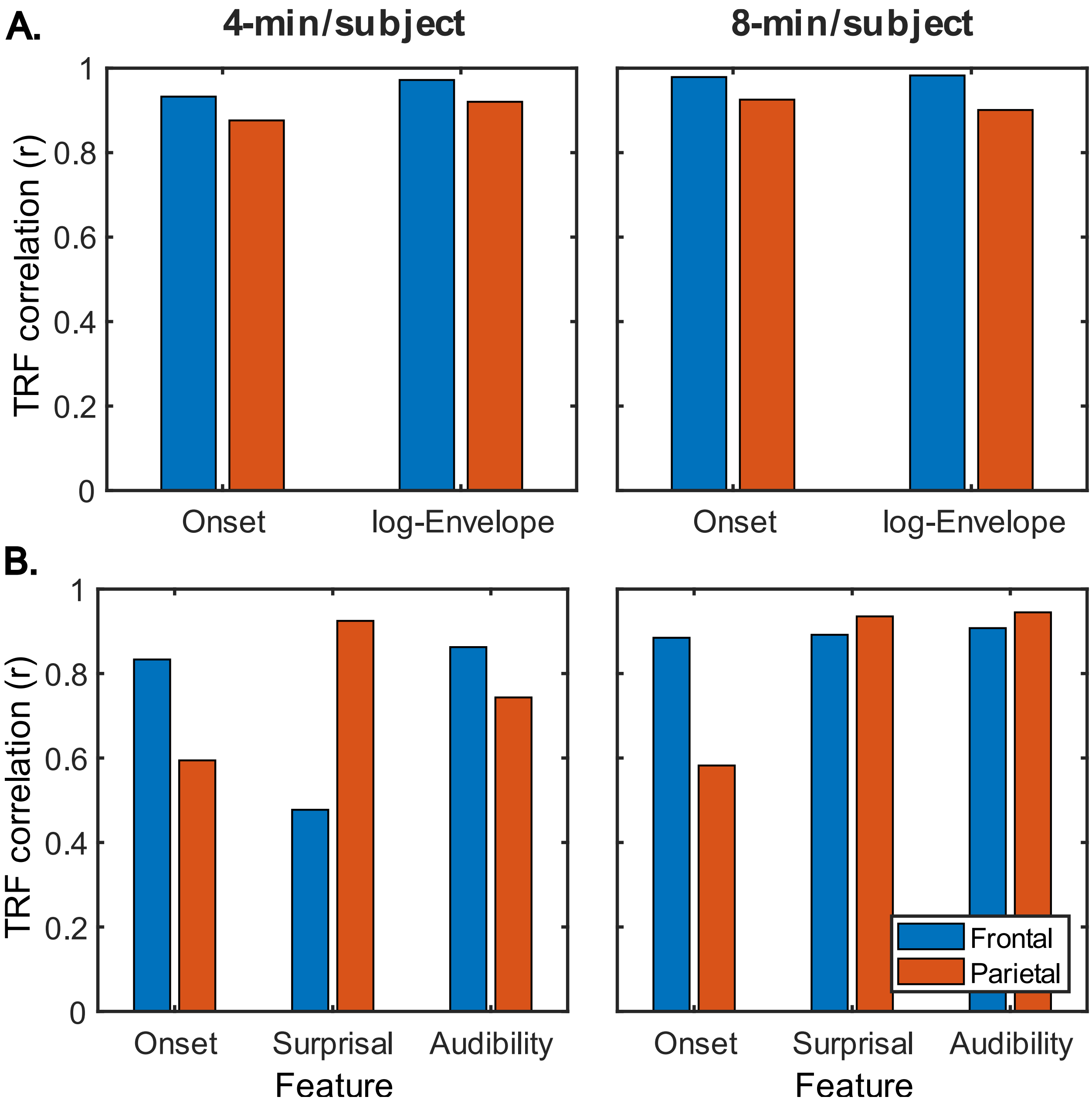

Figure 13. Similarity between average TRFs derived using subject-specific and generic analyses for features in the sparser (A) and denser (B) models. For generic analyses, both the results from analyses utilizing 4-min (left column) and 8-min (right column) per subject are shown. Different bar colors represent the frontal and parietal ROIs (see legend in lower left plot). Note that these analyses utilized TRFs derived using models trained on largest amounts of data in each type of analysis, i.e., most data per subject for subject-specific analyses, and most subjects for generic analyses.

4. Discussion

TRF analyses (Lalor and Foxe, 2010; Crosse et al., 2016) of EEG and MEG data are increasingly popular in studies of cortical processing of continuous, naturalistic stimuli such as speech (e.g., Di Liberto et al., 2015; Broderick et al., 2018; Weissbart et al., 2019) and music (e.g., Di Liberto et al., 2020; Marion et al., 2021). However, relatively few informational and educational resources demonstrating the behavior of these analyses under various constraints exist. Such resources allow researchers who are new to TRF analyses or considering adopting them to gain key intuition and insight to guide their study design. The goal of the present work was to demonstrate how quantity of collected data, a key parameter in experimental design, influences TRF analyses of attended speech representations in the context of a dual-talker continuous speech paradigm. We addressed this question using a previously collected dataset (Mesik et al., 2021) using two types of analyses: 1) Subject-specific analyses in which TRF models are independently fit to each participant's data, and 2) Generic analyses in which data from multiple participants is jointly used to fit a TRF model. For each analysis type we fit two different models, one of which had temporally dense features (acoustic envelope model), while the other had temporally sparse features (surprisal and audibility model). These models were fit repeatedly to explore how the amount of data per participant influences model prediction accuracies in the subject-specific approach and how prediction accuracies are influenced by the number of participants in the generic analyses. Finally, we used correlation analysis to compare the similarity of the TRFs derived in the two analysis approaches.

In addition to demonstrating the unsurprising general pattern whereby fitting models to more data provides monotonically improving prediction accuracies and more reliable TRF estimates, a closer examination of our results revealed several noteworthy phenomena. First, across both types of analyses, significant prediction accuracies could be achieved with just minutes of data per participant (Figs 1 and 8), although the denser model with envelope features had on average higher prediction accuracies than the sparser model with word-level features. While the denser model showed signs of performance saturation in both analysis approaches, this was less apparent in the sparser model where only the generic analyses utilizing 8 minutes of data per subject showed signs of saturation. Second, in analyses for capturing the contributions of individual features to the overall prediction accuracies (Figs 2 and 9), we observed a marked dissociation between denser and sparser models. Specifically, feature-specific model contributions were generally much smaller for the sparse model, and in subject-specific analyses, capturing these model contributions for word-level features required much greater amount of training data. However, it is noteworthy that in generic analyses, even 4 minutes of data per participant could reveal these feature-specific model contributions (Fig 9). This demonstrates that signals related to sparse features may have sufficient signal-to-noise ratio to be detected even with relatively small amounts of data, provided that the model is trained on sufficiently large dataset. Third, although TRFs derived with different amounts of data generally showed a high degree of time-domain consistency, we observed a systematic increase in TRF amplitudes with increasing data quantity (Figs 3-4, and 10-11). This pattern was mirrored by systematic decreases in the regularization parameter (Figs 5 and 12), which reflects the degree to which larger TRF amplitudes are penalized during model fitting. Finally, TRF patterns derived using subject-specific and generic analyses were highly similar (Fig 13), demonstrating that the two analyses reveal largely identical signatures of cortical speech processing.

While the majority of existing works utilizing TRF methods have used these tools to address specific questions about the nature of speech and music processing, only a handful of studies have explored the methodology itself. In general, the latter works focused on bigger picture overview of TRF methods and their utility in speech processing (Crosse et al., 2016; Sassenhagen, 2019), as well as on best practices in utilizing these methods in studies of special and clinical populations (Crosse et al., 2021). Additionally, Wong et al. (2018) explored performance of a range of regularization approaches for fitting forward and backward models in the context of a growing body of attention decoding literature. Within this body of work, a number of studies further explored effect of data quantity on attention decoding performance (e.g., O'Sullivan et al., 2015; Fuglsang et al., 2017; Wong et al., 2018). However, while this work, largely utilizing backward modeling, may appear related to the present efforts, a key distinction is that our work was the focused on the impact of data quantity on model training, rather than the performance of highly trained models in classification tasks (but see Mirkovic et al., 2015).

Most closely related to the present work, Di Liberto and Lalor (2017) investigated the effect of data quantity on performance of subject-specific and generic forward models in the context of phoneme-level speech processing. Similar to our results, Di Liberto demonstrated that the ability of subject-specific models to capture aspects of speech-evoked responses related to the phonemic processing improved with greater amount of data, with models requiring about 30 minutes of data to reliably capture phonemic responses. On the other hand, their generic model derived via averaging of subject-specific models was able to capture phonemic responses with 10-min of data, with no further improvement when more data per subject was used. The latter result deviates from our findings, which showed monotonic improvements of generic models as data from more participants was utilized in fitting. However, this apparent difference may stem from several factors. First, at the lowest end of the spectrum, our generic models were trained on data from as few as 2 participants, with 4-min of data per participant being used, while the Di Liberto's analyses always utilized the average of 9 participant models, each trained on 10-min of data or more. In other words, our analyses sampled the space of data quantities used to train models substantially more densely at this low end of the spectrum. Second, methodological differences, including Di Liberto's use of a single-talker paradigm (vs our use of dual-talker stimuli), modelled speech features (spectrogram and phonemic features vs envelope and word-level semantic and audibility features), and model performance quantification (differential between a more and less complex model performances vs overall prediction accuracy and feature-specific model contributions) preclude direct comparison of patterns of model performance. Importantly, however, our results agree with those of Di Liberto and Lalor in that generic models can provide significant predictive power even when they are trained and evaluated on relatively small amounts of data per participant (4-min and 10-min in our and Di Liberto's studies, respectively).

4.1 Utility of subject-specific and generic TRF analyses

Given that generic analyses demonstrated superior performance to subject-specific analyses when small amounts of data per participant was available, it is important to consider the scenarios for which each analysis approach may be appropriate. Despite requiring more data per participant, subject-specific analyses have been overwhelmingly more popular than generic analyses in studies of speech and music processing. A key advantage of these analyses is that for each participant, subject-specific fits provide independent estimates of both prediction accuracies, and the TRFs themselves, allowing for traditional approaches to group-level statistics. Additionally, subject-specific modelling is critically important for studies seeking to characterize individual differences within a population, and/or their relationship to behavioral performance or other subject-level characteristics.

Conversely, while cross-validation used during generic model fitting also provides independent prediction accuracies for all participants, the TRFs from different cross-validation folds are non-independent. Moreover, the interpretation of prediction accuracies for individual subjects in the context of generic analyses differs from subject-specific approach, as they reflect the predictability of a given participant's neural representations by a generic model, as opposed to the overall strength of speech representations in that participant. In other words, it may be the case that due to individual differences (e.g., due to anatomical variability) a particular participant's data may be poorly predicted by a generic model even if their individual model could perform substantially better. However, by capturing the shared aspects of neural processing within a larger group, generic analyses may be particularly useful for categorizing participants, or their mental states, which may have important applications both in clinical diagnostics, and for practical tools such as neuro-steered hearing aid devices. Indeed, several studies utilizing *backward* TRF models to decode attention have demonstrated the utility of generic models, albeit with a performance deficit relative to subject-specific models (e.g., Mirkovic et al., 2015; O'Sullivan et al., 2015). With respect to their utility for clinical diagnostics, Di Liberto and Lalor (2017) pointed out that generic models implicitly assume within-group homogeneity in neural representations, which may be particularly questionable within clinical populations (e.g., Levy et al., 1997; Happé et al., 2006). As such, researchers need to be cognizant about this limitation, and the extent to which TRF methods could be useful for diagnostic purposes for various conditions remains to be determined. Finally, on a practical note, one disadvantage of generic analyses, as implemented in the present work, is that fitting them to large datasets requires large amounts of computational resources and time, especially when utilizing a resampling approach to fitting. However, efficient use of resources (e.g., downsampling and reducing data dimensionality via methods such as denoised source separation, DSS; de Cheveigné and Simon, 2008), and alternative model estimation approaches, such as averaging of subject-specific models (e.g., Di Liberto & Lalor, 2017) or models fitted to subsets of multi-subject data, can mitigate these technical challenges.

4.2 Limitations

Although the present work provides a detailed exploration of TRF model performance as a function of training data quantity, caution should be taken in generalizing our results to other TRF studies. Specifically, choices in the experimental design (e.g., single vs multi-talker stimuli), data pre-processing (e.g., denoising algorithms), the applied model's feature space, and statistical analyses (e.g., use of cluster-based permutation tests; Maris and Oostenveld, 2007) could all substantially influence the performance of TRF analyses. For example, the use of dimensionality reduction techniques, such as DSS, could significantly reduce the amount of data needed to achieve significant prediction accuracy. On the other hand, utilizing high-dimensional feature patterns derived from deep neural networks (i.e., hundreds or thousands of features vs 2-3 features in the present study) may require substantially increased amount of data to yield significant prediction accuracies. Further work will be needed to provide estimates of data requirements, particularly for the latter comparison of low- vs high-dimensional models.

4.3 Conclusions

The goal of this work was to develop an informational resource for the growing field to TRF analyses of continuous speech processing, demonstrating the behavior of TRF analyses as a function of data quantity used in TRF fitting. In the context of relatively simple models of lower-level envelope processing, as well as higher-order processing of word-level features, we demonstrate that given a large-enough participant pool, small amounts of data (< 5 min) can be sufficient to train subject-specific models that predict significant variance in EEG responses to speech-masked speech. At the same time, substantially more data (15+ min) may be needed to capture aspects of data that systematically vary with word-level features. On the other hand, generic models can support significant prediction accuracy even for feature-specific variance with as little as 4-min of data per participant, while providing highly similar TRF estimates to those seen in subject-specific analyses. As such, despite their infrequent use, generic models have potential to be particularly useful for applications in clinical diagnostics, and multi-task studies with low per-task time budgets. While the present work is not, on its own, intended to be prescriptive about experimental duration, it may be a useful resource for informing selection of experimental duration, especially in conjunction with other tools, such as simulations and piloting.

Data availability

Data and analysis scripts will be made available upon reasonable request. Requests should be directed to JM, mesik002@umn.edu, or MW, wojtc001@umn.edu.

Ethics statement

The Institutional Review Board of the University of Minnesota approved the procedures in this study. All participants provided written informed consent to participate.

Author contributions

JM and MW designed the original experiment, analyzed the data, and wrote the manuscript. JM implemented experimental procedures and collected the data. All authors commented on the manuscript and approved the submitted version.

Funding

Financial support for this work was provided by NIH grant R01 DC015462 to MW.

Acknowledgments

We thank Lucia Ray for help with carrying out the original study that yielded the data used in the present work.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., and Robbins, K. A. (2015). The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Front. Neuroinform.* 9, 1–20. doi:10.3389/fninf.2015.00016.

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi:10.1163/156856897X00357.

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* 28, 3976-3983.e5. doi:10.1016/j.cub.2018.10.042.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803-809.e3. doi:10.1016/j.cub.2018.01.080.

Broderick, M. P., Liberto, G. M. Di, Anderson, A. J., Rofes, A., and Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Sci. Rep.* 11, 4963. doi:10.1038/s41598-021-84597-9.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing* 36, 287–

314. doi:10.1016/0165-1684(94)90029-9.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10, 1–14. doi:10.3389/fnhum.2016.00604.

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., and Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research. *Front. Neurosci.* 15. doi:10.3389/fnins.2021.705621.

de Cheveigné, A., and Simon, J. Z. (2008). Denoising based on spatial filtering. *J. Neurosci. Methods* 171, 331–339. doi:10.1016/j.jneumeth.2008.03.015.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi:10.1016/j.jneumeth.2003.10.009.

Di Liberto, G. M., and Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: methodological considerations for applied research. *Hear. Res.* 348, 70–77. doi:10.1016/j.heares.2017.02.015.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi:10.1016/j.cub.2015.08.030.

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., et al. (2020). Cortical encoding of melodic expectations in human temporal cortex. *Elife* 9, 1–26. doi:10.7554/eLife.51784.

Di Liberto, G. M., Wong, D., Melnik, G. A., and de Cheveigné, A. (2019). Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage* 196, 237–247. doi:10.1016/j.neuroimage.2019.04.037.

Ding, N., and Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi:10.1152/jn.00297.2011.

Donhauser, P. W., and Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron* 105, 385-393.e9. doi:10.1016/j.neuron.2019.10.019.

Fiedler, L., Wöstmann, M., Herbst, S. K., and Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *Neuroimage* 186, 33–42. doi:10.1016/j.neuroimage.2018.10.057.

Fuglsang, S. A., Dau, T., and Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage* 156, 435–444. doi:10.1016/j.neuroimage.2017.04.026.

Happé, F., Ronald, A., and Plomin, R. (2006). Time to give up on a single explanation for autism. *Nat. Neurosci.* 9, 1218–1220. doi:10.1038/nn1770.

Jutten, C., and Herault, J. (1991). Blind separation of sources, part i: an adaptive algorithm based on neuromimetic architecture. *Signal Processing* 24, 1–10. doi:10.1016/0165-1684(91)90079-X.

Kleiner, M., Brainard, D. H., and Pelli, D. G. (2007). What's new in psychtoolbox-3. in *Perception 36 ECVP Abstract Supplement*.

Kong, Y. Y., Mullangi, A., and Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hear. Res.* 316, 73–81. doi:10.1016/j.heares.2014.07.009.

Lalor, E. C., and Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.* 31, 189–193. doi:10.1111/j.1460-9568.2009.07055.x.

Levy, F., Hay, D. A., McStephen, M., Wood, C., and Waldman, I. (1997). Attention-deficit hyperactivity disorder: a category or a continuum? genetic analysis of a large-scale twin study. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 737–44. doi:10.1097/00004583-199706000-00009.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. MIT Press.

Marion, G., Di Liberto, G. M., and Shamma, S. A. (2021). The music of silence. part i: responses to musical imagery encode melodic expectations and acoustics. *J. Neurosci.* 41, JN-RM-0183-21. doi:10.1523/JNEUROSCI.0183-21.2021.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of eeg- and meg-data. *J. Neurosci. Methods* 164, 177–190. doi:10.1016/j.jneumeth.2007.03.024.

Mesik, J., Ray, L., and Wojtczak, M. (2021). Effects of age on cortical tracking of word-level features of continuous competing speech. *Front. Neurosci.* 15, 1–21. doi:10.3389/fnins.2021.635126.

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications. *J. Neural Eng.* 12. doi:10.1088/1741-2560/12/4/046007.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cereb. Cortex* 25, 1697–1706. doi:10.1093/cercor/bht355.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442. doi:10.1163/156856897X00366.

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? a late locus of selective attention to natural speech. *Eur. J. Neurosci.* 35, 1497–1503. doi:10.1111/j.1460-9568.2012.08060.x.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*. Available at:

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Sassenhagen, J. (2019). How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Lang. Cogn. Neurosci.* 34, 474–490. doi:10.1080/23273798.2018.1502458.

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2019). Cortical tracking of surprisal during continuous speech comprehension. *J. Cogn. Neurosci.*, 1–12. doi:10.1162/jocn_a_01467.

Wong, D. D. E., Fuglsang, S. A., Hjortkjær, J., Ceolini, E., Slaney, M., and de Cheveigné, A. (2018). A comparison of regularization methods in forward and backward models for auditory attention decoding. *Front. Neurosci.* 12, 531. doi:10.3389/fnins.2018.00531.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Atten. Percept. Psychophys.* 72, 2031–2046. doi:10.3758/APP.72.8.2031.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron* 77, 980–991. doi:10.1016/j.neuron.2012.12.037.