Behavioral/Cognitive

Generalizable EEG Encoding Models with Naturalistic Audiovisual Stimuli

Maansi Desai,[1] **Jade Holder,**[1] **Cassandra Villarreal,**[1] **Nat Clark,**[1] **Brittany Hoang,**[1] and **Liberty S. Hamilton**[1,2]
[1]Department of Speech, Language, and Hearing Sciences, Moody College of Communication, University of Texas at Austin, Austin, Texas 78712, and [2]Department of Neurology, Dell Medical School, University of Texas at Austin, Austin, Texas 78712

In natural conversations, listeners must attend to what others are saying while ignoring extraneous background sounds. Recent studies have used encoding models to predict electroencephalography (EEG) responses to speech in noise-free listening situations, sometimes referred to as "speech tracking." Researchers have analyzed how speech tracking changes with different types of background noise. It is unclear, however, whether neural responses from acoustically rich, naturalistic environments with and without background noise can be generalized to more controlled stimuli. If encoding models for acoustically rich, naturalistic stimuli are generalizable to other tasks, this could aid in data collection from populations of individuals who may not tolerate listening to more controlled and less engaging stimuli for long periods of time. We recorded noninvasive scalp EEG while 17 human participants (8 male/9 female) listened to speech without noise and audiovisual speech stimuli containing overlapping speakers and background sounds. We fit multivariate temporal receptive field encoding models to predict EEG responses to pitch, the acoustic envelope, phonological features, and visual cues in both stimulus conditions. Our results suggested that neural responses to naturalistic stimuli were generalizable to more controlled datasets. EEG responses to speech in isolation were predicted accurately using phonological features alone, while responses to speech in a rich acoustic background were more accurate when including both phonological and acoustic features. Our findings suggest that naturalistic audiovisual stimuli can be used to measure receptive fields that are comparable and generalizable to more controlled audio-only stimuli.

Key words: electroencephalography; encoding models; natural stimuli; spectrotemporal receptive field; speech perception

Significance Statement

Understanding spoken language in natural environments requires listeners to parse acoustic and linguistic information in the presence of other distracting stimuli. However, most studies of auditory processing rely on highly controlled stimuli with no background noise, or with background noise inserted at specific times. Here, we compare models where EEG data are predicted based on a combination of acoustic, phonetic, and visual features in highly disparate stimuli—sentences from a speech corpus and speech embedded within movie trailers. We show that modeling neural responses to highly noisy, audiovisual movies can uncover tuning for acoustic and phonetic information that generalizes to simpler stimuli typically used in sensory neuroscience experiments.

Introduction

Sound and speech perception rarely occur in isolation. Understanding speech in natural environments involves the detection and parsing of acoustic and linguistic cues within overlapping talkers and background noise. To understand how the brain performs this task, many researchers have started incorporating naturalistic stimuli in their experimental paradigms (Hamilton and Huth, 2018; Fiedler et al., 2019). Such work has demonstrated comparable or even better results for responses from more controlled paradigms (Lerner et al., 2011; Wehbe et al., 2014; Huth et al., 2016; Hamilton and Huth, 2018) and are more representative of our daily environment. What is considered "naturalistic" may vary: some studies use more naturalistic continuous or full sentence stimuli, while others use consonant–vowel syllables (Shankweiler and Studdert-Kennedy, 1966). Often the sentences used in these studies are presented in isolation and are far less "natural" than those used in everyday communication. Many studies have successfully used controlled stimuli to understand speech perception, such as sentences from the Texas Instruments Massachusetts Institute of Technology (TIMIT) corpus (Chang et al., 2010; Mesgarani et al., 2014a,b;

Received Nov. 13, 2020; revised Aug. 24, 2021; accepted Aug. 29, 2021.

Author contributions: M.D. and L.S.H. designed research; M.D., J.H., C.V., N.C., B.H., and L.S.H. performed research; L.S.H. contributed unpublished reagents/analytic tools; M.D., J.H., C.V., N.C., B.H., and L.S.H. analyzed data; M.D. and L.S.H. wrote the paper.

This work was supported by a grant from the Texas Speech-Language Hearing Foundation (The Elizabeth Wiig Doctoral Student Research Award, to M.D.). We thank Garret Kurteff and Alexander Huth for helpful comments on the manuscript. We also thank Natalie Miller, Rachel Sorrells, Jacob Cheek, Garret Kurteff, and Ian Griffith for assistance in data collection.

Correspondence should be addressed to Liberty S. Hamilton at liberty.hamilton@austin.utexas.edu.

https://doi.org/10.1523/JNEUROSCI.2891-20.2021

Tang et al., 2017; Hamilton et al., 2018; Akbari et al., 2019). Others have used audiobooks (Broderick et al., 2019; Hausfeld et al., 2018), which arguably are more natural than TIMIT sentences, but can lack the natural variation of pitch, timbre, and other suprasegmental features of speech present in natural communication, and may often be read by only one talker. In addition, sentences from speech corpora like TIMIT are often repetitive and tedious to listen to for electroencephalography (EEG) tasks >1 h. Part of our motivation for this study was to use stimuli that were more engaging for participants and to investigate whether neural responses can still be modeled robustly. A secondary aim was to investigate whether the observed receptive fields were similar across different stimulus types.

Numerous electrophysiological studies have demonstrated neural tracking of the acoustic envelope (Horton and D'Zmura, 2011; Kubanek et al., 2013; Di Liberto et al., 2015; Fuglsang et al., 2017; Vanthornhout et al., 2018), phoneme and phonological features (Ding and Simon, 2012; Di Liberto et al., 2015; Khalighinejad et al., 2017; Brodbeck et al., 2018; Di Liberto et al., 2019), pitch (Krishnan et al., 2005; Teoh et al., 2019), and even semantic information (Broderick et al., 2019) in speech. We expand on these studies by investigating acoustic and linguistic features encoding in both controlled and noisy, naturalistic stimuli.

Part of the current study assesses whether and how responses to audiovisual stimuli may generalize to dissimilar audio-only contexts. Speech perception involves both auditory and visual cues, especially when a listener must comprehend speech in noisy environments. Integrating visual and auditory information enables the deciphering of speech from noise, particularly for those with hearing impairments (Altieri and Wenger, 2013; Maglione et al., 2015; Manfredi et al., 2018; Hendrikse et al., 2019; Puschmann et al., 2019). Visual information also has been shown to modulate responses to auditory speech using electrocorticography (ECoG) and fMRI (Ozker et al., 2018; Karas et al., 2019), with stronger modulation when audiovisual speech is clear compared with when one modality is corrupted. These and other studies show that incorporating visual information, including lipreading, can enhance speech perception.

Here we used EEG to model neural responses to speech to two entirely different stimulus sets—controlled sentences from the TIMIT corpus, and audiovisual stimuli from children's movie trailers (MTs). Our first goal was to determine whether acoustic and phonological encoding in EEG are stimulus dependent. One motivation was to quantitatively assess whether it is possible to replace some of the more monotonous stimulus sets with more engaging stimuli. In addition, by exploring how well encoding models trained on one stimulus set can generalize to another, we can determine the robustness of observed feature selectivity in EEG. Finally, we demonstrate that visual and auditory information may be encoded separately for some stimuli, and that the influence of visual information on auditory input is likely stimulus specific.

Materials and Methods

Participants. Seventeen participants with typical hearing (8 males; age, 20–35 years; mean age, 25.5 ± 4.5 years) were recruited from the University of Texas at Austin community. The ethnicity of our participants are as follows: 68% white, 13% Asian, 13% Hispanic, and 6% African American. All participants were native English speakers, but 88% of participants spoke one or more languages other than English. Pure tone and speech-in-noise hearing tests were performed using standard clinical procedures (American Speech-Language-Hearing Association, 2005) to ensure typical hearing ranges across all participants. Typical hearing responses for the pure tone test consisted of hearing thresholds <25 dB bilaterally for all frequency tones between 125 and 8000 Hz, tested separately for each ear. The QuickSIN test (Duncan and Aarts, 2006) was administered to assess typical hearing in noise [not >0–3 dB signal-to-noise (SNR) loss]. Participants provided their written consent for all portions of the experiment and were compensated $15/h for their participation. All experimental procedures were approved by the Institutional Review Board at the University of Texas at Austin.

Experimental design and statistical analyses. Two contrasting stimulus types were used in this study. The first set consisted of sentence stimuli from the TIMIT corpus, which included continuous sentences spoken in English by multiple male and female talkers with no background noise or overlapping sounds [Garofolo et al., 1993.] These stimuli also included transcriptions of the precise timing of the onset and offset of each phoneme and word. The second set of stimuli were children's movie trailers, which contained overlapping speakers, music, and background noise (https://trailers.apple.com/). While these stimuli were entirely unrelated to the TIMIT sentences, members of the laboratory similarly transcribed the onset and offset of each word and phoneme, alongside a high-level description of the auditory environment (e.g., speech, speech with background noise, or background noise only) using ELAN transcription software followed by automatic alignment using FAVE-align (https://zenodo.org/record/9846), a modified version of the Penn Phonetics forced aligner. These timings were then manually corrected using Praat software [https://uvafon.hum.uva.nl/praat/ (Praat: doing phonetics by computer, version 5.74)]. Each stimulus was transcribed by two authors (J.H., C.V., N.C.) to verify the reliability of the transcribed boundaries. Although TIMIT and movie trailer stimuli were qualitatively very different in terms of the types of sounds present, we verified that the distribution of phoneme counts was comparable across TIMIT and movie trailers (two-sample Kolmogorov–Smirnov test: $d = 0.1$, $p = 0.99$).

During the task, participants listened to 23 unique movie trailer stimuli alternated with TIMIT sentence stimuli in four blocks of 125 sentences each, and a fifth block of 100 sentences. The first four TIMIT blocks consisted of unique sentences with the exception of the final sentence. The fifth block contained 10 unique sentences with 10 repeats of each in a randomized order. The average length of each TIMIT sentence varied between 1.5 and 2 s long with an interstimulus interval of 1 s between each stimulus presentation. For the movie trailer stimuli, 23 unique movie trailers were used, and each was presented once; however, two unique stimuli (Inside Out and Paddington 2) were presented twice. The TIMIT sentences and movie trailers were presented through an iPad running custom software written in Swift (version 4; https://developer.apple.com/swift/), presented via an external monitor (see Data acquisition). While the precise recording time for the EEG experience differed across participants, participants heard an average of 1184 s of TIMIT sentences and 3055.11 s of movie trailers. The average length of the EEG recording time was 4856.27 s (~81 min).

During the task, participants were asked to watch and listen to the movie trailers but were not asked to attend to any particular speaker. For TIMIT, the participants were instructed to listen to the sentences while staring at a fixation cross. The overall task alternated between presenting five unique movie trailers and then one block of TIMIT sentences (125 sentences in the first four blocks and 100 in the fifth block). One participant (MT0007) was excluded because of the poor quality of the data, and another participant (MT0017) was excluded for the TIMIT data analysis only because of the poor quality of the data.

Data acquisition. Neural responses were continuously recorded from a 64-electrode scalp EEG cap at a sampling rate of 25 kHz using the BrainVision actiChamp system (Brain Products). The impedance level for the EEG signal was kept at <15 kΩ. Eye movements were measured through electrooculography (EOG), with vertical EOG and horizontal EOG measurements taken to aid in removing ocular artifacts from the neural data. The auditory stimuli were directly synchronized with the EEG data using a StimTrak stimulus processor (Brain Products). These stimuli were controlled by the experimenters outside of the EEG suite, with visual stimuli projected on a ViewPixx monitor inside the EEG suite. Audio levels were tested before the start of the task and were presented through insert earbuds (E-A-RTone Gold 10 Ω, 3M) at a comfortable volume.

Preprocessing. EEG and EOG data were downsampled to 128 Hz using BrainVision Analyzer software. The remaining neural preprocessing steps were conducted using customized Python scripts and functions from the MNE-python software package (Gramfort et al., 2013). First, EEG data were rereferenced offline to the average of the mastoid electrodes (TP9 and TP10) and notch filtered at 60 Hz to remove any electrical artifact. Data were then bandpass filtered between 1 and 15 Hz using a zero-phase, noncausal bandpass FIR (finite impulse response) filter (Hamming window, 0.0194; passband ripple with 53 dB stopband attenuation, −6 dB falloff). This filtering approach has been used in previous studies of neural tracking of speech using EEG (Di Liberto et al., 2015; O'Sullivan et al., 2015; Broderick et al., 2019). Raw data were visually inspected, and specific time points were manually rejected based on any nonbiological sources of movement, such as if the participant moved or clenched their jaw and created electromyographic noise. Less than 10% of the data were manually rejected. An independent component analysis was conducted to identify the response of components for eye blinks and saccade artifacts as the EOG responses were recorded in separate channels alongside the 64 channels of scalp EEG data. Components reflecting ocular movements were subsequently removed from the data.

EEG data were epoched according to the onset and offset of acoustic stimuli to analyze the EEG signals, which corresponded with the TIMIT and movie trailer stimuli. The onset of each trial was identified through a customized script using a match filter procedure (Turin, 1960), where the sound waveform of individual stimuli was convolved with the audio signal recorded to the EEG system, and the peak of the convolution was used to determine the offset and onset of each trial. Once data were epoched according to specific sentences or movie trailer stimuli, we then used the stimulus transcription textgrids to identify the timing of specific auditory and visual features.

Auditory and visual feature extraction. The auditory features extracted from our stimuli included phonological features, the acoustic envelope, and the pitch of each stimulus. For the phonological features, we created a binary phoneme feature matrix to indicate the timing of place and manner of articulation features for all phonemes for a given TIMIT sentence or for the movie trailer. Each element of the matrix was labeled with a 1, for the presence of a feature, and 0, for the absence of a given feature (Hamilton et al., 2018). Previous work using ECoG has demonstrated that speech-sensitive regions of the superior temporal gyrus respond to phonological features as opposed to the phonemes alone (Mesgarani et al., 2014a). Further research shows that these features are well tracked in EEG data as well (Di Liberto et al., 2015; Khalighinejad et al., 2017). Thus, we included the following place and manner of articulation features into the binary feature matrix: sonorant, voiced, obstruent, back, front, low, high, dorsal, coronal, labial, syllabic, plosive, fricative, and nasal. For example, the feature matrix would include a value of "1" at the onset of the "obstruent," "fricative," and "voiced" categories to indicate the onset of hearing the phoneme, '/v/'.

The acoustic envelope of each speech stimulus was extracted using the Hilbert transform followed by a low-pass filter (third-order Butterworth filter; cutoff frequency, 25 Hz). The envelope, which represents the dynamic temporal changes in speech (Raphael et al., 2007), was extracted for each of the individual TIMIT and movie trailer audio files. Prior research has also shown that the auditory cortex tracks the pitch of a given sound (Chung and Bidelman, 2016; Tang et al., 2017; Teoh et al., 2019; Hall and Planck, 2002). To fit encoding models to predict neural responses to pitch, we first computed the absolute pitch of each stimulus using the PraatIO package in Python (Jadoul et al., 2018), which provides a Python-based interface to Praat, the linguistics software. To ensure that the absolute pitch (fundamental frequency) of each stimulus yielded better model performance, we compared the fundamental frequency with binned representations of the pitch, as in the study by Tang et al. (2017). The binned pitches were calculated by extracting 10 log-spaced bins from 50 to 300 Hz. To determine whether a spectrogram representation might further improve model performance, we also fit a model incorporating the mel-band spectrogram with 15 frequencies spaced between 75 Hz and 8 kHz.

A major difference between the TIMIT stimuli and movie trailer stimuli is the presence of visual information in the movie clips. Since concurrent visual information can also affect encoding of auditory features (Grant and Seitz, 2000; Molholm et al., 2004; Beauchamp, 2005; Holcomb et al., 2005; Kaiser et al., 2005; Schneider et al., 2008; Besle et al., 2009; Chandrasekaran et al., 2009; Kayser et al., 2010; Başkent and Bazo, 2011; Crosse et al., 2015; Atilgan et al., 2018; Di Liberto et al., 2019; Puschmann et al., 2019), we wanted to control for this potential difference in modeling EEG responses to our auditory features. This analysis also relates to our overall goal of understanding how generalizable encoding models can be when derived from acoustically rich, audiovisual, naturalistic stimulus sets compared with TIMIT. Visual features were calculated for the movie trailer condition using a nonlinear Gabor motion energy filter bank (Nishimoto et al., 2011). Briefly, each frame of the movie was zero padded with black pixels at the top and bottom (to convert a 720 × 1280 pixel frame into 1280 × 1280 pixels), and then each image frame was downsampled to 96 × 96 pixels. These frames were then converted to grayscale by first transforming RGB pixel values into L*A*B* color space and retaining only the luminance channel. Next, each grayscale movie was decomposed into 2139 3D Gabor wavelet filters. These filters are created by multiplying a 3D spatiotemporal sinusoid by a 3D spatiotemporal Gaussian envelope. We used filters with five spatial frequencies, log spaced from 1.5 to 24 cycles/image, three temporal frequencies (0, 1.33, and 2.667 Hz), and eight directions (0–315° in 45° steps). Velocities were calculated over 10 frames of the movie at a time. We also included zero temporal frequency filters at 0°, 45°, 90°, and 135°, and one zero spatial frequency filter. The filters are positioned on each frame of the movie. Adjacent Gabor wavelets were separated by 4 SDs of the spatial Gaussian envelope. Each filter was also computed at two quadratic phases (0° and 90°), as in the study by Nishimoto et al. (2011). The Gabor features were then log transformed to scale down very large values. Finally, we took the first 10 principal components of this stimulus matrix to reduce the dimensionality of the Gabor basis function matrix. Reducing the dimensionality from 2139 to 10 principal components explained ∼60% of the variance in the data. This also corresponded to the point at which the second derivative of the variance explained curve approached zero (the "elbow" of the curve).

In addition to calculating the Gabor wavelet filters, we also identified time segments with contained scene cuts in the movie trailers. We wanted to assess whether these scene cuts contained comparable visual information. Authors M.D. and B.H. watched all 23 unique movie trailers and hand annotated the onset and offset timing information in ELAN (version 6.0 2020) every time a scene changed in all of the trailers. The neural data were epoched using the single-scene cut features. We assessed model performance by comparing the wavelet filters and scene cuts and found that the Gabor wavelets and scene cuts contained complementary information, with the best model performance incorporating both [Wilcoxon signed-rank test, full visual model with scene cut and Gabor wavelet versus Gabor only ($W = 22\,244.0$, $p = 5.44 \times 10^{-144}$) and full visual model versus scene cut ($W = 152030.0$, $p = 2.06 \times 10^{-31}$). By incorporating these visual features into our model, we were able to regress out any EEG activity related to both static and moving aspects of the visual stimulus in the movie trailer stimuli. This also allowed us to assess whether including visual feature information significantly changes the measured auditory feature encoding weights.

Encoding models for neural tracking of acoustic, linguistic, and visual features. To model EEG responses to both audio and audiovisual stimuli, we used a linear regression approach with different sets of acoustic, linguistic, and/or visual features. This approach is sometimes referred to as an encoding model, a spectrotemporal receptive field, a multivariate temporal response function, or simply a linear model (Theunissen et al., 2000; Mesgarani et al., 2014a,b; Di Liberto et al., 2015; Hamilton et al., 2018; Holdgraf et al., 2017). The goal of the forward-modeling TRFs is to describe the statistical relationship

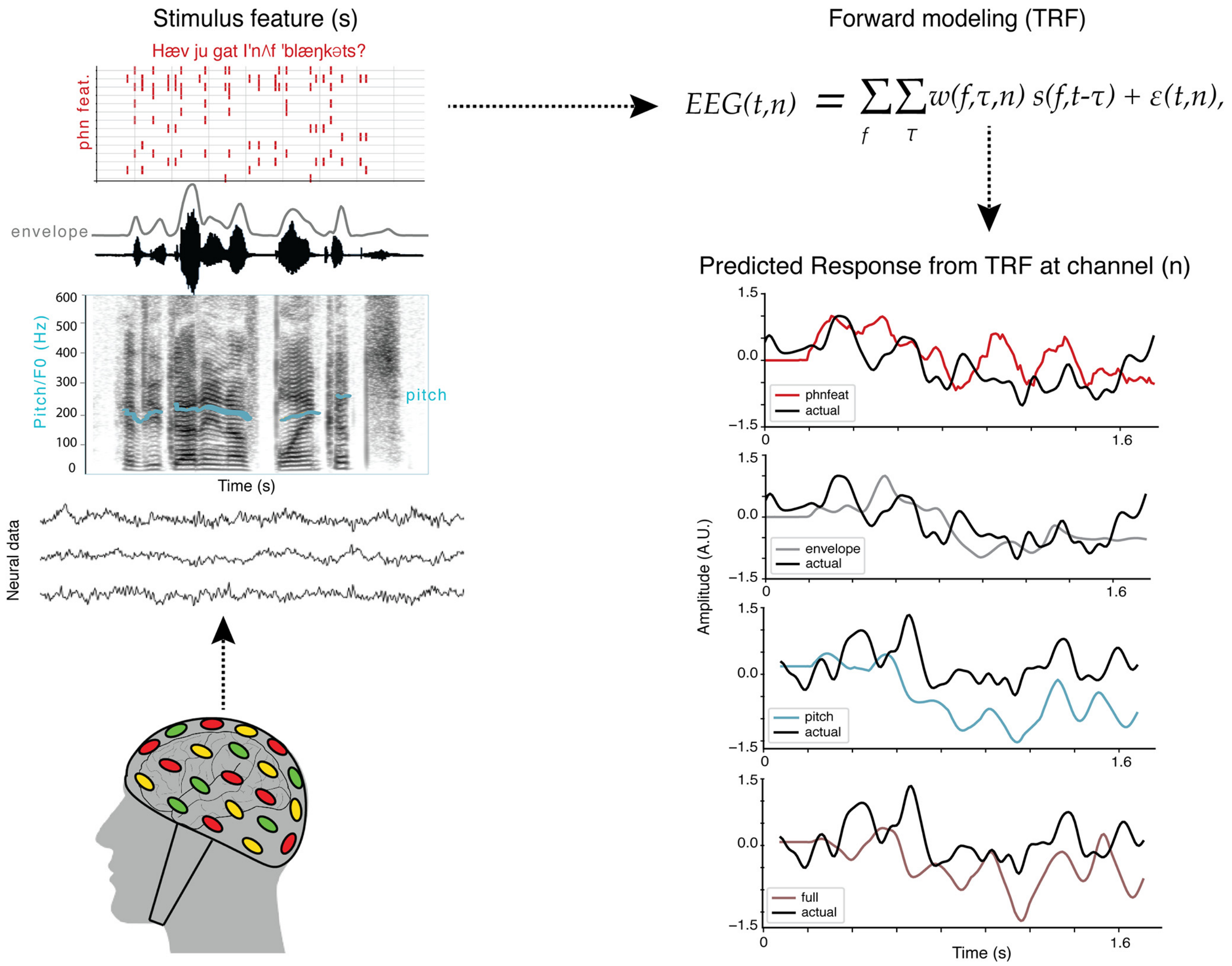

Figure 1. Analysis schematic showing encoding model framework for predicting EEG responses to a given speech feature. We fit encoding models to neural data collected from participants as they listened to sentences from the TIMIT corpus and as they watched movie trailers, which contained speech in the presence of background noise. Speech features included the acoustic envelope, phonological features, pitch, and a combined or full model consisting of all the aforementioned features. Forward modeling was used to compute the temporal receptive field (TRF), which was then used to predict the neural response to a specific speech feature (pitch, acoustic envelope, phonological feature) or a combination of features from a given EEG channel in both conditions from the EEG task (TIMIT and movie trailers). See Crosse et al., 2016 for a similar modeling paradigm.

between the input (auditory speech feature or visual feature) and output (the predicted EEG response based on the stimulus features).

All 64 channels were used for all EEG participants, and separate models were fit to predict the activity in each EEG channel. The equation for the forward model TRF is shown as follows:

$$\mathrm{EEG}(t,n) = \sum_{\mathrm{f}}\sum_{\tau} w(f,\tau,n)s(f,t-\tau) + \varepsilon(t,n).$$

This model calculates the instantaneous neural response EEG at time t from electrode n and is expressed as a convolution between an input speech stimulus property, $s(f,t-\tau)$, with the EEG TRF weights $w(f,\tau,n)$. The TRF is thus the mathematical transformation of a stimulus feature, f, into the EEG signal at different time lags τ. $\varepsilon(t,n)$ are the residual values from the linear regression otherwise not accounted for the input values. Figure 1 depicts all components of the forward modeling paradigm. The schematic for the predicted response from the TRF at a specific channel can be seen for the three unique features (pitch, acoustic envelope, and phonological features) as well as the combined model, which incorporates all three unique features as individual (f) values into the TRF model. For our audiovisual analysis, we could also simultaneously model responses to auditory and visual features of the stimulus. This framework allowed us to test different hypotheses about which features (acoustic, linguistic, or visual) were represented in the EEG data, and whether we could model how the brain tracks these features in a stimulus that contains background noise, music, or overlapping speech, with comparable fidelity to speech in isolation. For a schematic including visual features, see Figure 7.

For all acoustic and linguistic feature models, we fit multivariate temporal receptive fields (mTRFs) using time delays from 0 to 600 ms, which encompasses the temporal integration times for such responses as found in prior work (Hamilton et al., 2018). These analyses were performed in Python using custom scripts that implement cross-validated ridge regression. The weights (w) were fit using ridge regression on a subset of the data (the training set), with the regularization parameter chosen by a bootstrap procedure ($n = 100$ bootstraps) and tested on a validation set that was separate from our final test set. The ridge parameter was chosen as the value that resulted in the highest average correlation performance across all bootstraps and was set to the same value across electrodes. We tested ridge parameters of 0 as well as from 102 to 108, in 20 log-spaced steps. For TIMIT, the training set consisted of 489 of 499 sentences (the TIMIT blocks 1–4, described above). The 10 unique sentences that were heard in TIMIT block 5 and were also heard in blocks 1–4 were used as the test set, so no identical sentences were used in training and testing. For movie trailers, the training set consisted of 21 of the 23 movie trailers. The remaining two movie trailers were averaged and used as the test set to evaluate the model performance was then used to evaluate the model performance in which these repetitions were averaged and used as the test set for the model for the clean speech condition. The performance of the model was assessed by calculating the predicted EEG response for held-out data using separate model features, and then calculating the correlation between this predicted EEG and the actual held-out EEG. Noise-ceiling correction was applied using methods detailed in the study by Schoppe et al. (2016), where each of these correlations was divided by the maximum possible correlation given the trial-to-trial noise in the EEG data. In a separate analysis, we also tested the effect on model performance when repeating the movie trailers in the test set up to 10 times.

To statistically evaluate the model performance in TIMIT, we conducted a Friedman ANOVA, which is a nonparametric version of the repeated-measures ANOVA test. This was used rather than a standard ANOVA because our data violated the normality assumption. The dependent variable was the model performance, while the independent variable was the model type (all individual feature models and the full combined model).

To compare performance across models, we plotted the individual model performance against the full model for each possible feature. To visualize the general distribution of correlation values, we plotted each point separately in addition to the convex hull surrounding those points (see Fig. 3). The significance of each model was determined by randomly shuffling the stimulus labels in 2 s chunks and computing a model based on this randomized data (using the ridge parameter from the true, unshuffled data), calculating the correlation between the predicted and held-out data, and comparing that correlation to the unshuffled data. This was performed 100 times, which corresponded to a bootstrap p value < 0.01.

Variance partitioning. A consequence of using natural speech is that the stimulus features may be correlated with one another. For example, in the absence of background noise, the acoustic envelope is correlated with when phonological features occur in speech. To determine the shared versus unique contributions of each feature set, we used a variance partition analysis (de Heer et al., 2017). The unique variance for a given feature represents the amount of additional variance that is added when including those features in a model. The purpose of this analysis was to identify the individual contribution of each feature, the variance shared by pairs of features, and the variance explained by a combination of all features. R^2 values were calculated directly from the mTRF linear models. A total of seven unique features and intersections was used in this analysis.

The unique variance for a given feature was calculated by subtracting the R^2 for a paired model from the R^2 for a total model. For example, the unique variance for pitch was calculated by fitting the full model (envelope, phonological features, and pitch), and fitting a model with only envelope and phonological features. Taking the R^2 value for the full model and subtracting the R^2 value for the pairwise model would generate the unique contribution that was explained by adding pitch to the model. Unique individual features are as follows:

$$r^2_{\text{unique}_{\text{phnfeat}}} = r^2_{\text{full}} - r^2_{\text{pitch+envelope}}$$

$$r^2_{\text{unique}_{\text{envelope}}} = r^2_{\text{full}} - r^2_{\text{pitch+phnfeat}}$$

$$r^2_{\text{unique}_{\text{pitch}}} = r^2_{\text{full}} - r^2_{\text{phnfeat+envelope.}}$$

Equations for variation partitioning are shown below. In brief, we calculated model fits for each of the following features by using individual feature sets or the union of all pairwise and triplet combinations, as follows:

$$\text{Single models: Phonological feature (phnfeat), envelope, pitch}$$
$$\text{Pairwise models: phnfeat} \cup \text{envelope, envelope} \cup$$
$$\text{pitch, phnfeat} \cup \text{pitch} \quad \text{Full model: phnfeat} \cup \text{envelope} \cup \text{pitch.}$$

We then obtained the shared variance for each pair of models from the following equations:

$$\begin{aligned}
&\text{phn feat} \cap \text{envelope} = \text{phn feat} + \text{envelope} - \text{phn feat} \cup \text{envelope}\\
&\text{phn feat} \cap \text{pitch} = \text{phn feat} + \text{pitch} - \text{phn feat} \cup \text{pitch}\\
&\text{envelope} \cap \text{pitch} = \text{envelope} + \text{pitch} - \text{envelope} \cup \text{pitch.}
\end{aligned}$$

We then use these values to determine the intersection of all three models (the shared variance), as follows:

$$\begin{aligned}
&\text{phn feat} \cap \text{envelope} \cap \text{pitch} = \text{phn feat} \cup \text{envelope} \cup\\
&\text{pitch} + \text{phn feat} + \text{envelope} + \text{pitch} - \text{phn feat} \cup \text{envelope}\\
&- \text{phn feat} \cup \text{pitch} - \text{envelope} \cup \text{pitch.}
\end{aligned}$$

Finally, we could calculate the shared variance for each of the pairs of models without including the intersection of the combination of all three feature types:

$$\begin{aligned}
&(\text{phn feat} \cap \text{envelope})\backslash\text{pitch} = \text{phn feat} + \text{envelope} -\\
&\text{phn feat} \cap \text{envelope} - \text{phn feat} \cap \text{envelope}\\
&\cap \text{pitch}(\text{phn feat} \cap \text{pitch})\backslash\text{envelope} = \text{phn feat} + \text{pitch}\\
&- \text{phn feat} \cup \text{pitch} - \text{phn feat} \cap \text{envelope}\\
&\cap \text{pitch}(\text{envelope} \cap \text{pitch})\backslash\text{phn feat} = \text{envelope} + \text{pitch}\\
&- \text{envelope} \cup \text{pitch} - \text{phn feat} \cap \text{envelope} \cap \text{pitch.}
\end{aligned}$$

For some of the variance partition estimates, a correction needed to take place as certain models (single and pairwise) contained noise, some of which could be attributed to overfitting. In such cases, the variance partition calculations resulted in negative values where a *post hoc* correction was used to remove the biased values (de Heer et al., 2017).

In addition to the variance partitioning calculations, we conducted an additional correlation analysis for both TIMIT and movie trailer stimuli. For all acoustic features, we calculated the instantaneous correlation between the different sound feature representations (Fig. 2). This shows that the correlation structure for TIMIT and movie trailers is relatively similar, with the strongest differences in the spectrogram correlations, which are likely driven by the presence of multiple simultaneous sound sources in movie trailers but not TIMIT.

mTRF generalization across stimulus types. We next wished to assess the degree to which models trained on TIMIT sentences or movie trailers could generalize to the other stimulus set. This analysis allowed us to assess whether stimulus selectivity was similar across conditions, or whether neural tracking measures were stimulus specific. We conducted the same mTRF analysis for all speech features in both conditions from the EEG experiment to generate our weights and correlation values. We then used the weights calculated from using TIMIT as the training set to predict the neural responses of the respective speech features in the movie trailer EEG data. That is, using pretrained models from TIMIT, we then assessed model performance using movie trailers as the test set. We then compared the correlation between the new predicted and actual EEG response to the correlation values generated by our original analysis, where training and test data came from the same stimulus type. We then performed the same analysis using pretrained models from the movie trailer data, evaluated on the TIMIT test set. The purpose of this analysis was to see whether responses from TIMIT were generalizable to the responses from the movie trailers and vice versa. For example, to predict the EEG in response to the movie trailers from mTRFs calculated from TIMIT, as follows:

$$\text{EEG}_{\text{pred MT from TIMIT}}(t,n) = \sum_{\text{f}} \sum_{\tau} w_{\text{TIMIT}}(f,\tau,n)\, s_{\text{MT}}(f,t-\tau) + \varepsilon(t,n).$$

And to calculate EEG in response to TIMIT from mTRFs calculated from movie trailer stimuli:

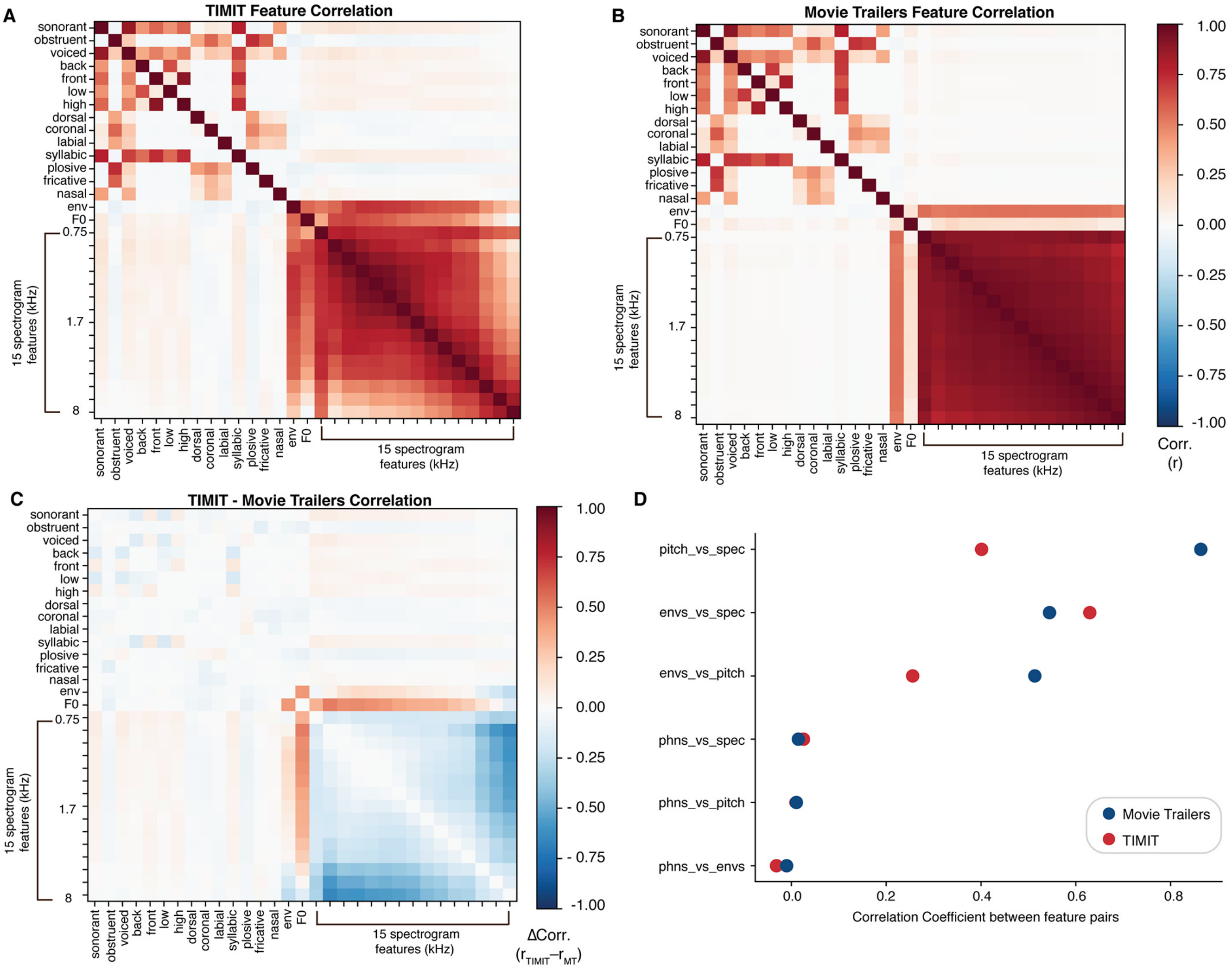

Figure 2. Speech feature correlation analysis for TIMIT and movie trailers. ***A***, Stimulus correlations within TIMIT show the co-occurrence of features within the TIMIT stimuli. For example, sonorant phonemes are likely to be voiced and syllabic. The co-occurrence of specific phonological features with spectral and pitch features was low, since sentences were spoken by a variety of speakers. ***B***, Stimulus correlations within the movie trailers. Co-occurrence of phonological features was highly similar to TIMIT. The spectrogram features were more correlated with one another, likely because of the presence of background noise, music, and other sounds in tandem with speech. ***C***, The difference in stimulus correlation values between TIMIT and movie trailers. Overall, phonological feature correlations were very similar, but differences were observed in the co-occurrence of low- and high-spectrotemporal information, with TIMIT showing separate epochs with low- or high-frequency content (but not both), and movie trailers showing epochs with frequencies across the spectrum. ***D***, Average stimulus correlations for acoustic and linguistic features in TIMIT and movie trailers. Acoustic features were generally more correlated than the phonological features, but the degree of correlation across stimulus sets was relatively similar.

$$\mathrm{EEG}_{\mathrm{pred\ \ TIMIT\ \ from\ \ MT}}(t,n) = \sum_{f}\sum_{\tau} w_{\mathrm{MT}}(f,\tau,n)s_{\mathrm{TIMIT}}(f,t-\tau) + \varepsilon(t,n).$$

Finally, we determined whether there was a relationship between the performance on the cross-stimulus models and the original models by computing the linear correlation between the r values when the test set was held constant. We also assessed the correlation between the weight matrices themselves by calculating the linear correlation between w_{TIMIT} and w_{MT} for each channel and participant.

Segmented correlation analysis. An additional difference between the movie trailers and TIMIT stimuli is that the trailers contained more acoustic information with multiple overlapping sound sources. While the mTRF and the variance partition analysis demonstrated correlation values between the predicted and actual EEG and the unique feature contribution to model performance, respectively, they did not reveal whether specific time points within the stimulus were more reliably predicted by our model. We thus implemented a segmented correlation analysis to assess the correlation between the predicted and actual EEG for given acoustic features based on whether the stimulus was speech alone, speech with background noise, or background only (no speech). The purpose of this analysis was to determine whether responses during speech alone were better modeled than speech in background noise. We segmented the data for one test set movie trailer (Inside Out) into intervals with speech only, speech and background, or background only using the software, Praat. We then calculated EEG predictions for only each interval of interest and assessed the correlation between actual and predicted EEG using each of the individual models (envelope, phonological features, pitch) or the full model.

The segmented correlations were then averaged within each of the individual auditory environments for models using each speech feature (phonological feature, pitch, and envelope) in isolation and in the full model. For the “background sounds only” auditory environment, the phonological feature model correlations were set to zero as there were no linguistic features from which to predict. We hypothesized that the phonological features during the clean speech only should generate higher correlation values overall than speech with background noise or background only.

Effect of number of repeats of test set on model evaluation performance. The purpose of this analysis was to identify whether increasing the number of repetitions for the movie trailer test set stimuli improved the overall model performance through increased EEG SNR. Rather than averaging all possible repetitions of the test set, we tested how the performance of the model varied as a result of averaging between 1 and 10 repetitions of the test set. Since most participants heard 10 repeats of the test set for TIMIT but only 2 repeats for movie trailers, we collected additional movie trailer data from one participant. This participant heard the two unique movie trailer test set stimuli a total of 10 times each, in addition to the original training set trailers.

We performed the same mTRF analysis on the full model (containing all acoustic and linguistic features) and compared encoding model performance between TIMIT and the movie trailers. To choose the repetition subsets, we used a bootstrapping analysis such that different random subsets of 1–10 repetitions were used in the test set, with a maximum of 10 bootstraps. This analysis examined how the model performance affected the average correlation value between the actual and predicted EEG data for each increase in test set stimuli in the movie trailer dataset.

Data availability. The code for reproducing the figures in this article is available at https://github.com/HamiltonLabUT/generalizable_EEG_manuscript. The EEG data are available at https://osf.io/p7qy8/?view_only=bd1ca019ba08411fac723d48097c231d.

Results

Acoustic and linguistic representations of speech

Speech occurs in many different environments in the real world. People may listen to speech in a quiet environment (e.g., listening to an audiobook), in the presence of light background noise (e.g., working at home with home appliance sounds, other individuals talking in common spaces), or even in highly noisy environments with a variety of background sounds (e.g., attempting to converse in a restaurant or at a music festival). Many encoding model approaches rely on responses recorded in relatively controlled stimuli in the absence of noise, or in specific controlled noise conditions. While such models have been helpful in improving our understanding of the neural representations of natural speech, it is unclear whether these findings might generalize to a more naturalistic and acoustically rich stimulus set, such as movie trailers. Here we investigated whether the brain tracks similar acoustic and phonological features in acoustically diverse stimuli (TIMIT sentences and movie trailers). We also investigated whether comparable model performance was observed across both conditions, and whether this depended on the specific features included in the model. We chose the acoustic envelope, phonological features, pitch, and a combination of these three models to encompass a broad range of features that are involved in speech processing (Mesgarani et al., 2014a, b; Di Liberto et al., 2018; Hamilton et al., 2018; Oganian and Chang, 2019).

We trained on 64-channel EEG data responses using a full model (combining the acoustic and phonological features of envelope, phonological features, and pitch into one model) and individual speech features in both speech conditions (envelope only, phonological features only, or pitch only). Model performance was assessed by correlating the EEG responses predicted by our model and the actual EEG responses for held-out data. We hypothesized that the combination of acoustic and linguistic features would contribute to higher model performance for movie trailers, since speech occurs in the presence of other unrelated sounds, and modeling these features separately can effectively lead to "denoising" of the data. In contrast, we expected that models based on predicting EEG from individual features (e.g., pitch or acoustic envelope or phonological features) should still perform relatively well when predicting responses to TIMIT.

We first compared the noise ceiling-corrected correlation values between the individual models against the full model (Fig. 3*A*,*B*). Model performance across all participants in TIMIT for the individual features was more similar to the full model performance (Fig. 3*A*). The average correlation value across all subjects for TIMIT was highest for the full model ($r = 0.35$) and phonological features ($r = 0.34$), with lower performance for the acoustic envelope ($r = 0.26$) and pitch ($r = 0.31$). Despite relatively similar performance, there was a significant effect of model type for TIMIT (Friedman's ANOVA χ^2 test $= 17$; df = 3, $p = 0.0005$). *Post hoc* Wilcoxon signed-rank tests showed that this effect was driven by significant differences between the full model and pitch ($W = 6.0$, $p = 0.0009$), along with the full model and the acoustic envelope ($W = 0.0$, $p = 0.00006$). In contrast, the full model and phonological features models were not statistically different from each other for TIMIT data ($W = 37.0$, $p = 0.17$). Performance differences between the phonological features model and pitch were not significant ($W = 27.0$, $p = 0.06$), but the phonological feature model was significantly better than the acoustic envelope model ($W = 18.0$, $p = 0.015$). The envelope model performed worse than the pitch model ($W = 19.0$, $p = 0.02$). Overall, these results demonstrate that the linguistic content from the phonological features seems to drive model performance.

We next performed the same comparisons for models fit on the movie trailer stimulus set, which included overlapping talkers and visual information (Fig. 3*B*). We found a significant effect of model type across participants (Friedman's ANOVA, $\chi^2 = 31.8$, df = 3, $p = 5.77 \times 10^{-7}$). *Post hoc* Wilcoxon signed-rank tests showed that the full model consistently outperformed the pitch model ($W = 0.0$, $p = 0.00003$), the acoustic envelope ($W = 1.0$, $p = 0.000061$), and the phonological feature model ($W = 1.0$, $p = 0.00001$). *Post hoc* tests demonstrated that, for movie trailers, correlations for the phonological feature model were statistically higher than for the pitch model ($W = 3.0$, $p = 0.0002$), the envelope model was statistically higher than pitch ($W = 26.0$, $p = 0.03$), and performance did not differ between the acoustic envelope and phonological feature models ($W = 43.0$, $p = 0.21$). Similar to TIMIT, the average correlation performance across all subjects for movie trailers was highest for the full model ($r = 0.10$) and phonological features ($r = 0.08$), with the acoustic envelope ($r = 0.07$) and pitch yielding lower average correlations ($r = 0.05$).

This first analysis shows how individual feature models compare to a joint model with acoustic and phonological features, but it does not explain how individual features may uniquely contribute to model performance. Thus, we used a variance partitioning analysis in which we computed the unique variance explained by each individual feature, all pairwise combinations of features, and the variance shared by all acoustic and phonological features (Fig. 3*C*). Overall, we saw higher shared variance across features for TIMIT and more unique variances contributed by each feature for movie trailers.

Individual features contributed relatively more unique information to the overall model performance for movie trailers (Fig. 3*D*, Venn diagram). As with TIMIT, the phonological features contributed the most unique variance across participants (Fig. 3*D*). Unlike TIMIT, the amount of shared information was lower for the movie trailers compared with TIMIT, as indicated by a smaller relative area for shared variance in the Venn diagram.

While these results showed strong tracking of acoustic and phonological features, we also tested whether other feature

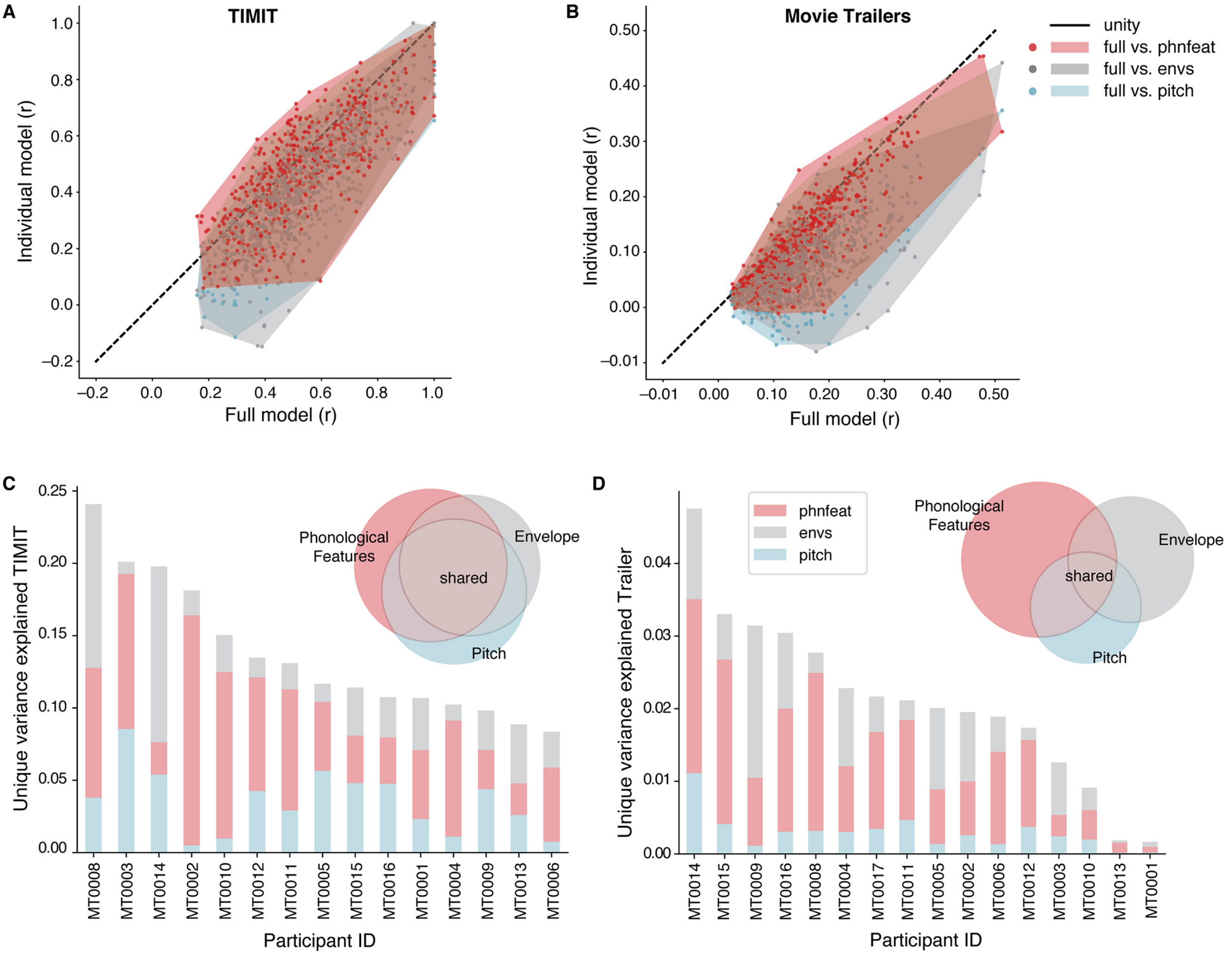

Figure 3. Contributions of phonological and acoustic representations in predicting EEG. Comparing individual speech features (pitch, acoustic envelope, phonological features) with the full model (combination of all auditory features). ***A***, Prediction performance (noise ceiling-corrected correlation between predicted and actual EEG data) of significant electrodes for each condition in models fit using TIMIT responses. Each individual dot is a single electrode. The shaded regions indicate the convex hull around the scatter points for each comparison, to indicate how the points are distributed along, above, or below the unity line. ***B***, Same as ***A***, for movie trailers. ***C***, Variance partition analysis shows the unique variance explained by individual features (phonological features, pitch, and envelope) for each participant separately (bar chart) and across all participants (pie chart) when fit on TIMIT data. ***D***, Same as ***C***, for movie trailers condition.

representations for pitch and acoustic information might yield better model performance. For example, we previously used binned pitch representations rather than the single fundamental frequency (F0) value to uncover pitch tuning in intracranial EEG (Tang et al., 2017). In this study, however, the single F0 model outperformed the binned pitch model (single F0 TIMIT, r_{avg} = 0.31; movie trailers, r_{avg} = 0.05; binned pitch: TIMIT, r_{avg} = 0.27; movie trailers, r_{avg} = 0.04). This was the case both for TIMIT (Wilcoxon signed-rank test: $W = 17575.0$, $p = 1.88 \times 10^{-7}$) and for movie trailers (Wilcoxon signed-rank test: $W = 199624.0$, $p = 3.32 \times 10^{-11}$). Thus, for all subsequent analyses, we used the single F0 feature. In addition to testing different representations for pitch, we tested whether spectrogram features would improve the performance of the full model. This was not the case—the overall correlations for both the full and individual models were lower when using the spectrogram instead of the envelope, suggesting that fine-grained spectrotemporal information is not as strongly represented in the EEG (Fig. 4).

Altogether, our results suggest that neural tracking occurs in response to multiple individual features. While the unique variance was highest for phonological features, we were also able to identify robust model performance for the acoustic envelope and additional unique variance explained by this feature. This suggests that the brain may not only track the envelope of speech, but also of the background sounds and music from the movie trailers. Overall, our results suggest that neural tracking for phonological features occurs both in noise-free as well as more uncontrolled, naturalistic settings despite the presence of varied background noise.

Are receptive field models from each condition generalizable to the other?

Our previous analysis showed that we were able to predict EEG in noise-free continuous speech and in naturalistic, noisy conditions using linear models that incorporated acoustic and linguistic features. We next asked whether models fit on one stimulus set would generalize so that they could predict neural responses for another stimulus type. This would allow us to answer whether naturalistic stimuli are a feasible replacement for receptive field analyses. We used the weights from the individual feature models and the full model calculated from responses to TIMIT to predict responses to untrained TIMIT and movie

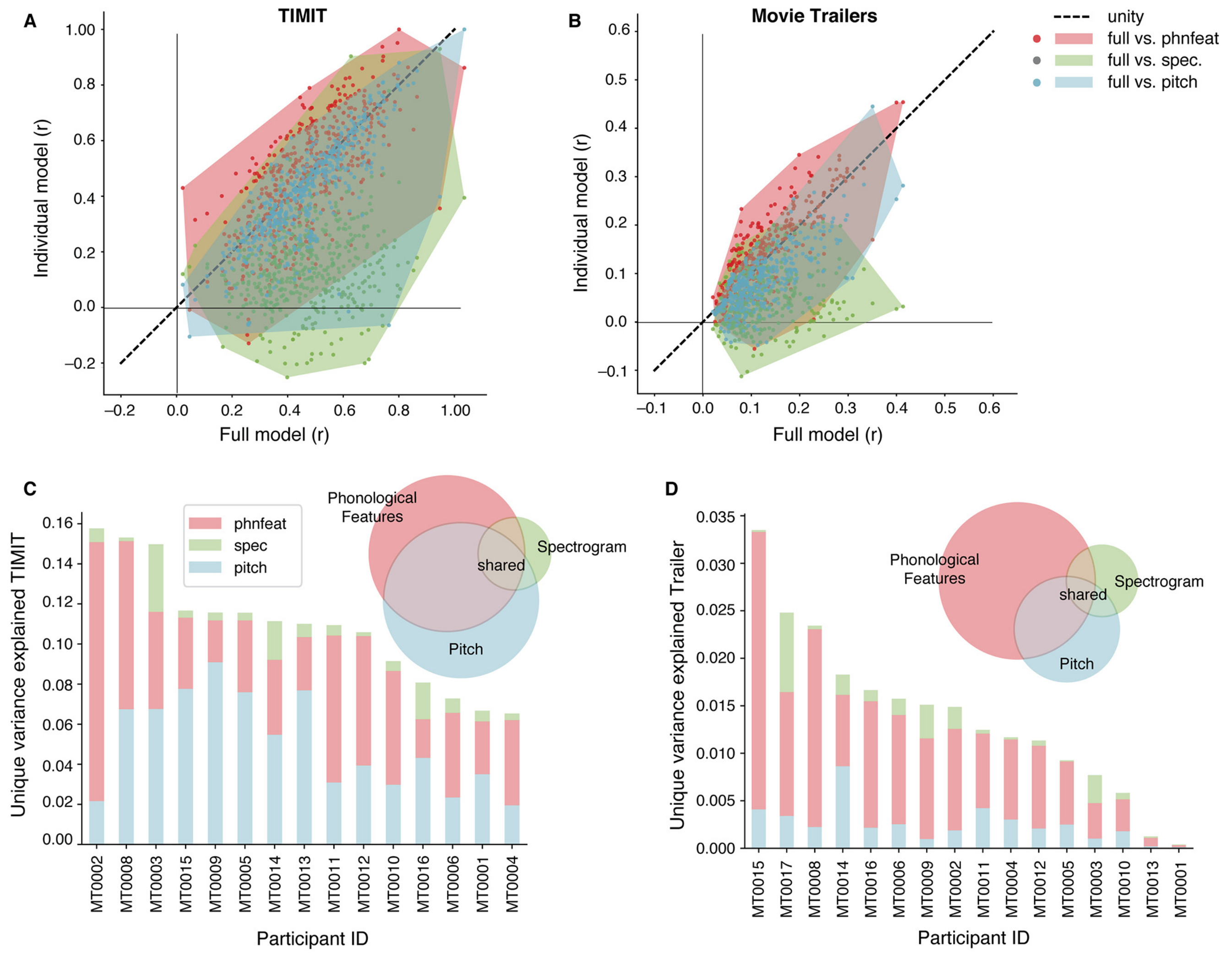

Figure 4. Model performance when incorporating spectrogram features in addition to phonological features and pitch. ***A***, Prediction performance (correlation between predicted and actual EEG data) of significant electrodes for each condition in models fit using TIMIT. Each individual dot is a single electrode. Electrode color indicates the individual model type. The shaded regions indicate the convex hull around the scatter points for each comparison, to indicate how the points are distributed along, above, or below the unity line. The average correlation value for each individual performance for TIMIT is as follows: phonological features, *r* = 0.33; spectrogram, *r* = 0.09; and pitch, *r* = 0.29. ***B***, Same as ***A***, for movie trailers. The average correlation value for each individual performance for movie trailers is as follows: phonological features, *r* = 0.07; spectrogram, *r* = 0.02; and pitch, *r* = 0.05. ***C***, Variance partition analysis shows the average unique variance explained by individual features (phonological features, pitch, and spectrogram) for each participant separately (bar chart) and across all participants (pie chart) when fit on TIMIT data. ***D***, Same as ***C***, for movie trailers condition.

trailer neural data. For each of these cross-predictions, we compared the performance of the model trained on one dataset and tested on the same stimulus type (e.g., predict TIMIT responses from TIMIT training data) or trained on a different stimulus type (e.g., predict TIMIT responses from movie trailer training data). This cross-prediction schematic is depicted in Figure 5*A*.

Results of this cross-prediction comparison are shown in Figure 5, *B* and *C*. Noise ceiling-corrected correlation values below the unity line indicate electrodes for which the within-stimulus model performance was better, and those above the unity line indicate better cross-stimulus performance. Overall, using the same stimulus type for training and testing (e.g., predict TIMIT from TIMIT or predict MT from MT) tends to result in better model performance. Still, the response to one stimulus could be modeled from the other, although with slightly worse performance. We saw a statistically significant correlation between performance in one condition (i.e., predict TIMIT from TIMIT or predict MT from MT) to the generalization condition (i.e., predict TIMIT from MT or predict MT from TIMIT; Fig. 5*B*,*C*, red regression lines). This was the case across all feature condition types, whether this be the individual models or the full model.

Finally, to determine whether models trained on TIMIT or movie trailer stimuli yielded similar encoding model weights, and thus inferred selectivity, we directly compared the TRF weights derived from TIMIT or movie trailer training data (Fig. 5*D*–*F*). Figure 5, *D* and *E*, shows the weights for one example channel using TIMIT (Fig. 5*D*) or movie trailers (Fig. 5*E*), which are highly visually similar and show the same pattern of positive and negative weights for each of the acoustic and phonological features. When compared quantitatively, weights derived from the different datasets were significantly correlated (with a maximum of up to 0.35 when averaged across 16 subjects; Fig. 5*F*).

Overall, we found that the model performance was better for TIMIT (Fig. 5*B*) when we trained and tested on TIMIT stimuli. While unsurprising, these responses were still able to be predicted from models using movie trailers as the training set,

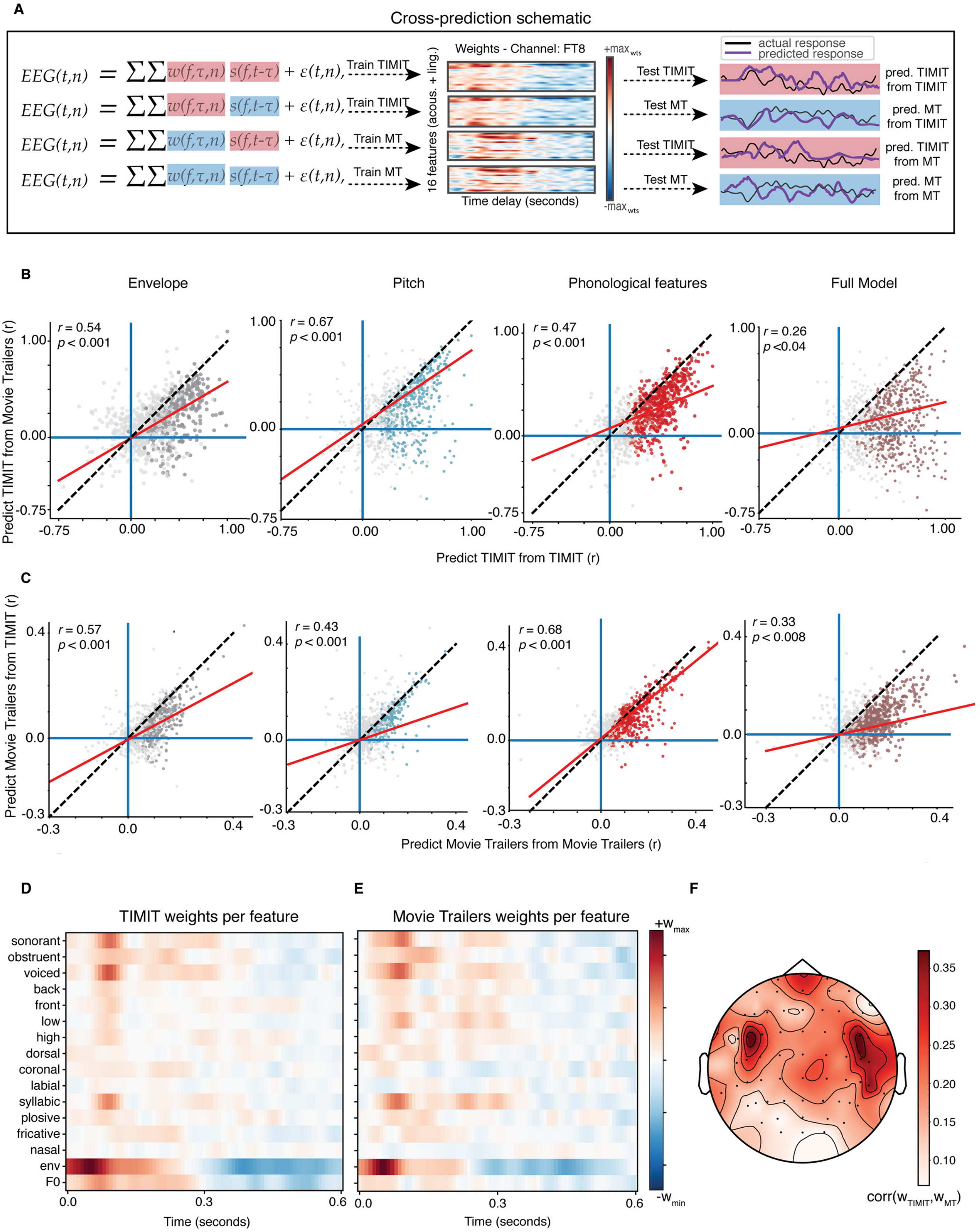

Figure 5. Cross-prediction analysis shows that models fit to TIMIT and movie trailer stimuli are generalizable to the other stimulus set. ***A***, The cross-prediction analysis schematic is shown for the following conditions: training on TIMIT and testing on TIMIT, training on TIMIT and testing on MT, training on MT and testing on MT, and training on MT and testing on TIMIT. In each of these training and testing conditions, the weights for the training condition were used for a given stimulus set and the stimulus feature for either the same or a different stimulus was used for testing. ***B***, Model performance for TIMIT test sets with TIMIT training data (*x*-axis) or movie trailer training data (*y*-axis). Each dot in the individual convex hull plots represents an individual electrode, where clusters of dots of the same color represent electrodes for a corresponding feature across participants. Dots in gray show nonsignificant models ($p > 0.05$, bootstrap test).

suggesting that feature encoding derived from neural responses to highly uncontrolled, noisy stimuli, can generalize to a more controlled stimulus set (and vice versa). The high model performance for predicting held-out TIMIT EEG data from models also trained on TIMIT data could partially be attributed to the fact that listeners do not have to filter extraneous background sounds while attending to the primary speech source, so the signal-to-noise ratio of responses is higher. In addition, the correlation structure of the training/test stimuli may play a role (Fig. 2). Prediction correlations were generally lower for the movie trailers but were still significantly higher than chance (Fig. 5*C*).

Although models fit on the movie trailers generalized to predicting responses to TIMIT, the model performance correlations for movie trailers were still lower on average. Thus, we next examined whether these results were because of differences in the amount of testing data for both TIMIT and movie trailers, or whether this was because of inherent differences in the amount of speech alone versus speech in background noise.

The effect of stimulus repetitions on model evaluation and performance

We observed overall lower model performance for all feature types in the movie trailer condition compared with TIMIT (Figs. 3-5). This led us to ask whether this discrepancy was because of the different number of test set repeats for TIMIT and movie trailers, or whether other factors were responsible. We hypothesized that more repetitions of test set stimuli would improve the average correlation performance of our encoding models, because averaging more EEG repetitions should increase the SNR of neural responses. We used the same TRF models in a single participant who had two separate recording sessions. In the first recording session, the participant listened to all blocks of TIMIT and listened to and watched all of the movie trailers. In this session, the participant watched and listened to two repetitions of the test set (Paddington and Inside Out) stimuli. In the second recording session, the same participant watched and listened to all of the movie trailers. However, during this session, the participant also heard Paddington and Inside Out a total of 10 times each. No additional TIMIT sentences were played in this second session. We assessed model performance for TIMIT in session 1 (Fig. 6*A*) and the repeated movie trailers in session 2 (Fig. 6*B*).

We observed an increase in model performance with increasing numbers of repeats for the TIMIT test set stimuli (Fig. 6*A*) and movie trailer test set stimuli (Fig. 6*B*). However, the average correlations for two repetitions of TIMIT were still greater than those of the movie trailers, even with 10 repetitions. This suggests that the overall signal-to-noise ratio was not the only factor driving poorer model performance for movie trailers compared with TIMIT. Including 10 repetitions of the movie trailer test set stimuli significantly improved model performance (Fig. 6*B*; $W = 1.0$, $p = 3.79 \times 10^{-12}$).

As with TIMIT, adding more test set repetitions for the movie trailer stimuli increased the observed model performance (Fig. 6*B*,*C*). This improvement may be attributed to averaging over more trials of EEG data, which decreases trial-specific noise in the response. However, because the correlation performance never reached that of TIMIT, other factors must also be at play.

As the TIMIT condition consisted of continuous speech without any background noise, we hypothesized that prediction performance for the movie trailer dataset might be higher for time points where no background noise was present. Thus, we parsed the movie trailer testing stimulus Inside Out into the following three different auditory environments: speech (~34% of the trailer presented), speech with background noise (~35%), and background noise only (~18%). Of note, the remaining 13% of Inside Out was silence, and the segmented analysis was conducted on all participants except for MT0001, as this participant did not hear this trailer as a part of the EEG experiment. We then calculated EEG predictions for each of these environments separately. We observed the highest correlations for speech without background noise, followed by speech with background noise, and finally background noise alone (Fig. 6*D*). The phonological features models outperformed the other models in speech alone and speech in background noise, but obviously could not adequately predict activity during background noise only (when no speech is present). While the magnitude of these correlations across all auditory environments was still lower than TIMIT, this suggests that the presence of background noise in the stimulus can also lead to less robust tracking of acoustic and phonetic features.

When comparing the performance of models within each of these auditory conditions quantitatively, we found a significant effect of model type for speech with noise (Friedman's ANOVA, χ^2 test $= 18.92$, $\text{df} = 3$, $p = 0.0003$), speech only (Friedman's ANOVA, χ^2 test $= 14.6$, $\text{df} = 3$, $p = 0.002$), and noise only (Friedman's ANOVA, χ^2 test $= 27.56$, $\text{df} = 3$, $p < 0.0001$). For speech only, *post hoc* Wilcoxon signed-rank tests showed that this effect was driven by significant differences between the full model and pitch model ($W = 2.0$, $p = 0.0001$), the full model and phonological features model ($W = 10.0$, $p = 0.002$), and the full model and acoustic envelope model ($W = 8.0$, $p = 0.002$). In addition, the phonological features and pitch models were significantly different ($W = 24.0$, $p = 0.04$), but the acoustic envelope was not statistically different from pitch ($W = 36.0$, $p = 0.19$), and phonological features were not statistically different from the acoustic envelope ($W = 47.0$, $p = 0.49$). Overall, this suggests that the joint information of phonological features and acoustic information is helpful in predicting EEG responses to clear speech without background noise, but the phonological feature model still performs well.

When comparing model performance for speech in noise, the full model was not significantly different from the phonological features model ($W = 26.0$, $p = 0.06$). In contrast, the full model significantly outperformed the envelope ($W = 0.0$, $p < 0.00006$) and pitch models ($W = 10.0$, $p = 0.003$). Phonological feature models showed higher performance than the individual pitch ($W = 15.0$, $p = 0.008$) and acoustic envelope models ($W = 25.0$, $p = 0.048$). The acoustic envelope and pitch models did not show statistically significant differences in performance ($W = 45.0$, $p = 0.42$). Similar to the clear speech condition, the phonological features model appears to drive performance even in the presence of background noise.

Finally, when assessing noise only, the full model and acoustic envelope showed similar performance ($W = 26.0$, $p = 0.06$), but the full model significantly outperformed the phonological features ($W = 1.0$, $p = 0.0001$) and pitch models

←

Dashed black line, Unity line; red line, regression line. ***C***, Same as ***B***, but with model performance on movie trailer test set for movie trailer training (*x*-axis) or TIMIT training data (*y*-axis). ***D***, Weight matrix for the TIMIT full model in one example channel 27 (FC6) from one participant (MT0008). ***E***, Same as ***D*** for movie trailers. Note that the weights are highly similar to those in ***D*** despite training the model on separate stimuli. ***F***, Average correlation between TIMIT and movie trailer weights for all participants. Across all participants, correlations among the receptive field weights were highest in central/temporal electrodes.

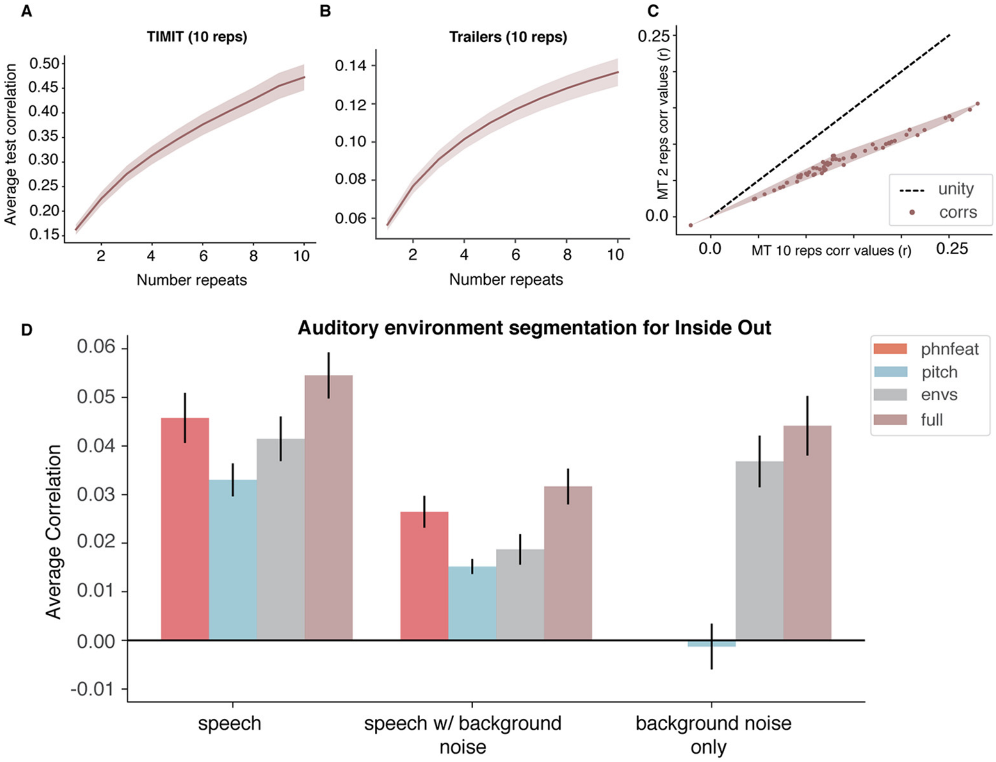

Figure 6. Effects of the number of repetitions and auditory environment on prediction performance. ***A***, The average correlation value between the predicted and actual EEG data in response to TIMIT increases as repetitions are added to the test set. ***B***, Same as ***A*** for movie trailers. ***C***, Convex hull showing correlation values in one subject from two separate recording sessions, where the *x*-axis consists of 10 repetitions of the test set stimuli, whereas the correlations shown on the *y*-axis were evaluated using two repetitions of the test set. Each brown dot represents a single electrode. Averages are across different partitions of the test set using a bootstrap procedure without replacement. ***D***, Average correlation performance for one test set trailer (Inside Out) in which the stimulus was segmented as speech only, speech with background noise, and background noise alone. Models were fit on individual speech features (phonological features, acoustic envelope, pitch) or a combination of all three (full model). Performance was highest during speech-only epochs.

($W = 1.0$, $p = 0.0001$). The poor performance of the model for phonological features is to be expected here, as these features are all zeros during "noise only" epochs. On the other hand, the background sounds may have associated pitch, but this did not appear to model responses well, and in fact the pitch model performed similarly to phonological features ($W = 55.0$, $p = 0.80$). We found that the acoustic envelope and pitch ($W = 2.0$, $p < 0.0002$) and phonological features and acoustic envelope ($W = 0.0$, $p < 0.0001$) were significantly different. The similar performance of the envelope and full model indicates that, in the presence of background sounds that do not include speech, the envelope is adequate for modeling EEG responses to auditory information.

Audiovisual components for speech tracking in a noisy environment

Up to this point, we have examined tracking of specific acoustic or phonological features in TIMIT and movie trailer stimuli. However, one obvious difference between the two stimuli is that the movie trailers also include visual information, which has been shown to influence auditory perception (Beauchamp, 2005; Holcomb et al., 2005; Schneider et al., 2008; Di Liberto et al., 2018). If we wish to replace a more controlled stimulus like TIMIT with an uncontrolled stimulus like movie trailers, we must also examine to what extent the visual features influence auditory feature selectivity. As a final analysis, we examined how the visual components were involved in speech tracking of the movie trailer stimuli. We built the same linear regression models, which now included all of the audio features (pitch, phonological features, and acoustic envelope) and added an additional set of visual features. The visual features were calculated from a motion energy model in which movie stimuli were decomposed into a set of spatiotemporal Gabor wavelet basis functions (Nishimoto et al., 2011) and manually annotated scene cuts.

The spatiotemporal Gabor wavelets allow us to investigate visual feature selectivity using Gabor wavelets that capture both static and moving visual aspects of the movie trailer. Figure 7*A* illustrates some example spatiotemporal features, with each row representing a feature, and each column representing the evolution of that feature over time. The relative weights for each of these Gabor wavelets are used to construct a visual feature matrix (Fig. 7*B*) that describes visual motion parameters over time. To make the problem more tractable and to include a comparable number of auditory and visual features, we reduced the dimensionality of the Gabor feature matrix using a principal component analysis (PCA). A final example of the features that are used in the full audiovisual model is shown in Figure 7*C*, with some video stills to illustrate the visual scene at four example time points in our test set stimulus. We also used an additional "scene cut" visual feature alongside the Gabor wavelets for the visual and combined auditory and visual analysis, which improved model performance over the Gabor filters alone (see Materials and Methods).

Next, we determined whether including visual features in the movie trailer model fitting influenced our ability to assess model performance between audio- and visual-only responses (Fig. 7*D*). We used the noise ceiling-corrected correlation values and found that the average performance scores across all subjects for

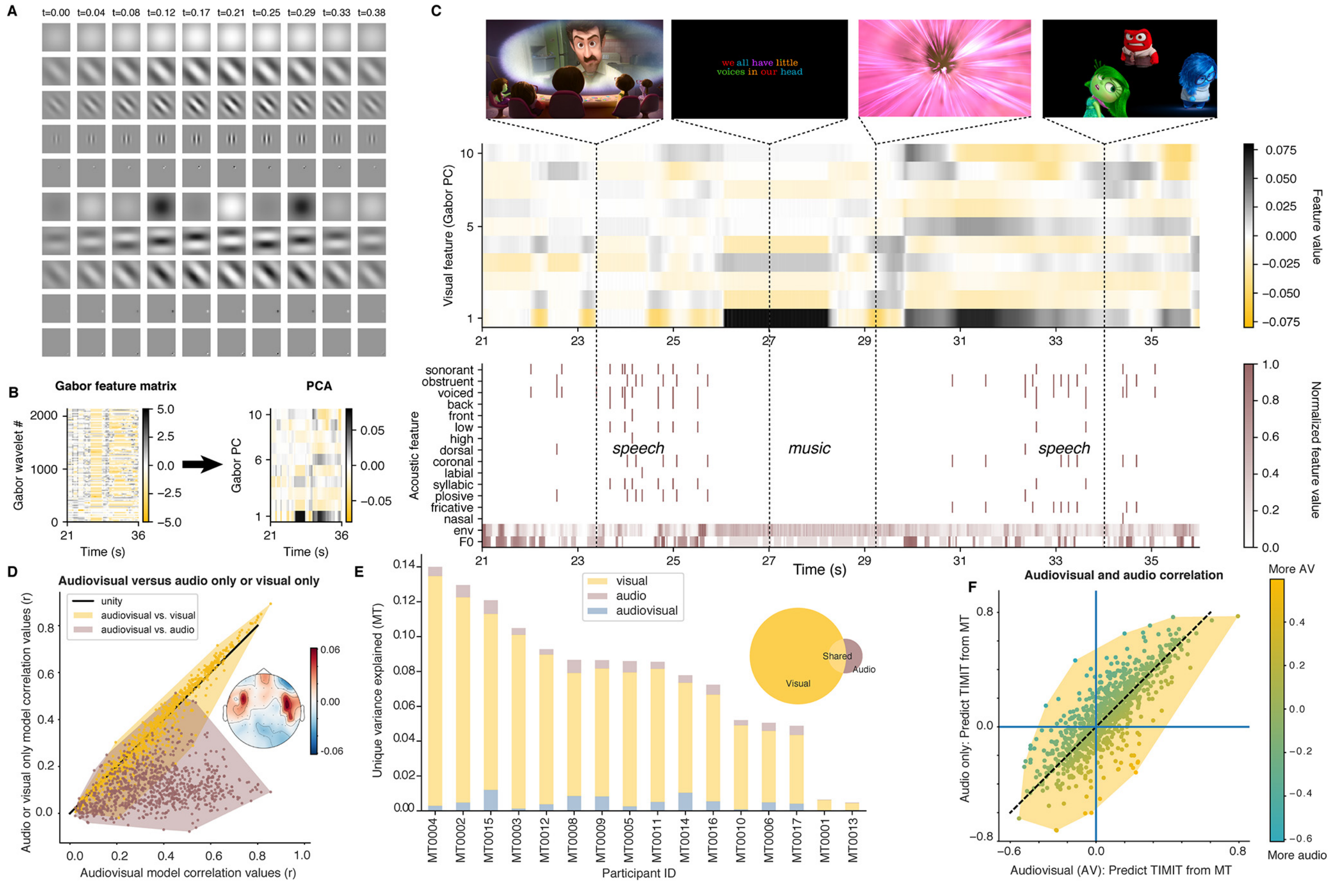

Figure 7. Unique contribution of visual features for the noisy (movie trailer condition) to assess model performance. ***A***, Visual stimuli in the movie trailers were decomposed into a set of Gabor wavelet features using a motion energy model. These features are static or drifting gratings at different spatial and temporal frequencies. Ten example spatiotemporal Gabors are shown, where each row represents one spatiotemporal feature set, and each column represents the evolution of that feature over time. In our experiment, we used a total of 2139 features, so these represent only a small fraction of the total set. ***B***, The 2139 Gabor features are decomposed into their first 10 principal components using PCA on the entire Gabor feature matrix (2139 features over time). For illustration purposes, only 15 s of data are shown. This reduced dimensionality matrix then serves as the visual input to our mTRF models. ***C***, Example combined visual and acoustic/linguistic features for 15 s of our test stimulus Inside Out. The acoustic features are identical to those used in the previous model fits, while the visual features include the Gabors shown in ***A–C***. Example frames are shown for four time points in the stimulus. ***D***, Model performance for the audiovisual combined model versus visual-only or audio-only models. Each dot represents an individual EEG channel. The topographic map shows the difference between the audiovisual correlations and visual-only correlations, averaged across all participants. Red indicates increased variance explained when adding auditory information. ***E***, Unique variance explained by visual, auditory, or combined audiovisual information. The visual features contribute a large amount of variance of the EEG responses to movie trailers. Such results are further corroborated by the pie chart, demonstrating unique variance among the visual-only, audio-only, and shared audiovisual features. ***F***, A comparison of the cross-prediction analysis where both an audio-only model or an audiovisual model was used to derive feature weights for movie trailers. The auditory weights for each analysis were then used to predict TIMIT responses. Overall, these models are highly correlated, showing that partialing out the visual information does not strongly affect cross-prediction performance. Each dot represents an individual EEG channel.

the full auditory model was $r = 0.10$, whereas adding the visual information into the auditory model yielded a correlation value of $r = 0.33$, across all 64 EEG channels. Combining auditory and visual information resulted in more robust model performance than just having the auditory features alone (Wilcoxon signed-rank statistical test, $W = 6201.0$, $p < 0.0001$; Fig. 7*D*, mauve points). Surprisingly, the visual information alone (maximum, $r = 0.35$) had a better model performance compared with the combined audiovisual model, and we found that that these two models were not statistically significant from one another (Wilcoxon signed-rank statistical test, $W = 255416.0$, $p = 0.46$). However, this was mainly the case for EEG electrodes outside the typical auditory ROI (Fig. 7*D*, inset, topographic map). We found that the visual stimulus information contributed a significant proportion of variance in the audiovisual speech condition (maximum, $r = 0.8$). In fact, the visual components occupied a significantly larger subspace in the bar plot compared with the audio-only or the joint audiovisual components (Fig. 7*E*). Despite this large contribution of visual information to overall explained variance, the shared region of the Venn diagram shows that each of these features (audio vs audiovisual) can be modeled separately, suggesting that this particular audiovisual stimulus does not significantly alter how the auditory features of speech are tracked in the movie trailers. To test this directly, we conducted a cross-prediction analysis (Fig. 5) to predict neural responses to TIMIT from models fit using movie trailer training data with audio features only, and to predict TIMIT from models fit using movie trailer data with the combined audiovisual features (Fig. 7*F*). Notably, no visual information was presented during TIMIT stimulus presentation, so the visual features are set to zero in the TIMIT stimulus matrix used for the prediction. We found that correlation values for both models were concentrated long the unity line, suggesting that the model performance was comparable for both conditions, and that regressing out the visual information did not significantly alter the predictability of the response to auditory information. Finally, we demonstrated that the auditory-related weights fit on audio-only and audiovisual stimuli (including Gabor wavelet features plus scene cuts)

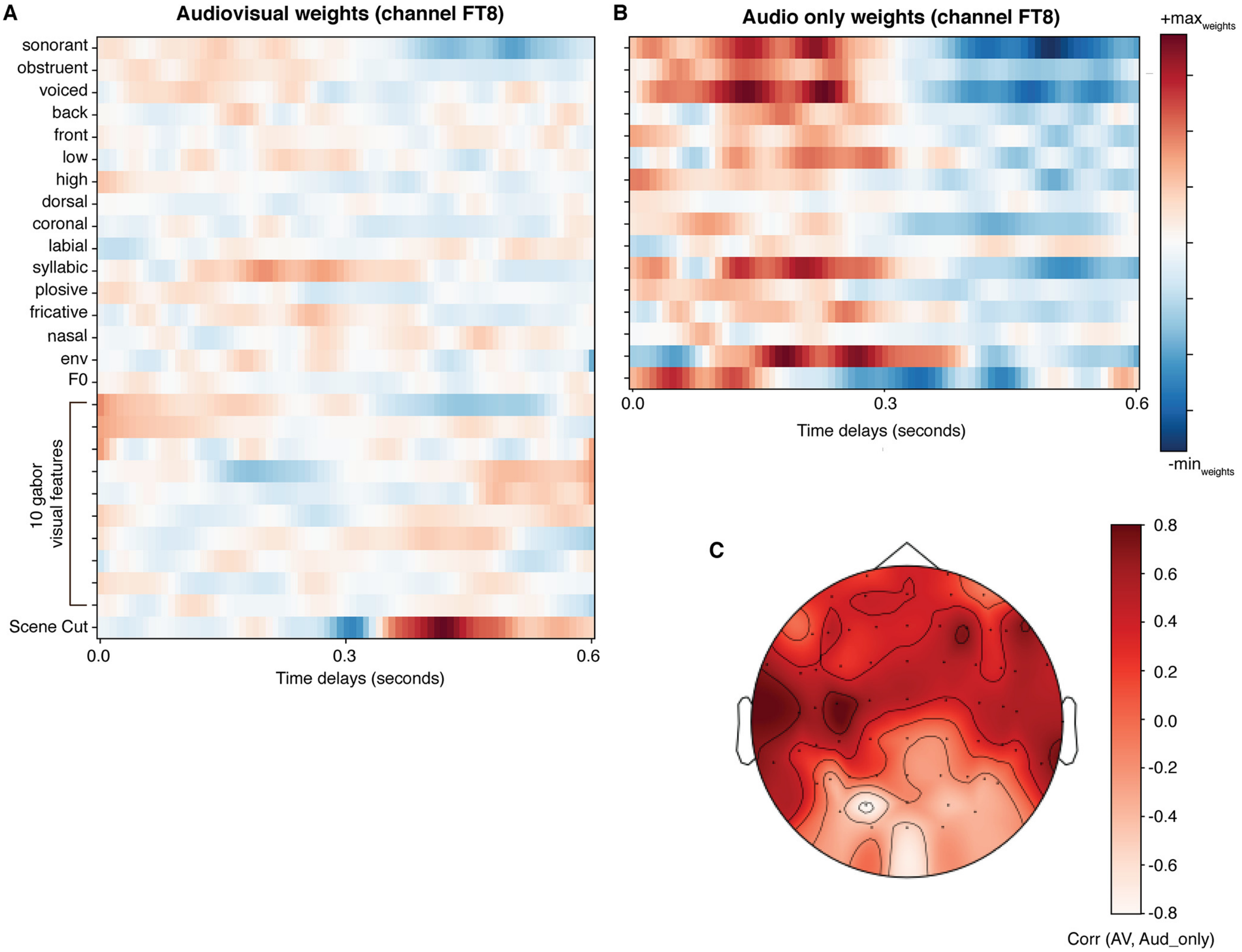

Figure 8. Weights from the full auditory encoding model are similar to the auditory information in the audiovisual model. ***A***, Weights from the audiovisual model that included all 16 auditory features along with 10 Gabor wavelet features and one visual feature for scene cuts. ***B***, Heat map of weights from the full auditory model without visual information (16 features). ***C***, Topographic map of the correlation between auditory feature weights from the audiovisual model and audio-only model averaged across participants. Model weights were highly correlated, especially over temporal and central electrodes.

were similar, suggesting that regressing out the visual information does not alter auditory encoding information (Fig. 8).

Discussion

Understanding speech in real-world scenarios involves parsing noisy mixtures of acoustic information that may include speech and nonspeech sources. Natural environments are also inherently multisensory. Much of our understanding of how the brain processes speech relies on tightly controlled stimuli in the absence of noise or parametrically controlled noise added to continuous speech stimuli. While these endeavors have been highly fruitful in uncovering acoustic and phonological tuning in the brain, it is not clear to what extent models based on controlled stimuli can generalize to more complex stimuli. Taken further, in some experimental environments, it may be desirable to investigate speech processing using stimuli that are more enjoyable to listen to, so more data could be acquired without participants becoming bored or frustrated. This could include, for example, clinical populations or research involving children. Still, absent is an evaluation of how such models generalize to more controlled datasets, where it is difficult to interpret whether models based on naturalistic stimuli reflect the same processes involved in parsing more controlled stimuli.

In this study, we addressed these questions by collecting neural responses to acoustic and linguistic features in clean speech (TIMIT) and naturalistic noisy speech (movie trailers) conditions using EEG. We were able to show that tracking of phonological features, pitch, and the acoustic envelope were both achievable and generalizable using a multisensory stimulus.

In our first analysis, we described how individual acoustic or phonological features could predict brain data, and how that compared with a full model incorporating all of these features. Generally, the acoustic envelope and phonological features were more predictive of EEG data than pitch, which contributed a small but significant proportion of unique variance for both TIMIT and movie trailers. A possible reason for lower correlation values in the individual pitch models in both speech conditions (TIMIT and movie trailers) could be attributed to the fact that pitch tracking may be more robustly identified in higher frequencies of the EEG, for example, the following response (Krishnan, 1999; Galbraith et al., 2000; Zhu et al., 2013).

While our model performance correlation values were substantially lower in the movie trailers compared with TIMIT, it may be that the background noise corrupted measurement of the acoustic envelope from the individual sources. This effect on the envelope may have been partially mitigated by including additional features for the pitch and phonological features in the movie trailer dataset. The full model always outperformed any one feature set alone, and the shared variance was relatively low for this dataset (Fig. 2*D*). This suggests that including these additional features allows for more robust assessment of neural tracking.

Perhaps most promising for moving toward more natural stimulus experiments is our finding that encoding models for TIMIT and movie trailers are generalizable and not stimulus specific. It was possible to use the weights fitted from the movie trailer condition to predict responses to the TIMIT sentences and vice versa. This indicates that using more naturalistic stimuli is a valid approach for characterizing auditory receptive fields. This corroborates findings by Jessen et al. (2019), who built encoding models to analyze EEG responses to naturalistic audiovisual stimuli in infants, and found robust encoding of both the auditory envelope and visual motion features. Di Liberto et al. (2018) also showed largely separable responses to auditory and visual information in an EEG study on speech entrainment in participants with dyslexia. In our dataset, the visual components did not affect the auditory encoding model weights to any significant degree. While others have shown that visual information can significantly affect encoding of auditory information, much of the visual information in the movie trailers was not directly related to the speech. For example, some scenes were narrated, so the speaker was not visible, and in many scenes the action of the characters went beyond speech-related mouth movements. For a movie clip where much of the speech is accompanied by a view of the person talking (and their mouth moving), we might expect more overlap in the variance explained by auditory and visual features (O'Sullivan et al., 2017, 2020; Ozker et al., 2018). Future studies could investigate this directly by comparing movies where the auditory and visual information are either correlated (as in a "talking head" interview) or decorrelated (narration of a visual scene).

Finally, the segmentation analysis demonstrated that times in the movie trailers where speech occurred without background noise could be more robustly predicted than times where background noise was present. While the prediction performance was still worse compared with using TIMIT as the input speech feature, model performance improved with more test set averaging. In future studies, it may be useful to separate the speech and nonspeech stimuli and fit the same encoding models on multiple tiers of speakers and nonspeech sounds to assess model performance. Others have demonstrated that multitalker separation is possible with deep neural networks (Luo et al., 2018), but separating music, speech, and sound effects present in movie trailers is significantly more complex.

While we were able to predict EEG responses to phonological and acoustic features in continuous speech, a caveat is that some nonlinear effects may not be well modeled. For example, in traditional auditory "odd-ball" tasks, the response to the same stimulus may differ based on prior expectations of its appearance (Squires et al., 1975). This would not be captured by our mTRF modeling, which assumes the same response to acoustic and phonological features for every presentation. Future studies could incorporate contextual sensitivity into these models to better explain such nonlinearities.

Finally, an additional caveat of this study is that participants were not asked about the semantic content of either stimulus; nor was attention assessed directly. However, other studies incorporating naturalistic stimuli with mixed noise have shown strong entrainment to speech even in the presence of strong background noise (Ding and Simon, 2013), separable responses to mixtures of music stimuli (Treder et al., 2014), and that the degree of speech intelligibility can be predicted from envelope entrainment (Vanthornhout et al., 2018).

Our study demonstrates that movie trailer stimuli can be used to identify acoustic and phonological feature tuning that can still predict responses to more controlled stimuli. This suggests that researchers can use stimuli that may be both more representative of our daily environment and more enjoyable to listen to or watch. Our results also provide some intuition for how mTRF model performance changes based on stimulus characteristics as well as the amount of data and number of repetitions. Last, while visual responses added significant variance to the EEG responses to the movie trailers, this visual information did not significantly change auditory tuning. Overall, our results provide evidence that robust auditory and audiovisual selectivity can be uncovered using more naturalistic, multimodal stimuli. This is promising for clinical research or research on populations who may not tolerate or become fatigued by traditional psychoacoustics paradigms.

References

Akbari H, Khalighinejad B, Herrero J, Mehta A, Mesgarani N (2019) Towards reconstructing intelligible speech from the human auditory cortex. Sci Rep 9:874.

Altieri N, Wenger MJ (2013) Neural dynamics of audiovisual speech integration under variable listening conditions: an individual participant analysis. Front Psychol 4:615.

American Speech-Language-Hearing Association (2005) Guidelines for manual pure-tone threshold audiometry. Rockville, MD: Author. Available from https://www.asha.org/policy/gl2005-00014/.

Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, Lee AKC, Bizley JK (2018) Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. Neuron 97:640–655.e4.

Başkent D, Bazo D (2011) Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. Ear Hear 32:582–592.

Beauchamp MS (2005) See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. Curr Opin Neurobiol 15: 145–153.

Besle J, Bertrand O, Giard M-H (2009) Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. Hear Res 258:143–151.

Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.

Broderick MP, Anderson AJ, Lalor EC (2019) Semantic context enhances the early auditory encoding of natural speech. J Neurosci 39: 7564–7575.

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5: e1000436.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432.

Chung W-L, Bidelman GM (2016) Cortical encoding and neurophysiological tracking of intensity and pitch cues signaling English stress patterns in native and nonnative speakers. Brain Lang 155-156:49–57.

Crosse MJ, Butler JS, Lalor EC (2015) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J Neurosci 35:14195–14204.

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. Front Hum Neurosci 10:604.

de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The hierarchical cortical organization of human speech processing. J Neurosci 37:6539–6557.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Di Liberto GM, Peter V, Kalashnikova M, Goswami U, Burnham D, Lalor EC (2018) Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. Neuroimage 175:70–79.

Di Liberto GM, Wong D, Melnik GA, de Cheveigné A (2019) Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196:237–247.

Ding N, Simon JZ (2012) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89.

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735.

Duncan KR, Aarts NL (2006) A comparison of the HINT and quick SIN tests. J Speech Lang Pathol Audiol 30:86–94.

Fiedler L, Wöstmann M, Herbst SK, Obleser J (2019) Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. Neuroimage 186:33–42.

Fuglsang SA, Dau T, Hjortkjær J (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage 156:435–444.

Galbraith GC, Threadgill MR, Hemsley J, Salour K, Songdej N, Ton J, Cheung L (2000) Putative measure of peripheral and brainstem frequency-following in humans. Neurosci Lett 292:123–127.

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS (1993) DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1, Vol 93. NASA STI/Recon Technical Report N.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R (2013) MEG and EEG data analysis with MNE-Python. Front Neurosci 7:267.

Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection of spoken sentences. J Acoust Soc Am 108:1197–1208.

Hall DA, Plack CJ (2009) Pitch processing sites in the human auditory brain. Cereb Cortex 19:576–585.

Hamilton LS, Huth AG (2018) The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci 35:573–582.

Hamilton LS, Edwards E, Chang EF (2018) A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr Biol 28:1860–1871.e4.

Hausfeld L, Riecke L, Valente G, Formisano E (2018) Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. Neuroimage 181:617–626.

Hendrikse MME, Llorach G, Grimm G, Hohmann V (2019) Realistic audiovisual listening environments in the lab: analysis of movement behavior and consequences for hearing aids. In: Proceedings of the 23rd International Congress on Acoustics integrating 4th EAA Euroregio 2019: 9-13 September 2019 in Aachen, Germany (Ochmann M, Vorländer M, Fels J, eds). (pp 7616–7622). Berlin: Deutsche Gesellschaft für Akustik.

Holcomb PJ, Anderson J, Grainger J (2005) An electrophysiological study of cross-modal repetition priming. Psychophysiology 42:493–507.

Holdgraf CR, Rieger JW, Micheli C, Martin S, Knight RT, Theunissen FE (2017) Encoding and decoding models in cognitive electrophysiology. Front Syst Neurosci 11:61.

Horton C, D'Zmura M (2011) EEG reveals divergent paths for speech envelopes during selective attention. Int J Bioelectromagnetism 13:217–222.

Huth AG, Lee T, Nishimoto S, Bilenko NY, Vu AT, Gallant JL (2016) Decoding the semantic content of natural movies from human brain activity. Front Syst Neurosci 10:81.

Jadoul Y, Thompson B, de Boer B (2018) Introducing Parselmouth: a Python interface to Praat. J Phon 71:1–15.

Jessen S, Fiedler L, Münte TF, Obleser J (2019) Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie. NeuroImage 202:116060.

Kaiser J, Hertrich I, Ackermann H, Mathiak K, Lutzenberger W (2005) Hearing lips: gamma-band activity during audiovisual speech perception. Cereb Cortex 15:646–653.

Karas PJ, Magnotti JF, Metzger BA, Zhu LL, Smith KB, Yoshor D, Beauchamp MS (2019) The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. Elife 8:e48116.

Kayser C, Logothetis N, Panzeri S (2010) Visual influences on information representations in auditory cortex. In Front. Neurosci. Conference Abstract: Computational and Systems Neuroscience.

Khalighinejad B, Cruzatto da Silva G, Mesgarani N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. J Neurosci 37:2176–2185.

Krishnan A (1999) Human frequency-following responses to two-tone approximations of steady-state vowels. Audiol Neurootol 4:95–103.

Krishnan A, Xu Y, Gandour J, Cariani P (2005) Encoding of pitch in the human brainstem is sensitive to language experience. Brain Res Cogn Brain Res 25:161–168.

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G (2013) The tracking of speech envelope in the human cortex. PLoS One 8:e53398.

Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915.

Luo Y, Chen Z, Mesgarani N (2018) Speaker-independent speech separation with deep attractor network. IEEE/ACM Trans Audio Speech Lang Process 26:787–796.

Maglione AG, Scorpecci A, Malerba P, Marsella P, Giannantonio S, Colosimo A, Babiloni F, Vecchiato G (2015) Alpha EEG frontal asymmetries during audiovisual perception in cochlear implant users. Methods Inf Med 54:500–504.

Manfredi M, Cohn N, De Araújo Andreoli M, Boggio PS (2018) Listening beyond seeing: event-related potentials to audiovisual processing in visual narrative. Brain Lang 185:1–8.

Mesgarani N, Cheung C, Johnson K, Chang EF (2014a) Phonetic feature encoding in human superior temporal gyrus. Science 343:1006–1010.

Mesgarani N, David SV, Fritz JB, Shamma SA (2014b) Mechanisms of noise robust representation of speech in primary auditory cortex. Proc Natl Acad Sci U S A 111:6792–6797.

Molholm S, Ritter W, Javitt DC, Foxe JJ (2004) Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study. Cereb Cortex 14:452–465.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646.

O'Sullivan AE, Crosse MJ, Di Liberto GM, Lalor EC (2017) Visual cortical entrainment to motion and categorical speech features during silent lipreading. Front Hum Neurosci 10:679.

O'Sullivan AE, Crosse MJ, D Liberto GM, de Cheveigné A, Lalor EC (2020) Neurophysiological indices of audiovisual speech integration are enhanced at the phonetic level for speech in noise. bioRxiv. Advance online publication. Retrieved September 9, 2021. .

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706.

Oganian Y, Chang EF (2019) A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci Adv 5:eaay6279.

Ozker M, Yoshor D, Beauchamp MS (2018) Converging evidence from electrocorticography and BOLD fMRI for a sharp functional boundary in superior temporal gyrus related to multisensory speech processing. Front Hum Neurosci 12:141.

Puschmann S, Daeglau M, Stropahl M, Mirkovic B, Rosemann S, Thiel CM, Debener S (2019) Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise. Neuroimage 196:261–268.

Raphael LJ, Borden GJ, Harris KS (2007) Speech science primer: physiology, acoustics, and perception of speech. Baltimore: Lippincott Williams and Wilkins.

Schneider TR, Debener S, Oostenveld R, Engel AK (2008) Enhanced EEG gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. Neuroimage 42:1244–1254.

Schoppe O, Harper NS, Willmore BD, King AJ, Schnupp JW (2016) Measuring the performance of neural models. Front Comput Neurosci 10:10.

Shankweiler D, Studdert-Kennedy M (1966) Lateral differences in perception of dichotically presented synthetic CV syllables and steady-state vowels. J Acoust Soc Am 39:1256.

Squires NK, Squires KC, Hillyard SA (1975) Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr Clin Neurophysiol 38:387–401.

Tang C, Hamilton LS, Chang EF (2017) Intonational speech prosody encoding in the human auditory cortex. Science 357:797–801.

Teoh ES, Cappelloni MS, Lalor EC (2019) Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. Eur J Neurosci 50:3831–3842.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331.

Treder MS, Purwins H, Miklody D, Sturm I, Blankertz B (2014) Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. J Neural Eng 11:026009.

Turin G (1960) An introduction to matched filters. IEEE Trans Inform Theory 6:311–329.

Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T (2018) Speech intelligibility predicted from neural entrainment of the speech envelope. J Assoc Res Otolaryngol 19:181–191.

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T (2014) Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One 9:e112575.

Zhu L, Bharadwaj H, Xia J, Shinn-Cunningham B (2013) A comparison of spectral magnitude and phase-locking value analyses of the frequency-following response to complex tones. J Acoust Soc Am 134:384–395.