Electrophysiological indices of hierarchical speech processing differentially reflect the comprehension of speech in noise

Abbreviated title: EEG speech indices differentially reflect comprehension

Shyanthony R. Synigal[1], Andrew J. Anderson[2], Edmund C. Lalor[1,2]

[1]Department of Biomedical Engineering, University of Rochester, 201 Robert B. Goergen Hall, P.O. Box 270168, Rochester, NY 14627, USA.

[2]Department of Neuroscience and Del Monte Institute for Neuroscience, University of Rochester, 201 Robert B. Goergen Hall, P.O. Box 270168, Rochester, NY 14627, USA.

Corresponding Author: Edmund C. Lalor PhD, Department of Biomedical Engineering, University of Rochester, 201 Robert B. Goergen Hall, P.O. Box 270168, Rochester, NY 14627, USA; Edmund_Lalor@urmc.rochester.edu

Number of pages: 38

Number of figures: 6

Number of tables: 3

Number of words for abstract: 230

Number of words for introduction: 660

Number of words for discussion: 1,492

Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements and Support

This work was supported by a grant from the Simons Foundation Autism Research Initiative (SFARI). Additional support was provided by the Del Monte Institute for Neuroscience. The authors thank Ms. Xueying Wang and Dr. Aaron Nidiffer for some assistance with data collection, and Dr. Aaron Nidiffer and Dr. Madeline Cappelloni for helpful comments on the manuscript.

ABSTRACT

The past few years have seen an increase in the use of encoding models to explain neural responses to natural speech. The goal of these models is to characterize how the human brain converts acoustic speech energy into different linguistic representations that enable everyday speech comprehension. For example, researchers have shown that electroencephalography (EEG) data can be modeled in terms of acoustic features of speech, such as its amplitude envelope or spectrogram, linguistic features such as phonemes and phoneme probability, and higher-level linguistic features like context-based word predictability. However, it is unclear how reliably EEG indices of these different speech representations reflect speech comprehension in different listening conditions. To address this, we recorded EEG from neurotypical adults who listened to segments of an audiobook in different levels of background noise. We modeled how their EEG responses reflected different acoustic and linguistic speech features and how this varied with speech comprehension across noise levels. In line with our hypothesis, EEG signatures of context-based word predictability and phonetic features were more closely correlated with behavioral measures of speech comprehension and percentage of words heard than EEG measures based on low-level acoustic features. EEG markers of the influence of top-down, context-based prediction on bottom-up acoustic processing also correlated with behavior. These findings help characterize the relationship between brain and behavior by comprehensively linking hierarchical indices of neural speech processing to language comprehension metrics.

SIGNIFICANCE STATEMENT

Acoustic and linguistic features of speech have been shown to be consistently tracked by neural activity even in noisy conditions. However, it is unclear how signatures of low- and high-level features covary with one another and relate to behavior across these listening conditions. Here, we find that categorical phonetic feature processing is more affected by noise than acoustic and word probability-based speech features. We also find that phonetic features and word probability-based features better correlate with measures of intelligibility and comprehension. These results extend our understanding of how various speech features are comparatively reflected in electrical brain activity and how they relate to perception in challenging listening conditions.

INTRODUCTION

Given the importance of speech communication in human life, tremendous amounts of research have focused on characterizing the neurophysiology of language comprehension (Hickok 2015). This research has revealed a network of brain areas that are functionally specialized for processing different hierarchical levels of speech and language (Hickok and Poeppel 2007). For example, work has shown that low level acoustic and spectrotemporal features of speech are chiefly processed in early auditory cortex (de Heer et al. 2017), with various phonological features being processed in secondary areas like the superior temporal gyrus (Hamilton, Edwards, and Chang 2018; Mesgarani et al. 2014) and some prefrontal areas (Burton 2009; de Heer et al. 2017), and meaning being represented across large areas of cortex (Huth et al. 2016; Anderson et al. 2017; Pereira et al. 2018).

While much of the above knowledge has been obtained from functional neuroimaging and invasive recordings in neurosurgical patients, parallel efforts have been made to obtain noninvasive magneto- and electrophysiological (MEG/EEG) markers reflecting hierarchical speech features. This includes modeling how EEG and MEG track the amplitude envelope of natural speech (Lalor and Foxe 2010a) and how neural responses reflect the spectrotemporal (Di Liberto, O'Sullivan, and Lalor 2015; Daube, Ince, and Gross 2019), phonetic (Di Liberto, O'Sullivan, and Lalor 2015), phoneme-level probability (Di Liberto et al. 2019; Gwilliams et al. 2020; Brodbeck, Hong, and Simon 2018), lexical (Heilbron et al. 2022), prosodic (Teoh, Cappelloni, and Lalor 2019), and semantic (Heilbron et al. 2022; Broderick et al. 2018) features of natural speech.

Some advantages of EEG are that it is significantly cheaper and easier to use in applied research in different cohorts (Peck et al. 2021; Salisbury et al. 2002). Consequently, there has been considerable interest in exploring how different EEG markers of speech processing reflect speech intelligibility (Verschueren, Vanthornhout, and Francart 2021) and language comprehension (Broderick et al. 2022; Ahissar et al. 2001). Many of these studies altered the intelligibility or comprehensibility of speech by adding background noise (Iotzov and Parra 2019) or by degrading the speech signal itself (Viswanathan et al. 2021). For example, some work has shown that cortical tracking of the speech envelope decreases as noise levels increase (Etard and Reichenbach 2019; Lesenfants et al. 2019; Vanthornhout et al. 2018; Zou et al. 2019), although others have suggested such tracking remains robust until the background noise is more than twice as loud as the speech it masks (Ding and Simon 2013). Meanwhile, experiments that have explored EEG indices of semantic processing appear to show a strong correlation with speech intelligibility and/or understanding (Broderick et al. 2018). Very few studies, however, have systematically explored how EEG indices of both low- and high-level speech processing covary across different levels of speech comprehension (Strauß et al. 2022; Yasmin et al. 2023). This is the goal of the present study.

We explore how EEG markers of speech envelope, spectrogram, acoustic onsets, phonetic features, and lexical surprisal processing vary with subjective measures of the percentage of words heard (as a proxy for intelligibility) and objective measures of comprehension across different background noise conditions. We hypothesize that EEG measures of higher-level processing (e.g., lexical surprisal and phonetic features) will more strongly correlate with behavior than lower-level measures (e.g., envelope tracking). We also test how listeners exploit linguistic context to process noisy speech and how that effect might manifest in EEG. Here, we leverage a recently introduced measure of predictive speech perception that quantifies how the tracking of low-level speech features varies as a function of the context-based semantic content of that speech (Broderick, Anderson, and Lalor 2019), but this time using lexical surprisal. We hypothesize that this measure strengthens for speech in moderate levels of noise (when speech is still intelligible) relative to speech in quiet, before falling off at high levels of background noise (when speech is no longer intelligible). With this study we seek to extend our understanding of the hierarchical processing of continuous speech under a range of realistic listening conditions.

METHODS

Participants

28 healthy adults (9 males, 18-35 years old) participated in this study. One subject was excluded due to an insufficient amount of data and two were excluded due to technical issues, resulting in a dataset of 25 participants. Each participant provided written informed consent and reported having normal hearing, normal or corrected-to-normal vision, English as their first and main language, and no history of neurological disorders. Participants were also compensated for their participation. All procedures were approved by the University of Rochester Human Subjects Review Board.

Stimuli and experimental procedure

Participants listened to 70 minutes of *A Wrinkle in Time* by Madeleine L'Engle which was read by an American female speaker. Each trial was one minute long and was presented at one of five noise levels: quiet (no noise) and +3 dB, -3 dB, -6 dB, and -9 dB signal-to-noise ratios (SNRs). The background noise was spectrally matched stationary noise, which was estimated from the clean speech using a 46th order forward linear predictive model. The prediction order was calculated based on the sampling rate of the audio clips (Crosse, Di Liberto, and Lalor 2016; Ding and Simon 2013). There were 14 minutes' worth of audio for each of the five noise conditions. The storyline was preserved from trial to trial, but the conditions were pseudorandomized such that no noise level occurred consecutively. Participants rated how many words they heard (on a scale of 0-100%) and answered two multiple choice comprehension questions after each trial. The comprehension questions used here were the same questions used from a previous study (Maddox and Lee 2018), except we presented only two out of the four original questions created for each trial. The stimuli were presented through Sennheiser HD650 headphones at a sampling rate of 44.1 kHz using Psychtoolbox (Kleiner, Brainard, and Pelli 2007) and custom MATLAB scripts (MATLAB 2019).

Data acquisition and preprocessing

EEG data were recorded from 128 scalp electrodes (plus two mastoid channels that were not analyzed in this work). The data were acquired at a 1024 Hz sampling rate with the BioSemi Active Two system. The data were preprocessed using the PREP pipeline and its default parameters (Bigdely-Shamlo et al. 2015). This pipeline first used detrending to high pass filter the data at 1 Hz followed by 60 Hz line noise removal. Afterwards, robust re-referencing was applied which allows the data to be referenced to an average of all channels except those contaminated with noise. This function identifies and interpolates noisy channels in an iterative manner such that the re-referencing itself is not affected by the noise. The cleaned data was then low pass filtered at 8 Hz, using a filter with an 8.5 Hz cutoff frequency and 80 dB stopband attenuation. Next, the data were epoched and independent component analysis (ICA) was applied using EEGLAB's picard function (Delorme and Makeig 2004; Pion-Tonachini, Kreutz-Delgado, and Makeig 2019) to remove muscle and eye artifacts. Lastly, the data were downsampled to 128 Hz.

Speech stimulus characterization

Speech is organized in a hierarchical manner where sounds can form syllables, syllables form words, words form sentences, and so on. To assess how our brains might concurrently process speech across levels of this hierarchy, we chose to model EEG responses to speech based on several different representations, all of which were computed on the clean versions of each trial.

Envelope. We first calculated the speech envelope, a well-established feature shown to be robustly tracked by cortical activity (Aiken and Picton 2008; Destoky et al. 2019; Di Liberto, O'Sullivan, and Lalor 2015; Ding and Simon 2013; Etard and Reichenbach 2019; Lalor and Foxe 2010b; Nourski et al. 2009; Pasley et al. 2012) and to be important for speech recognition and intelligibility (Ahissar et al. 2001; Drullman, Festen, and Plomp 1994; Shannon et al. 1995). The speech signal was first lowpass filtered at 20 kHz (22.05 kHz cutoff frequency, 1 dB passband attenuation, 60 dB stopband attenuation). The broadband speech envelope was calculated using a gammachirp auditory filterbank to mimic the filtering properties of the cochlea (Irino and Patterson 2006). This filterbank was used to filter the speech into 16 bands from 250 Hz to 8 kHz with an equal loudness contour (essentially creating a spectrogram). Lastly, the frequency bands were averaged together.

Acoustic onsets and spectrogram. We chose to model two additional acoustic features, acoustic onsets and spectrogram, which were shown to be reflected in cortical activity above and beyond the speech envelope (Brodbeck et al. 2020; Di Liberto, O'Sullivan, and Lalor 2015; Sohoglu and Davis 2020). Acoustic onsets were approximated by computing the first derivative of the speech envelope and then half-wave rectifying the result. A 16-band spectrogram was calculated using the same filterbank and parameters as the speech envelope, just without the final averaging step.

Phonetic features. To calculate phonetic features, the Montreal Forced Aligner (McAuliffe et al. 2017) was first used to partition and time align each word in the story into phonemes according to the International Phonetic Alphabet for American English. Then, each phoneme was linearly mapped onto a set of 19 binary phonetic features based on the University of Iowa's phonetics project (http://www.uiowa.edu/~acadtech/phonetics/english/english.html/).

Lexical surprisal. Lastly, we calculated the surprisal of each word based on its preceding context using the Transformer-XL model (Dai et al., 2019). This model contains a recurrence mechanism that allows it to build and reuse memory from previous segments and learn longer-term dependencies, while preserving the temporal information of previous word embeddings. This model was chosen because it can predict the probability of an upcoming word using the context from all preceding words. The softmax of the values from the output layer of the model were taken to estimate the probability of each word, and the negative log of a word's probability was computed to estimate lexical surprisal (Dai et al. 2019).

Modeling the relationship between speech features and EEG responses

One goal of the present study was to find how the encoding of individual speech features changes with SNR and find how those changes relate to comprehension and what participants reported hearing. To index the encoding of individual speech features, we used a forward model which acted as a filter or kernel that described the transformation from those features to the EEG responses recorded at each electrode. Here, we modeled acoustic onsets, the spectrogram, phonetic features, and lexical surprisal. An additional vector with impulses placed at each word's onset was included to capture any acoustic related onset responses that the acoustic onset predictor may have missed. All five features were modeled together to control for variance explained by the competing features and to find which feature best explained certain EEG responses (Brodbeck, Presacco, and Simon 2018; Gillis et al. 2021; Brodbeck, Hong, and Simon 2018).

The data from each of the five experimental (SNR) conditions were modeled separately using 14-fold leave-one-out cross-validation and ridge regression. Each feature was normalized between 0-1 and the EEG were z-scored. The features were then concatenated and partitioned into training and test sets. The stimuli were lagged from -100-700ms to capture both short and long latency responses to acoustic and linguistic features. Cross-validation was conducted on the clean condition to select the optimal regularization parameter, λ, which ranged from 10^{-1}–10^{8}. We identified the regularization parameter that resulted in the highest prediction accuracy for each individual test fold. We then selected the parameter that produced the highest reconstruction accuracy most often (across all test folds) so that we could use one parameter to train the models for each condition. Using the same parameter for all folds and conditions (within a participant) allowed for a fairer comparison of model performance since each participant's trials would be on the same scale and it minimized model overfitting.

A temporal response function (TRF), $\mathrm{w}\,(\tau, \mathrm{n})$, was trained using the selected regularization parameter and the training data to predict the neural responses, $r(t, n)$, from the set of concatenated speech features, $s(t-\tau)$. Then, we separated the model into the segments that corresponded to acoustic onsets, spectrogram, phonetic features, and surprisal. The forward modeling procedure can be expressed as follows, including the residual response, $\varepsilon(t, n)$, not explained by the model:

$$r(t,n) = \sum_{\tau} w(\tau, n)\, s(t-\tau) + \varepsilon(t, n)$$

Each model segment was tested on the held-out data, and its performance was assessed by quantifying the correlation between the predicted EEG and the actual EEG. Forward model performance, or prediction accuracy, was averaged across all folds. The results from the acoustic onset, spectrogram, and phonetic feature model segments were then averaged across 12 well-predicted frontotemporal electrodes (6 bilaterally symmetric pairs), and the results from the surprisal model were averaged across 12 well-predicted parieto-occipital channels. There may be some positive bias in the prediction accuracies (i.e., higher accuracies) since we have selected well predicted electrodes based on the current dataset, but we assumed that this bias is present in all conditions and does not vary systematically across conditions.

Backward modeling, where EEG is used to reconstruct an estimate of the speech envelope was also employed to enable the current results to be directly referenced to other studies in which backward modeling is more common. The backward model or decoder, $g(\tau, n)$, describes the transformation from lagged EEG responses at all electrodes, $r(t+\tau, n)$, to an estimate of the speech envelope, $\hat{s}(t)$. As detailed elsewhere (Crosse et al. 2016), the modeling procedure can be expressed as:

$$\hat{s}(t) = \sum_{n} \sum_{\tau} r(t+\tau, n)\, g(\tau, n)$$

Similar to the forward models, this analysis was conducted separately for each experimental (SNR) condition using 14-fold leave-one-out cross-validation. First, the stimuli and responses were normalized and then partitioned into train and test sets. Here the EEG data were lagged, τ, from -100–300ms since we were interested in short latency responses to the speech envelope, $s(t)$. We employed the same method of cross-validation and regularization parameter (10^{-1}–10^{8}) selection as before. A decoder was trained using the selected regularization parameter and the training data to reconstruct an estimate of the speech envelope. After testing the decoder on held-out data, we assessed model performance by computing the correlation between the actual speech envelope and the reconstructed speech envelope. This model performance, also known as reconstruction accuracy, was averaged across folds and participants (**Figure 1**).

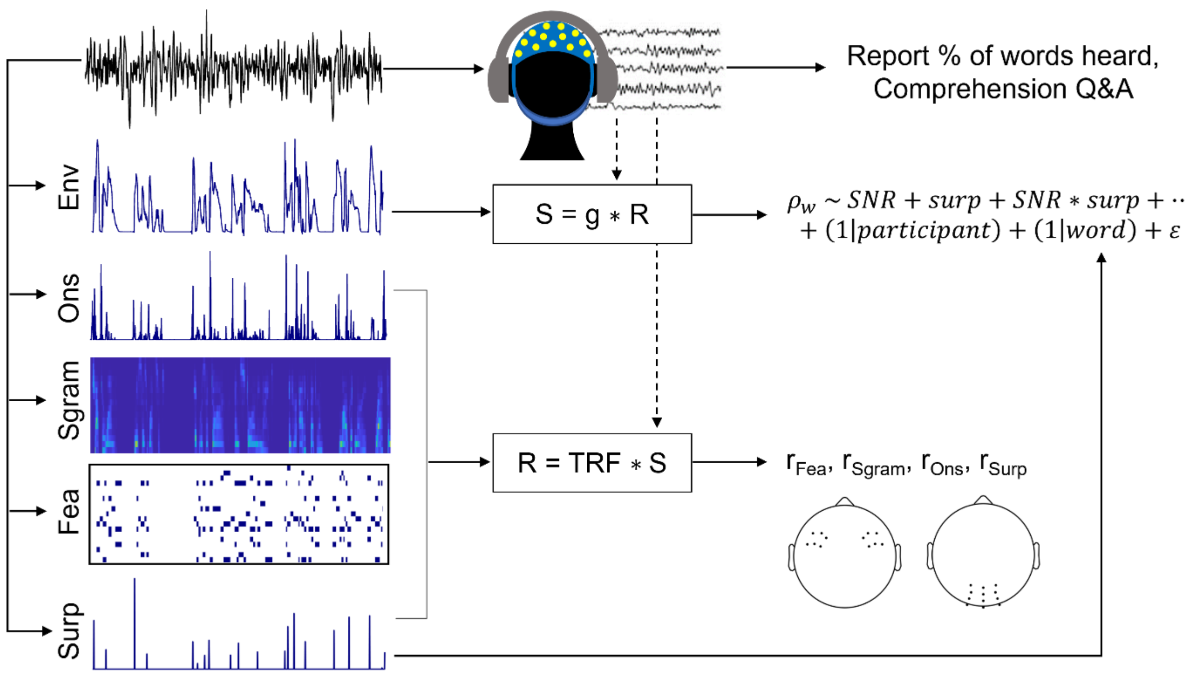

Figure 1. Methods. EEG data were recorded while participants listened to an audiobook in different levels of noise. Forward modeling was used to estimate EEG responses (R) from the clean representations of the acoustic onsets, spectrogram, phonetic features, and word surprisal (S being a concatenation of the features). Model performance (r) was assessed by calculating the correlation between the predicted EEG and the actual EEG and then averaged across the selected channels. Backward modeling was also used to reconstruct an estimate of the clean speech envelope. Model performance (reconstruction accuracy, ρ_w) was assessed by calculating the correlation between the original speech envelope and the reconstructed

speech envelope. A linear mixed-effects model (LME) was then used to determine the influence of word surprisal (surp) on envelope reconstruction accuracy.

Assessing the role of context on acoustic encoding in different levels of background noise

Another major goal of the present study was to test how listeners might rely more on context to encode speech that is masked by moderate levels of background noise. To do this, we used a variant of a recently introduced approach that involves quantifying how the tracking of low-level speech features varies as a function of the context-based semantic content of that speech (Broderick, Anderson, and Lalor 2019). Specifically, we used a linear mixed-effects (LME) model to explore the extent to which word predictability in the form of word surprisal influences how the envelope of that word was reflected in EEG and how that influence changes across SNRs. Using an LME in this way, one can measure the relationship between the main variable(s) of interest while controlling for variability caused by random factors. We used the lmerTest (version 3.1-3) and lme4 (version 1.1-30) packages in R to model the following equation:

$$\rho_w \sim 1 + SNR + surp + envStd + f_{rel} + res + f_{rel} * envStd + f_{rel} * res + surp * envStd + surp * res + surp * SNR + (1|participant) + (1|word)$$

The dependent variable is word reconstruction accuracy which was calculated as the Spearman's correlation between the actual word envelope and the predicted word envelope for the first 100 ms of each word. The independent variables are SNR, lexical surprisal (*surp*), envelope variability (*envStd*), relative frequency (f_{rel}), resolvability (*res*), and various interactions such as the interaction between surprisal and SNR.

Envelope variability, relative pitch, and resolvability were selected as nuisance regressors. This was because these measures can correlate with surprisal, with one another, and with envelope tracking; so, they are included here to ensure that they aren't inherently contaminating the lexical surprisal effects. Relative pitch is pitch normalized to the vocal range of the speaker (Tang, Hamilton, and Chang 2017) and resolvability measures whether a sound's harmonics are processed between distinct (resolved) or within the same (unresolved) filters of the cochlea (Shackleton and Carlyon 1994). Prosodic cues such as relative pitch and resolvability can uniquely predict EEG activity even after accounting for other acoustic and phonetic features (Teoh, Cappelloni, and Lalor 2019).

Relative pitch was extracted using Praat (Boersma and Weenink 2013). Once the software estimates absolute pitch, this result is then z-scored, resulting in relative pitch. Resolvability was extracted using custom scripts based on a model of the human auditory periphery (McDermott and Simoncelli 2011; Teoh, Cappelloni, and Lalor 2019). Envelope variability was shown to correlate with relative pitch and resolvability and all three features have been shown to influence envelope reconstruction accuracy, so we too included these features to control for acoustic related changes in the speaker's voice (Broderick, Anderson, and Lalor 2019). Envelope variability is represented as the standard deviation of the speech envelope.

The LME model also included by-word and by-participant random intercepts as some words may be easier to reconstruct than others and some participants may, on average, have higher reconstructions than other participants. No random slopes were included, as they caused the model to not converge even with the addition of an optimizer. Like Broderick et al. 2019, word reconstruction accuracy and the nuisance regressors were measured in the first 100ms following each word's onset (Broderick, Anderson, and Lalor 2019).

Statistical analyses

All statistical analyses were performed in R (version 4.2.0) and in MATLAB R2021b (MATLAB 2021). Due to the skewed distribution of the behavioral results, comparisons were calculated using a nonparametric Friedman's test, followed by a Wilcoxon Rank Sum test. All corrections for multiple comparisons were performed using false discovery rate (FDR), specifically Benjamini & Yekutieli (BY) correction, unless otherwise stated. FDR (BY) corrected pairwise t-tests were used to determine differences in envelope reconstruction accuracy and EEG prediction accuracy between conditions. Permutation testing was performed to test the significance of the EEG-speech model predictions. A null model was created for each SNR by shuffling the stimulus of interest between trials (except for the surprisal vectors which were shuffled within trial) and calculating a new model with each shuffle. This procedure, including regularization, was repeated 30 times. The null prediction accuracies were averaged across folds, permutations, and electrodes to result in one value per person. Pairwise t-tests were performed between the actual and null prediction accuracies across participants.

Unaggregated prediction accuracies (each accuracy for each trial per person) were used in an LME model (lmerTest version 3.1-3 and lme4 version 1.1-30 in R) to test how prediction accuracies changed across SNRs. This model included SNR, speech feature type, and the interaction between the two as fixed effects, and a by-participant random intercept. We then computed estimated marginal means (emmeans version 1.8.2 in R) with Tukey adjustment to perform multiple comparisons tests between each feature. Each LME model in this study was calculated using the default parameters which included fitting the models with restricted maximum likelihood (REML) and using Satterthwaite's method for the t-tests. LME models were also used to model the relationship between behavior and reconstruction/prediction accuracy. Permutation tests were used to test the significance of the final LME analysis where we tested the relationship between lexical surprisal and envelope tracking. Surprisal values were shuffled 5000 times while all other variables remained fixed, and an LME model was calculated for each shuffle. We calculated the proportion of coefficients that were greater than the observed values. All data and scripts are available upon request.

RESULTS

Speech perception decreases as SNR decreases

Behavioral scores were collected to show how participants' perception of the story changed as listening conditions became more challenging. After each one-minute-long trial, participants rated an estimate of the number of words they were able to hear on a scale from 0-100% and answered two multiple-choice comprehension questions, each of which had four possible answers. As expected, there was a significant reduction in percentage of words heard

($X^2(4) = 98.894$, $p = 2.20 \times 10^{-16}$) and comprehension scores ($X^2(4) = 83.44$, $p = 2.20 \times 10^{-16}$, Freidman test followed by Wilcoxon rank sum test with FDR correction (BY), **Figure 2**, **Table 1**) with the addition of noise. After the experiment, participants reported hearing very few words in the -9 dB SNR condition but that they had attempted to use context clues from previous trials to answer the questions in the -9 dB SNR trials. To control for this potential confound in our measure of comprehension, we used a procedure similar to Orf and colleagues, where nine additional volunteers answered the comprehension questions for each trial without listening to the story (Orf et al. 2022). Based on the performance of those new participants, we determined a new empirical chance level of 28% rather than 25%. All participants performed above chance in the quiet, +3 dB SNR, and -3 dB SNR conditions. Comprehension scores were not above chance for five participants in the -6 dB SNR condition (p = 0.066 and above) and 13 participants in the -9 dB SNR condition (p = 0.066 and above). Although participants reported hearing fewer words in the +3 dB SNR condition compared to speech in quiet, they performed similarly in terms of comprehension (p = 1.000).

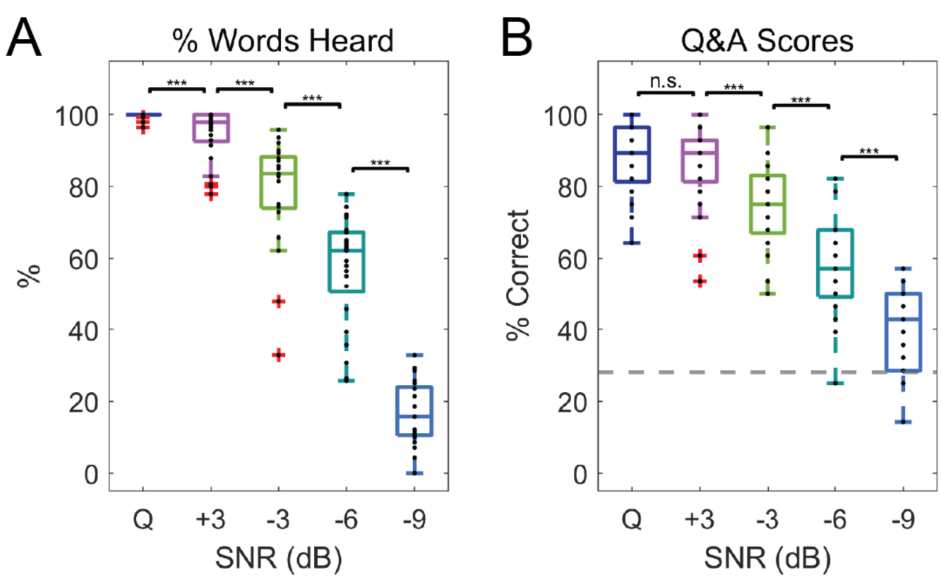

Figure 2. Behavioral results. **A**. Average percentage of words participants reported hearing in each condition. **B**. Average percentage of correctly answered comprehension questions for each condition. The

dotted gray line is the chance level at 28% which was calculated using a separate set of participants who answered the comprehension questions without listening to the audiobook. For both plots, significance is indicated by * if $p < 0.05$, ** if $p < 0.01$, and *** if $p < 0.001$ using pairwise Wilcoxon rank sum tests. The black points in each are individual participants.

Table 1. Wilcoxon rank sum test results comparing behavioral scores between all conditions.

% Words Heard					Q&A Scores				
	quiet	+3 dB	-3 dB	-6 dB		quiet	+3 dB	-3 dB	-6 dB
+3 dB	0.00093	-	-	-	+3 dB	1.00000	-	-	-
-3 dB	4.2e-05	4.2e-05	-	-	-3 dB	0.00020	0.00080	-	-
-6 dB	4.2e-05	4.2e-05	4.2e-05	-	-6 dB	8.5e-05	8.5e-05	0.00014	-
-9 dB	4.2e-05	4.2e-05	4.2e-05	4.2e-05	-9 dB	8.5e-05	8.5e-05	8.5e-05	0.00027

Hierarchical feature encoding declines across SNRs

Studies have typically modeled brain responses to isolated speech features. Recent work, however, is increasingly beginning to model acoustic and linguistic features simultaneously (Brodbeck, Hong, and Simon 2018; Brodbeck, Presacco, and Simon 2018; de Heer et al. 2017; Gillis et al. 2021; Heilbron et al. 2022) to find how the brain uniquely encodes a feature of interest when accounting for others, as some features may be correlated and explain similar neural activity (Daube, Ince, and Gross 2019). Since few studies have modeled the encoding of simultaneous features in challenging listening conditions (Brodbeck et al. 2020), we were interested in how a range of acoustic and linguistic features were encoded in noise.

To test this, we first fit one complete model comprised of acoustic onsets, the speech spectrogram, phonetic features, and word surprisal. Word onset (not pictured in the subsequent figures) was also included to ensure the surprisal measure was not reflecting variance that would be better explained by onset responses to individual words. The full model was separated into its constituent model pieces (where, for example, constituent models may have one feature in the case of surprisal, or many features in the case of phonemes) and tested on left out data. The acoustic onset, spectrogram, and phonetic feature prediction accuracies were averaged across

12 well-predicted frontotemporal electrodes and the surprisal prediction accuracies were averaged across 12 well-predicted parieto-occipital electrodes (electrodes shown in **Figure 1**).

Each model performed above chance for each SNR ($p < 0.01$, permutations followed by one-tailed, paired t-tests). As expected, the prediction accuracies for each constituent model decreased with increasing levels of noise (**Figure 3A**, paired t-tests with FDR [BY] correction). We then elected to run an LME model analysis exploring how the encoding of different speech features falls off with decreasing SNR. This helps account for the fact that EEG prediction accuracies can vary greatly between subjects (based on, for example, cortical folding or skull/scalp geometry). We fit the following LME model:

$$r \sim SNR + Feature + SNR * Feature + (1|participant)$$

Since we were interested in trends across noise levels, SNR was treated as a continuous variable rather than a categorical variable. EEG predictions based on the speech spectrogram had the shallowest slope across noise conditions ($\beta_{sgram} = -0.009$ which is the difference between SNR and SNR:Sgram, **Table 2**)—suggesting that the spectrogram was relatively well reflected in the EEG even as noise levels increased. This trend in prediction accuracy, however, was similar to both the acoustic onset ($p = 0.708$) and surprisal trends ($p = 0.395$). The spectrogram trend was significantly greater than the phonetic feature trend ($p = 4 \times 10^{-4}$), suggesting that phonetic feature representations are more sensitive to noise. Altogether, high- and low-level features, with the exception of phonetic features in some cases, declined similarly across SNRs.

Another way to examine the effect of SNR on speech feature encoding is to visualize the TRF models themselves. In fact, recent work has shown that the amplitude and latency of acoustic onset and semantic TRFs are influenced by speech SNR (Yasmin et al. 2023). Given that the encoding of specific frequency bands or phonetic properties may decrease differentially with noise, we examined the change in the spectrogram and phonetic feature TRF weights at each SNR. In quiet, we saw the strongest TRF weights in the lower frequency bands, suggesting the importance of low frequency spectral tracking when no noise is present. Overall, we found a decrease and narrowing in time of the TRF weights across a broad range of frequencies as SNR decreased. Notably, model weights below 1,800 Hz seem to completely diminish in the -9 dB SNR condition (**Figure 3B top**). Altogether, these results support the importance of low frequency speech signals in speech encoding which is reasonable given that others have shown that this range is important for speech intelligibility in noise (Chang, Bai, and Zeng 2006; Turner et al. 2004) and given that the type of noise used in the present study matches the speech spectrum.

Table 2. LME model results showing how prediction accuracies changed across SNRs and post-hoc contrasts between each feature

Marginal R^2 = 0.204, Conditional R^2 = 0.457					
Fixed effects	Estimate	Std. Error	t value	p-value	
Onsets (intercept)	8.488×10^{-2}	4.513×10^{-3}	18.809	$< 2 \times 10^{-16}$	***
SNR	-1.013×10^{-2}	5.171×10^{-4}	-19.581	$< 2 \times 10^{-16}$	***
Sgram	2.875×10^{-3}	2.426×10^{-3}	1.185	0.236	
Fea	-1.112×10^{-2}	2.426×10^{-3}	-4.585	4.62×10^{-6}	***
Surp	-2.035×10^{-2}	2.426×10^{-3}	-8.391	$< 2 \times 10^{-16}$	***
SNR:Sgram	7.829×10^{-4}	7.313×10^{-4}	1.070	0.284	
SNR:Fea	-2.113×10^{-3}	7.313×10^{-4}	-2.889	0.004	***
SNR:Surp	-3.668×10^{-4}	7.313×10^{-4}	-0.501	0.616	
Contrasts	Estimate	Std. Error	z ratio	p-value	
Ons – Sgram	-7.830×10^{-4}	7.310×10^{-4}	-1.070	0.708	
Ons – Fea	2.113×10^{-3}	7.310×10^{-4}	2.889	0.020	*
Ons – Surp	3.670×10^{-4}	7.310×10^{-4}	0.501	0.958	
Sgram – Fea	2.896×10^{-3}	7.310×10^{-4}	3.960	4×10^{-4}	***
Sgram – Surp	1.150×10^{-3}	7.310×10^{-4}	1.572	0.395	
Fea – Surp	-1.746×10^{-3}	7.310×10^{-4}	-2.388	0.079	

In the phonetic features case, although there is a decrease across practically all features around 100 ms in the +3 dB SNR condition, voicing and some manner of articulation (plosive and fricative) and vowel backness features (front, back, and diphthong) remained largely intact in this condition (**Figure 3B bottom**). As noise levels continued to increase, we saw even more of a decrease in each feature. In the -6 dB SNR condition, only some plosive, fricative, and vowel backness feature weights remained whereas all other features here and in the -9 dB SNR condition diminished. Although TRF weights for individual features decreased in the +3 dB SNR, we began to see larger reductions in TRF weight amplitudes in the -6 dB SNR similar to

Swaminathan and Heinz who found their greatest lapse in phonetic feature reception around -5 dB SNR (Swaminathan and Heinz 2012).

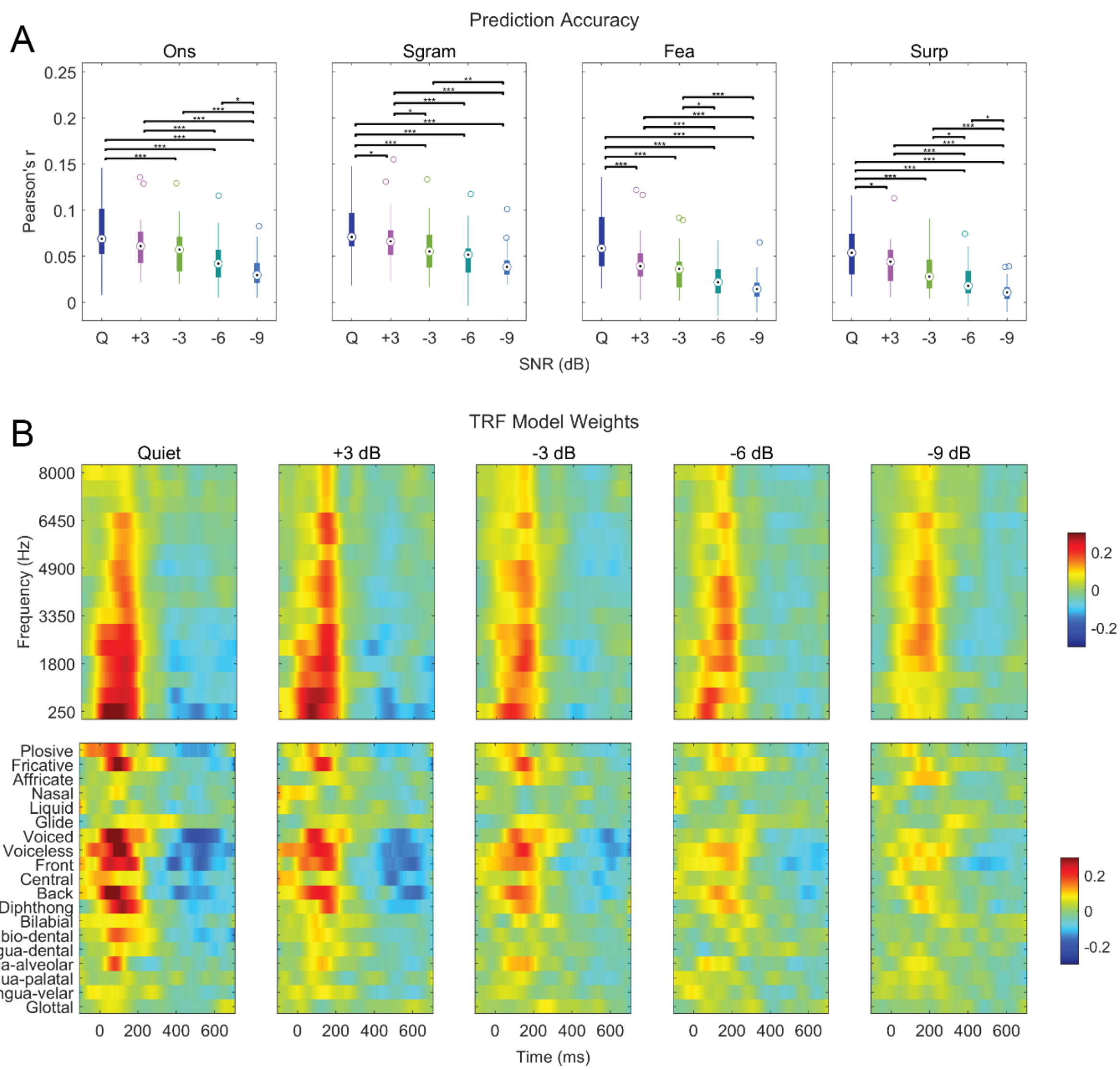

Figure 3. Forward modeling results. **A**. Prediction accuracies which were calculated using acoustic onsets, spectrogram, phonetic feature, and surprisal from left to right. Significance is indicated by * if $p < 0.05$, ** if $p < 0.01$, and *** if $p < 0.001$ using pairwise t-tests with FDR (BY) correction. Spectrogram (**B top**) and phonetic feature (**B bottom**) TRF model weights across the experimental conditions. The model weights were averaged across 12 frontotemporal electrodes and originated from a model that included acoustic onsets, spectrogram, phonetic features, and surprisal.

Linguistic features are highly predictive of behavior

One of the main hypotheses we had in this study was that, across SNRs, EEG signatures of linguistic processing would be more closely correlated with behavior than EEG measures of low-level acoustic processing. To test this, we modeled the relationship between each speech feature's model prediction accuracy with the percentage of words heard and comprehension scores using LME models. We used this method so that we could pool data across SNRs which would result in 5 data points per participant. As such, we included fixed effects that corresponded to the model accuracies for each feature and a random effect for participant. We modeled both behavioral metrics separately.

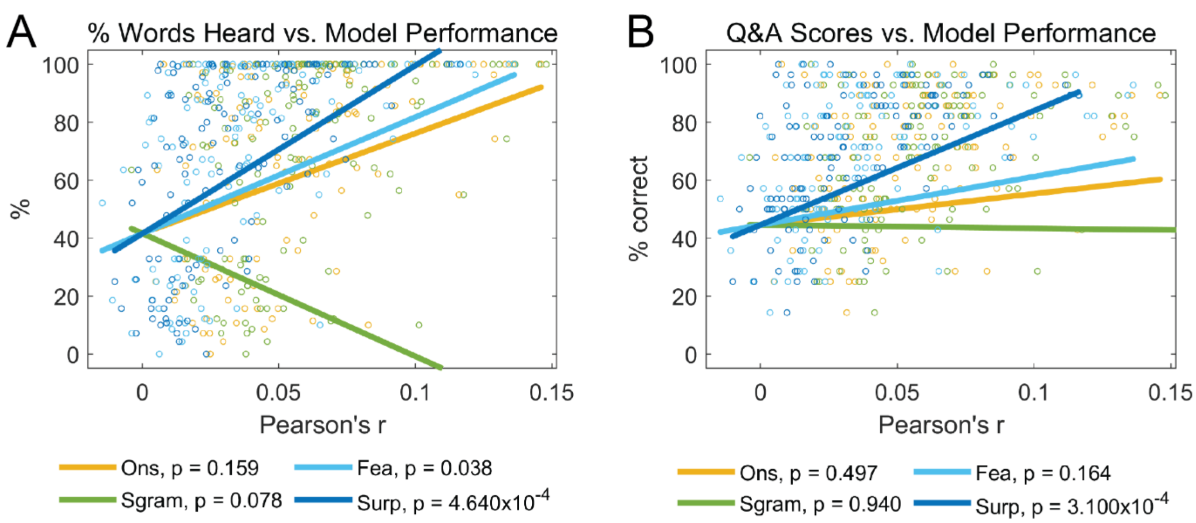

Figure 4. The relationship between forward model performance (prediction accuracy) and behavior. LME models were used to relate the model performance for each speech feature to the percentage of words heard (**A**) and comprehension scores (**B**). The marginal and conditional R^2s are 0.444 and 0.583 for A and 0.429 and 0.701 for B.

In line with our hypothesis, these analyses show that lexical surprisal is highly predictive of both the percentage of words heard (β = 580.101, p = 4.640 x 10^{-4}, **Figure 4A**) and comprehension scores (β = 394.314, p = 3.100 x 10^{-4}, **Figure 4B**). Phonetic feature performance was significantly predictive of percentage of words heard (β = 401.892, p = 0.038), but not comprehension (β = 166.587, p = 0.164). The model did not produce any significant effects for the acoustic onset (β = 345.738, $p = 0.159$ for % words heard; β = 107.556, $p = 0.497$ for comprehension) or spectrogram (β = -424.404, $p = 0.078$ for % words heard; β = -11.802, $p = 0.940$ for comprehension) fixed effects. The spectrogram fixed effect may have negative trends because the spectrogram and acoustic onset model performances are highly (yet negatively) correlated; the percentage of words heard LME model reported a correlation of -0.623 and the comprehension LME model reported a correlation of -0.602. The spectrogram fixed effect may be fitting to the noise that remains after the acoustic onsets have explained all it can behavior-wise. Nevertheless, these results support that linguistic features are more predictive of subjective and objective measures of speech perception than acoustic features.

Envelope reconstruction accuracy maps well to behavior

Numerous studies have shown that cortical activity tracks the temporal modulations of the speech envelope (Aiken and Picton 2008; Destoky et al. 2019; Di Liberto, O'Sullivan, and Lalor 2015; Ding and Simon 2013; Etard and Reichenbach 2019; Lalor and Foxe 2010b; Nourski et al. 2009; Pasley et al. 2012). Speech envelope reconstruction is a powerful method of indexing speech tracking given its overall advantage of using all scalp data to provide a better SNR. It is unknown, however, exactly what information envelope reconstruction indexes because the envelope has been shown to capture syllabic boundaries (Hertrich et al. 2012; Oganian and Chang 2019), phonetic feature information (Rosen 1992), and prosodic cues (Myers, Lense, and Gordon 2019). Nevertheless, due to its common usage in speech research and its improved SNR, we were interested in how speech envelope tracking would compare to our acoustic onset, spectrogram, phonetic feature, and surprisal encoding results and how it would relate to behavior.

The first step in this analysis was to use a backward modeling procedure to find how speech envelope tracking is affected by different levels of background noise. Speech in quiet and the +3 dB SNR condition shared similar reconstruction accuracies ($p = 0.497$, paired t-test with FDR [BY] correction) and were reconstructed more reliably than all other SNRs ($p < 0.05$). The -3 dB SNR and -6 dB SNR accuracies were not significantly different from each other ($p = 0.547$)

but were both higher than the -9 dB SNR condition ($p = 2.2 \times 10^{-7}$ and $p = 2.5 \times 10^{-6}$, **Figure 5A**). We compared the changes in reconstruction accuracy across SNRs to the other features by including it in the original LME model. Interestingly, the speech envelope feature had the steepest slope in model performance across SNRs compared to all other features (marginal/conditional $R^2 = 0.521/0.644$, $\beta = -0.017$, $p < 0.0001$). In other words, the speech envelope caused the greatest change in prediction accuracy across SNRs when all other features were fixed.

Next, we tested if the fidelity of envelope tracking was indicative of how well participants could hear and understand the story in comparison to the forward modeled features. This was tested using an LME model that contained fixed effects for envelope reconstruction accuracy and prediction accuracies for models trained on acoustic onsets, spectrograms, phonetic features, and lexical surprisal (and word onset). As expected, envelope reconstruction scores were highly predictive of percent words heard (marginal/conditional $R^2 = 0.512/0.695$, $\beta = 264.240$, $p = 6.730 \times 10^{-4}$) and comprehension (marginal/conditional $R^2 = 0.476/0.705$, $\beta = 149.070$, $p = 3.570 \times 10^{-3}$, **Figure 5B**). Surprisal was still highly predictive of both measures ($\beta = 431.332$, $p = 8.342 \times 10^{-3}$ for % words heard; $\beta = 326.367$, $p = 2.450 \times 10^{-3}$ for comprehension). Even though the phonetic feature predictors had greater or similar slopes to the envelope predictor, the LME models did not report these results as significant ($\beta = 130.305$, $p = 0.089$ for % words heard; $\beta = 147.956$, $p = 0.206$ for comprehension). To our surprise, the spectrogram model performance now significantly (negatively) predicted percentage of words heard in this model as well ($\beta = -587.122$, $p = 0.014$). Furthermore, we calculated EEG prediction accuracy based on a full model trained on the concatenation of acoustic onsets, spectrogram, phonetic features, surprisal, and word onset. We then related the full model accuracy and envelope reconstruction accuracies to behavior. This full model was more predictive of behavior (% words heard, marginal/conditional $R^2 = 0.497/0.731$, $\beta = 486.279$, $p = 5.900 \times 10^{-5}$; comprehension, marginal/conditional $R^2 = 0.470/0.729$, $\beta = 311.894$, $p = 3.180 \times 10^{-5}$) than envelope reconstruction accuracy (% words heard, $\beta = 296.395$, $p = 2.500 \times 10^{-4}$; comprehension, $\beta = 167.173$, $p = 1.190 \times 10^{-3}$).

The main caveat of the results above is that we're comparing two different modeling methods. Backward models have the ability to take advantage of all EEG channels and up-weigh more informative channels, resulting in higher model performance compared to forward models. As such, we reperformed the analysis with a forward model that included acoustic onsets, spectrograms, phonetic features, word surprisal, and the speech envelope (and word onsets). In this case, we found that the change in the envelope encoding across SNRs is similar to all other features modeled except the phonetic features (marginal/conditional R^2 = 0.199/0.491, p = 0.0485, LME model).

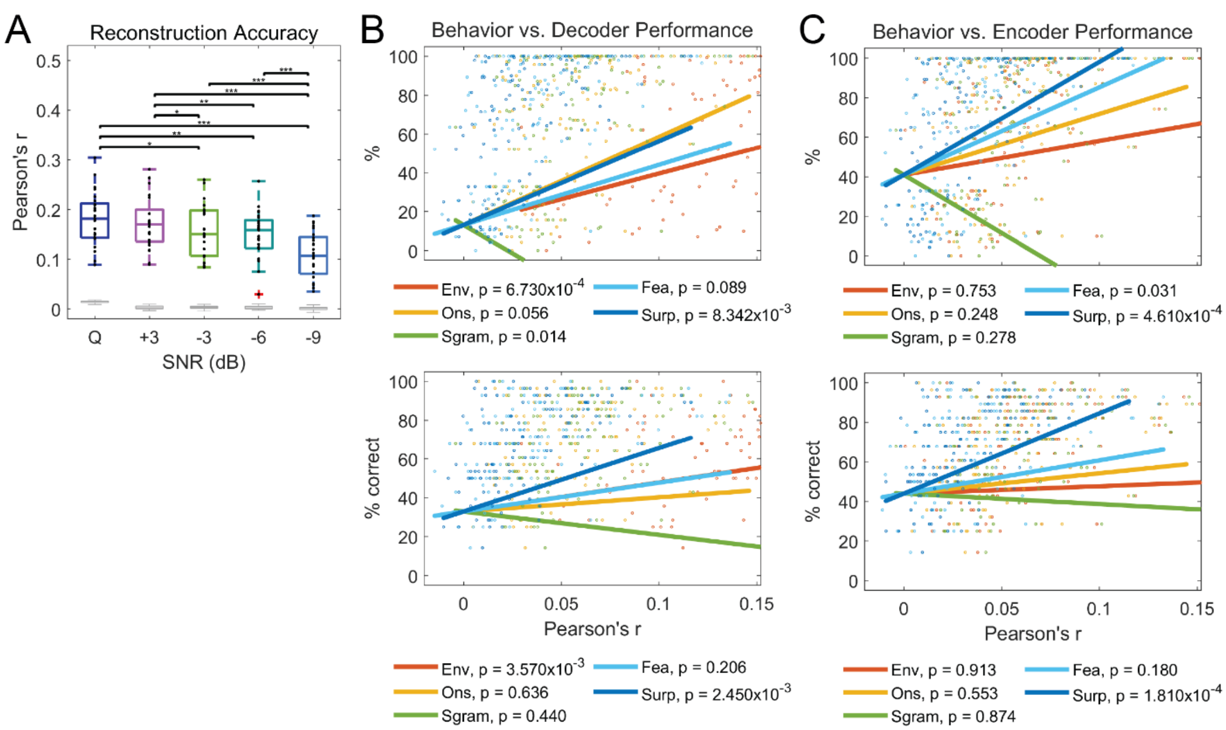

Figure 5. Envelope modeling results. **A.** The colored boxplots are the envelope reconstruction accuracies for each condition, and the black points represent each participant. The gray boxplots represent mean reconstruction accuracies (per person) based on shuffled permutations. Significance is indicated by * if p < 0.05, ** if p < 0.01, and *** if p < 0.001 using pairwise t-tests with FDR (BY) correction. **B**. We then used LME models to determine the relationship between envelope decoder performance and behavior (percentage of words heard on top and comprehension scores on the bottom). Each circle represents a

participant, so there are 125 circles (5 conditions x 25 participants). **C**. Same as B, except using results from a forward model trained on all features including the envelope.

This reanalysis also showed that higher-level linguistic features are most correlated with behavior (surprisal vs. % words heard, β = 570.767, p = 4.610 x 10^{-4}; surprisal vs. comprehension, β = 405.383, p = 1.810 x 10^{-4}; phonetic features vs. % words heard, β = 442.387, p = 0.031). The envelope (% words heard, β = 171.309, p = 0.753; comprehension, β = 37.104, p = 0.913), acoustic onset (% words heard, β = 307.682, p = 0.248; comprehension, β = 102.556, p = 0.553), and spectrogram (% words heard, β = -587.997, p = 0.278; comprehension, β = -53.038, p = 0.874) model effects were not significant in addition to phonetic features in the comprehension model (β = 167.955, p = 0.180, **Figure 5C**). The marginal and conditional R^2 are 0.445 and 0.568 for the % words heard model and 0.434 and 0.697 for the comprehension model. In short, when modeling the envelope in the forward direction, its encoding decreases at a similar rate to other low-level acoustic features and it predicts behavior similar to them as well. Conversely, when using stimulus reconstruction, envelope models perform best across SNRs and predict comprehension scores similar to higher level features.

Lexical surprisal's influence on early auditory encoding decreases at high noise levels

Recent work by Broderick and colleagues has shown that the cortical tracking of an individual word's acoustic and phonetic representations was enhanced the more semantically similar it was to its preceding context (Broderick, Anderson, and Lalor 2019). Since higher-level representations can bias perception when incoming stimuli are noisy (de Lange, Heilbron, and Kok 2018), perhaps in the present study, participants relied more on a higher-level feature (i.e., next word probability/surprisal) when noise levels slightly increased—thereby strengthening the relationship between word surprisal and lower-level feature encoding. This influence of lexical surprisal would then decrease the noisier the speech became, i.e., as people begin to fail to understand the speech. We aimed to test these hypotheses using a two-stage regression analysis that consists of stimulus reconstruction followed by LME modeling.

Stimulus reconstruction was used (due to its increased SNR) to reconstruct an estimate of the speech envelope for each individual word. We then compared the predicted and actual envelopes using Spearman's correlation. All words in the story were also scored on their lexical surprisal, which was calculated as the negative logarithm of a word's probability given the words that came before it. Envelope variability, relative pitch, and resolvability (the nuisance regressors) were also calculated to control for acoustic related properties of the stimuli, many of which could correlate with word surprisal and word reconstruction accuracy. Surprisal, SNR, and the nuisance regressors were included in an LME model to predict word reconstruction accuracy for the first 100 ms of each word.

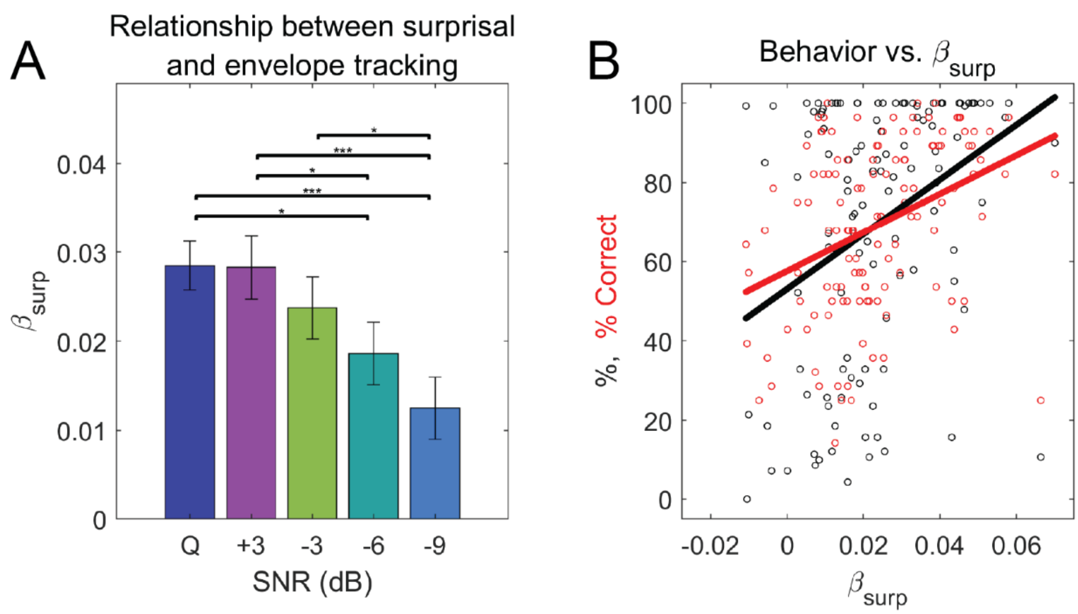

Figure 6. The relationship between lexical surprisal and envelope tracking in different levels of noise. **A**. Surprisal coefficient in quiet and the interaction between surprisal and SNR using a 100 ms word window. The error bars are the standard errors of each measure, calculated using the LME model. Significance is indicated by * if $p < 0.05$, ** if $p < 0.01$, and *** if $p < 0.001$ using R's "emtrends" function to compare estimated marginal means of linear trends. **B**. The relationship between the surprisal coefficients across SNRs and behavior. Percentage of words heard are in black, and the comprehension scores are in red. There are 125 circles (5 SNRs x 25 participants).

There was a significantly positive relationship between lexical surprisal and word reconstruction accuracy in quiet (marginal/conditional R^2 = 0.017/0.030, $\beta = 2.850 \times 10^{-2}$, t = 10.519, $p < 2.00 \times 10^{-16}$, **Table 3**). In other words, the more surprising a word (or the less probable that word was given its preceding context), the greater that word's envelope was reflected in the EEG. The interaction between surprisal and SNR shows how this coefficient changes in the other noise levels (**Figure 6A**). The influence of lexical surprisal on word reconstruction accuracy in quiet was similar to the +3 dB ($\beta = -2.172 \times 10^{-4}$, $t = -0.062$, $p = 0.951$) and -3 dB ($\beta = -4.767 \times 10^{-3}$, $t = -1.363$, $p = 0.173$) SNR conditions and greater than the -6 dB ($\beta = -9.900 \times 10^{-3}$, $t = -2.817$, $p = 0.005$) and -9 dB ($\beta = -1.600 \times 10^{-2}$, $t = -4.570$, $p = 4.890 \times 10^{-6}$) SNR conditions. This shows that lexical surprisal impacts envelope tracking in low levels of background noise (+3 dB and -3 dB SNRs) similar to when speech is in quiet, but this influence significantly decreases with moderate to high levels of noise. It is also important to note that the nuisance regressors and additional interactions we've included also significantly influenced word reconstruction accuracy, reinforcing the value of controlling for those variables and ensuring our results were not confounded by other acoustic related measures.

Table 3. LME model showing the relationship between lexical surprisal and envelope tracking in each SNR.

	Estimate	Std. Error	t value	Pr(>|t|)	
Quiet (Intercept)	8.992×10^{-2}	5.451×10^{-3}	16.496	$< 2 \times 10^{-16}$	***
SNR +3	-4.703×10^{-3}	3.514×10^{-3}	-1.338	0.18077	
SNR -3	-1.671×10^{-2}	3.524×10^{-3}	-4.724	2.12×10^{-6}	***
SNR -6	-3.196×10^{-2}	3.496×10^{-3}	-9.141	$< 2 \times 10^{-16}$	***
SNR -9	-5.822×10^{-2}	3.511×10^{-3}	-16.586	$< 2 \times 10^{-16}$	***
surp	2.850×10^{-2}	2.709×10^{-3}	10.519	$< 2 \times 10^{-16}$	***
surp:SNR +3	-2.172×10^{-4}	3.528×10^{-3}	-0.062	0.95091	
surp:SNR -3	-4.767×10^{-3}	3.497×10^{-3}	-1.363	0.17287	
surp:SNR -6	-9.900×10^{-3}	3.514×10^{-3}	-2.817	0.00484	**
surp:SNR -9	-1.600×10^{-2}	3.501×10^{-3}	-4.570	4.89×10^{-6}	***
envStd	6.042×10^{-2}	1.430×10^{-3}	42.251	$< 2 \times 10^{-16}$	***
f_{rel}	8.697×10^{-3}	1.576×10^{-3}	5.519	3.42×10^{-8}	***
res	-1.605×10^{-2}	1.453×10^{-3}	-11.051	$< 2 \times 10^{-16}$	***
envStd:f_{rel}	6.692×10^{-3}	1.148×10^{-3}	5.832	5.49×10^{-9}	***
f_{rel}:res	-6.327×10^{-3}	1.450×10^{-3}	-4.363	1.28×10^{-5}	***
surp:envStd	1.624×10^{-2}	1.305×10^{-3}	12.448	$< 2 \times 10^{-16}$	***
surp:res	-6.872×10^{-3}	1.325×10^{-3}	-5.186	2.17×10^{-7}	***

Permutation tests were conducted to test if each surprisal estimate was significantly different from chance. The surprisal values were shuffled 5000 times while keeping all other variables intact, and a new LME model was computed with each shuffle. This procedure resulted in none of the null coefficients being greater than the observed values. Despite studies showing that low SNRs benefit from the utilization of context (Mayo, Florentine, and Buus 1997), we found the permutation results for the -6 dB and -9 dB conditions to be surprising. Participants reported hearing either no words or very few words in, for instance, the -9 dB SNR condition, yet we still see a significant influence of lexical surprisal on envelope tracking. We did not control for any other features (including surprisal) in our envelope reconstruction model, so these unaccounted-for features may have contributed to the significantly positive interaction coefficients in the -6 dB and -9 dB SNR conditions. Or the fact that participants were able to hear any words in both conditions may have been enough to cause this significant effect. Lastly, we found that single subject surprisal coefficients also predicted their self-reported percentage of words heard (marginal/conditional R^2 = 0.124/NA, β = 691.434, p = 5.290 x 10^{-5}) and comprehension scores (marginal/conditional R^2 = 0.138, 0.177, β = 488.401, p = 2.140 x 10^{-5}) across SNRs (**Figure 6B**). That is to say, the stronger the influence of surprisal on envelope tracking, the better the participants were able to hear and comprehend the story.

DISCUSSION

This study sought to establish how well indices of hierarchical neural speech processing reflect language comprehension—advancing on prior work that has typically tested specific hierarchical levels without controlling for the others. We first characterized how the encoding of a range of hierarchical speech features diminished in noise and if those changes in encoding were predictive of behavior. We found that the encoding of acoustic and surprisal features declined similarly as noise levels increased, and that phonetic feature encoding was more affected by noise than the acoustic features. In addition, lexical surprisal and phonetic feature encoding were the most predictive of participants' behavioral scores across SNRs. Speech envelope models were predictive of behavior, but only when employing decoding models. Lastly, we investigated how lexical surprisal influenced the neural tracking of the speech envelope. In general, we found that the envelopes of more unexpected words were better reflected in the EEG. This was true in quiet and in low levels of background noise, but this relationship weakened as noise levels increased.

We hypothesized that acoustic features would be the most invariant to noise, but this was only partially supported by our results. The degree to which envelope and spectrogram features were reflected in EEG decreased at a slower rate only in comparison to phonetic features. However, when we analyzed envelope reconstruction accuracies, rather than EEG predictions based on the speech envelope, decoding accuracies decreased at a faster rate than all other features. This was surprising, first, due to a previous finding that the synchronization between neural activity and the speech envelope remained unaffected until the speech signal had an SNR of -9 dB (Ding and Simon 2013). These stark differences may have been due to a combination of factors: neural recording modality, data preprocessing, model training and testing procedures between conditions, or the regularization method used (e.g., boosting versus ridge regression). Instead, our results show a gradual decrease in envelope tracking across SNRs similar to Vanthornhout and colleagues (Vanthornhout et al. 2018).

Although the rate at which our acoustic and linguistic model accuracies declined did not completely support our hypothesis, these results may not be surprising given recent work. Kell and McDermott measured primary and non-primary auditory cortices' invariance to background noise using fMRI. Invariance was measured by correlating voxel responses to natural sounds in quiet with the voxel response to those same sounds in noise. They found that primary and non-primary auditory cortices were similarly invariant to natural sounds in spectrally matched background noise tested at a 0 dB SNR. However, non-primary areas became more robust to noise than primary areas when sounds were presented in real-world noise (Kell and McDermott 2019). So, our model performances may result from how the brain represents speech in the type of synthetic noise we used. Models in the present study could have also been affected by attention. Participants may have allocated less attention to the -9 dB SNR trials due to the large amount of noise, in turn skewing the encoding/decoding of the different speech features.

Our findings also showed that phonetic features and lexical surprisal were most predictive of subjective behavioral metrics (percentage of words heard) and lexical surprisal was most predictive of objective metrics (comprehension). Previously published work has shown that neural measures of lexical surprisal is highly predictive of behavior (Mesik, Ray, and Wojtczak 2021). Many studies have also shown that the speech envelope (using stimulus reconstruction or cross-correlation) contributes and relates to speech intelligibility and comprehension (Ahissar et al. 2001; Decruy et al. 2020; Iotzov and Parra 2019; Lesenfants et al. 2019; Muncke, Kuruvila, and Hoppe 2022; Vanthornhout et al. 2018). Once we included envelope reconstructions in our analysis, it also proved to be an accurate predictor of behavior. However, backward modeling greatly improves overall model performance due to its ability to utilize all recorded neural channels, thereby increasing neural signal-to-noise ratio. Spectrogram and phonetic features have previously been shown to better predict EEG than the speech envelope (Di Liberto, O'Sullivan, and Lalor 2015), so we believe that our behavior-prediction accuracy correlations were due to how the envelope was modeled, rather than the information the speech envelope itself carries or how well it is reflected in the brain.

Interestingly, we found that phonetic features uniquely predicted neural activity even when controlling for the speech spectrogram and acoustic onsets. This is in line with previous studies showing that the addition of phonetic features to spectrotemporal representations improve EEG prediction (Di Liberto, O'Sullivan, and Lalor 2015; Sohoglu and Davis 2020) and its correlation with speech intelligibility (Lesenfants et al. 2019) and that phonetic features uniquely predict EEG responses even when attending to a specific talker (Teoh, Ahmed, and Lalor 2022). However, the present phonetic feature results contrast with previous work which suggested that responses to articulations could be explained by simpler acoustic features (Daube, Ince, and Gross 2019). Nevertheless, our findings that phonetic feature encoding declines at a different rate and better predicts behavior compared to the spectrogram, provides further evidence that the two features are dissociable.

Another one of our key hypotheses was that participants would use lexical context to predict and encode the acoustic features of each word. This was found to be true: our LME analysis (stage two of the two-stage regression) revealed that the more unexpected a word, the better we were able to reconstruct that word's envelope. However, we had also hypothesized that participants would rely more on these predictions for speech in moderate levels of noise (when speech is still intelligible) relative to speech in quiet, before falling off at high levels of background noise (when speech is no longer intelligible). This result was only partially borne out. Specifically, the use of lexical context in processing the speech acoustics did decrease as the speech became noisier, but there was no evidence to support a stronger reliance on context in moderate levels of noise. In particular, while there was no difference in comprehension scores between the quiet and +3 dB SNR conditions, there was no increase in the influence of surprisal on envelope tracking for the latter condition compared to speech in quiet. We did notice a larger spread of the percentage of words heard scores across subjects in the +3 dB SNR condition. So, we explored the possibility that the subjects who were starting to struggle might put forth more effort to understand and process the story by relying more on context (and thus might have a higher surprisal weight in **Figure 6**) than those who remained at ceiling. But we found no significant difference. This was a little surprising given that context has be known to affect behavior (Golestani et al. 2013) and neural activity (Boulenger et al. 2011; Koskinen et al. 2020; Strauß et al. 2022) in challenging listening conditions. Future work with larger subject numbers and perhaps even lower levels of background noise (e.g., + 6 dB SNR) might reveal such an effect.

Our LME model analysis based on word surprisal seems on face value to be at odds with Broderick and colleagues who found that the envelopes of words that were more semantically *similar* to their context were better reflected in the EEG. That is to say, envelope tracking is enhanced for words that share a similar meaning with their context (Broderick, Anderson, and

Lalor 2019). However, semantic similarity and lexical surprisal tend to share a moderate, and sometimes weak negative correlation (Frank and Willems 2017). Indeed, a re-analysis of Broderick et al.'s original EEG data has revealed that both semantic similarity and lexical surprisal play complementary (positive) roles in estimating when envelope tracking is enhanced (Broderick and Lalor 2020). Nevertheless, the nature of this duality remains mysterious, and we hope it will provide the grounds for an exciting body of future work. We anticipate it will take a substantial battery of future experiments to shape a unifying explanation, with stimuli that can disentangle correlations between semantic similarity, lexical surprisal, and other linguistic factors that could come into play (e.g., semantic content, next-word entropy, phonetic surprisal, next-phoneme entropy). In any case, what seems clear in the present results is that lexical context influences the neural tracking of speech acoustics on a word-by-word basis, and this influence drops as speech becomes unintelligible.

In summary, the current results show that phonetic features are more susceptible to noise than acoustic speech features. While linguistic features are more predictive of behavior than acoustic features, envelope decoding models can be used to improve this relationship. We have also found that the encoding of certain phonetic features decreases in even low levels of noise, and that the encoding of frequencies below 1.3k essentially disappears in high noise levels. Lastly, we show support that context influences a word's acoustic encoding. This influence lessens in high background noise levels. Future work will aim to further characterize how people might rely more or less on top-down context to process bottom-up speech input as a function of stimulus type, task, and listening conditions.

REFERENCES

Ahissar, Ehud, Srikantan Nagarajan, Merav Ahissar, Athanassios Protopapas, Henry Mahncke, and Michael M Merzenich. 2001. 'Speech comprehension is correlated with temporal response patterns recorded from auditory cortex', *Proceedings of the National Academy of Sciences*, 98: 13367-72.

Aiken, S. J., and T. W. Picton. 2008. 'Human cortical responses to the speech envelope', *Ear Hear*, 29: 139-57.

Anderson, A. J., J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, M. Aguilar, X. Wang, D. Doko, and R. D. S. Raizada. 2017. 'Predicting Neural Activity Patterns Associated with Sentences Using a Neurobiologically Motivated Model of Semantic Representation', *Cereb Cortex*, 27: 4379-95.

Bigdely-Shamlo, N., T. Mullen, C. Kothe, K. M. Su, and K. A. Robbins. 2015. 'The PREP pipeline: standardized preprocessing for large-scale EEG analysis', *Front Neuroinform*, 9: 16.

Boersma, Paul, and David Weenink. 2013. 'Praat: doing phonetics by computer [Computer program]. Version 5.3. 51', *Online: http://www. praat. org/retrieved, last viewed on*, 12.

Boulenger, Véronique, Michel Hoen, Caroline Jacquier, and Fanny Meunier. 2011. 'Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study', *Brain and language*, 116: 51-63.

Brodbeck, C., L. E. Hong, and J. Z. Simon. 2018. 'Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech', *Curr Biol*, 28: 3976-83 e5.

Brodbeck, C., A. Jiao, L. E. Hong, and J. Z. Simon. 2020. 'Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers', *PLoS Biol*, 18: e3000883.

Brodbeck, C., A. Presacco, and J. Z. Simon. 2018. 'Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension', *Neuroimage*, 172: 162-74.

Broderick, M. P., A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor. 2018. 'Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech', *Curr Biol*, 28: 803-09 e3.

Broderick, M. P., A. J. Anderson, and E. C. Lalor. 2019. 'Semantic Context Enhances the Early Auditory Encoding of Natural Speech', *J Neurosci*, 39: 7564-75.

Broderick, Michael P, and Edmund C Lalor. 2020. 'Co-existence of prediction and error signals in electrophysiological responses to natural speech', *bioRxiv*: 2020.11. 20.391227.

Broderick, Michael P, Nathaniel J Zuk, Andrew J Anderson, and Edmund C Lalor. 2022. 'More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative', *European Journal of Neuroscience*, 56: 5201-14.

Burton, M. W. 2009. 'Understanding the role of the prefrontal cortex in phonological processing', *Clin Linguist Phon*, 23: 180-95.

Chang, J. E., J. Y. Bai, and F. G. Zeng. 2006. 'Unintelligible low-frequency sound enhances simulated cochlear-implant speech recognition in noise', *IEEE Trans Biomed Eng*, 53: 2598-601.

Crosse, M. J., G. M. Di Liberto, A. Bednar, and E. C. Lalor. 2016. 'The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli', *Front Hum Neurosci*, 10: 604.

Crosse, M. J., G. M. Di Liberto, and E. C. Lalor. 2016. 'Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration', *J Neurosci*, 36: 9888-95.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. 'Transformer-xl: Attentive language models beyond a fixed-length context', *arXiv preprint arXiv:1901.02860*.

Daube, Christoph, Robin AA Ince, and Joachim Gross. 2019. 'Simple acoustic features can explain phoneme-based predictions of cortical responses to speech', *Current Biology*, 29: 1924-37. e9.

de Heer, W. A., A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen. 2017. 'The Hierarchical Cortical Organization of Human Speech Processing', *J Neurosci*, 37: 6539-57.

de Lange, F. P., M. Heilbron, and P. Kok. 2018. 'How Do Expectations Shape Perception?', *Trends Cogn Sci*, 22: 764-79.

Decruy, L., D. Lesenfants, J. Vanthornhout, and T. Francart. 2020. 'Top-down modulation of neural envelope tracking: The interplay with behavioral, self-report and neural measures of listening effort', *Eur J Neurosci*, 52: 3375-93.

Delorme, A., and S. Makeig. 2004. 'EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis', *J Neurosci Methods*, 134: 9-21.

Destoky, F., M. Philippe, J. Bertels, M. Verhasselt, N. Coquelet, M. Vander Ghinst, V. Wens, X. De Tiege, and M. Bourguignon. 2019. 'Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope', *Neuroimage*, 184: 201-13.

Di Liberto, G. M., J. A. O'Sullivan, and E. C. Lalor. 2015. 'Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing', *Curr Biol*, 25: 2457-65.

Di Liberto, Giovanni M, James A O'Sullivan, and Edmund C Lalor. 2015. 'Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing', *Current Biology*, 25: 2457-65.

Di Liberto, Giovanni M, Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné. 2019. 'Low-frequency cortical responses to natural speech reflect probabilistic phonotactics', *Neuroimage*, 196: 237-47.

Ding, N., and J. Z. Simon. 2013. 'Adaptive temporal encoding leads to a background-insensitive cortical representation of speech', *J Neurosci*, 33: 5728-35.

Drullman, R., J. M. Festen, and R. Plomp. 1994. 'Effect of reducing slow temporal modulations on speech reception', *J Acoust Soc Am*, 95: 2670-80.

Etard, O., and T. Reichenbach. 2019. 'Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise', *J Neurosci*, 39: 5750-59.

Frank, Stefan L, and Roel M Willems. 2017. 'Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension', *Language, Cognition and Neuroscience*, 32: 1192-203.

Gillis, M., J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck. 2021. 'Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics', *J Neurosci*, 41: 10316-29.

Golestani, N., A. Hervais-Adelman, J. Obleser, and S. K. Scott. 2013. 'Semantic versus perceptual interactions in neural processing of speech-in-noise', *Neuroimage*, 79: 52-61.

Gwilliams, Laura, Jean-Remi King, Alec Marantz, and David Poeppel. 2020. 'Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content', *bioRxiv*.

Hamilton, L. S., E. Edwards, and E. F. Chang. 2018. 'A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus', *Curr Biol*, 28: 1860-71 e4.

Heilbron, Micha, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. 2022. 'A hierarchy of linguistic predictions during natural language comprehension', *Proceedings of the National Academy of Sciences*, 119: e2201968119.

Hertrich, I., S. Dietrich, J. Trouvain, A. Moos, and H. Ackermann. 2012. 'Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal', *Psychophysiology*, 49: 322-34.

Hickok, Greg. 2015. *Neurobiology of language* (Elsevier: Boston, MA).

Hickok, Gregory, and David Poeppel. 2007. 'The cortical organization of speech processing', *Nature Reviews Neuroscience*, 8: 393-402.

Huth, Alexander G, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. 'Natural speech reveals the semantic maps that tile human cerebral cortex', *Nature*, 532: 453-58.

Iotzov, I., and L. C. Parra. 2019. 'EEG can predict speech intelligibility', *J Neural Eng*, 16: 036008.

Irino, T., and R. D. Patterson. 2006. 'A Dynamic Compressive Gammachirp Auditory Filterbank', *IEEE Trans Audio Speech Lang Process*, 14: 2222-32.

Kell, A. J. E., and J. H. McDermott. 2019. 'Invariance to background noise as a signature of non-primary auditory cortex', *Nat Commun*, 10: 3958.

Kleiner, Mario, David Brainard, and Denis Pelli. 2007. 'What's new in Psychtoolbox-3?'.

Koskinen, M., M. Kurimo, J. Gross, A. Hyvarinen, and R. Hari. 2020. 'Brain activity reflects the predictability of word sequences in listened continuous speech', *Neuroimage*, 219: 116936.

Lalor, E. C., and J. J. Foxe. 2010a. 'Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution', *European Journal of Neuroscience*, 31: 189–93.

Lalor, Edmund C, and John J Foxe. 2010b. 'Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution', *European journal of neuroscience*, 31: 189-93.

Lesenfants, D., J. Vanthornhout, E. Verschueren, L. Decruy, and T. Francart. 2019. 'Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations', *Hear Res*, 380: 1-9.

Maddox, R. K., and A. K. C. Lee. 2018. 'Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners', *eNeuro*, 5.

MATLAB. 2019. Natick, Massachusetts: The MathWorks Inc.

———. 2021. Natick, Massachusetts: The MathWorks Inc.

Mayo, L. H., M. Florentine, and S. Buus. 1997. 'Age of second-language acquisition and perception of speech in noise', *J Speech Lang Hear Res*, 40: 686-93.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In *Interspeech*, 498-502.

McDermott, J. H., and E. P. Simoncelli. 2011. 'Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis', *Neuron*, 71: 926-40.

Mesgarani, N., C. Cheung, K. Johnson, and E. F. Chang. 2014. 'Phonetic feature encoding in human superior temporal gyrus', *Science*, 343: 1006-10.

Mesik, J., L. Ray, and M. Wojtczak. 2021. 'Effects of Age on Cortical Tracking of Word-Level Features of Continuous Competing Speech', *Front Neurosci*, 15: 635126.

Muncke, J., I. Kuruvila, and U. Hoppe. 2022. 'Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise', *Front Neurosci*, 16: 876421.

Myers, B. R., M. D. Lense, and R. L. Gordon. 2019. 'Pushing the Envelope: Developments in Neural Entrainment to Speech and the Biological Underpinnings of Prosody Perception', *Brain Sci*, 9.

Nourski, K. V., R. A. Reale, H. Oya, H. Kawasaki, C. K. Kovach, H. Chen, M. A. Howard, 3rd, and J. F. Brugge. 2009. 'Temporal envelope of time-compressed speech represented in the human auditory cortex', *J Neurosci*, 29: 15564-74.

Oganian, Y., and E. F. Chang. 2019. 'A speech envelope landmark for syllable encoding in human superior temporal gyrus', *Sci Adv*, 5: eaay6279.

Orf, Martin, Malte Wöstmann, Ronny Hannemann, and Jonas Obleser. 2022. 'Auditory neural tracking reflects target enhancement but not distractor suppression in a psychophysically augmented continuous-speech paradigm', *bioRxiv*: 2022.06.18.496558.

Pasley, B. N., S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. 2012. 'Reconstructing speech from human auditory cortex', *PLoS Biol*, 10: e1001251.

Peck, Fleming C, Laurel J Gabard-Durnam, Carol L Wilkinson, William Bosl, Helen Tager-Flusberg, and Charles A Nelson. 2021. 'Prediction of autism spectrum disorder diagnosis using nonlinear measures of language-related EEG at 6 and 12 months', *Journal of neurodevelopmental disorders*, 13: 1-13.

Pereira, F., B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko. 2018. 'Toward a universal decoder of linguistic meaning from brain activation', *Nat Commun*, 9: 963.

Pion-Tonachini, L., K. Kreutz-Delgado, and S. Makeig. 2019. 'ICLabel: An automated electroencephalographic independent component classifier, dataset, and website', *Neuroimage*, 198: 181-97.

Rosen, S. 1992. 'Temporal information in speech: acoustic, auditory and linguistic aspects', *Philos Trans R Soc Lond B Biol Sci*, 336: 367-73.

Salisbury, D. F., M. E. Shenton, C. B. Griggs, A. Bonner-Jackson, and R. W. McCarley. 2002. 'Mismatch negativity in chronic schizophrenia and first-episode schizophrenia', *Arch Gen Psychiatry*, 59: 686-94.

Shackleton, T. M., and R. P. Carlyon. 1994. 'The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination', *J Acoust Soc Am*, 95: 3529-40.

Shannon, R. V., F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. 1995. 'Speech recognition with primarily temporal cues', *Science*, 270: 303-4.

Sohoglu, E., and M. H. Davis. 2020. 'Rapid computations of spectrotemporal prediction error support perception of degraded speech', *Elife*, 9.

Strauß, A., T. Wu, J. M. McQueen, O. Scharenborg, and F. Hintz. 2022. 'The differential roles of lexical and sublexical processing during spoken-word recognition in clear and in noise', *Cortex*, 151: 70-88.

Swaminathan, J., and M. G. Heinz. 2012. 'Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise', *J Neurosci*, 32: 1747-56.

Tang, C., L. S. Hamilton, and E. F. Chang. 2017. 'Intonational speech prosody encoding in the human auditory cortex', *Science*, 357: 797-801.

Teoh, E. S., F. Ahmed, and E. C. Lalor. 2022. 'Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment', *J Neurosci*, 42: 682-91.

Teoh, E. S., M. S. Cappelloni, and E. C. Lalor. 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: 3831-42.

Turner, C. W., B. J. Gantz, C. Vidal, A. Behrens, and B. A. Henry. 2004. 'Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing', *J Acoust Soc Am*, 115: 1729-35.

Vanthornhout, J., L. Decruy, J. Wouters, J. Z. Simon, and T. Francart. 2018. 'Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope', *J Assoc Res Otolaryngol*, 19: 181-91.

Verschueren, E., J. Vanthornhout, and T. Francart. 2021. 'The effect of stimulus intensity on neural envelope tracking', *Hear Res*, 403: 108175.

Viswanathan, V., H. M. Bharadwaj, B. G. Shinn-Cunningham, and M. G. Heinz. 2021. 'Modulation masking and fine structure shape neural envelope coding to predict speech intelligibility across diverse listening conditions', *J Acoust Soc Am*, 150: 2230.

Yasmin, Sonia, Vanessa Irsik, Ingrid S Johnsrude, and Björn Herrmann. 2023. 'The Effects of Speech Masking on Neural Tracking of Acoustic and Semantic Features of Natural Speech', *bioRxiv*: 2023.02. 10.527537.

Zou, J., J. Feng, T. Xu, P. Jin, C. Luo, J. Zhang, X. Pan, F. Chen, J. Zheng, and N. Ding. 2019. 'Auditory and language contributions to neural encoding of speech features in noisy environments', *Neuroimage*, 192: 66-75.