Neural Dynamics of the Processing of Speech Features: Evidence for a Progression of Features from Acoustic to Sentential Processing

I.M Dushyanthi Karunathilake[1], Christian Brodbeck[2], Shohini Bhattasali[3], Philip Resnik[4], Jonathan Z. Simon[1,5,6]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA,

[2]Department of Computing and Software, McMaster University, Hamilton, ON, Canada,

[3]Department of Language Studies, University of Toronto, Scarborough, Canada,

[4]Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA,

[5]Department of Biology, University of Maryland, College Park, MD, USA, [6]Institute for Systems Research, University of Maryland, College Park, MD, USA

Abstract

When we listen to speech, our brain's neurophysiological responses "track" its acoustic features, but it is less well understood how these auditory responses are modulated by linguistic content. Here, we recorded magnetoencephalography (MEG) responses while subjects listened to four types of continuous-speech-like passages: speech-envelope modulated noise, English-like non-words, scrambled words, and narrative passage. Temporal response function (TRF) analysis provides strong neural evidence for the emergent features of speech processing in cortex, from acoustics to higher-level linguistics, as incremental steps in neural speech processing. Critically, we show a stepwise hierarchical progression of progressively higher order features over time, reflected in both bottom-up (early) and top-down (late) processing stages. Linguistically driven top-down mechanisms take the form of late N400-like responses, suggesting a central role of predictive coding mechanisms at multiple levels. As expected, the neural processing of lower-level acoustic feature responses is bilateral or right lateralized, with left lateralization emerging only for lexical-semantic features. Finally, our results identify potential neural markers of the computations underlying speech perception and comprehension.

INTRODUCTION

Human language is known for its hierarchical structure, and in the course of speech perception and understanding, the brain first performs computations on the acoustic waveform, the results of which further undergo processing through various intermediate stages, integrating both bottom-up and top-down mechanisms, leading ultimately to semantic processing. Prior research has shown that these many neural processing stages align with at least some levels in the speech and linguistic hierarchy[1–3], including acoustic analysis, phonological analysis, morphemic analysis, lexical (word-level) processing, syntactic structures, and semantic (meaning-level) processing. However, the specific temporal dynamics and how these processes emerge during discourse level speech processing are still not well understood. When the information passed to neural intermediate processing stages is insufficient or incompatible with the subsequent processing, it can disrupt the information flow, indirectly affecting both lower and higher-level stages. Identifying the neural bases underlying these stages can provide insights about the intricate nature of auditory language processing. In this study, we aim to investigate the progression of the temporal dynamics of speech processing, and its reorganization in response to changes in the linguistic content of the sensory input.

Previous research has shown numerous brain regions that are sensitive to specific aspects of language understanding[4–6]. However, focus on speech-based language understanding has been limited, compared to text-based language understanding. The inherently limited temporal resolution of functional magnetic resonance imaging (fMRI) poses challenges in investigating quickly varying auditory responses and understanding the fast temporal dynamics of speech comprehension. Studies using imaging modalities with higher temporal resolution, such as magnetoencephalography (MEG) and electroencephalography (EEG), face different issues, such as confounds due to different stimulus lengths (durations), which has led many investigators to instead focus on processing individual words, rather than capturing the broader aspects of full spoken language processing (for a review see [7]). Recently, however, advances in neural speech-tracking measures such as the temporal response function (TRF) paradigm, have allowed investigators to study time-locked neural responses to many different speech features, and in more ecologically valid settings including long duration continuous speech. These neural speech-tracking measures are well established for acoustic properties of the speech such as the speech envelope (and envelope onset), and much is known about how they are modulated separately by top-down and bottom-up mechanisms[8,9]. Furthermore, recent research has revealed that many linguistic (non-acoustic) elements of speech, e.g., sub-lexical, lexical, lexico-semantic and context-based properties, also demonstrate neural tracking[1,2,10] above and beyond auditory neural tracking. How these tracking measures depend on the level of linguistic content of the speech, however, is still poorly understood, e.g., for very different levels of available semantic information. Furthermore, roles of top-down vs. bottom-up processing mechanisms, provide complementary insights into how the brain processes speech and language, may carry significant clinical implications for listeners who experience difficulty processing speech, e.g., older adults.

To answer these questions, we employed magnetoencephalography (MEG) to record the neural responses of listeners presented with four different kinds of speech material (Fig. 1(A)). In addition to ordinary narrative speech, we also presented word-scrambled narrative speech (with word-level semantic content but no more), and narrated non-words (which sounds like speech but with no semantic information whatsoever). Finally, we also employed speech envelope-modulated noise, which has the same prosody and rhythm of ordinary narrative speech, but does not sound like speech in any other way, and is entirely unintelligible even at the phoneme level. In this way each type of passage was designed to neurally progress through the brain up to a specific level in the hierarchy of speech processing and stop there: acoustic processing (for speech modulated noise), phoneme and word-boundary identification (narrated non-words), word meaning (scrambled narration), and full construction and processing of structured meaning (narrative), respectively. All four stimulus types were generated using Google text to speech synthesizer API[11] and exhibited similar accent, speech-like prosody, and rhythm across passages (Figure S6). The speech envelope-modulated noise was prepared by using the envelope of the synthesized speech to then modulate noise (with a speech-shaped spectrum), giving the noise a speech-like, varying rhythmicity. The non-word passages lack both lexical meaning and syntactic structure, somewhat resembling listening to a different language but maintaining the same accents, phonotactics, and prosody as the narrative and scrambled word passages. The scrambled word passages were constructed from narrative passages but with the words randomly permuted to eliminate both syntactic and contextual relationship between words, and then spoken with a natural prosody. The narrative passages, on the other hand, were linguistically well constructed with full structured meaning.

Audio examples of these passage types can be listened to at https://dushk88.github.io/progression-of-neural-features/.

Using multiple TRF (mTRF) analysis, with different TRFs contributing simultaneously for each respective speech feature, from acoustic levels to contextual levels (gammatone envelope spectrogram, gammatone onset spectrogram, phoneme onset, word onset, phoneme surprisal, cohort entropy, unigram (context-free) word surprisal, and contextual word surprisal), we investigate how different feature representations evolve as the brain steps through the processing levels (Fig. 1(B)). Analogous to event related potentials (ERPs), the TRF is a continuous signal which exhibits the neural encoding of speech feature over time (typically over hundreds of milliseconds), where peaks of different latencies indicate separate processing stages. We hypothesize that the ascending brain processing stages will show emergent features, from acoustic to sentence-level linguistic, as incremental steps in the processing of the speech occurs. Our findings support these hypotheses but additionally find that many speech features require more than one stage of processing: early processing which is primarily bottom-up, and late processing which is primarily top-down (consistent with the corrections that may be required for predictive coding models, analogous to generalized N400 ERP responses). We also confirm that hemispheric lateralization varies with speech feature: lower-level (more acoustic) processing generally manifests bilaterally or with a weak right hemisphere advantage, whereas left-lateralization dominates for lexico-semantic processing. Lastly, we demonstrate how the temporal dynamics of each feature processing is modulated by the linguistic content of the stimuli.

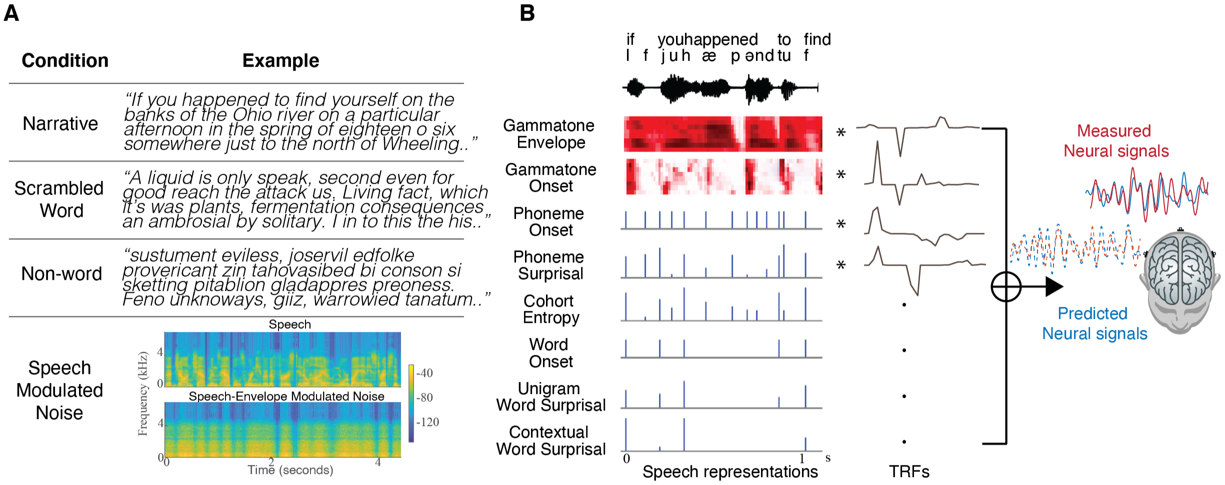
A
Condition
Example
Narrative
"If you happened to find yourself on the banks of the Ohio river on a particular afternoon in the spring of eighteen o six somewhere just to the north of Wheeling.."
Scrambled Word
"A liquid is only speak, second even for good reach the attack us, Living fact, which it's was plants, fermentation consequences an ambrosial by solitary. I in to this the his.."
Non-word
"sustument eviless, joservil edfolke provericant zin tahovasibed bi conson si sketting pitablion gladappres preoness. Feno unknoways, giiz, warrowied tanatum.."
Speech Modulated Noise
Speech
Speech-Envelope Modulated Noise
Frequency (kHz)
Time (seconds)
-40
-80
-120
B
if you happened to find
Gammatone Envelope
Gammatone Onset
Phoneme Onset
Phoneme Surprisal
Cohort Entropy
Word Onset
Unigram Word Surprisal
Contextual Word Surprisal
Speech representations
TRFs
Measured Neural signals
Predicted Neural signals

Fig. 1. Overview of the study design and analysis framework. (A). Examples of the four stimulus types. Participants (30 younger adults) listened to 1-minute-long speech passages of each passage type (32 passages total) while magnetoencephalography (MEG) brain activity was recorded. All stimuli had similar prosody and rhythm. Speech-modulated noise (bottom) is unintelligible and its spectro-temporal characteristics are shown in the bottom row. (B). Multivariate temporal response functions (mTRFs) were used to model the brain activity at different levels of speech representations and at each current dipole. Orthographic and phonemic transcriptions aligned with a sample acoustic waveform are shown for reference. Speech representations includes acoustic features (8-band auditory gammatone spectrogram; acoustic envelope and acoustic onset), sub-lexical features (phoneme onset, phoneme surprisal and cohort entropy) and lexical features (word onset, unigram word surprisal and contextual word surprisal).

RESULTS

Emergent features of speech processing

The present study first aimed to investigate the emergence of neural speech processing in response to varying levels of speech and linguistic information in the sensory input, by testing which speech representations are tracked by the brain response for each passage type. The test for significance of each speech representation (predictor) was done by comparing explained variance within pairs of models, one with all predictors included and the other for which the test predictor (speech representation of interest) was excluded; the test predictor was denoted as significant if the difference in explained variance was statistically significant. The full model employed for passages using speech-modulated noise and non-words included predictors for: gammatone envelope spectrogram, gammatone onset spectrogram, phoneme onset, word onset, phoneme surprisal, and cohort entropy. The model for scrambled word passages additionally included unigram word surprisal, and for narrative passages additionally incorporated both unigram word surprisal and contextual word surprisal. For the non-word passages, neither unigram nor contextual word surprisal could be applied as there were no real words. In the scrambled word passages, where context does not provide meaningful cues, contextual word surprisal collapsed to the unigram word surprisal (see Methods, predictor variables); therefore only the unigram word surprisal was used, since the explained variance by contextual word surprisal in the absence of coherent meaning is more conservatively ascribed to that of unigram surprisal. Statistical summary tables are reported in Table S1.

Model comparison results for all passage types are illustrated in Fig. 2. In the modulated noise condition (first column), only the acoustic features, specifically the gammatone envelope spectrogram ($t_{max} = 6.92, p < 0.001$) and gammatone onset spectrogram ($t_{max} = 5.79, p < 0.001$), contributed significantly to the observed neural data variance explained, i.e., significantly improving the model fit over the test model. Conversely, none of the linguistic predictors, phoneme onset ($t_{max} = 3.30, p = 0.07$), word onset ($t_{max} = 2.46, p = 0.91$), phoneme surprisal ($t_{max} = 1.51, p = 1.0$), and cohort entropy($t_{max} = 1.82, p = 0.99$) showed a significant contribution to the model's predictive power. However, in the presence of low-content speech stimuli, whether non-words (second column) or scrambled words (third column), in addition to these acoustic features, linguistic segmentation responses (phoneme and word onset) and statistically based linguistic features (phoneme surprisal and cohort entropy) also significantly contributed to the model's predictive power (non-words: gammatone envelope ($t_{max} = 11.90, p < 0.001$), gammatone onset ($t_{max} = 9.37, p < 0.001$), phoneme onset ($t_{max} = 7.25, p < 0.001$), phoneme surprisal ($t_{max} = 5.60, p < 0.001$), cohort entropy ($t_{max} = 6.83, p < 0.001$), word onset ($t_{max} = 6.90, p < 0.001$); scrambled words: gammatone envelope ($t_{max} = 10.97, p < 0.001$), gammatone onset ($t_{max} = 10.68, p < 0.001$), phoneme onset ($t_{max} = 6.43, p < 0.001$), phoneme surprisal ($t_{max} = 7.13, p < 0.001$), cohort entropy ($t_{max} = 8.60, p < 0.001$), word onset ($t_{max} = 6.17, p < 0.001$)). These results indicate that the acoustic features represented by the gammatone envelope and onset spectrograms are encoded in the brain regardless of the intelligibility of the sensory input, whereas linguistic features are tracked by the brain only when the linguistic units or linguistic unit boundaries are intelligible, regardless of any higher-level meaning.

Furthermore, model comparisons conducted on both scrambled ($t_{max} = 6.67, p < 0.001$) and narrative ($t_{max} = 6.48, p < 0.001$) passages revealed that when the words are individually meaningful, and irrespective of the structured coherence of the passages, the brain significantly tracked unigram (absent of context) word surprisal. This suggests that the brain is sensitive to the overall predictability of individual words, regardless of the overall coherence of the passage. Additionally, in narrative passages (fourth column) where structured contextual meaning was present, the brain exhibited substantial additional tracking of contextual word surprisal ($t_{max} = 5.48, p < 0.001$), over and beyond unigram word surprisal. Model comparison between unigram and contextual word surprisal in narrative passages additionally verified that contextual word surprisal is better encoded in the brain than unigram surprisal ($t_{max} = 4.70, p = 0.02$). These results indicate that the brain integrates both context-free and contextual level information during speech understanding, but contextual-level information is more strongly represented.

The anatomical distribution of the neural sources processing this hierarchy of speech processing was observed in locations consistent with an origin in Heschl's gyrus (HG), spreading to the superior temporal gyrus (STG) and much of temporal lobe (see **Fig. *2***). For higher-level linguistic features including phoneme surprisal, cohort entropy, word onset, unigram word surprisal, and contextual word surprisal, the feature representations additionally extended to left frontal regions.

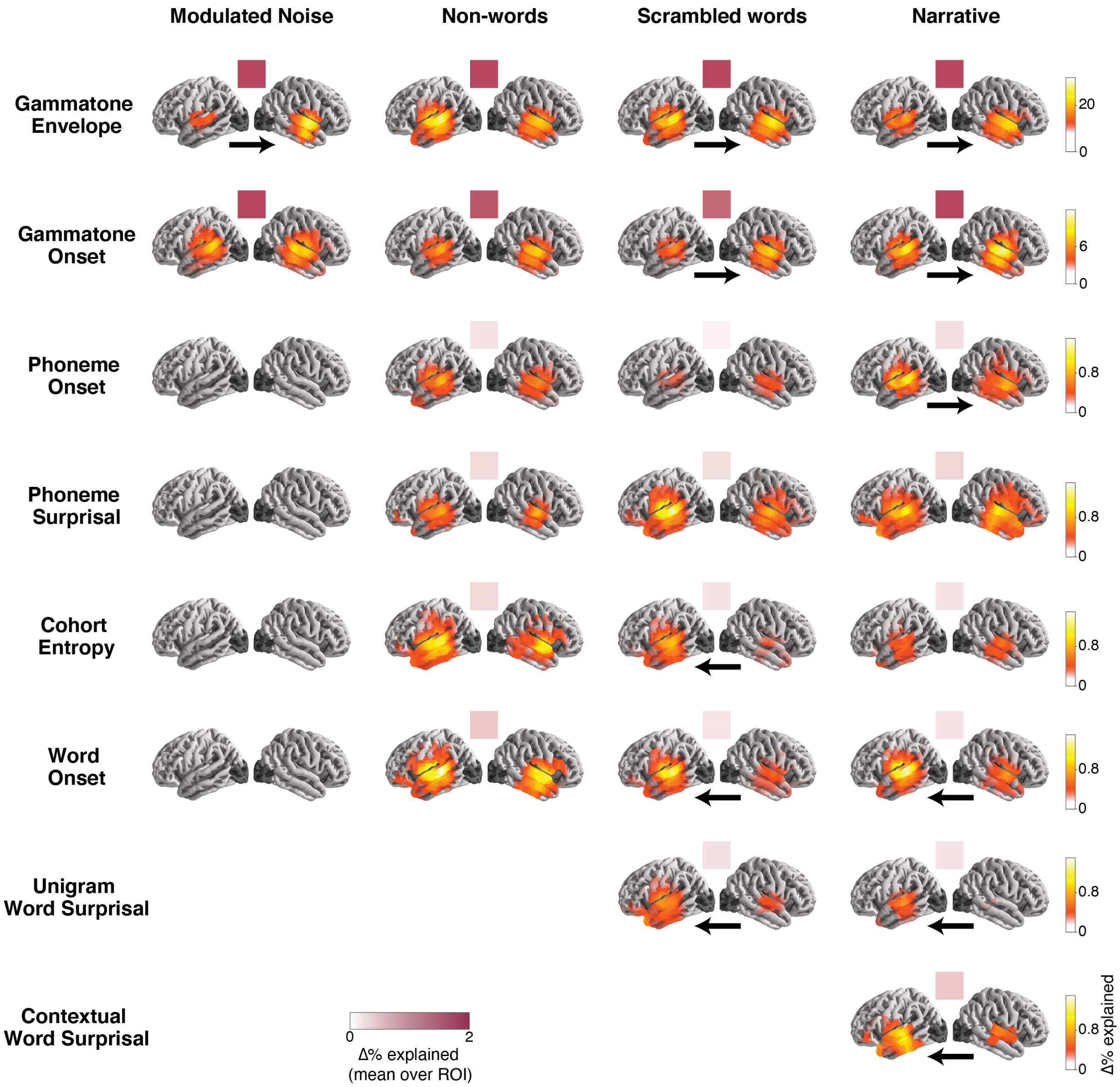

Fig. 2. Emergence of hierarchical speech processing. Anatomical brain plots visualize the cortical regions where each respective predictor significantly contributes to the model fit. Colored squares above the anatomical plots indicate average explained variance over frontal, temporal, and parietal regions. Black arrows below anatomical plots indicate significant hemispheric asymmetry. The first two rows show that acoustic features are represented in the brain irrespective of the passage type and intelligibility. Later rows show that linguistic features are tracked only when the linguistic feature boundaries are intelligible, irrespective of any higher-level (e.g., sentential meaning). When the context supports higher-level meaning above and beyond that of individual words, contextual word

surprisal is additionally represented in the brain. Lower-level feature processing is more right-lateralized, while higher level feature processing is more left-lateralized.

Lateralization of speech feature processing

We also examined the lateralization of neural speech feature processing for each passage type and speech feature. Instances of statistically significant lateralization are indicated by arrows in Fig. 2. Lateralization varied depending on the passage type and specific speech feature. Overall, lower-level speech feature processing exhibited a bilateral and right lateralized pattern (narrative: gammatone envelope ($t_{max} = -5.04, p < 0.001$), envelope onset ($t_{max} = -4.36, p = 0.02$), phoneme onset ($t_{max} = -4.57, p = 0.005$)) in the sources spanning in most of the temporal lobe, whereas higher-level speech feature processing were more left lateralized (narrative: word onset ($t_{max} = 3.21, p = 0.02$), unigram surprisal ($t_{max} = 3.23, p = 0.03$), contextual surprisal ($t_{max} = 3.30, p = 0.02$)) in superior temporal gyrus (STG), anterior temporal lobe and extending into frontal cortex. On the other hand, phoneme-level feature processing displayed a more bilateral pattern (narrative: phoneme surprisal ($t_{max} = -2.01, p = 0.82$), cohort entropy ($t_{max} = 2.38, p = 0.63$)). These results suggest distinct specialization of hemispheric regions for the processing of lower-level acoustic information vs. higher-level linguistic analysis.

Interestingly, the non-word passages showed predominantly bilateral responses across the different speech features (gammatone envelope ($t_{max} = 4.33, p = 0.06$), envelope onset ($t_{max} = -4.17, p = 0.04$), phoneme onset ($t_{max} = 3.58, p = 0.08$), phoneme surprisal ($t_{max} = 2.72, p = 0.29$), cohort entropy ($t_{max} = 3.92, p = 0.06$), word onset ($t_{max} = 2.58, p = 0.39$)), suggesting a more symmetrical hemispheric engagement of neural resources in non-word processing.

Effect of context on progression of neural speech processes: early and late

Neural responses obtained using MEG, with its fine-grained time resolution, may provide even greater insight from the temporal progression of cascading neural processes than from their anatomical locations. Having tested which types of speech-feature processing occurs in different contexts and in different anatomical regions, we then investigated how these contextual factors also influence the underlying neural mechanisms, associated with each the processing of speech feature, in the time domain. To this end, we utilized temporal response function (TRF) analysis that describes how the brain responds to each predictor over a range of latencies. To compare the

TRFs between passage types, TRFs magnitudes over the brain sources were aggregated. Analogous to event related potential (ERP) responses to punctate sounds, that exhibit distinct peaks at specific latencies characterized by their current polarity, so also do these TRFs, representing the direction and strength of the neural current response to each predictor, at various latencies. The dominant TRF peaks were identified and compared across passage types using repeated measures ANOVA (post hoc paired sample *t*-tests corrected for multiple comparisons using the false discovery rate method). To ensure unbiased TRF comparison across passage types, TRFs were generated from the same number of predictors. Peak latencies were also compared, and unless otherwise mentioned, no significant differences were found for latencies. Fig. 3, Fig***. 4***Fig***. 5*** illustrate average TRFs and their main peaks, and the accompanying bar plots provide a comprehensive comparison across the different speech passage types. The results presented in these figures show either only left or right hemisphere responses, so as not to overwhelm the figures; full analysis results, however, are included in the supplementary materials (Table S2, S3, S4, S5, S6, S7, S8, S9 and Figure S2, S2, S3).

Neural responses to acoustic features (Fig. 3) showed two prominent peaks: an early peak with a positive current polarity, and a late peak with a negative current polarity. These two peak latencies for the gammatone envelope were ~60 ms and ~120 ms respectively, while for the envelope onset feature, peak latencies were ~70 ms and ~150 ms (c.f. the early (P1) and late (N1) peaks of an auditory ERP). The late responses showed a predominantly right hemispheric lateralization ($p <$ 0.001). When comparing these two neural responses across passage types, we found that neural responses to speech passages were stronger compared to the non-speech modulated-noise ($p <$ 0.001). This effect was smaller for the right hemisphere early responses (left: early: $d = 1.06$, late: $d = 1.120$; right: early: $d = 0.47$, late: $d = 1.20$), and it was observed that the late peak was nearly absent in the modulated noise responses. When comparing the envelope onset responses among the speech passages, no significant differences were observed ($p > 0.2$). However, for envelope responses significant differences were found across speech passages in the left hemisphere. Early responses were smaller in narrative passages compared to scrambled and non-words ($p < 0.001$), whereas late responses were stronger in non-words compared to meaningful words ($p < 0.02$).

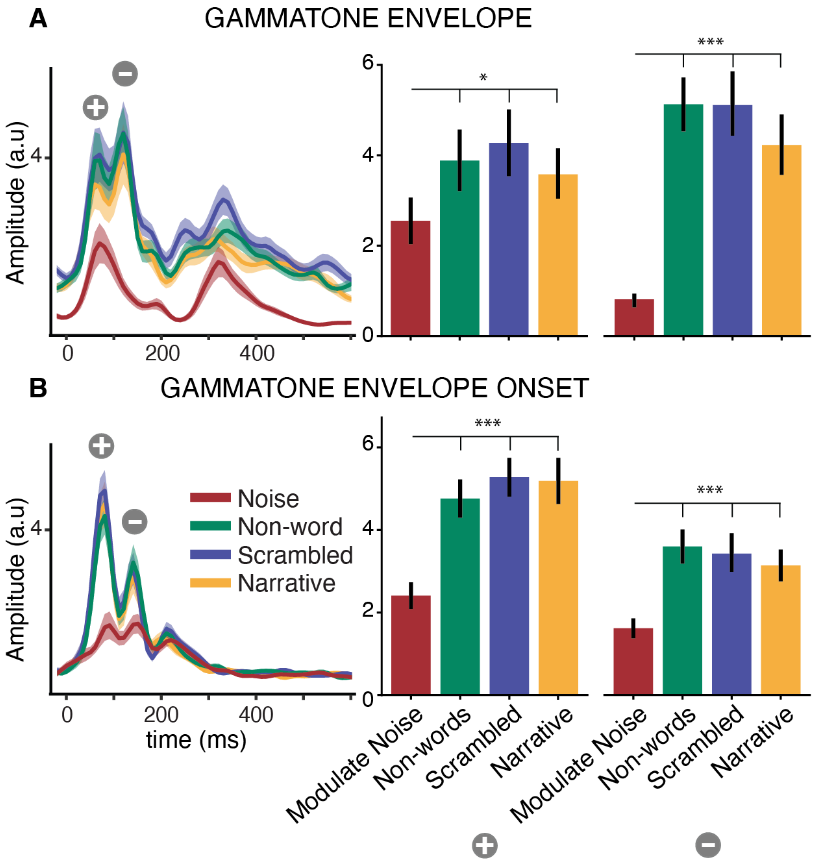

Fig. 3. Neural responses to acoustic features. (A). Gammatone envelope and (B). gammatone envelope onset responses. Left panels show the TRF magnitude aggregated over sources and subjects, by passage type. The TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. The right panel bar plots compare the peak amplitudes, first early then late, across passage types. Both early and late responses are stronger for speech compared to non-speech (noise). Only right hemisphere results shown (see supplemental Figure S2 for both hemispheres and individual data points). **p*<0.05, ***p*<0.01, ****p*<0.001

The analysis of phoneme onset responses (Fig. 4A) also revealed a robust early positive polarity peak with ~70 ms latency; the substantially later peak at ~250 ms latency was noisy and not robust across subjects. When comparing the peak amplitudes across passage types, no significant differences were observed in the right hemisphere for late responses. In the left hemisphere, early responses were stronger for non-words compared to scrambled passages (*p* = 0.002).

Phoneme surprisal (Fig. 4B) also showed two prominent peaks: an early positive polarity peak at ~70 ms and a late negative polarity peak at ~300 ms. Similar to phoneme onset responses, significant differences between passage types were found only in the left hemisphere. Both the early ($p = 0.03$) and late ($p < 0.03$) peaks were stronger in response to scrambled words compared to narrative and non-word passages.

For cohort entropy responses (Fig. 4C), two main processing mechanisms were observed for the scrambled and narrative passages: an early positive peak at ~70 ms and a late negative peak at ~380 ms. However, non-word passages showed a robust intermediate positive polarity peak at ~200 ms. Therefore, three peaks were identified as early, middle and late responses. The early peak was stronger for non-words compared to scrambled ($p = 0.01$) and to narrative ($p = 0.02$), while the middle peak was stronger in non-words compared to meaningful words ($p < 0.001$). In contrast, the late peak was stronger in scrambled words compared to narrative ($p = 0.009$); additionally, this peak was delayed for non-words compared to meaningful words ($p < 0.001$). Finally, the early cohort entropy responses were left lateralized for meaningful words ($p = 0.002$), middle non-word responses ($p = 0.03$) and late scrambled word responses ($p = 0.001$).

Analogous to cohort entropy responses, word onset responses (Fig. 4D) displayed two main peaks for both scrambled and narrative passages, while a middle peak was evident for non-words. Both early and middle peaks, occurring at ~100 ms and at ~200 ms respectively, exhibited a positive polarity. In contrast, the broad late peak at ~450 ms showed a negative polarity, resembling a characteristic N400 response. The early peak was stronger for meaningful words compared to non-words ($p < 0.001$), whereas this effect was reversed for the middle peak ($p < 0.001$). Interestingly, no significant differences were observed between the scrambled and narrative passages for both early ($p = 0.09$) and middle ($p = 0.07$) peaks. Remarkably, the late peak exhibited greater strength in response to scrambled words compared to non-words and narrative passages ($p = 0.003$). Moreover, the late peak latency was significantly delayed in the progression from narrative to scrambled (by ~30 ms, $p = 0.02$) to non-words (by ~50 ms, $p = 0.002$). Additionally, consistent with the explained variance lateralization comparisons, the non-words early and middle responses showed bilateral response ($p = 1.0$), while in meaningful words, the early responses were left lateralized ($p = 0.001$).

In the above analysis, we conservatively separated the early and middle peaks in cohort entropy and word onset responses into different processing stages, due to the considerable temporal separation between them. However, because of the strong similarity between the peak amplitudes and polarity, we also performed a separate analysis where the positive peaks (early and middle) were grouped together. In this analysis no significant differences in peak amplitudes were observed across the passage types (cohort entropy: $p > 0.08$, word onset: $p > 0.06$); as expected, latency comparisons revealed that the peak is delayed in non-words compared to meaningful words ($p < 0.001$).

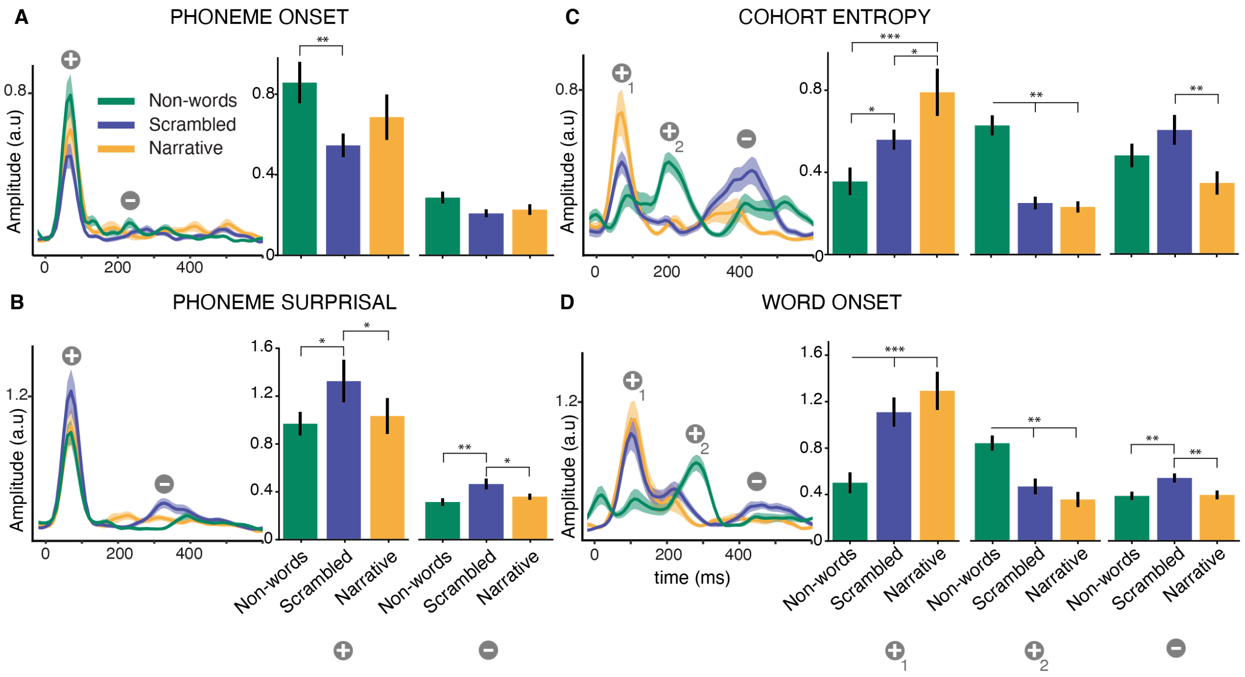

Fig. 4. Neural responses to sub-lexical and word onset speech features. (A). Phoneme onset, (B). phoneme surprisal, (C). cohort entropy, and (D) word onset (TRF magnitude plots and TRF peak bar plots as in Fig. 3). TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. For both word onset and cohort entropy responses, non-words showed a robust positive polarity peak between early and late peaks. These early, middle, and late peaks are indicated by $\oplus_1$, $\oplus_2$, and ⊖ respectively. The bar plots compare the peak amplitudes across passage types. Only left hemisphere results are shown here (see supplemental Figure S3 for both hemispheres and individual data points). Overall, the early responses were very differently modulated by the linguistic content. The

middle peak (second positive polarity peak) was strongest for non-words, while the late peak (negative polarity) was strongest for scrambled passages.

Unigram word surprisal TRFs (Fig. 5A) showed two main peaks, comparable to the early and late peaks observed in the word onset responses. Consistent with the explained variance lateralization, both peaks showed left hemispheric dominance. When comparing the peak strength between the scrambled and narrative passages, no significant differences were found for the early peak (p = 0.16). However, interestingly, the late peak in the scrambled word passages TRF was stronger (p < 0.001) and delayed by ~30 ms (p = 0.04) compared to narrative passages.

The TRFs between unigram and contextual word surprisal within the narrative passage were also compared (Fig. 5B). Both predictors represent word surprisal and exhibit a similar range of values, facilitating a direct comparison. Both TRFs showed similar peaks at comparable latencies and were left lateralized (p < 0.001). In contrast to the similarity in peak timing, contextual word surprisal showed stronger amplitudes for both early (p < 0.001) and late (p < 0.001) peaks in both hemispheres when compared to unigram surprisal, indicating contextual information is more robustly tracked.

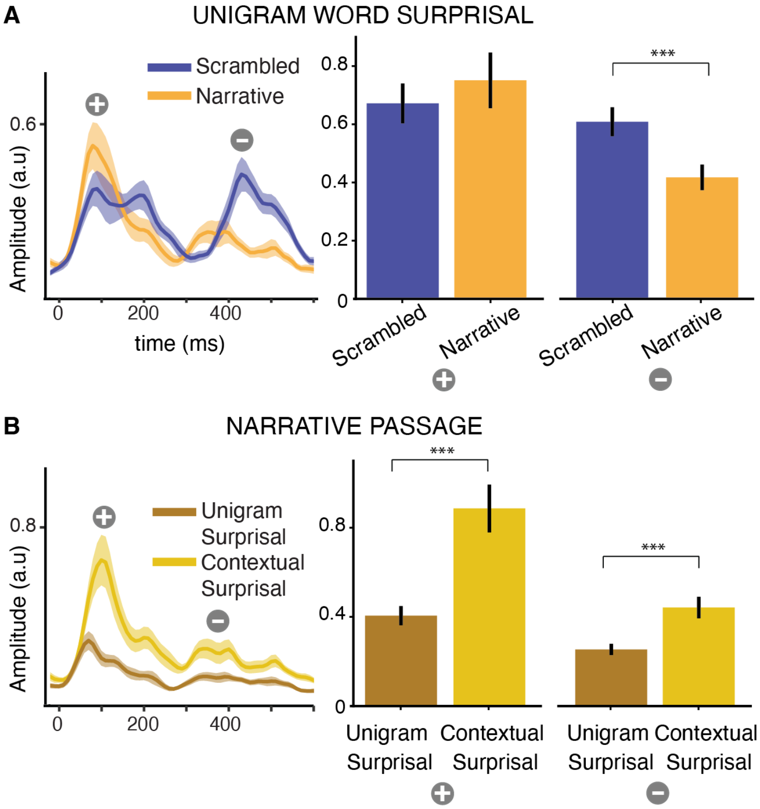

Fig. 5. Neural responses to lexico-semantic features. (A). Unigram surprisal (B). Unigram and contextual word surprisal for the narrative passage (TRF magnitude plots and TRF peak bar plots as in Fig. 3). The TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. Only left hemisphere results are shown here (see supplemental Figure S4 for both hemispheres and individual data points). The late unigram surprisal responses (N400-like) are stronger for scrambled passages compared to narrative passage. Contextual word surprisal responses are stronger compared to unigram surprisal responses. Note that the peak amplitudes for unigram surprisal in (A) and (B) are different, as the TRF model in (A) does not include a separate predictor for contextual surprisal.

In summary, TRF peak amplitude and latency comparisons revealed that, as the speech features become more abstract and less directly related to the acoustics, both early and late neural mechanisms tend to be delayed. Acoustic feature responses to speech were stronger compared to non-speech. Notably, the early and late peaks exhibited different modulations by linguistic content, consistent with representing different neural mechanisms. Furthermore, the TRFs showed quite different peak latencies for non-words compared to meaningful words. For linguistic level features, the late TRF peak amplitudes were stronger for scrambled words compared to non-words and narrative passages. Additionally, cohort entropy and word level late processing were delayed from narrative to scrambled to non-words. Peak lateralization analysis was consistent with explained variance lateralization analysis: lower-level feature processing was more right-lateralized, while higher level feature processing was more left-lateralized.

DISCUSSION

Using acoustic stimuli with similar prosody and rhythm but progressing from lacking any linguistic information (speech modulated noise) to possessing well-formed phonemes but no more (non-words), to possessing well-formed words but no larger scale context (scrambled), to fully well-formed linguistic information (narrative), allowed us to trace hierarchical neural processing as speech and speechlike sounds are eventually turned into language with full meaning in an ecologically valid setting. The TRF analysis enabled investigation of millisecond-level processing, revealing the timing and location of processes involved in the neural hierarchy as speech and language stages unfold. Critically, the TRF analysis also revealed distinctions between early processing stages, primarily driven by bottom-up processing. and later stages, allowing access to top-down driven processing.

We first showed evidence that the brain separately represents hierarchical speech and linguistic structures, with emergence of these features from acoustics to contextual processing arising with the increasing contextual information necessary for language comprehension. When the acoustic stimuli were unintelligible, only acoustic information was processed; sub-lexical and lexical processing were not activated. As the stimuli progressed from non-speech to speech, from meaningless non-words to meaningful words, both sub-lexical and lexical level linguistic feature processing emerged[12]. These linguistic features included both segmentation and statistically based linguistic features. The processing differences between non-words and meaningful words were evident from both lateralization and TRF latency analysis, pointing to different neural mechanisms involved in lexico-semantic processing. Thus, consistent with previous work, our results demonstrate that regardless of the stimulus type, acoustic features such as acoustic envelope and envelope onsets, are represented as such in the brain[9,13–16], reflecting a lower-level, initially bottom-up, sensory processing mechanism[17]. (Sub)-lexical features processes are activated only when (sub)-lexical units are recognizable and intelligible for linguistic process activation. Moving from non-words to meaningful words, our findings show the emergence of lexico-semantic processes, all while avoiding multiple inherent confounds of using an incomprehensible foreign language[18,19]. While these studies using comprehensible and incomprehensible language have shown that higher level word features (word unigram surprisal, word entropy, and contextual word surprisal) are not encoded for incomprehensible language, the explicit quantification of differences between non-words and meaningful words was not conducted in our models. This is due to the unavailability of unigram surprisal for non-words. Any unigram surprisal defined for non-words would be uniform across all non-words, and therefore identical to word onsets, and could not significantly account for more variance over and beyond the word onset predictor already included. Moving from scrambled words to narrative passages, our results also show the emergence of context-based word surprisal processing, indicating that the brain incorporates context to predict the structured meaning in line with the predictive coding theories[20,21]. This context-based word surprisal processing represents a higher-level processing that involves integration of linguistic and syntactic information to construct a structured meaning[10,20].

Hemispheric lateralization of auditory and speech processing has been widely studied and is of great interest, but results still show much variability across different studies[22]. Our lateralization results reveal distinct patterns of brain processing depending on the level of speech processing and stimulus type, under very similar listening conditions, e.g., all stimuli have natural prosody, were of extended duration (60 s). For the conditions with well-formed words, acoustic level processing is strongly bilateral or with a right hemispheric advantage. For sub-lexical level processing, the activation is bilateral, whereas lexical level processing shows a pronounced left hemispheric dominance. While the lower-level acoustic processing has been identified as a bilateral process, where lower-level acoustic processing involves both hemispheres [23,24], the right hemisphere's extra involvement in acoustic level processing aligns with its specialization in acoustic analysis, including extraction of spectral and temporal features from auditory input[8,25,26]. The left lateralization in higher level responses is consistent with the well-established left hemisphere specialization for language functions, including lexical representation and combinatorial syntactic and contextual processing[27,28,22,29]. Indeed, it is crucial to emphasize that numerous studies have reported different patterns of lateralization with tasks and language processes[28,30–32]. The modulated noise condition showed bilateral responses for acoustic onsets while envelope responses showed right hemispheric dominance. Previous studies have shown right lateralization of slow acoustic modulation for tasks involving low language demands[33] while left hemisphere envelope tracking has been associated with speech intelligibility (when higher-level features are not explicitly modeled)[34]. Interestingly, during non-word processing, the brain exhibited bilateral responses at every level of processing, suggesting non-word processing engages both hemispheres in their speech processing. This suggests that non-word processing utilizes more brain resources from both hemispheres, consistent with the brain not being specialized for non-word understanding[35,36]. These results indicate that pre-lexical auditory input analysis occurs in both hemispheres, and left lateralization emerges when the lexical-semantic processes are involved[37]. The observed lateralization patterns underscore the specialized contribution of each hemisphere to different aspects of speech comprehension and emphasizes the brain's flexibility in adapting to various linguistic and acoustic demands.

Critically, the TRF analysis provided valuable insights into the fast temporal dynamics and multiple neural mechanisms associated with each speech feature processing, and how they are influenced by linguistic complexity. This analysis reveals multiple processing stages associated with each speech feature (distinct peak polarities and latencies suggest that they arise from distinct neural sources[28,38]) and modulation of both acoustic and language-based feature processing by linguistic content. The results also suggest distinct top-down and bottom-up mechanisms for each feature, as will be discussed next.

Consistent with previous findings, acoustic feature (envelope and envelope onset) responses, showed two main peaks: e.g., an envelope early peak at ~60 ms and late peak at ~120 ms (and comparable peak latencies for the envelope onset). In line with previous work, our results show that acoustic responses are stronger for speech compared to non-speech[15,23,35,39]. The observed difference between speech and non-speech may mainly attributed to the underlying acoustic differences[17]. Additionally, envelope encoding may be expressed as both an acoustic feature and also a linguistic feature, thereby utilizing additional brain regions beyond acoustic processing for speech comprehension[15,40]. Moreover, previous studies have shown that envelope and envelope onset tracking can be modulated by intelligibility[41–44], attention[9,45,46] and linguistic content[35], all of which may contribute to the observed differences here between speech and non-speech.

Here, the acoustic envelope onset responses are not affected by the level of linguistic content, though they are known to be affected by other cognitive factors such as selective attention[9,46]. In contrast, however, acoustic envelope responses were stronger for non-words compared to meaningful words, specifically in the left hemisphere. These stark differences suggest that envelope and envelope onset tracking arise from quite different neural sources, even though they are temporally related[13]. The stronger activity observed for non-words over meaningful words in the left hemisphere is indeed consistent with previous studies[32,44], indicating engagement of more resources and higher-level processing mechanisms in the left hemisphere. During our listening tasks, subjects were actively engaged and required to perform a probe task; the presence of linguistic content, and syntactic and semantic structures, plausibly reduces any extra lower-level processing demands, thereby reducing the need for envelope processing from non-words to scrambled words to narrative passages.

The two distinct acoustic TRF peaks, early and late, indicate distinct underlying processing mechanisms. Indeed, previous studies have associated early auditory cortical responses with low-level acoustic (encoded at the periphery) processing, while associating the late response with top-down mechanisms affected by selective attention and task[45,47]. The consistent presence here of an early peak, irrespective of the passage type, reflects lower-level processing of acoustic information and its latency suggests a dominantly bottom-up driven mechanism. Conversely, the late peak, almost absent for non-speech and modulated by linguistic content, suggests a strongly top-down influenced mechanism. Thus, we emphasize here that the early peak is primarily driven by bottom-up mechanisms, while the late peak is strongly influenced by top-down mechanisms.

Phoneme level features also showed two main peaks[48], at ~70 ms and at ~380 ms, delayed compared to acoustic feature peaks, and non-words showed an additional peak at ~200 ms. Phoneme onset responses were found to be right lateralized for narrative, whereas they were bilateral or left lateralized for other linguistic features; this aligns with previous results that the response to phoneme onset may reflect more of a mixed acoustic-linguistic measure rather than a purely linguistic measure[17]. The early responses for non-words were enhanced for phoneme onset but were smaller for phoneme surprisal and cohort entropy compared to scrambled passages. It is also possible that differences in predictor distributions between words and non-words (see Fig. S1) may have influenced the statistics-based phoneme features, which in turn may have indirectly affected the phoneme onset responses due to their concurrent timing. Additional activation of brain regions in in the processing of non-words may also have contributed to the phoneme onset difference.

The temporal structre of cohort entropy TRFs for non-words closely resembled the word onset responses, especially compared to those of phoneme surprisal. While one might expect similar trends between the phoneme surprisal and cohort entropy for non-words, it is important to note that these measures quantify different aspects of sub-lexical processing: phoneme surprisal represents *phonological* uncertainty, whereas cohort entropy reflects *lexical* uncertainty[49] (see[50] for a review). Therefore, phoneme surprisal is more strictly sub-lexical, but cohort entropy is lexically based as well. In this sense, cohort entropy likely reflects word-level feature processing more than just phoneme level processing, and this is supported by our results.

Indeed, the considerable temporal separation between the early and middle peaks of word onset and cohort entropy may suggest additional mechanisms associated with non-word processing. A key difference between segmenting non-words vs. words is that boundaries between non-words are not clearly defined, and identifying them relies entirely on indirect cues such as implicit pauses and prosody changes. This seems likely to contribute substantially to differences in word onset TRF morphology between words and non-words. When early and middle peaks were combined, no amplitude differences were observed between passage types, only latency, indicating that they indeed represent a single source that is linked to word segmentation, a bottom-up mechanism for lexical activation, but the latency of which depends upon the difficulty of the segmentation problem.

Our findings showed that lexical level predictors (word onset, unigram surprisal and contextual surprisal) elicit neural responses at similar latencies, characterized by an early peak at ~100 ms and a late peak at ~450 ms, considerably delayed compared to acoustic and sub-lexical responses. The early peak which showed no difference between the scrambled and narrative passages, indicates that this early neural component is primarily driven by bottom-up processing mechanisms involved in initial analysis and activation of meaningful words, irrespective of the context. Phoneme and word surprisal early components are so early that they qualify as bottom-up processing even though linguistic surprisal is often associated with predictive coding models which themselves may be treated as requiring a top-down contribution[50]. Critically, however, predictive coding models can be dominantly bottom-up, for example, a recent model of predictive processing in auditory midbrain [51]

In general, the late peaks for phoneme and lexical level features were stronger in scrambled words compared to non-words and narrative passages, and were delayed from narrative to scrambled to non-words, suggesting the late responses are affected by linguistic content. Even though different patterns were observed for early stage between passage types, the later stage trends were consistent for both phoneme level and lexical level, suggesting the late stage may represent similar neural mechanisms. Additionally, the left lateralization of both early and late peaks highlights the involvement of language-level mechanisms, even at the early stage of lexical level neural processing. Remarkably, the late peak resembles the characteristics of the ERP N400 response, a well-known brain response associated with being modulated by predictability, often used to investigate semantic processing, and also modulated by intelligibility and comprehension (for a review see [52,53]). Thus, our results suggest that context-based predictability facilitates the pre-activation of word identification or semantic integration, thereby reducing the strength and latency of N400-like response in narrative passages compared to scrambled words. These results are consistent with the previous work that has shown differences between scrambled and narrative passages in the late response[54,55]. Some studies have also reported weaker late response with scrambling[18,56], though this difference may be related to the variations in the experimental design (e.g., EEG vs MEG, contextual measure employed). Conversely, the smaller N400-like responses for non-words aligns with non-words being unpredictable, and thus not activating the N400 mechanism[57]. However, in the current study the non-word passages did include non-words that resembled real words (e.g., "sustument" and "bi"), which could lead to possible lexical activation of root words, and, consequently, elicit some N400 response. Some lexical activation for non-words could diminish the difference between narrative and non-words. Therefore, the N400-like response seen here could arise from both semantic and non-semantic violations of expectation[57]. These interpretations are further supported by the latency analysis, which showed that peaks are delayed from narrative to scrambled to non-words. The earliest processing of the narrative stimulus suggests that rapid access to the mental lexicon is facilitated by the contextual information. These results are consistent with previous studies showing a reduction in N400 for semantically congruent as well as pseudo-words or non-word processing[52,57]. Other studies have shown that the N400 is stronger for non-words compared to words[58,59], however, but in paradigms where the non-words were presented between meaningful words, which alters the experimental design, behavioral expectations, and, likely, the neural processing form the current work. These results suggest that the late responses to both phoneme and lexical features are influence by top-down driven mechanisms, that facilitate both phonological and lexical processing necessary in speech understanding.

Moreover, our results further highlight that the contextual word predictions are robustly represented compared to non-contextual (unigram) word predictions, suggesting that when context-based predictions better are supported, they align with the predictive coding mechanisms compared to non-contextual word predictions[60,61]. Of course these two measures represent different cognitive operations, where context-based surprisal involves word retrieval based on contextual and syntactic information, whereas unigram surprisal retrieve words from the mental lexicon based solely on sensory cues[2,59,62]. Our results support this distinction, as we observed dissociable effects of local and contextual features, with both representing in the neural responses over and beyond the other feature.

The current analysis does have its limitations. Specifically, more fine-grained stages within the speech and language processing hierarchy, such as syntactic-only processing and semantic-only processing, were not included (due to experimental constraints related to limiting the duration of the recording sessions). Additionally, other speech features, including but not limited to morphemes, function words, and content words, were not incorporated into the analysis and analysis of these aspects are considered topics for future work.

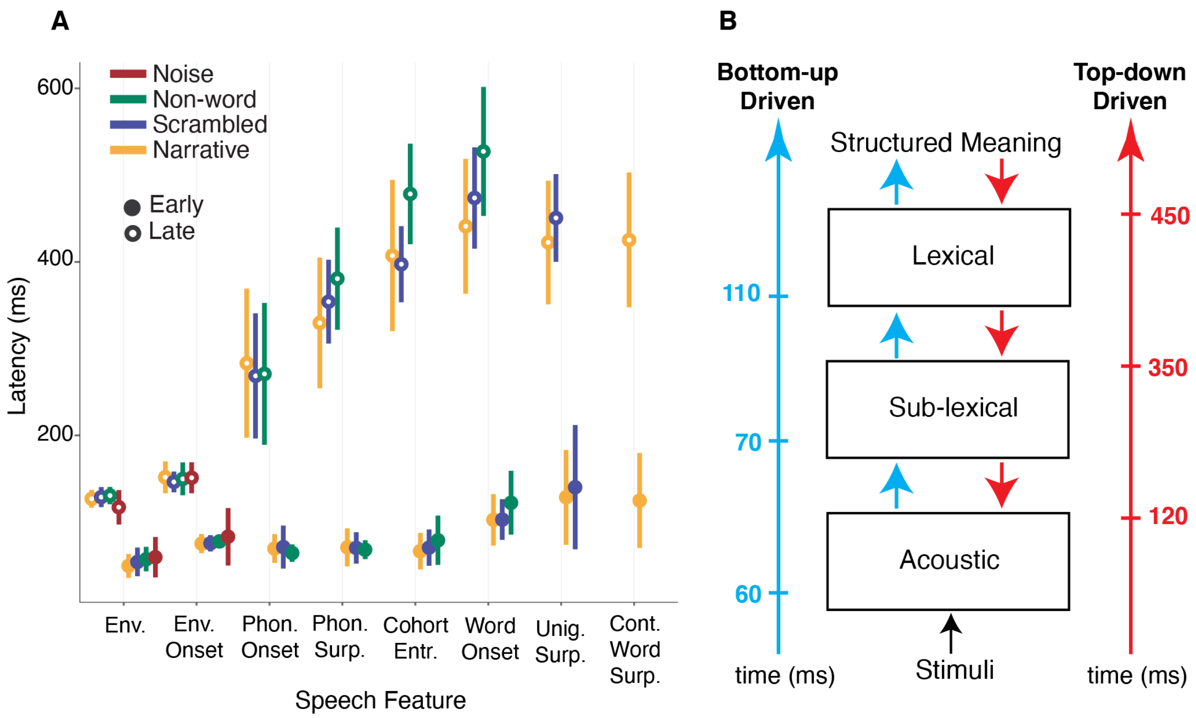

Fig. 6. Temporal profile of speech feature processing. (A). Latency of both early and late processing stages associated with each feature processing. As the features go up in hierarchical acoustic and linguistic structures both early and late peak processing show longer latencies (B). Schematic summary of the bottom-up and top-down temporal profiles at each processing level. Acoustic = [Envelope, Envelope Onset], Sub-lexical = [Phoneme Onset, Phoneme Surprisal], Lexical = [Cohort Entropy, Word Onset, Unigram Surprisal, Contextual Word Surprisal].

In summary, our TRF analysis revealed that the brain processes the hierarchy of acoustic and linguistic structures (from acoustics to context-based features) in a progression of neural stages, and with a characteristic temporal dynamic associated with each feature processing. As we ascend the hierarchy, processing of features shows longer latencies for both early and late mechanisms (Fig. 6A), suggesting a graded computation of features, over time, in the cortex[3], starting as early as ~50 ms and extending to ~500 ms. These mechanisms accumulate sounds features, analyze for lexical-semantic information, and integrate with the semantic context. Typical feature processing has both an early and late stage, with the early processing stage being driven by bottom-up activation and the late processing being influenced by top-down mechanisms, as inferred based on latency and modulation by linguistic content. These findings are summarized in Fig. 6B, illustrating the hierarchical processing of acoustic and linguistic structures in speech comprehension and timing associated with earliest bottom-up and top-down mechanisms at each level. While bottom-up driven mechanisms are less intriguing, top-down driven mechanisms demonstrate involvement in predictive coding mechanisms, making them better neural markers of cognitive decoding. Our results complement previous fMRI studies[5,6,23] by leveraging the temporal dynamics of feature processing and electrophysiological studies by investigating effects of linguistic content on neural tracking measures[1,2].

In conclusion, using multiple stimulus types with varying linguistic content (modulated noise, non-words, scrambled words and narrative), we provided neural evidence for the progression of different speech features along the speech and linguistic hierarchy, with increasing the semantic information in the sensory input. Our findings highlighted hemispheric lateralization, temporal dynamics and neural mechanisms associated with each level and how they are further modulated by linguistic content. Our analysis reveals the bottom-up and top-down mechanisms associated at each stage processing. These insights deepen our understanding of the neural markers that might be utilized to evaluate cognitive decoding and the construction of sentence meaning, particularly in different clinical populations.

METHODS

Participants

34 native English speaking younger adults (17 females, mean age 22 y, age range 18-29 y, 3 left-handed) participated in this experiment. Data from four subjects were excluded from the analysis because of technical issues during data acquisition (1 subject) and poor performance on the behavioral tasks (see experimental procedure) (3 subjects), leaving thirty participants in the analysis (15 females, mean age 22 y, age range 18-29 y, 1 left-handed). All participants reported normal hearing and no history of neurological or hearing-related disorders. All experimental procedures were approved by the Internal Review Board of the University of Maryland, College Park. The participants gave their written informed consent before the experiment and received monetary compensation, or course credit (1 subject).

Speech stimuli

Four types of speech stimuli: narrative, scrambled word, non-words and speech-modulated noise were generated as described below (sample materials are shown in Fig. 1(a) and can be listened to at https://dushk88.github.io/progression-of-neural-features/). Text used for speech stimuli were excerpts from the book "The Botany of Desire" by Michael Pollan[63]. Speech stimuli were computer synthesized using Google text to speech API[11] (gTTS) (see example: https://cloud.google.com/text-to-speech). The use of modern text-to-speech synthesizers provides human-like, natural-sounding speech[64,65], and ensures acoustic parameters like speech rate, rhythm, and emphasis are consistent across passage types, which is crucial for comparing neural responses across passage types in the current study.

The narrative (structured and meaningful) passages were excerpts from the first section of the book. A separate section of the book was used where the words were randomly permuted to create the scrambled word (structured intermediate) passages. Another section, non-overlapping with the previous passages, was used to generate the speech-modulated noise (unintelligible speech) passages. For the non-word (gibberish) passages, nonsense words were extracted from https://www.soybomb.com/tricks/words/ and were randomly arranged to form a continuous passage. Initial versions of both scrambled and non-word passages lacked punctuation marks, but since silences and pauses between words and sentences create natural sounding and rhythmic speech, and in gTTS pauses and silences are cued by punctuation marks, punctuation marks were manually added to the scrambled and non-word passages (using the distribution of the number of words between punctuation marks in the original book).

Speech was synthesized with gTTS using the English US accent male voice and Google Wavenet voice type "en-US-Wavenet-J" (https://google.com/text-to-speech/docs/voices) at the default sampling rate 24 kHz. Once the speech passages were generated, audio files were lowpass filtered below 4 kHz since the MEG audio delivery (air tube) system has a lowpass cutoff of ~4 kHz. Then the silence segments were trimmed to 400 ms and the audio stimuli were resampled to 22.5 kHz. For each of the speech stimulus types, 1-minute-duration excerpts were extracted.

For construction of the modulated noise passage, the corresponding speech stimuli generated for modulated noise passages were further modified. First, stationary noise was generated with the same frequency spectrum as the speech by randomizing the phases of the stimulus frequency spectrum and inverting back to the time domain. In order to add back the lost rhythmicity to the noise, the stationary speech shaped noise was then modulated with the corresponding slow speech envelope of the original speech (See Fig. 1(a)). The slow speech envelope was extracted by low pass filtering (with a 5 Hz cutoff) the Hilbert envelope of the speech passage.

Experimental procedure

The experiment was conducted in four blocks. Each block comprised of one passage from each passage type, and each passage was repeated twice. The order of passage types was counterbalanced across subjects. The narrative passages were presented in chronological order to preserve the story line to increase the subjects' attention. In total, each participant listened to a total of 32 trials (4 blocks × 4 types × 2 repetitions = 32 trials) and 8 trials from each passage type (4 blocks × 2 repetitions), where a trial is defined as a presentation of 1-minute-long stimulus passage. At the start of each passage type, subjects were instructed which passage type they were about to listen to. A probe question (depending on the type of passage) was for each passage (counting occurrences of a probe word; a contextual question based on the story passage; judging which emotion was conveyed in the speech-modulated noise passage) to help maintain participant's attention to the listening task. Participants who correctly answered at least 70% of the questions (excluding the emotion judgement) were included in the analysis.

The subjects lay supine during the entire experiment and were asked to minimize body movements. Subjects kept their eyes open and fixated at a center of a grey screen. The stimuli were delivered bilaterally at ~70 dB SPL with E-A-RTONE 3 A tubes (impedance 50 Ω) which severely attenuate frequencies above 3 – 4 kHz, and E-A-RLINK (Etymotic Research, Elk Grove Village, United States) disposable earbuds inserted into ear canals.

Data acquisition and preprocessing

Neuromagnetic data were recorded inside a dimly lit, magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany) with a whole head 157-channel MEG system (KIT, Kanazawa, Japan), installed at the Maryland Neuroimaging Center. The data were recorded with a sampling rate of 1 kHz along with an online low-pass filter (< 200 Hz) and a 60 Hz notch filter. Three of the additional sensor channels were employed as environment reference channels.

All data analyses were performed in mne-python 0.23.0[66,67] and eelbrain 0.36[68]. Flat channels were excluded and the environmental magnetic interferences were suppressed using temporal signal space separation (tSSS)[69]. Then MEG data were filtered between 1 and 60 Hz using a zero-phase FIR filter (mne-python 0.23.0 default settings). Artifacts such as ocular, cardiac, and muscle artifacts were reduced using independent component analysis (ICA)[70]. The cleaned data were then low pass filtered between 1 and 10 Hz and downsampled to 100 Hz for further analysis.

Neural source localization

The scalp surface (> 2000 points), five head position indicator (HPI) coils (three placed on the forehead, left and right ear), and anatomical landmarks (nasion, left and right periauricular) of each participant was digitized using Polhemus 3SPACE FASTRAK three-dimensional digitizer. The position of the participant's head relative to the sensors was determined before and after the experiment using HPI coils attached to the scalp surface and the two measurements were averaged. The digitized head shape and the HPI coils locations were used to co-register the template FreeSurfer "fsaverage"[71] brain to each participant's head shape using rotation, translation, and uniform scaling.

A neural source space was generated by four-fold icosahedral subdivision of the white matter surface of the fsaverage brain, with the constraint that all source dipoles be oriented perpendicular to the cortical surface. The source space data and the noise covariance estimated from empty room data were used to compute inverse operator via minimum norm current estimation[72,73]. The subsequent analysis were limited to frontal, temporal, and parietal brain regions based on the 'aparc' FreeSurfer parcellation[74].

Predictor variables

The speech signal was analyzed in distinct feature spaces that represent various levels of the language hierarchy. These features were grouped into four primary categories: acoustic properties (i.e., acoustic envelope and acoustic onsets), sub-lexical properties (i.e., phoneme onset, phoneme surprisal, and cohort entropy), lexical properties (i.e., word onset and unigram word surprisal), and contextual features (i.e., contextual word surprisal). The methodology for generating each of these predictors is detailed below. Overall, these predictors were generated using a combination of signal processing techniques, automatic speech recognition (ASR) systems, and probabilistic models. All predictor variables were downsampled to 100 Hz.

The acoustic envelope predictor is a measure of the amplitude modulation of the speech signal, and reflects the acoustic power/energy of the speech signal over time. In contrast, the acoustic envelope onset predictor is a measure of the salient transients of the speech signal, which are particularly prominent at the beginning of syllables or phonemes. The acoustic envelope and acoustic onsets were computed based on the human auditory system inspired gammatone filters computed by Gammatone Filterbank Toolkit 1.0[75], using 256 center frequencies with cut-off frequencies ranging logarithmically from 20 to 5000 Hz. Each frequency band's envelope was resampled to 1000 Hz and transformed to log scale. The resulting envelope spectrogram was then averaged into 8 logarithmically spaced frequency bands to obtain the final acoustic envelope predictor. Eight bands were chosen as a trade-off between computational efficiency and the ability to capture detailed information about the amplitude modulation. The acoustic onset representations were computed using the above gammatone acoustic envelope 256-band spectrogram, by applying an auditory edge detection algorithm[76]. The onset spectrogram was also subjected to a processing step involving averaging across eight logarithmically spaced frequency bands. The distributions of the acoustic envelope and onset predictor were found to be comparable across speech conditions, non-words, scrambled and narrative passages. However, some variations were observed between the speech stimuli and the speech modulated noise stimuli, as evidenced by the comparisons shown in Figure S5(A). This discrepancy may be attributed to the diminishment of formants and/or sharp onsets in the non-speech (due to its modulation being induced only by the broad band envelope of the speech stimuli).

Preliminary speech audio alignment for the occurrence of discrete words and phonemes was accomplished using the Montreal Forced Aligner[77]. Grapheme to phoneme conversion was done using the pre-trained 'english-g2p' model available within the Montreal Forced Aligner. The pronunciation lexicon, transcriptions, and audio file were aligned using the pre-trained 'english' acoustic model. The resulting annotations were visually examined in PRAAT[78] and manually adjusted when necessary. Phoneme onsets and word onsets predictors were modeled as impulses at the onset of each phoneme and word respectively. Phoneme surprisal and cohort entropy, which reflect separate information-theoretic properties of the sub-lexicon in its lexical context, are widely used in neural word processing analysis[1,2,49]. Phoneme surprisal quantifies the level of probabilistic surprisal associated with the current phoneme, given the occurrence of the sequence of phonemes prior to it within the current word. On the other hand, cohort entropy captures the level of uncertainty of remaining lexical candidates that match the observed phoneme sequence. Mathematically, phoneme surprisal for a given position i within a word is defined as the $-\log_2 \frac{\sum_{word}^{cohort_i} p_{word}}{\sum_{word}^{cohort_{i-1}} p_{word}}$ and cohort entropy is defined as $-\sum_{word}^{cohort_i} p_{word}\, log_2(p_{word})$. Here, *cohort*$_i$ refers to the set of words that are compatible with the phoneme sequence from the beginning of the word to the i^{th} phoneme, and p_{word} is the probability of the word derived from the *wordfreq* Python library[79]. The *wordfreq* python library is based on the Exquisite Corpus and covers a broad range of words that appear at least once per 100 million words. The phonetic lexicon for each word was extracted from the CMU pronouncing dictionary, available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict. The corpus comprised of all the words that were included in both the CMU dictionary and the *wordfreq*. Cohort entropy and phoneme surprisal values were computed for each phoneme onset and represented as impulses, scaled by its corresponding value. These two predictors were similar in the non-speech, scrambled, and narrative passages as they included meaningful words. However, they showed different distributions between meaningful words and non-words as illustrated in Figure S5(B). As expected, phoneme surprisal exhibited a greater proportion of highly surprising phonemes for non-words, whereas cohort entropy displayed more zeros for non-words, since the potentially available lexicon must become empty after some number of phonemes.

Analogous to the phoneme level surprisal predictor, two different measures of word level surprisal were estimated: unigram word surprisal and contextual word surprisal. Unigram word surprisal measures how surprising a word is independent of the context and is based on the probability distribution of individual words computed from *wordfreq*. Unigram word surprisal for each word is calculated by $-\log_2(p_{word})$ and represented as an impulse at each word onset, scaled by the unigram word surprisal value. In contrast, contextual word surprisal depends on the preceding context and reflects how surprising the current word is given the previous context. Contextual word surprisal was estimated using the open source, pre-trained, and transformer-based[80] large language model GPT-2, implemented in the Hugging Face environment[81]. Each 1-minute-long passage was preprocessed (removing punctuation and converting to lower case, with the exception of proper nouns), tokenized using byte-pair encoding[82], and provided to the neural network model. The tokens could represent either complete words or sub-words. The final layer of the model was utilized to calculate the word surprisal. This final layer outputs prediction scores for each token in the vocabulary, indicating the likelihood of it being the next word given the preceding tokens (context) that extends all previous tokens, extending to a maximum of 1024 tokens. The prediction scores were subjected to a SoftMax transformation to compute probabilities. The current word probability was determined by the probability associated with its corresponding token. In cases where words span over multiple tokens, word probability was computed by the joint probability of those tokens. Contextual word surprisal was computed as $-\log_2(P_{word}|context)$ and represented as an impulse at each word onset, scaled by the corresponding contextual word surprisal of that word. The unigram and contextual word surprisal values were calculated only for the scrambled and narrative passages since they were not defined for non-words. However, as can be seen from the Figure S5(C), a high correlation between contextual and unigram word surprisal was observed for the scrambled word condition $(r(741) = 0.91, p < 0.001)$, suggesting that contextual word surprisal collapses to unigram word surprisal when the context fails to provide informative cues for predicting the next word, as would be expected. Due to this very strong correlation between these two predictors in the scrambled passages, the contextual word surprisal predictor was excluded from the TRF modelling there and only the more conservative unigram word surprisal was used.

Forward models (Temporal Response Functions)

The forward model approach referred to as temporal response function analysis[83] was used to estimate how a set of predictor variables relates to the source localized MEG data. The model for each neural source is defined as:

$$r(t) = \sum_{i}^{N} \sum_{\tau}^{T} h(i,\tau)x(i,t-\tau) + \varepsilon(t)$$

Where $r(t)$ is the neural response at time *t*, $x(i,t)$ is the *i*th predictor time series, and $\varepsilon(t)$ is the residual neural response not explained by the model. The TRF, $h(i,\tau)$, is a filter that describes the linear relationship between the predictor time series and neural source time series (input and output) at different time lags within the integration window $[\tau, T]$. In this model, each time lag of each predictor competes against each other to explain variance of the neural response, which results in larger TRF model weights associated with greater contributions to the explained variance. The TRF model weights were estimated by minimizing the mean absolute difference between actual $(r(t))$ and predicted $(\widehat{r(t)} = r(t) - \varepsilon(t))$ neural response. Model performance is evaluated by the prediction accuracy measured by the correlation coefficient between the actual and predicted neural response, a measure of neural tracking of the predictors.

To compute TRFs for each subject, condition, and at each source dipole, the eight trials per condition (total 8 minutes) were concatenated and the boosting algorithm[84] was employed. Prior to boosting, L1 standardization was performed on both the predictors and neural responses by subtracting the mean and dividing by the mean absolute value. TRF lags from -20 ms to 800 ms were used, with a basis of 50 ms Hamming windows employed to smooth the otherwise overly sparse TRFs. TRF estimation used four-fold cross-validation, where two folds were allocated for training, one-fold for validation and one-fold for testing. For each testing fold, each of the remaining three partitions served as a validation set, resulting in three TRFs per testing fold. These three TRFs were averaged to generate one average TRF per testing fold, which was then used to compute the prediction accuracy against the testing set. The TRFs and corresponding prediction accuracies from each of the testing folds were further averaged to generate a single TRF and single prediction accuracy per source dipole.

Phonetic feature modelling

Before starting, we first analyzed how the phonetic features, phoneme onset, phoneme surprisal and cohort entropy should best be modeled, since different previous studies have used used different approaches: modeling word-initial phonemes as separate features[1]; including word-initial phonemes only in phoneme surprisal and cohort entropy[85]; and including word-initial phoneme only in phoneme onset[2]. We compared models with and without word-initial phoneme onset on a base model with envelope spectrogram, envelope onset and word onset. The model with the word-initial phoneme onset showed better prediction accuracy compared to a model without the word-initial phoneme onset ($t_{max} = 5.31, p < 0.001$). To test for the phoneme surprisal and cohort entropy, we compared the three models by including, excluding, or separately modelling the word-initial phoneme, using a base model with gammatone envelope spectrogram, onset spectrogram and word onset. Model comparisons with adjusted r-squared revealed that including the word-initial phoneme yield the best prediction accuracies for both phoneme surprisal ($1\ vs\ 2:\ t_{max} = 4.38, p < 0.001, 1\ vs\ 3:\ t_{max} = 3.81, p = 0.02$) and cohort entropy ($1\ vs\ 2:\ t_{max} = 5.07, p < 0.001, 1\ vs\ 3:\ t_{max} = 4.78, p = 0.02$). We therefore opted to include the word-initial phoneme in the phonetic feature modelling.

TRF peak extraction

TRFs showed prominent peaks with a distinct polarity at distinct latencies, reflecting major processing stages along the speech and language processing pathway. The amplitudes and latencies of these peaks served as the strength of neural processing at the corresponding stage. To investigate how neural auditory processing stages differ based on the linguistic content of the stimuli, the peak amplitudes and latencies were compared across passage types.

First, we identified the time windows for the main peaks associated with each predictor and their respective polarities. The time windows for each predictor were 1) Envelope: Early (20-130 ms), Late (70-180 ms); 2) Envelope onset: Early (20-170 ms), Late (70-240 ms); 3) Phoneme onset: Early (40-200 ms), Late (120-410 ms); 4) Phoneme surprisal: Early (40-200 ms), Late (110-470 ms); 5) Cohort entropy: Early (40-120 ms), Middle (140-350 ms), Late (260-600 ms); 6) Word onset: Early (40-200 ms), Middle (220-350 ms), Late (310-650 ms); 7) Unigram word surprisal: Early (40-300 ms), Late (310- 610 ms); 8) Contextual word surprisal: Early (40-300 ms), Late (310-610 ms). Early and middle peaks have positive current polarity while the late peak is a negative current polarity peak (respectively, directed out of, or into, the cortical surface).

A peak-picking algorithm was developed to pick the maximum peaks with the corresponding polarity within the given time window. The algorithm followed these steps: 1) TRFs were aggregated across the source ROIs by taking the absolute sum; 2) Peaks within the given time window were identified; 3) Selection of the maximum peak aligned with the given current polarity was achieved by checking the source current polarity relative to cortical surface in the transverse temporal region in the original source TRFs; 4) If none of the peaks satisfied the polarity constraint, the minimum of the average TRFs in the given time window was used as the peak amplitude, and the latency was set to NaN (not a number). A small number of peaks (<1.5 %) were further manually adjusted where appropriate.

Statistical analysis

Statistical analysis was performed in R[86] version 4.0 and Eelbrain. The significance level was set at $\alpha = 0.05$.

Significance of each speech feature over and beyond other features was evaluated by comparing full and reduced models. The full models for modulated noise and non-words included: gammatone envelope, envelope onset, phoneme onset, phoneme surprisal, cohort entropy and word onset. Additionally, the scrambled passages also included unigram surprisal; the narrative passages included both unigram surprisal and context-based word surprisal. Each reduced model included all the features of the full model, except excluding the single predictor under investigation. The proportion of explained variance between the full and reduced model at each current source dipole were tested using mass-univariate one-tailed paired sample t-test with threshold-free cluster enhancement (TFCE)[87] with a null distribution based on 10,000 permutations of model labels.

Hemispheric lateralization of each feature was performed to examine the lateralization of each speech feature processing. The explained variance maps for each feature were transformed to a common space by first morphing to a symmetric brain template 'fsaverage_sym' and consecutively morphed the right hemisphere to the left hemisphere. The explained variance between left and right hemispheres were tested using mass-univariate two-tailed paired sample t-test with TFCE.

TRF amplitude comparison were performed using repeated samples ANOVA and using post hoc paired samples t-test with corrections for multiple comparisons using false discovery rate (fdr) corrections. To ensure unbiased TRF comparison across passage types, TRFs were generated from a similar number of predictors across passage types.

Statistical summary tables are reported in supplementary materials.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health grant R01-DC019394 and National Science Foundation grant SMA 1734892. We thank Vinoj Jayasundara for his assistance in context-based linguistic feature generation and Ciaran Stone for excellent technical assistance in MEG data recordings.

Author Contributions

I.M.D.K. and J.Z.S. conceived and designed the experiments; I.M.D.K. collected data; I.M.D.K., C.B., S.B., P.R., and J.Z.S. designed the analysis, analyzed data and interpreted data; I.M.D.K. drafted manuscript; I.M.D.K., C.B., S.B., P.R., and J.Z.S. edited and revised manuscript.

Competing Interests

The authors declare no competing interests

Additional Information

Code and dataset supporting the findings of this paper will be shared once the paper is accepted. Correspondence and requests for additional materials should be addressed to J.Z.S or I.M.D.K

REFERENCES

1. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Curr. Biol.* 28, 3976-3983.e5 (2018).
2. Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T. & Brodbeck, C. Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. *J. Neurosci.* 41, 10316–10329 (2021).
3. Keshishian, M. *et al.* Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nat. Hum. Behav.* 7, 740–753 (2023).
4. Deniz, F., Tseng, C., Wehbe, L., Dupré La Tour, T. & Gallant, J. L. Semantic Representations during Language Comprehension Are Affected by Context. *J. Neurosci.* 43, 3144–3158 (2023).

5. Xu, J., Kemeny, S., Park, G., Frattali, C. & Braun, A. Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage* 25, 1002–1015 (2005).
6. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *J. Neurosci.* 31, 2906–2915 (2011).
7. Alday, P. M. M/EEG analysis of naturalistic stories: a review from speech to language processing. *Lang. Cogn. Neurosci.* 34, 457–473 (2019).
8. Ding, N. & Simon, J. Z. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89 (2012).
9. Brodbeck, C., Jiao, A., Hong, L. E. & Simon, J. Z. Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLOS Biol.* 18, e3000883 (2020).
10. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl. Acad. Sci.* 119, e2201968119 (2022).
11. Oord, A. van den *et al.* WaveNet: A Generative Model for Raw Audio. in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)* 125 (2016).
12. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* 18, 903–911 (2015).
13. Hamilton, L. S., Edwards, E. & Chang, E. F. A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Curr. Biol.* 28, 1860-1871.e4 (2018).

14. Oganian, Y. & Chang, E. F. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Sci. Adv.* 5, eaay6279 (2019).

15. Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D. & Schalk, G. The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE* 8, e53398 (2013).

16. Steinschneider, M., Nourski, K. V. & Fishman, Y. I. Representation of speech in human auditory cortex: Is it special? *Hear. Res.* 305, 57–73 (2013).

17. Karunathilake, I. M. D., Kulasingham, J. P. & Simon, J. Z. Neural tracking measures of speech intelligibility: Manipulating intelligibility while keeping acoustics unchanged. *Proc. Natl. Acad. Sci.* 120, e2309166120 (2023).

18. Gillis, M., Vanthornhout, J. & Francart, T. Heard or Understood? Neural Tracking of Language Features in a Comprehensible Story, an Incomprehensible Story and a Word List. *eneuro* 10, ENEURO.0075-23.2023 (2023).

19. Tezcan, F., Weissbart, H. & Martin, A. E. A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. *eLife* 12, e82386 (2023).

20. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* 7, 430–441 (2023).

21. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* 118, e2105646118 (2021).

22. Peelle, J. E. The hemispheric lateralization of speech processing depends on what “speech” is: a hierarchical perspective. *Front. Hum. Neurosci.* 6, (2012).

23. Binder, J. R. Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cereb. Cortex* 10, 512–528 (2000).

24. Aiken, S. J. & Picton, T. W. Human Cortical Responses to the Speech Envelope. *Ear Hear.* 29, 139–157 (2008).
25. Ross, E. D., Thompson, R. D. & Yenkosky, J. Lateralization of Affective Prosody in Brain and the Callosal Integration of Hemispheric Language Functions. *Brain Lang.* 56, 27–54 (1997).
26. Poeppel, D. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun.* 41, 245–255 (2003).
27. Hickok, G. & Poeppel, D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99 (2004).
28. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402 (2007).
29. Gow, D. W. The cortical organization of lexical knowledge: A dual lexicon model of spoken language processing. *Brain Lang.* 121, 273–288 (2012).
30. Price, C. J. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage* 62, 816–847 (2012).
31. Fedorenko, E., Nieto-Castañon, A. & Kanwisher, N. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* 50, 499–513 (2012).
32. Bradshaw, A. R., Thompson, P. A., Wilson, A. C., Bishop, D. V. M. & Woodhead, Z. V. J. Measuring language lateralisation with different language tasks: a systematic review. *PeerJ* 5, e3929 (2017).
33. Boemio, A., Fromm, S., Braun, A. & Poeppel, D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci.* 8, 389–395 (2005).

34. Peelle, J. E., Gross, J. & Davis, M. H. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cereb. Cortex* 23, 1378–1387 (2013).
35. Mai, G., Minett, J. W. & Wang, W. S.-Y. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *NeuroImage* 133, 516–528 (2016).
36. Bozic, M., Tyler, L. K., Ives, D. T., Randall, B. & Marslen-Wilson, W. D. Bihemispheric foundations for human speech comprehension. *Proc. Natl. Acad. Sci.* 107, 17439–17444 (2010).
37. Overath, T. & Paik, J. H. From acoustic to linguistic analysis of temporal speech structure: Acousto-linguistic transformation during speech perception using speech quilts. *NeuroImage* 235, 117887 (2021).
38. Rauschecker, J. P. & Scott, S. K. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724 (2009).
39. Nourski, K. V. *et al.* Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. *Hear. Res.* 371, 53–65 (2019).
40. Peelle, J. E. Hierarchical processing for speech in human auditory cortex and beyond. *Front. Hum. Neurosci.* (2010) doi:10.3389/fnhum.2010.00051.
41. Ahissar, E. *et al.* Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci.* 98, 13367–13372 (2001).
42. Peelle, J. E., Troiani, V., Wingfield, A. & Grossman, M. Neural Processing during Older Adults' Comprehension of Spoken Sentences: Age Differences in Resource Allocation and Connectivity. *Cereb. Cortex* 20, 773–782 (2010).

43. Ding, N. & Simon, J. Z. Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.* 8, (2014).
44. Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z. & Francart, T. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *J. Assoc. Res. Otolaryngol.* 19, 181–191 (2018).
45. Ding, N. & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.* 109, 11854–11859 (2012).
46. Fiedler, L., Wöstmann, M., Herbst, S. K. & Obleser, J. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* 186, 33–42 (2019).
47. O'Sullivan, J. A. *et al.* Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb. Cortex* 25, 1697–1706 (2015).
48. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr. Biol.* 25, 2457–2465 (2015).
49. Gwilliams, L., King, J.-R., Marantz, A. & Poeppel, D. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nat. Commun.* 13, 6606 (2022).
50. Gwilliams, L. & Davis, M. H. Extracting Language Content from Speech Sounds: The Information Theoretic Approach. in *Speech Perception* (eds. Holt, L. L., Peelle, J. E., Coffin, A. B., Popper, A. N. & Fay, R. R.) vol. 74 113–139 (Springer International Publishing, Cham, 2022).

51. De Cheveigné, A. *Predictive Coding in the Auditory Brainstem*. http://biorxiv.org/lookup/doi/10.1101/2023.12.31.573202 (2024) doi:10.1101/2023.12.31.573202.

52. Kutas, M. & Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annu. Rev. Psychol.* 62, 621–647 (2011).

53. Lau, E. F., Phillips, C. & Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933 (2008).

54. Slaats, S., Weissbart, H., Schoffelen, J.-M., Meyer, A. S. & Martin, A. E. Delta-Band Neural Responses to Individual Words Are Modulated by Sentence Processing. *J. Neurosci.* 43, 4867–4883 (2023).

55. Lam, N. H. L., Schoffelen, J.-M., Uddén, J., Hultén, A. & Hagoort, P. Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. *NeuroImage* 142, 43–54 (2016).

56. Broderick, M. P., Zuk, N. J., Anderson, A. J. & Lalor, E. C. More than words: Neurophysiological correlates of semantic dissimilarity depend on comprehension of the speech narrative. *Eur. J. Neurosci.* 56, 5201–5214 (2022).

57. Deacon, D., Dynowska, A., Ritter, W. & Grose-Fifer, J. Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology* 41, 60–74 (2004).

58. Holcomb, P. J. Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology* 30, 47–61 (2007).

59. Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F. & Pernier, J. ERP Manifestations of Processing Printed Words at Different Psycholinguistic Levels: Time Course and Scalp Distribution. *J. Cogn. Neurosci.* 11, 235–260 (1999).

60. Payne, B. R., Lee, C.-L. & Federmeier, K. D. Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials: Effects of context on world-level N400. *Psychophysiology* 52, 1456–1469 (2015).

61. Dambacher, M., Kliegl, R., Hofmann, M. & Jacobs, A. M. Frequency and predictability effects on event-related potentials during reading. *Brain Res.* 1084, 89–103 (2006).

62. Huizeling, E., Arana, S., Hagoort, P. & Schoffelen, J.-M. Lexical Frequency and Sentence Context Influence the Brain's Response to Single Words. *Neurobiol. Lang.* 3, 149–179 (2022).

63. Pollan, M. *The Botany of Desire : A Plant's Eye View of the World*. (Random House, New York, 2001).

64. Herrmann, B. The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. *Int. J. Speech Technol.* (2023) doi:10.1007/s10772-023-10027-y.

65. Aoki, N. B., Cohn, M. & Zellou, G. The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Lett.* 2, 045204 (2022).

66. Gramfort, A. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7, (2013).

67. Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *NeuroImage* 86, 446–460 (2014).

68. Brodbeck, C., Das, P., jpkulasingham, Reddigari, S. & Brooks, T. L. Eelbrain 0.36. Zenodo https://doi.org/10.5281/zenodo.5152554 (2021).

69. Taulu, S. & Simola, J. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys. Med. Biol.* 51, 1759–1768 (2006).

70. Bell, A. J. & Sejnowski, T. J. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.* 7, 1129–1159 (1995).

71. Fischl, B. FreeSurfer. *NeuroImage* 62, 774–781 (2012).

72. Dale, A. M. & Sereno, M. I. Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *J. Cogn. Neurosci.* 5, 162–176 (1993).

73. Hämäläinen, M. S. & Ilmoniemi, R. J. Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.* 32, 35–42 (1994).

74. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980 (2006).

75. Heeris, J. Gammatone Filterbank Toolkit. (2018).

76. Fishbach, A., Nelken, I. & Yeshurun, Y. Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients. *J. Neurophysiol.* 85, 2303–2323 (2001).

77. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. in *Interspeech 2017* 498–502 (ISCA, 2017). doi:10.21437/Interspeech.2017-1386.

78. Boersma, P. & Weenink, D. Praat: doing phonetics by computer. (2021).

79. Speer, R., Chin, J., Lin, A., Jewett, S. & Nathan, L. LuminosoInsight/wordfreq: v2.2. Zenodo https://doi.org/10.5281/ZENODO.1443582 (2018).

80. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

81. Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, Online, 2020). doi:10.18653/v1/2020.emnlp-demos.6.

82. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016). doi:10.18653/v1/P16-1162.

83. Lalor, E. C., Power, A. J., Reilly, R. B. & Foxe, J. J. Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *J. Neurophysiol.* 102, 349–359 (2009).

84. David, S. V., Mesgarani, N. & Shamma, S. A. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Netw. Comput. Neural Syst.* 18, 191–212 (2007).

85. Gaston, P., Brodbeck, C., Phillips, C. & Lau, E. Auditory Word Comprehension Is Less Incremental in Isolated Words. *Neurobiol. Lang.* 4, 29–52 (2023).

86. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2020).

87. Smith, S. & Nichols, T. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44, 83–98 (2009).

88. Ding, N. *et al.* Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* 81, 181–187 (2017).

Supplementary Materials

Table S1. Summary statistics for the model prediction comparisons. $t_{\max}$ and corresponding *p* values are reported. Second column summarizes the contribution of each feature to the model's predictive power. Third column summarizes the lateralization results.

	Contribution to model prediction (Full vs Reduced)				Lateralization (Left vs Right)			
	Modulated Noise	Non-words	Scrambled words	Narrative	Modulated Noise	Non-words	Scrambled words	Narrative
Gammatone Envelope	6.92 **(<0.001)**	11.9 **(<0.001)**	10.97 **(<0.001)**	10.47 **(<0.001)**	-4.56 **(<0.001)**	4.33 (0.06)	-4.5 **(0.01)**	-5.04 **(<0.001)**
Envelope onset	5.79 **(<0.001)**	9.37 **(<0.001)**	10.68 **(<0.001)**	9.9 **(<0.001)**	-3.06 (0.08)	-4.17 **(0.04)**	-5.4 **(<0.001)**	-4.36 **(0.02)**
Phoneme onset	3.3 (0.07)	7.25 **(<0.001)**	6.43 **(<0.001)**	5.08 **(<0.001)**	0 (1)	3.58 (0.08)	2.83 (0.23)	-4.57 **(0.005)**
Phoneme surprisal	1.51 (1.0)	6.83 **(<0.001)**	8.6 **(<0.001)**	6.99 **(<0.001)**	0 (1)	2.72 (0.29)	3.74 (0.05)	-2.01 (0.82)
Cohort Entropy	1.82 (0.99)	6.9 **(<0.001)**	6.17 **(<0.001)**	5.75 **(<0.001)**	0 (1)	3.92 (0.06)	4.91 **(<0.001)**	2.38 (0.63)
Word onset	2.46 (0.91)	5.6 **(<0.001)**	7.13 **(<0.001)**	5.28 **(<0.001)**	0 (1)	2.54 (0.48)	4.21 **(0.003)**	3.21 **(0.02)**
Unigram word surprisal			6.67 **(<0.001)**	6.48 **(<0.001)**			5.07 **(0.002)**	3.23 **(0.03)**
Contextual word surprisal				5.48 **(<0.001)**				3.30 **(0.02)**

Table S2. Summary Statistics for envelope TRF peak amplitude comparisons. *P*-values are corrected for multiple comparisons using false discovery rate (FDR). LH and RH represent left and right hemispheres respectively.

		Envelope - Early			Envelope - Late		
		Noise	Non-words	Scrambled	Noise	Non-words	Scrambled
LH	Narrative	t_{29}=4.4, *p*<**0.001**	t_{29}=-4.7, *p*<**0.001**	t_{29}=-4.7, *p*<**0.001**	t_{29}=7.1, *p*<**0.001**	t_{29}=-2.7, *p*=**0.015**	t_{29}=-0.3, *p* = 0.77
	Scrambled	t_{29}=6.2, *p*<**0.001**	t_{29}=-0.6, *p*=0.58		t_{29}=5.9, *p*<**0.001**	t_{29}=-3.4, *p*=**0.003**	
	Non-words	t_{29}=5.9, *p*<**0.001**			t_{29}=6.9, *p*<**0.001**		
RH	Narrative	t_{29}=2.4, *p*=**0.04**	t_{29}=-1.3, *p*=0.24	t_{29}=-1.7, *p*=0.14	t_{29}=5.2, *p*<**0.001**	t_{29}=-1.9, *p*=0.09	t_{29}=-1.9, *p*=0.09
	Scrambled	t_{29}=2.8, *p*=**0.04**	t_{29}= 1.0, *p*=0.33		t_{29}=6.3, *p*<**0.001**	t_{29}=-0.1, *p*=0.89	
	Non-words	t_{29}=2.5, *p*=**0.04**			t_{29}=7.2, *p*<**0.001**		

Table S3. Summary Statistics for envelope onset TRF peak amplitude comparisons. Other details as in Table S2.

		Envelope Onset- Early			Envelope Onset- Late		
		Noise	Non-words	Scrambled	Noise	Non-words	Scrambled
LH	Narrative	t_{29}=6.2, *p*<**0.001**	t_{29}=0.3, *p*=0.96	t_{29}=0.1, *p*=0.96	t_{29}=3.5, *p*=**001**	t_{29}=0.7, *p*=0.55	t_{29}=1.4, *p* = 0.25
	Scrambled	t_{29}=6.2, *p*<**0.001**	t_{29}=0.2, *p*=0.96		t_{29}=2.9, *p*=**0.02**	t_{29}=-0.6, *p*=0.55	
	Non-words	t_{29}=6.4, *p*<**0.001**			t_{29}=2.8, *p*=**0.02**		
RH	Narrative	t_{29}=6.7, *p*<**0.001**	t_{29}=1.7, *p*=0.13	t_{29}=-0.3, *p*=0.78	t_{29}=4.5, *p*<**0.001**	t_{29}=-1.8, *p*=0.12	t_{29}=-1.0, *p*=0.39
	Scrambled	t_{29}=5.6, *p*<**0.001**	t_{29}= 1.7, *p*=0.13		t_{29}=4.7, *p*<**0.001**	t_{29}=-0.5, *p*=0.63	
	Non-words	t_{29}=6.3, *p*<**0.001**			t_{29}=5.4, *p*<**0.001**		

Table S4. Summary Statistics for phoneme onset TRF peak amplitude comparisons. Other details as in Table S2.

		Phoneme onset - Early		Phoneme Onset - Late	
		Non-words	Scrambled	Non-words	Scrambled
LH	Narrative	t_{29}=-2.0, *p=0.08*	t_{29}=1.8, *p=0.08*	t_{29}=-1.8, *p=0.11*	t_{29}=0.7, *p=0.50*
	Scrambled	t_{29}=-3.9, *p=**0.002***		t_{29}=-2.3, *p=0.09*	
RH	Narrative	t_{29}=-0.6, *p=0.57*	t_{29}=0.8, *p=0.57*	t_{29}=-0.6, *p=0.67*	t_{29}=0.4, *p=0.67*
	Scrambled	t_{29}=-1.4, *p=0.52*		t_{29}=-1.1, *p=0.67*	

Table S5. Summary Statistics for phoneme surprisal TRF peak amplitude comparisons. Other details as in Table S2.

		Phoneme Surprisal - Early		Phoneme Surprisal - Late	
		Non-words	Scrambled	Non-words	Scrambled
LH	Narrative	t_{29}=0.8, p=0.46	t_{29}=-2.4, p=**0.03**	t_{29}=1.3, p=0.20	t_{29}=-2.4, p=**0.03**
	Scrambled	t_{29}=2.6, p=**0.03**		t_{29}=3.3, p=**0.008**	
RH	Narrative	t_{29}=1.3, p=0.33	t_{29}=-0.1, p=0.89	t_{29}=2.0, p=0.17	t_{29}=1.0, p=0.34
	Scrambled	t_{29}=1.9, p=0.22		t_{29}=1.0, p=0.34	

Table S6. Summary Statistics for cohort entropy TRF peak amplitude comparisons. Other details as in Table S2.

		Cohort Entropy - Early		Cohort Entropy - Middle		Cohort Entropy - Late	
		Non-words	Scrambled	Non-words	Scrambled	Non-words	Scrambled
LH	Narrative	t_{29}=4.1, p<**0.001**	t_{29}=2.5, p=**0.02**	t_{29}=-7.0, p<**0.001**	t_{29}=-0.3, p=0.75	t_{29}=-2.1, p=0.06	t_{29}=-3.3, p=**0.009**
	Scrambled	t_{29}=2.9, p=**0.01**		t_{29}=-7.3, p<**0.001**		t_{29}=1.9, p=0.06	
RH	Narrative	t_{29}=2.6, p=**0.04**	t_{29}=1.4, p=0.27	t_{29}=-5.6, p<**0.001**	t_{29}=0.81, p=0.42	t_{29}=-2.1, p=0.13	t_{29}=-1.5, p=0.20
	Scrambled	t_{29}=0.9, p=0.37		t_{29}=-5.2, p<**0.001**		t_{29}=-0.4, p=0.68	

Table S7. Summary Statistics for word onset TRF peak amplitude comparisons. Other details as in Table S2.

		Word Onset - Early		Word Onset - Middle		Word Onset - Late	
		Non-words	Scrambled	Non-words	Scrambled	Non-words	Scrambled
LH	Narrative	t_{29}=5.1, p<**0.001**	t_{29}=1.8, p=0.09	t_{29}=-5.2, p<**0.001**	t_{29}=-1.9, p=0.07	t_{29}=0.2, p=0.86	t_{29}=-3.6, p=**0.003**
	Scrambled	t_{29}=5.1, p<**0.001**		t_{29}=-4.6, p<**0.001**		t_{29}=3.3, p=**0.003**	
RH	Narrative	t_{29}=2.1, p=0.12	t_{29}=1.8, p=0.12	t_{29}=-6.0, p<**0.001**	t_{29}=-0.6, p=0.57	t_{29}=-0.2, p=0.82	t_{29}=-2.3, p=0.09
	Scrambled	t_{29}=1.1, p=0.30		t_{29}=-5.3, p<**0.001**		t_{29}=1.7, p=0.16	

Table S8. Summary statistics for combined early and middle peak amplitude comparisons for cohort entropy and word onsets. Other details as in Table S2.

		Cohort Entropy		Word Onset	
		Non-words	Scrambled	Non-words	Scrambled
LH	Narrative	t_{29}=1.2, p=0.25	t_{29}=2.3, p=0.08	t_{29}=-2.5, p=0.06	t_{29}=1.7, p=0.12
	Scrambled	t_{29}=-1.5, p=0.20		t_{29}=1.6, p=0.12	

Table S9. Summary Statistics for unigram surprisal TRF peak amplitude comparisons. Other details as in Table S2.

		Unigram Early	Unigram - Late
		Scrambled	Scrambled
LH	Narrative	t_{29}=1.4, p=0.16	t_{29}=-4.2, p<**0.001**
RH	Narrative	t_{29}=1.8, p=0.08	t_{29}=-2.1, p=**0.04**

Table S10. Summary Statistics for contextual vs unigram word surprisal TRF peak amplitude comparisons. Other details as in Table S2.

		Early	Late
		Unigram Surprisal	Unigram Surprisal
LH	Contextual word surprisal	t_{29}=5.2, p<**0.001**	t_{29}=5.0, p<**0.001**
RH	Contextual word surprisal	t_{29}=5.2, p<**0.001**	t_{29}=3.5, p=**0.001**

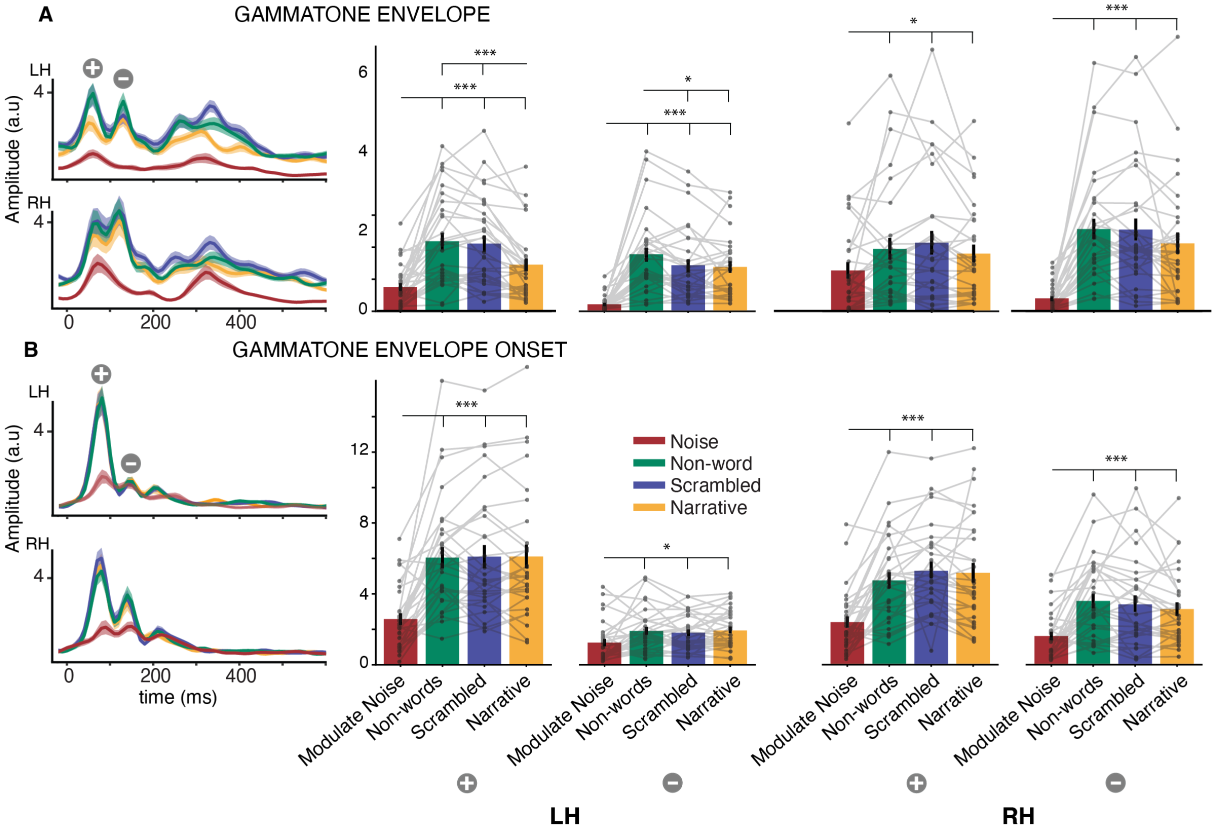

Figure S2. Neural Responses to acoustic features (A) Gammatone envelope and (B) Gammatone envelope onset in left (LH) and right (RH) hemispheres. This figure expands on the data shown in Fig. 3. The TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. Right panel bar plots compare the peak amplitudes across passage types. LH and RH denotes left and right hemisphere respectively. Both early and late responses are stronger for speech compared to non-speech. Differences between the speech passages were found only for the envelope responses and in the left hemisphere. *p<0.05, **p<0.01, ***p<0.001

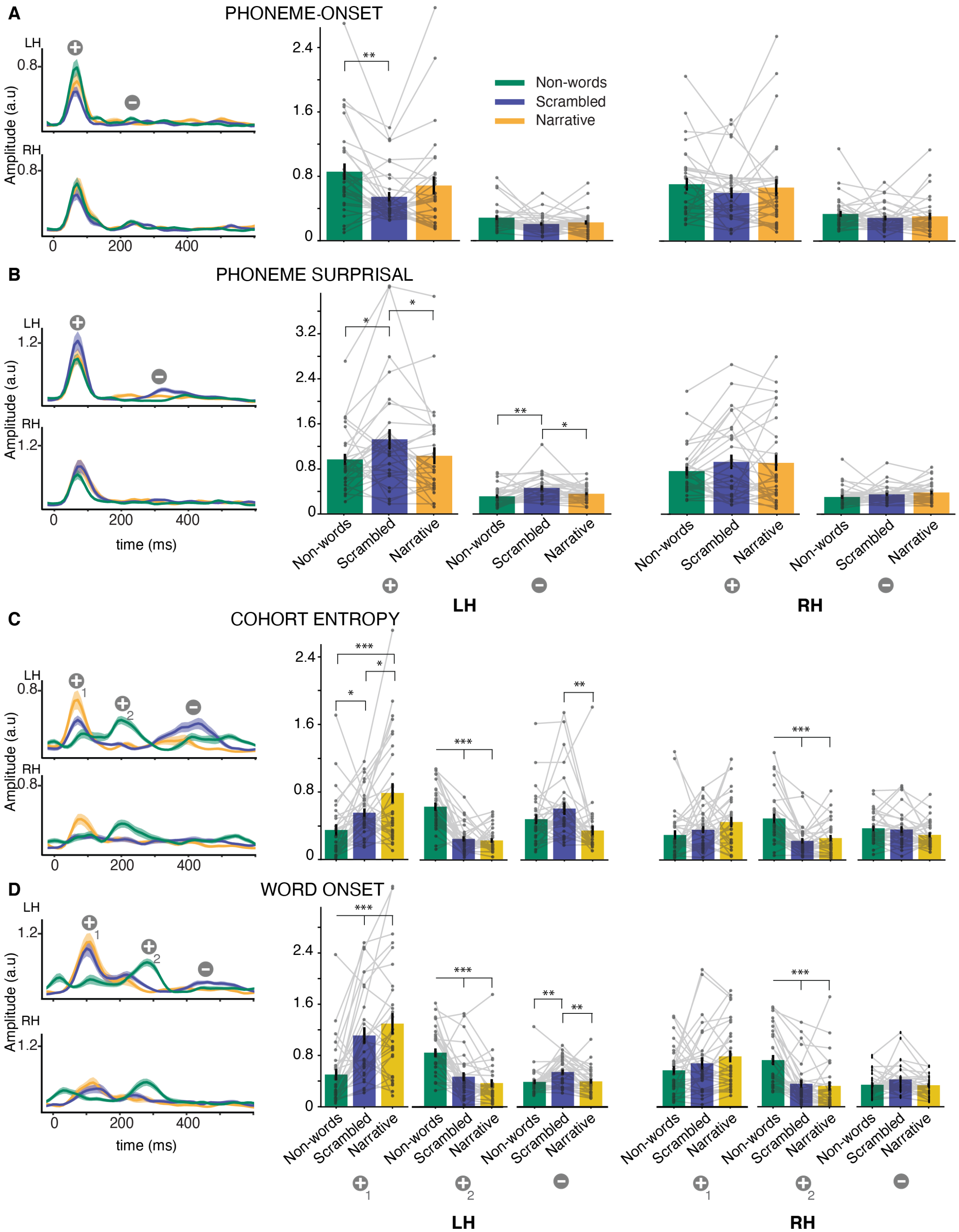

Figure S3. Neural responses to sub-lexical and word onset speech features (A). Phoneme onset, (B). word onset, (C). phoneme surprisal, and (D). cohort entropy (TRF magnitude plots and TRF peak bar plots as in Figure S2). This figure expands on the data

shown in Fig. 4. TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. For both word onset and cohort entropy responses, non-words showed a robust positive polarity peak between early and late peaks. These early, middle, and late peaks are indicated by $\oplus_1$, $\oplus_2$, and ⊖ respectively. The right column bar plots compare the peak amplitudes across passage types. LH and RH denotes left and right hemisphere respectively. Overall, the early responses were differently modulated by the linguistic content. The middle peak was stronger for non-words, while the late peak was stronger for scrambled passages. No differences, except the strong middle responses for non-words were found in the right hemisphere.

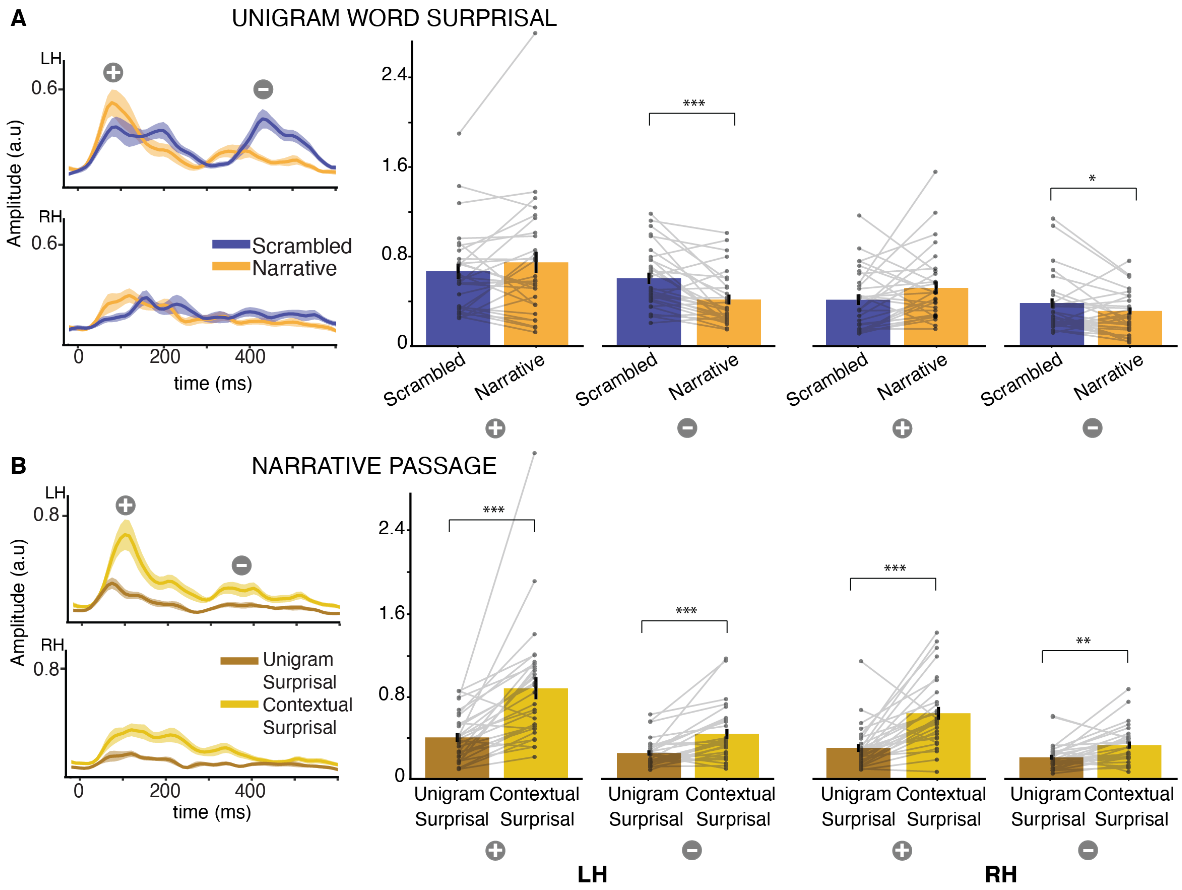

Figure S4. Neural responses to lexico-semantic features (A) Unigram word surprisal and (B) Unigram vs contextual word surprisal for the narrative passage (TRF magnitude plots and TRF peak bar plots as in Figure S2). This figure expands on the data shown in Fig. 5. The late peak in unigram surprisal responses is stronger for scrambled words compared to narrative passages. Contextual word surprisal is stronger compared to unigram (local) word surprisal. LH and RH denotes left and right hemisphere respectively. TRFs exhibit an early positive and a late negative polarity peak indicated by ⊕ and ⊖ respectively. The late unigram surprisal responses (N400-like) are stronger for scrambled passages compared to narrative passage. Contextual word surprisal responses are stronger compared to unigram surprisal responses. Note that the peak amplitudes for unigram surprisal in (A) and (B) are different, as the TRF model in (A) does not include contextual surprisal.

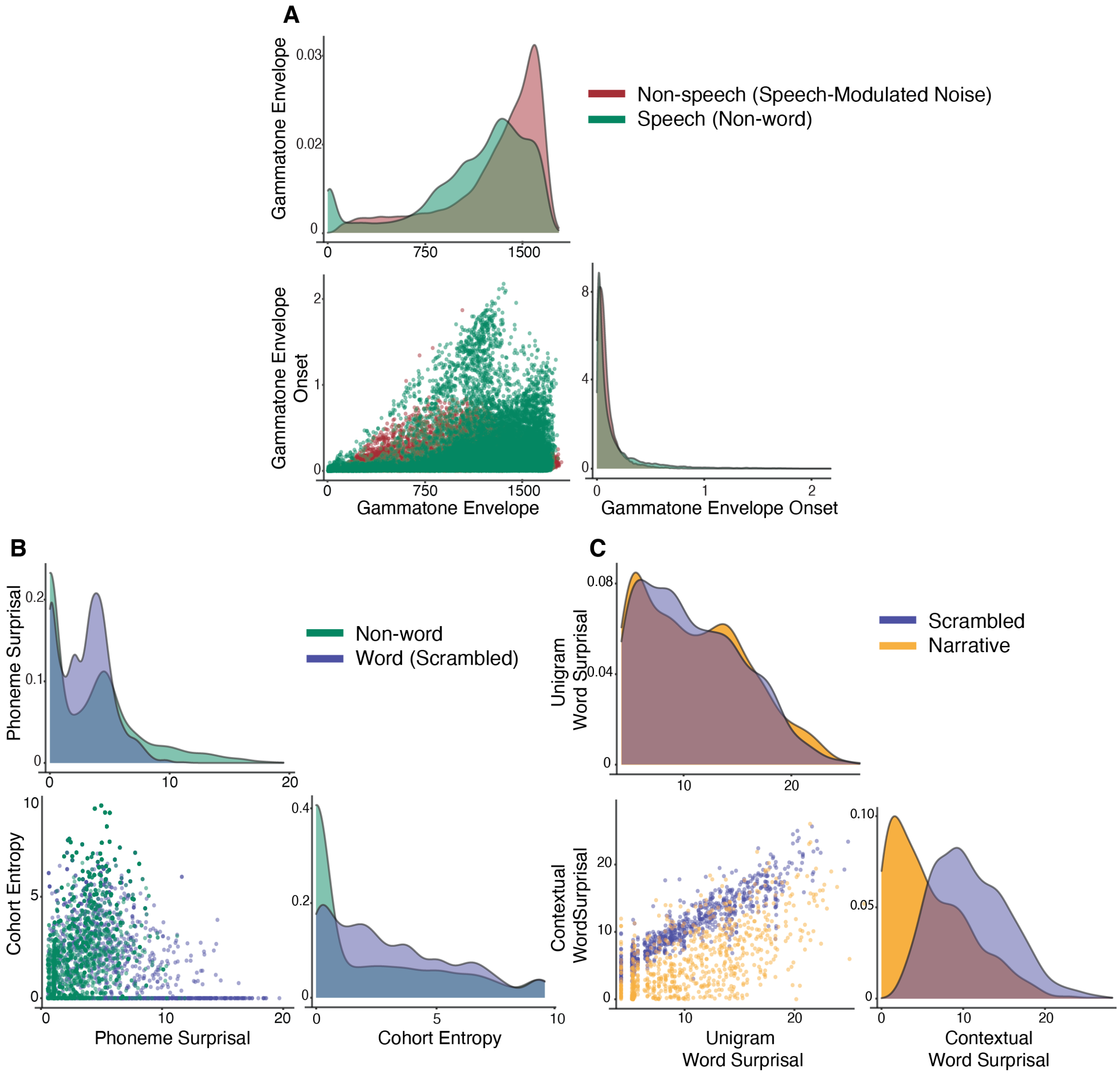

Figure S5. Comparison of predictor variables between passage types. (a). Acoustic feature comparisons between non-speech and (non-word) speech passage: they share similarities in the distribution of envelope onset predictor values, but not of envelope predictors. (b). Phoneme surprisal and cohort entropy comparison between non-words and meaningful words (scrambled passage): both predictor distributions depend strongly on the stimulus type. (c). Unigram and contextual word surprisal comparisons between scrambled and narrative passages: the two unigram word surprisal distributions are nearly identical, by design, but the contextual word surprisal distributions diverge strongly (the narrative case is strongly biased toward low surprisal, as expected; additionally, in the scrambled

word case, both forms of surprisal are highly correlated, collapsing into a narrow diagonal distribution. In each panel, the top and right plots show frequency histograms that present the distribution of each feature, where the y-axis represents the bin density of points, scaled to integrate to one.; the bottom left scatterplot shows a visualization of the correlation between the two predictor variables.

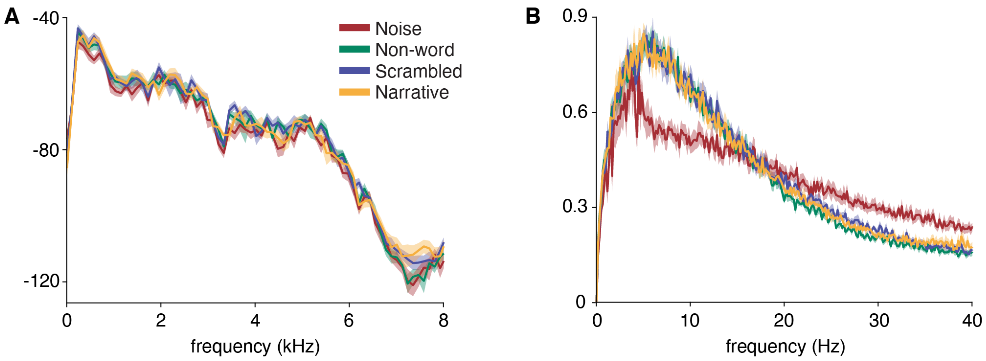

Figure S6. Comparison of Stimulus Acoustic Properties. (A). Periodograms and (B). Modulation spectrum obtained using the methods of [88]. Even though the spectral characteristics are similar between the stimulus types, the slow temporal modulation is different between speech and non-speech. There is no visible difference in acoustic properties between the speech passages. Periodograms and modulation spectra were computed for 10 chunks of 6 seconds each, per each passage type and then mean ± standard error is plotted to illustrate data.