Reviewed Preprint
v1 • September 2, 2024
Not revised

Reviewed Preprint
v2 • April 24, 2025
Revised by authors

Reviewed Preprint
v3 • January 23, 2026
Revised by authors

For correspondence:
jixingli@cityu.edu.hk
huangzw86@126.com

Author notes: See page 22

Reviewing editor: Nai Ding, Zhejiang University, China

Multi-talker speech comprehension at different temporal scales in listeners with normal and impaired hearing

Jixing Li[1], Qixuan Wang[2], Qian Zhou[3], Lu Yang[3], Yutong Shen[4], Shujian Huang[4], Shaonan Wang[5], Liina Pylkkänen[6], Zhiwu Huang[7]

[1]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, Hong Kong • [2]Department of Facial Plastic and Reconstructive Surgery, Eye & ENT Hospital, Fudan University, Shanghai, China; ENT institute, Eye & ENT Hospital, Fudan University, Shanghai, China • [3]Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China • [4]Department of Computer Science and Technology, Nanjing University, Nanjing, China • [5]Institute of Automation, Chinese Academy of Sciences, Beijing, China • [6]Department of Linguistics, Department of Psychology, New York University, New York, United States • [7]Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine; College of Health Science and Technology, Shanghai Jiao Tong University School of Medicine, Shanghai, China

eLife Assessment

This **valuable** study a computational language model, i.e., HM-LSTM, to quantify the neural encoding of hierarchical linguistic information in speech, and addresses how hearing impairment affects neural encoding of speech. Overall the evidence for the findings is **solid**, although the evidence for different speech processing stages could be strengthened by a more rigorous temporal response function (TRF) analysis. The study is of potential interest to audiologists and researchers who are interested in the neural encoding of speech.

https://doi.org/10.7554/eLife.100056.3.sa3

Abstract

Comprehending speech requires deciphering a range of linguistic representations, from phonemes to narratives. Prior research suggests that in single-talker scenarios, the neural encoding of linguistic units follows a hierarchy of increasing temporal receptive windows. Shorter temporal units like phonemes and syllables are encoded by lower-level sensory brain regions, whereas longer units such as sentences and paragraphs are processed by higher-level perceptual and cognitive areas. However, the brain's representation of these linguistic units under challenging listening conditions, such as a cocktail party situation, remains unclear. In this study, we recorded electroencephalogram (EEG) responses from both normal-hearing and hearing-impaired participants as they listened to individual and dual speakers narrating different parts of a story. The inclusion of hearing-impaired listeners allowed us to examine how hierarchically organized linguistic units in competing speech streams affect comprehension abilities. We leveraged a hierarchical language model to extract linguistic information at multiple levels—phoneme, syllable, word, phrase, and sentence—and aligned these model activations with the EEG data. Our findings showed distinct neural responses to dual-speaker speech between the two groups. Specifically, compared to normal-hearing listeners, hearing-impaired listeners exhibited poorer model fits at the acoustic, phoneme, and syllable levels as well as the sentence levels, but not at the word and phrase levels. These results suggest that hearing-impaired listeners experience disruptions at both shorter and longer temporal scales, while their processing at medium temporal scales remains unaffected.

Introduction

Human speech encompasses elements at different levels, from phonemes to syllables, words, phrases, sentences and paragraphs. These elements manifest over distinct timescales: Phonemes occur over tens of milliseconds, and paragraphs span a few minutes. Understanding how these units at different time scales are encoded in the brain during speech comprehension remains a challenge. In the visual system, it has been well-established that there is a hierarchical organization such that neurons in the early visual areas have smaller receptive fields, while neurons in the higher-level visual areas receive inputs from lower-level neurons and have larger receptive fields (Hubel & Wiesel, 1962; 1965). This organizing principle is theorized to be mirrored in the auditory system, where a hierarchy of temporal receptive windows (TRW) extends from primary sensory regions to advanced perceptual and cognitive areas (Hasson et al., 2008; Honey et al., 2012; Lerner et al., 2011; Murray et al., 2014). Under this assumption, neurons in the lower-level sensory regions, such as the core auditory cortex, support rapid processing of the ever-changing auditory and phonemic information, whereas neurons in the higher cognitive regions, with their extended temporal receptive windows, process information at the sentence or discourse level.

Recent functional magnetic resonance imaging (fMRI) studies have shown some evidence that different levels of linguistic units are encoded at different cortical regions (Blank & Fedorenko, 2020; Chang et al., 2022; Hasson et al., 2008; Lerner et al., 2011; Schmitt et al., 2021). For example, Schmitt et al. (2021) used artificial neural networks to predict the next word in a story across five stacked time scales. By correlating model predictions with brain activity while listening to a story in an fMRI scanner, they discerned a hierarchical progression along the temporoparietal pathway. This pathway identifies the role of the bilateral primary auditory cortex in processing words over shorter durations and the involvement of the inferior parietal cortex in processing paragraph-length units over extended periods. Studies using electroencephalogram (EEG), magnetoencephalography (MEG) and electrocorticography (ECoG) have also revealed synchronous neural responses to different linguistics units at different time scales (e.g., Ding et al., 2015; Ding & Simon, 2012; Honey et al., 2012; Luo & Poeppel, 2007). Notably, Ding et al. (2015) showed that the MEG-derived cortical response spectrum concurrently tracked the timecourses of abstract linguistic structures at the word, phrase and sentence levels.

Although there is a growing consensus on the hierarchical encoding of linguistic units in the brain, the neural representations of these units in a multi-talker setting remain less explored. In the classic “cocktail party” situation in which multiple speakers talk simultaneously (Cherry, 1953), listeners must separate a speech signal from a cacophony of other sounds (McDermott, 2009). Studies have shown that listeners with normal hearing can selectively attend to a chosen speaker in the presence of two competing speakers (Brungart, 2001; Shinn-Cunningham, 2008), resulting in enhanced neural responses to the attended speech stream (Brodbeck et al., 2018; Ding & Simon, 2012; Mesgarani & Chang, 2012; O’Sullivan et al., 2015; Zion Golumbic et al., 2013). For example, Ding and Simon (2012) showed that neural responses were selectively phase-locked to the broadband envelope of the attended speech stream in the posterior auditory cortex. Furthermore, when the intensity of the attended and unattended speakers is separately varied, the neural representation of the attended speech stream adapts only to the intensity of the attended speaker. Zion Golumbic et al. (2013) further suggested that the neural representation appears to be more “selective” in higher perceptual and cognitive brain regions such that there is no detectable tracking of ignored speech.

This selective entrainment to attended speech has primarily focused on the low-level acoustic properties of the speech, while largely ignoring the higher-level linguistic units beyond the phonemic levels. Brodbeck et al. (2018) were the first study to simultaneously compare the neural responses to the acoustic envelopes as well as the phonemes and words in two competing speech streams. They found that although the acoustic envelopes of both the attended and unattended speech could be decoded from brain activity in the temporal cortex, only phonemes

and words of the attended speech showed significant responses. However, as their study only examined two linguistic units, a complete model of how the brain tracks linguistic units, from phonemes to sentences, in two competing speech streams is still lacking.

In this study, we investigate the neural underpinnings of processing diverse linguistic units in the context of competing speech streams among listeners with both normal and impaired hearing. We included hearing-impaired listeners to examine how hierarchically organized linguistic units in competing speech streams impact comprehension abilities. The experiment design consisted of a multi-talker condition and a single-talker condition. In the multi-talker condition, participants listened to mixed speech from female and male speakers narrating simultaneously. Before each trial, instructions on the screen indicated which speaker to focus on. In the single-talker condition, the speeches from male and female speakers were presented separately (refer to Figure 1A for the experimental procedure). We employed a hierarchical multiscale Long Short-Term Memory network (HM-LSTM; Chung et al., 2017) to dissect the linguistic information of the stimuli across phoneme, syllable, word, phrase, and sentence levels (detailed model architecture is described in the "Hierarchical multiscale LSTM model" section in Materials and Methods). We then performed ridge regressions using these linguistic units as regressors, time-locked to the offset of each sentence at nine latencies (see Figure 1B for the analysis pipeline). This model-brain alignment method has been commonly employed in the literature (e.g., Caucheteux & King, 2022; Goldstein et al., 2022; Schmitt et al., 2021; Schrimpf et al., 2021). By examining how model alignments with brain activity vary across different linguistic levels between listeners with normal and impaired hearing, our goal is to identify specific levels of linguistic processing that pose comprehension challenges for hearing-impaired individuals.

Results

Behavioral results

A total of 41 participants (21 females, mean age=24.1 years, SD=2.1 years) with extended high frequency (EHF) hearing loss and 33 participants (13 females, mean age=22.94 years, SD=2.36 years) with normal hearing were included in the study. EHF hearing loss refers to hearing loss at frequencies above 8 kHz. It is considered a major cause of hidden hearing loss, which cannot be detected by audiometry (Bharadwaj et al., 2019). Although the phonetic information required for speech perception in quiet conditions is below 6 kHz, ample evidence suggests that salient information in the higher-frequency regions may also affect speech intelligibility (e.g., Apoux & Bacon, 2004; Badri et al., 2011; Collins et al., 1981; Levy et al., 2015). Unlike age-related hearing loss, EHF hearing loss is commonly found in young adults who frequently use earbuds and headphones for prolonged periods and are exposed to high noise levels during recreational activities (Motlagh Zadeh et al., 2019). Consequently, this demographic is a suitable comparison group for their age-matched peers with normal hearing. EHF hearing loss was diagnosed using the pure tone audiometry (PTA) test, thresholded at frequencies greater than 8 kHz. As shown in Figure 2A, starting at 10 kHz, participants with EHF hearing loss exhibit significantly higher hearing thresholds (M=6.42 dB, SD=7 dB) compared to those with normal hearing (M=3.3 dB, SD=4.9 dB), as confirmed by an independent two-sample one-tailed t-test (t(72)=2, p=0.02).

Figure 2B illustrates the distribution of intelligibility ratings for both mixed and single-talker speech across the two listener groups. The average intelligibility ratings for both mixed and single-talker speech were significantly higher for normal-hearing participants (mixed: M=3.89, SD=0.83; single-talker: M=4.64, SD=0.56) compared to hearing-impaired participants (mixed: M=3.38, SD=1.04; single-talker: M=4.09, SD=0.89), as shown by independent two-sample one-tailed t-tests (mixed: t(72)=2.11, p=0.02; single: t(72)=2.81, p=0.003). Additionally, paired two-sample one-tailed t-tests indicated significantly higher intelligibility scores for single-talker speech compared to mixed speech within both listener groups (normal-hearing: t(32)=4.58, p<0.0001; hearing-impaired: t(40)=4.28, p=0.0001). These behavioral results confirm that mixed speech presents greater comprehension challenges for both groups, with hearing-impaired participants experiencing more difficulty in understanding both types of speech compared to those with normal hearing.

Figure 1. **Methods and behavioral results.**

A. Experimental procedure. The experimental task consisted of a multi-talker condition followed by a single-talker condition. In the multi-talker condition, the mixed speech was presented twice with the female and male speakers narrating simultaneously. Before each trial, instructions appeared in the center of the screen indicating which of the talkers to attend to (e.g., ''Attend female''). In the single-talker condition, the male and female speeches were presented sequentially. **B.** Analyses pipeline. Hidden-layer activity of the HM-LSTM model, which represents each level of linguistic units for each sentence, was extracted and aligned with EEG data, time-locked to the offset of each sentence at nine different latencies.

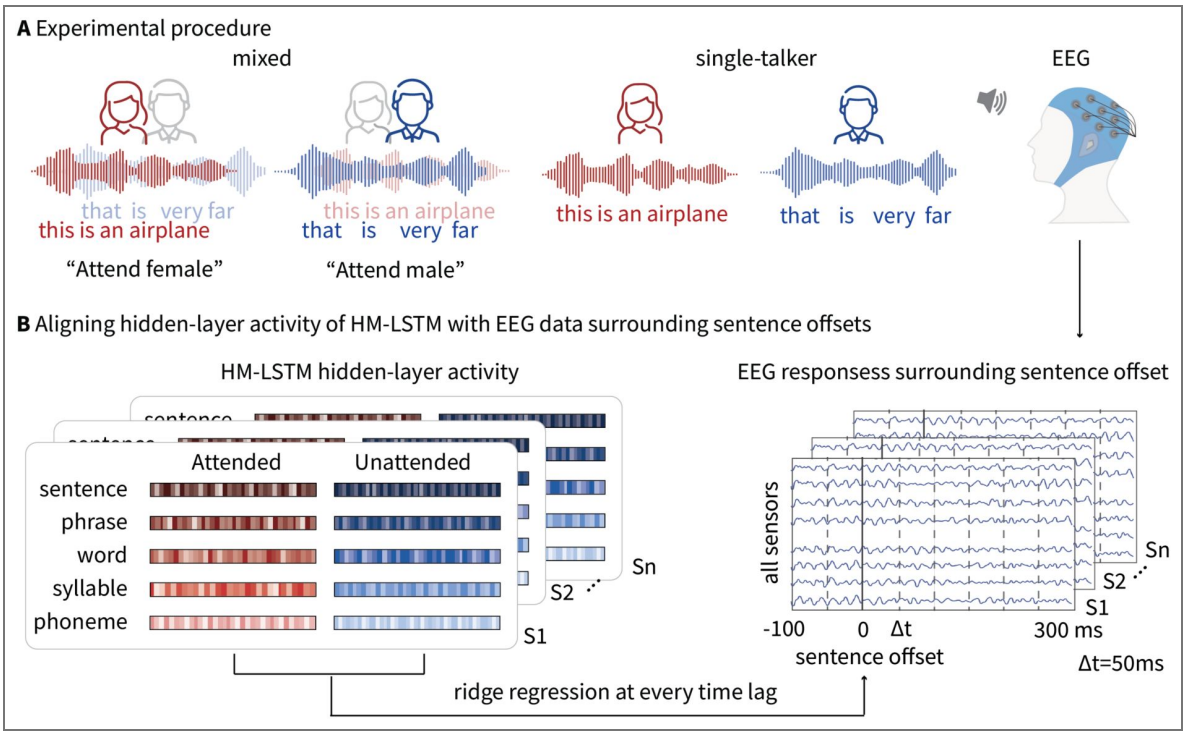
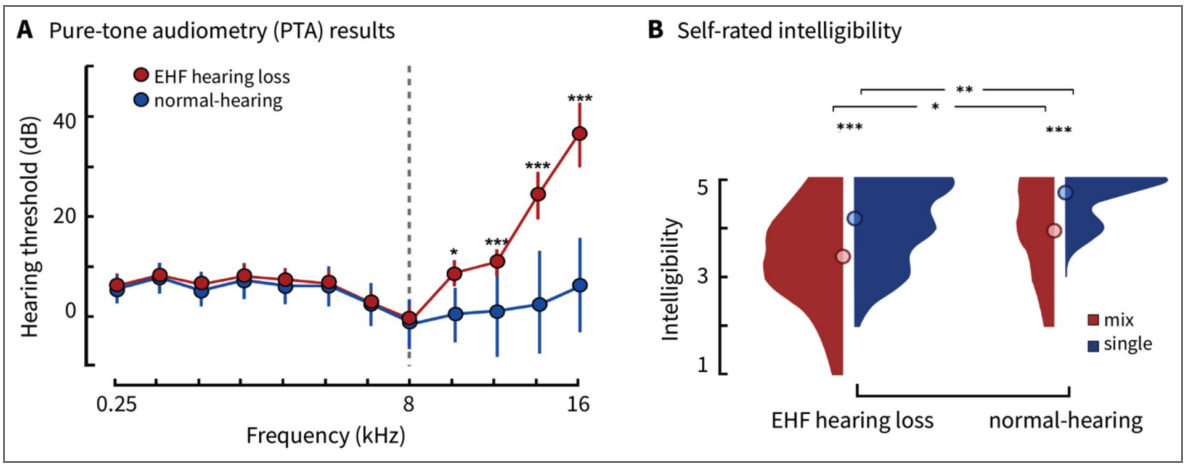


Figure 2. **Behavioral results.**

A. PTA results for participants with normal hearing and EHF hearing loss. Starting at 10 kHz, participants with EHF hearing loss have significantly higher hearing thresholds (M=6.42 dB, SD=7 dB) compared to normal-hearing participants (M=3.3 dB, SD=4.9 dB; t=2, p=0.02). **B.** Distribution of self-rated intelligibility scores for mixed and single-talker speech across the two listener groups. * indicates p < .05, ** indicates p <. 01 and *** indicates p < .001.



HM-LSTM model performance

To simultaneously estimate linguistic content at the phoneme, syllable, word, phrase and sentence levels, we adopted the HM-LSTM model originally developed by Chung et al. (2017). The original model consists of only two levels: the word level and the phrase level. We expanded its architecture to include five levels: phoneme, syllable, word, phrase, and sentence. Since our input consists of phoneme embeddings, we cannot directly apply their model, so we trained our model on the WenetSpeech corpus (Zhang et al., 2021), which provides phoneme-level transcripts. The inputs to the model were the vector representations of the phonemes in two sentences and the output of the model was the classification result of whether the second sentence follows the first sentence (see Figure 3A and the "Hierarchical multiscale LSTM model" section in "Materials and Methods" for the detailed model architecture). Unlike the Transformer-based language models that can predict only word- or sentence- and paragraph-level information, our HM-LSTM model can disentangle the impact of different informational content associated with phonemic and syllabic levels. After 130 epochs, the model achieved an accuracy of 0.87 on the training data and 0.83 on our speech stimuli, which comprise 570 sentence pairs. We subsequently extracted activity from the trained model's four hidden layers for each sentence in our stimuli to represent information at the phoneme, syllable, word, sentence, and paragraph levels, with the sentence-level information represented by the last unit of the fourth layer. We computed the correlations among the activations at the five levels of the HM-LSTM model (see "Correlations among LSTM model layers" section in "Materials and Methods" for the detailed analysis procedure). We did not observe very high correlations (all below 0.22) compared to prior model-brain alignment studies which report correlation coefficients above 0.5 for linguistic regressors (e.g., Gao et al., 2024; Sugimoto et al., 2024). In Chinese, a single syllable can also function as a word, potentially leading to higher correlations between regressors for syllables and words. However, we refrained from overinterpreting the results to suggest a higher correlation between syllable and sentence compared to syllable and word. A paired t-test of the syllable-word coefficients versus syllable-sentence coefficients across the 284 sentences revealed no significant difference ($t(28399)=-3.96$, $p=1$). This suggests that different layers of the model captured distinct patterns in the stimuli (see Figure 3B).

To verify that the hidden-layer activity indeed reflects information at the corresponding linguistic levels, we constructed a test dataset comprising 20 four-syllable sentences where all syllables contain the same vowels, such as "mā ma mà mǎ" (mother scolds horse), "shū shu shǔ shù" (uncle counts numbers). Table S1 in the Supplementary lists all four-syllable sentences with the same vowels. We hypothesized that the activity in the phoneme and syllable layer would be more similar than other layers for same-vowel sentences. The results confirmed our hypothesis: Hidden-layer activity for same-vowel sentences exhibited much more similar distributions at the phoneme and syllable levels compared to those at the word, phrase and sentence levels. Figure 3C displays the scatter plot of the model activity at the five linguistic levels for each of the 20 4-syllable sentences, post dimension reduction using multidimensional scaling (MDS). We used color-coding to represent the activity of five hidden layers after dimensionality reduction. Each dot on the plot corresponds to one test sentence. Only phonemes are labeled because each syllable in our test sentences contains the same vowels (see Table S1).The plot reveals that model representations at the phoneme and syllable levels are more dispersed for each sentence, while representations at the higher linguistic levels—word, phrase, and sentence—are more centralized. Additionally, similar phonemes tend to cluster together across the phoneme and syllable layers, indicating that the model captures a greater amount of information at these levels when the phonemes within the sentences are similar.

Regression results for single-talker speech versus attended speech

To examine the differences in neural activity between single- and dual-talker speech, we first compared the model fit of acoustic and linguistic features for the single-talker speech against the attended speech in the context of mixed speech across both listener groups. (see "Ridge regression at different time latencies" and "Spatiotemporal clustering analysis" in "Materials and Methods"

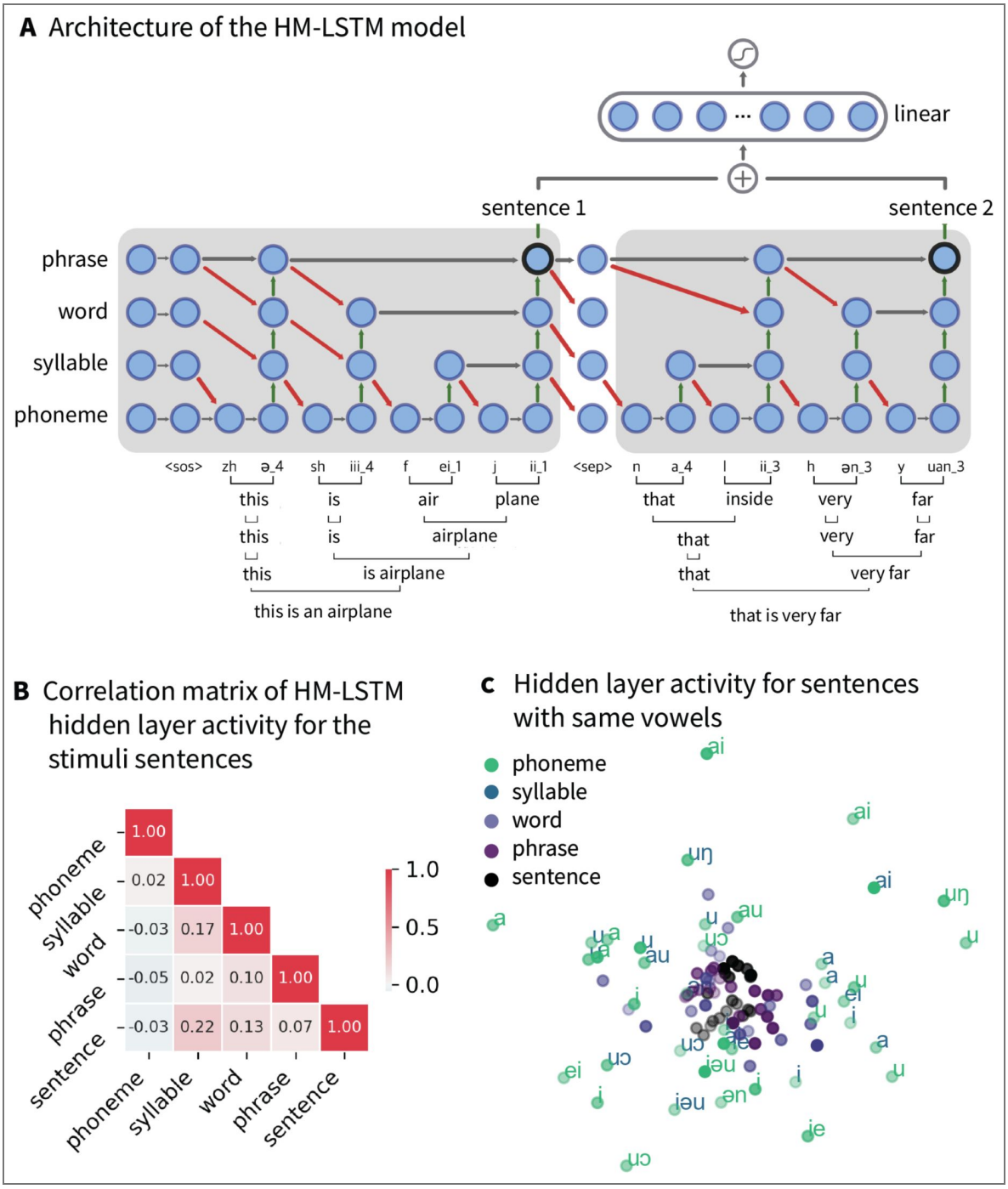


Figure 3. The HM-LSTM model architecture and hidden layer activity for the stimuli sentences and the 4-word Chinese sentences with same vowels.

A. The HM-LSTM model architecture. The model includes four hidden layers, corresponding to the phoneme-, syllable-, word- and phrase-level information. Sentence-level information was represented by the last unit of the 4th layer. The inputs to the model were the vector representations of the phonemes in two sentences and the output of the model was the classification result of whether the second sentence follows the first sentence. **B.** Correlation matrix for the HM-LSTM model's hidden layer activity for the sentences in the experimental stimuli. **C.** Scatter plot of hidden-layer activity at the five linguistic levels for each of the 20 4-syllable sentences after MDS.

for analysis details). We have also analyzed the epoched EEG data by decomposing it into different frequency bands (see "EEG Recording and Preprocessing" in "Materials and Methods"). We specifically examined the delta and theta bands, which are conventionally used in the literature for speech analysis. However, the results from these bands were very similar to those obtained using data from all frequency bands (see Supplementary Figures S2 and S3). Therefore, we opted to use the epoched EEG data from all frequency bands for our analyses. Figure 4 shows the sensors and time window where acoustic and linguistic features significantly better predicted EEG data in the left temporal region during single-talker speech as compared to the attended speech. It can be seen that for hearing-impaired participants, acoustic features showed a better model fit in single-talker settings as opposed to mixed speech conditions from −100 to 100 ms around sentence offsets (t=1.4, Cohen's d=1.5, p=0.002). However, no significant differences in the model fit between the single-talker and the attended speeches were observed for normal-hearing participants. Group comparisons revealed a significant difference in the model fit for the two conditions from −100 to 50 ms around sentence offsets (t=1.43, Cohen's d=1.28, p=0.011).

For the linguistic features, both the phoneme and syllable layers from the HM-LSTM model were more predictive of EEG data in single-talker speech compared to attended speech among hearing-impaired participants in the left temporal regions (phoneme: t=1.9, Cohen's d=0.49, p=0.004; syllable: t=1.9, Cohen's d=0.37, p=0.002). The significant effect occurred from approximately 0-100 ms for phonemes and 50-150 ms for syllables after sentence offsets. No significant differences in model fit were observed between the two conditions for participants with normal hearing. Comparisons between groups revealed significant differences in the contrast maps from 0-100 ms after sentence offsets for phonemes (t=2.39, Cohen's d=0.72, p=0.004) and from 50-150 ms after the sentence offsets for syllables (t=2.11, Cohen's d=0.78, p=0.001). The model fit to the EEG data for higher-level linguistic features—words, phrases, and sentences—does not show any significant differences between single-talker and attended speech across the two listener groups. This suggests that both normal-hearing and hearing-impaired participants are able to extract information at the word, phrase, and sentence levels from the attended speech in dual-speaker scenarios, similar to conditions involving only a single talker.

Regression results for single-talker versus unattended speech

We also compared the model fit for single-talker speech and the unattended speech under the mixed speech condition. As shown in Figure 5, the acoustic features showed a better model fit in single-talker settings as opposed to mixed speech conditions from −100 to 50 ms around sentence offsets for hearing-impaired listeners (t=2.05, Cohen's d=1.1, p=<0.001) and from −100 to 50 ms for normal-hearing listeners (t=2.61, Cohen's d=0.23, p=<0.001). No group difference was observed with regard to the contrast of the model fit for the two conditions.

All the five linguistic features were more predictive of EEG data in single-talker speech compared to the unattended speech for both hearing-impaired participants (phoneme: t=1.72, Cohen's d=0.79, p<0.001; syllable: t=1.94, Cohen's d=0.9, p<0.001; word: t=2.91, Cohen's d=1.08, p<0.001; phrase: t=1.4, Cohen's d=0.61, p=0.041; sentence: t=1.67, Cohen's d=1.01, p=0.023) and normal-hearing participants (phoneme: t=1.99, Cohen's d=0.31, p=0.02; syllable: t=1.78, Cohen's d=0.8, p<0.001; word: t=2.85, Cohen's d=1.55, p=0.001; phrase: t=1.74, Cohen's d=1.4, p<0.001; sentence: t=1.86, Cohen's d=0.81, p=0.046). The significant effects occurred progressively later from phoneme to sentence level for both hearing-impaired participants (phoneme: −100-100 ms; syllable: 0-200 ms; word: 0-250 ms; phrase: 200-300 ms; sentence: 200-300 ms) and normal-hearing participants (phoneme: −50-100 ms; syllable:0-200 ms; word: 50-250 ms; phrase: 100-300 ms; sentence: 200-300 ms.). No significant group differences in the model fit were observed between the two conditions for all the linguistic levels.

Regression results for attended versus unattended speech

Figure 6 depicts the model fit of acoustic and linguistic predictors against EEG data for both attended and unattended speech while two speakers narrated simultaneously. It can be seen that for normal-hearing participants, acoustic features demonstrated a better model fit for attended

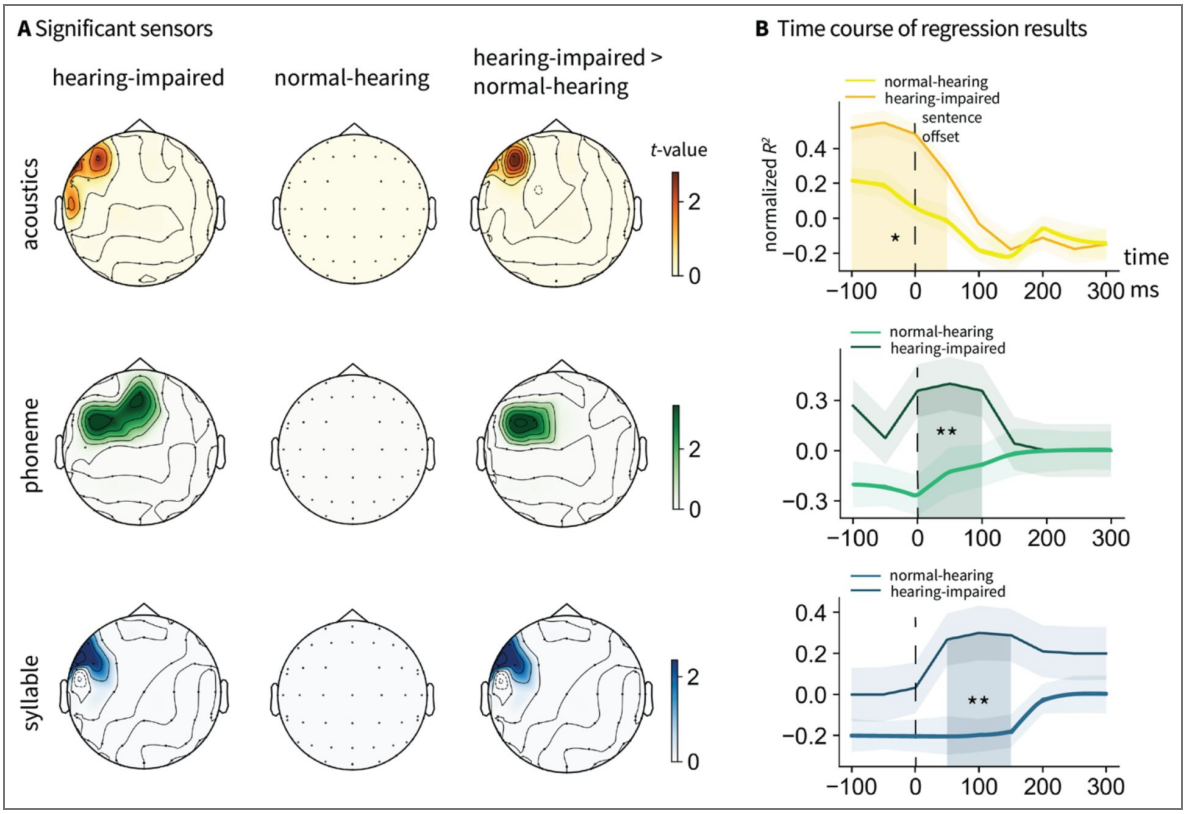


Figure 4. Significant sensor and time window for the model fit to the EEG data for the acoustic and linguistic features extracted from the HM-LSTM model between single-talker and attended speech across the two listener groups.

A. Significant sensors showing higher model fit for single-talker speech compared to the attended speech at the acoustic, phoneme, and syllable levels for the two listener groups and their contrast. **B.** Timecourses of mean model fit in the significant clusters where normal-hearing participants showed higher model fit at the acoustic, phoneme, and syllable levels than hearing-impaired participants. The coefficient of determination (R^2) were z-transformed. Shaded regions indicate significant time windows. * denotes p<0.05, ** denotes p<0.01 and *** denotes p<0.001.

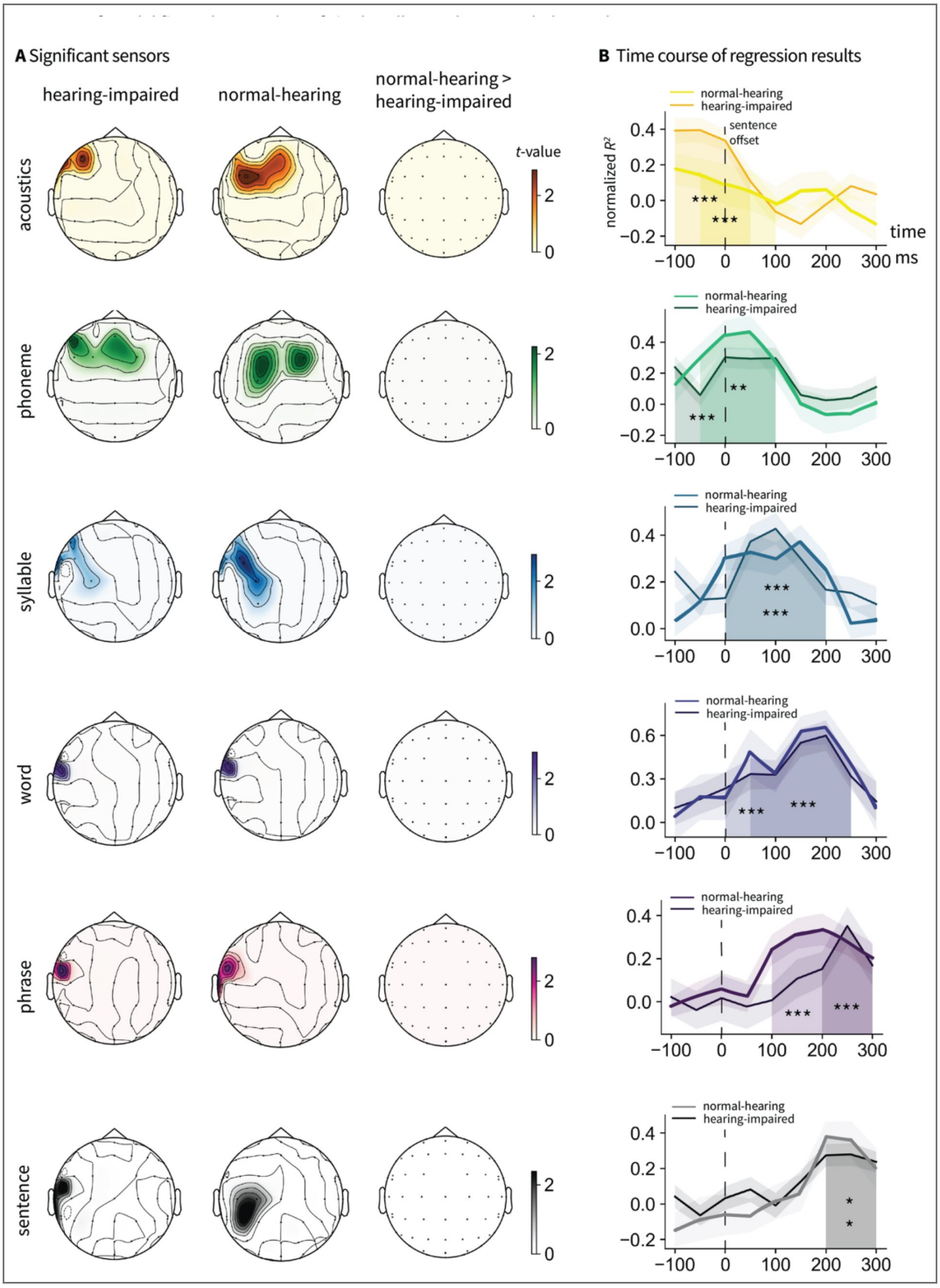


Figure 5. Significant sensor and time window for the model fit to the EEG data for the acoustic and linguistic features between the single-talker and unattended speech in the mixed speech condition across the two listener groups.

A. Significant sensors showing higher model fit for the single-talker speech compared to the unattended speech at the acoustic and linguistic levels for the two listener groups and their contrast. **B.** Timecourses of mean model fit in the significant clusters. The significant time windows for within-group comparisons. The coefficient of determination (R^2) were z-transformed. Shaded regions indicate significant time windows. * denotes p<0.05, ** denotes p<0.01 and *** denotes p<0.001.

speech compared to unattended speech from −100 ms to sentence offsets (t=3.21, Cohen's d=1.34, p=0.02). However, for hearing-impaired participants, no significant differences were observed in this measure. The difference between attended and unattended speech in normal-hearing and hearing-impaired participants was confirmed to be significant in the left temporal region from −100 ms to −50 ms before sentence offsets (t=2.24, Cohen's d=1.01, p=0.02) by a permutation two-sample t-test.

Both phoneme and syllable features significantly better predicted attended speech compared to unattended speech among normal-hearing participants (phoneme: t=1.58, Cohen's d=0.46, p=0.0006; syllable: t=1.05, Cohen's d=1.02, p=0.0001). The significant time window for phonemes was from −100 to 250 ms after sentence offsets, earlier than that for syllables, which was from 0 to 250 ms. No significant differences were observed in the hearing-impaired group. The contrast maps were significantly different across the two groups during the 0-100 ms window for phonemes (t=2.28, Cohen's d=1.32, p=0.026) and the 0-150 ms window for syllables (t=2.64, Cohen's d=1.04, p=0.022).

The word- and phrase-level features were significantly more effective at predicting EEG responses for attended speech than for unattended speech in both normal-hearing (word: t=2.59, Cohen's d=1.14, p=0.002; phrase: t=1.77, Cohen's d=0.68, p=0.027) and hearing-impaired listeners (word: t=3.61, Cohen's d=1.59, p=0.001; phrase: t=1.87, Cohen's d=0.71, p=0.004). The significant time windows for word processing were from 150-250 ms for hearing-impaired listeners and 150-200 ms for normal-hearing listeners. For phrase processing, significant time windows were from 150-300 ms for hearing-impaired listeners and 250-300 ms for normal-hearing listeners. No significant discrepancies were observed between the two groups regarding the model fit of words and phrases to the EEG data for attended versus unattended speeches. Surprisingly, we found a significantly better model fit for sentence-level features in attended speech for normal-hearing participants (t=1.52, Cohen's d=0.98, p=0.003) but not for hearing-impaired participants, and the contrast between the two groups was significant (t=1.7, Cohen's d=1.27, p<0.001), suggesting that hearing-impaired participants also struggle more with tracking information at longer temporal scales in multi-talker scenarios.

mTRF results for attended versus unattended speech

To validate the linguistic features from our HM-LSTM models, we also examined baseline models for the linguistic features using multivariate temporal response function (mTRF) analysis. Our predictors include phoneme, syllable, word, phrase, and sentence (i.e., marking a value of 1 at each unit boundary offset). Our EEG data spans the entire 10-minute duration for each condition, sampled at 10-ms intervals. The TRF results for our main comparison—attended versus unattended conditions—showed similar patterns to those observed using features from our HM-LSTM model (see Figure S2). At the phoneme and syllable levels, normal-hearing listeners showed marginally significantly higher TRF weights for attended speech compared to unattended speech at approximately −80 to 150 ms after phoneme offsets (t=2.75, Cohen's d=0.87, p=0.057), and 120 to 210 ms after syllable offsets (t=3.96, Cohen's d=0.73, p=0.083). At the word and phrase levels, normal-hearing listeners exhibited significantly higher TRF weights for attended speech compared to unattended speech at 190 to 290 ms after word offsets (t=4, Cohen's d=1.13, p=0.049), and around 120 to 290 ms after phrase offsets (t=5.27, Cohen's d=1.09, p=0.045). For hearing-impaired listeners, marginally significant effects were observed at 190 to 290 ms after word offsets (t=1.54, Cohen's d=0.6, p=0.059), and 180 to 290 ms after phrase offsets (t=3.63, Cohen's d=0.89, p=0.09). We also extracted the envelope at 10 ms intervals for both attended and unattended speech and computed mTRFs independently for each subject and sensor using a basis of 50 ms Hamming windows spanning –100 ms to 300 ms relative to envelope onset. The results showed that in hearing-impaired participants, attended speech elicited a significant cluster in the bilateral temporal regions from 270 to 300 ms post-onset (t = 2.40, p = 0.01, Cohen's d = 0.63). Unattended speech elicited an early cluster in right temporal and occipital regions from –100 ms to –80 ms (t = 3.07, p

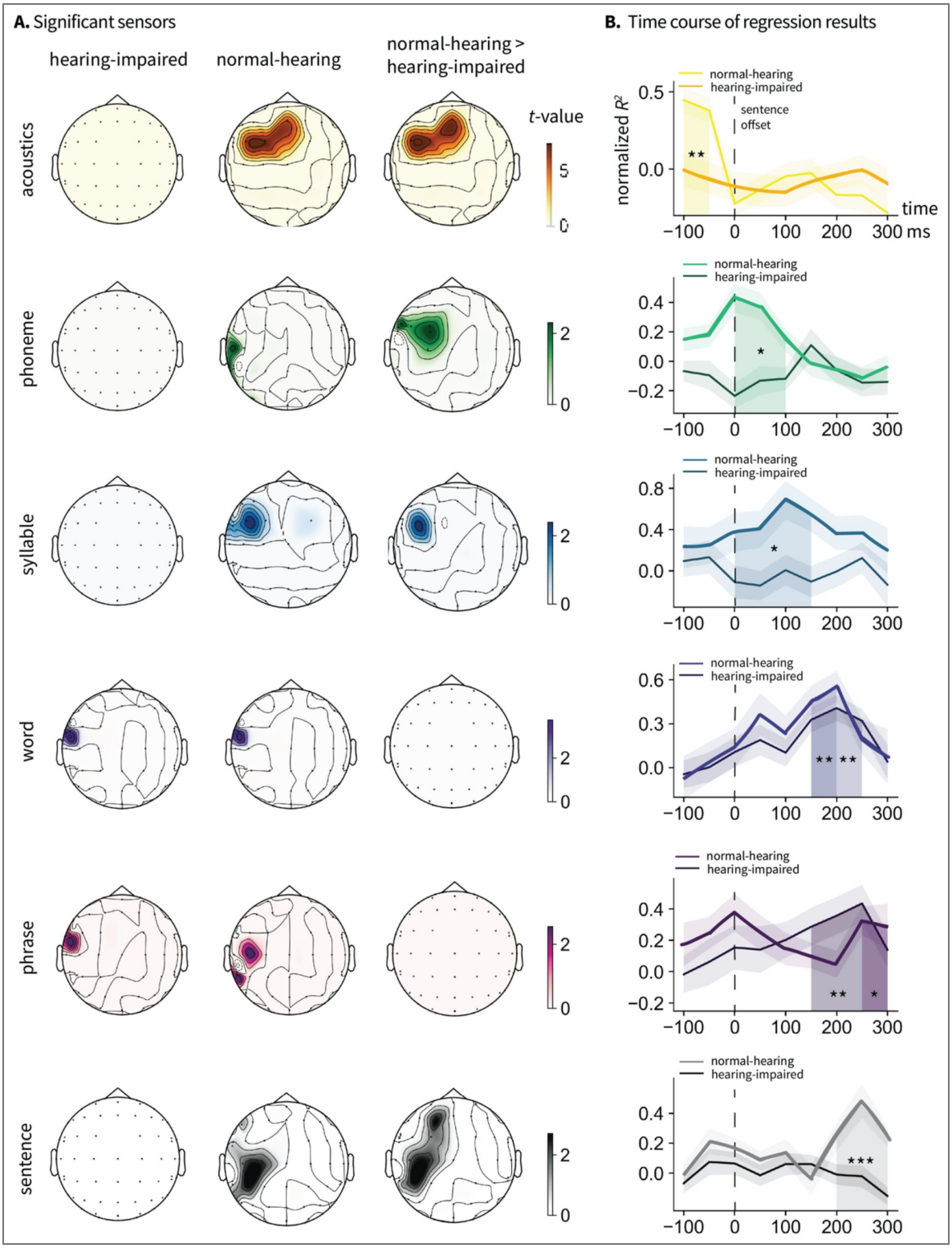


Figure 6. Significant sensor and time window for the model fit to the EEG data for the acoustic and linguistic features between the attended and unattended speech in the mixed speech condition across the two listener groups.

A. Significant sensors showing higher model fit for the attended speech compared to the unattended speech at the acoustic and linguistic levels for the two listener groups and their contrast. **B.** Timecourses of mean model fit in the significant clusters where normal-hearing participants showed higher model fit than hearing-impaired participants. The coefficient of determination (R^2) were z-transformed. Shaded regions indicate significant time windows. * denotes p<0.05, ** denotes p<0.01 and *** denotes p<0.001.

= 0.001, d = 0.83). Normal-hearing participants showed significant envelope tracking in the left temporal region at 280–300 ms after envelope onset (t = 2.37, p = 0.037, d = 0.48), with no significant cluster for unattended speech.

Discussion

Speech comprehension in a multi-talker environment is especially challenging for listeners with impaired hearing (Fuglsang et al., 2020). Consequently, exploring the neural underpinnings of multi-talker speech comprehension in hearing-impaired listeners could yield valuable insights into the challenges faced by both normal-hearing and hearing-impaired individuals in this scenario. Studies have reported abnormally enhanced responses to fluctuations in acoustic envelope in the central auditory system of older listeners (e.g., Goossens et al., 2016; Parthasarathy et al., 2019; Presacco et al., 2016) and listeners with peripheral hearing loss (Goossens et al., 2018; Millman et al., 2017). As older listeners also suffer from suppression of task-irrelevant sensory information due to reduced cortical inhibitory control functions (Gazzaley et al., 2005, 2008), it is possible that impaired speech comprehension in a cocktail party situation arises from an attentional deficit linked to aging (Du et al., 2016; Presacco et al., 2016). However, younger listeners with EHF hearing loss have also reported difficulty understanding speech in a multi-talker environment (Motlagh Zadeh et al., 2019). It remains unknown what information is lost during multi-talker speech perception, and how the hierarchically organized linguistic units in competing speech streams affect the comprehension ability of people with impaired hearing.

In this study, we show that for normal-hearing listeners, the acoustic and linguistic features extracted from an HM-LSTM model can significantly predict EEG responses during both single-talker and attended speech in the context of two speakers talking simultaneously. Interestingly, their intelligibility scores for mixed speech are lower compared to single-talker speech, suggesting that normal-hearing listeners are still capable of tracking linguistic information at these levels in a cocktail party scenario, although with potentially greater effort. The model fit of the EEG data for attended speech is significantly higher than that for unattended speech across all levels. This aligns with previous research on “selective auditory attention,” which demonstrates that individuals can focus on specific auditory stimuli for processing, while effectively filtering out background noise (Brodbeck et al., 2018; Brungart, 2001; Ding & Simon, 2012; Mesgarani & Chang, 2012; O’Sullivan et al., 2015; Shinn-Cunningham, 2008; Zion Golumbic et al., 2013). Expanding on prior research which suggested that phonemes and words of attended speech could be decoded from the left temporal cortex of normal-hearing participants (Brodbeck et al., 2018), our results demonstrate that linguistic units across all hierarchical levels can be tracked in the neural signals.

For listeners with hearing impairments, the model fit for attended speech is significantly poorer at the acoustic, phoneme, and syllable levels compared to that for single-talker speech. Additionally, there is no significant difference in model fit at the acoustic, phoneme, and syllable levels between attended and unattended speech when two speakers are talking simultaneously. However, the model fit for the word and phrase features do not differ between single-talker and attended speech, and is significantly higher for that of the unattended speech. These findings suggest that hearing-impaired listeners may encounter difficulties in processing information at shorter temporal scales, including the dynamic amplitude envelope and spectrotemporal details of speech, as well as phoneme and syllable-level content. This is expected as our EHF hearing loss participants all exhibit higher hearing thresholds at frequencies above 8 kHz. Although these frequencies exceed those necessary for phonetic information in quiet environments, which are below 6 kHz, they may still impact the ability to process auditory information at faster temporal scales more than at slower speeds. Surprisingly, hearing-impaired listeners did not demonstrate an improved model fit for sentence features of the attended speech compared to the unattended speech, indicating that their ability to process information at longer temporal scales is also compromised. One limitation to consider is the absence of a behavioral task, such as comprehension questions at the end of the listening sections. This raises the possibility that the

reduced cortical encoding of attended versus unattended speech across multiple linguistic levels in hearing-impaired listeners could stem from a different attentional strategy. For instance, they may focus on "getting the gist" of the story or intermittently disengage from the task, tuning back in only for selected keywords or word combinations. However, we would like to emphasize that our hearing-impaired participants have EHF hearing loss, with impairment limited to frequencies above 8 kHz. This condition is unlikely to be severe enough to induce a fundamentally different attentional strategy for this task. Furthermore, normal-hearing listeners may also employ varied attentional strategies, yet the comparison still revealed significant differences. Based on these findings, we hypothesize that hearing-impaired listeners may struggle to extract low-level information from competing speech streams. Such a disruption in bottom-up processing could impede their ability to discern sentence boundaries effectively, which in turn hampers their ability to benefit from top-down information processing.

The hierarchical temporal receptive window (TRW) hypothesis proposed that linguistic units at shorter temporal scales, such as phonemes, are encoded in the core auditory cortex, while information of longer duration is processed in higher perceptual and cognitive regions, such as the anterior temporal or posterior temporal and parietal regions (Hasson et al., 2008; Honey et al., 2012; Lerner et al., 2011; Murray et al., 2014). With the limited spatial resolution of EEG, we could not directly compare the spatial localization of these units at different temporal scales, however, we did observe an increasing latency in the significant model fit across different linguistic levels. Specifically, the significant time window for acoustic and phoneme features occurred around −100 to 100 ms relative to sentence offsets; syllables and words around 0-200 ms; and phrases and sentences around 200-300 ms. These progressively later effects from lower to higher linguistic levels suggest that these units may indeed be represented in brain regions with increasingly longer TRWs.

Our hierarchical linguistic contents were extracted using the HM-LSTM model adapted from (Chung et al., 2017). This model has been adopted by Schmitt et al. (2021) to show that a "surprisal hierarchy" based on the hidden layer activity correlated with fMRI blood oxygen level-dependent (BOLD) signals along the temporal-parietal pathway during naturalistic listening. Although their research question is different from ours, their results suggested that the model has effectively captured information at different linguistic levels. Our testing results further confirmed that the model representations at the phoneme and syllable levels are different from model representations at the higher linguistic levels when the phonemes within the sentences are similar. Compared to the increasingly popular "model-brain alignment" studies that typically use transformer architectures (e.g., (Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021), our HM-LSTM model is considerably smaller in parameter size and does not match the capabilities of state-of-the-art large language models (LLMs) in downstream natural language processing (NLP) tasks such as question-answering, text summarization, translation, etc. However, our model incorporates phonemic and syllabic level representations, which are absent in LLMs that operate at the sub-word level. This feature could provide unique insights into how the entire hierarchy of linguistic units is processed in the brain. It's important to note that we do not assert any similarity between the model's internal mechanisms and the brain's mechanisms for processing linguistic units at different levels. Instead, we use the model to disentangle linguistic contents associated with these levels. This approach has proven successful in elucidating language processing in the brain, despite the notable dissimilarities in model architectures compared to the neural architecture of the brain. For example, Nelson et al. (2017) correlated syntactic processing under different parsing strategies with the intracranial electrophysiological signals and found that the left-corner and bottom-up strategies fit the left temporal data better than the most eager top-down strategy; Goldstein et al. (2022) and Caucheteux & King (2022) also showed that the human brain and the deep learning language models share the computational principles as they process the same natural narrative.

In summary, our findings show that linguistic units extracted from a hierarchical language model better explain the EEG responses of normal-hearing listeners for attended speech, as opposed to unattended speech, when two speakers are talking simultaneously. However, hearing-impaired

listeners exhibited poorer model fits at the acoustic, phoneme, syllable, and sentence levels, although their model fits at the word and phrase levels were not significantly affected. These results suggest that processing information at both shorter and longer temporal scales is especially challenging for hearing-impaired listeners when attending to a chosen speaker in a cocktail party situation. As such, these findings connect basic research on speech comprehension with clinical studies on hearing loss, especially hidden hearing loss, a global issue that is increasingly common among young adults.

Materials and Methods

Participants

A total of 51 participants (26 females, mean age=24 years, SD=2.12 years) with EHF hearing loss and 51 normal-hearing participants (26 females, mean age=22.92 years, SD=2.14 years) took part in the experiment. 28 participants (18 females, mean age=23.55, SD=2.18) were removed from the analyses due to excessive motion, drowsiness or inability to complete the experiment, resulting in a total of 41 participants (21 females, mean age=24.1, SD=2.1) with EHF hearing loss and 33 participants (13 females, mean age=22.94, SD=2.36) with normal-hearing. All participants were right-handed native Mandarin speakers currently studying in Shanghai for their undergraduate or graduate degree, with no self-reported neurological disorders. EHF hearing loss was diagnosed using the PTA test, thresholded at frequencies above 8 kHz. The PTA was performed by experienced audiological technicians using an audiometer (Madsen Astera, GN Otometrics, Denmark) with headphones (HDA-300, Sennheiser, Germany) in a soundproof booth with background noise below 25 dB(A), as described previously (Wang et al., 2021). Air-conduction audiometric thresholds for both ears at frequencies of 0.5, 1, 2, 3, 4, 6, 8, 10, 12.5, 14 and 16 kHz were measured in 5-dB steps in accordance with the regulations of ISO 8253-1:2010.

Stimuli

Our experimental stimuli were two excerpts from the Chinese translation of “The Little Prince” (available at *http://www.xiaowangzi.org/*), previously used in fMRI studies where participants with normal hearing listened to the book in its entirety (Li et al., 2022). This material has been enriched with detailed linguistic predictions, from lexical to syntactic and discourse levels, using advanced natural language processing tools. Such rich annotation is critical for modeling hierarchical linguistic structures in our study. The two excerpts were narrated by one male and one female computer-synthesized voice, developed by the Institute of Automation, Chinese Academy of Sciences. The synthesized speech (available at *https://osf.io/fjv5n/*) is comparable to human narration, as confirmed by participants’ post-experiment assessment of its naturalness. Additionally, using computer-synthesized voice instead of human-narrated speech alleviates the potential issue of imbalanced voice intensity and speaking rate that can arise between female and male narrators. The two sections were matched in length (approximately 10 minutes) and mean amplitude (approximately 65 dB), and were mixed digitally in a single channel to prevent any biases in hearing ability between the left and right ears.

Experimental procedure

The experimental task consisted of a multi-talker condition and a single-talker condition. In the multi-talker condition, the mixed speech was presented twice with the female and male speakers narrating simultaneously. Before each trial, instructions appeared in the center of the screen indicating which of the talkers to attend to (e.g., “Attend Female”). In the single-talker condition, the male and female speeches were presented separately (see Figure 1A for the experiment procedure). The presentation order of the 4 conditions was randomized, and breaks were given between each trial. Stimuli were presented using insert earphones (ER-3C, Etymotic Research, United States) at a comfortable volume level of approximately 65 dB SPL. Participants were instructed to maintain visual fixation for the duration of each trial on a crosshair centered on the computer screen, and to minimize eye blinking and all other motor activities for the duration of

each section. The whole experiment lasted for about 65 minutes, and participants rated the intelligibility of the multi-talker and the single-talker speeches on a 5-point Likert scale after the experiment. The experiment was conducted at the Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth People's Hospital affiliated with the School of Medicine at Shanghai Jiao Tong University. The experimental procedures were approved by the Ethics Committee of the Ninth People's Hospital affiliated with Shanghai Jiao Tong University School of Medicine (SH9H-2019-T33-2). All participants provided written informed consent prior to the experiment and were paid for their participation.

Acoustic features of the speech stimuli

The acoustic features included the broadband envelopes and the log-mel spectrograms of the two single-talker speech streams. The amplitude envelope of the speech signal was extracted using the Hilbert transform. The 129-dimension spectrogram and 1-dimension envelope were concatenated to form a 130-dimension acoustic feature at every 10 ms of the speech stimuli.

Hierarchical multiscale LSTM model

We extended the original HM-LSTM model developed by Chung et al. (2017) to include not just the word and phrasal levels but also the sub-lexical phoneme and syllable levels. The model inputs were individual phonemes from two sentences, each transformed into a 1024-dimensional vector using a simple lookup table. This lookup table stores embeddings for a fixed dictionary of all unique phonemes in Chinese. This approach is a foundational technique in many advanced NLP models, enabling the representation of discrete input symbols in a continuous vector space. The output of the model was the classification result of whether the second sentence follows the first sentence. At each layer, the model implements a COPY or UPDATE operation at each time step *t*. The COPY operation maintains the current cell state of without any changes until it receives a summarized input from the lower layer. The UPDATE operation occurs when a linguistic boundary is detected in the layer below, but no boundary was detected at the previous time step *t-1*. In this case, the cell updates its summary representation, similar to standard RNNs. We trained our model on 10,000 sentence pairs from the WenetSpeech corpus (Zhang et al., 2021), a collection that features over 10,000 hours of labeled Mandarin Chinese speech sourced from YouTube and podcasts. We used 1024 units for the input embedding and 2048 units for each HM-LSTM layer.

Correlations among LSTM model layers

All the regressors are represented as 2048-dimensional vectors derived from the hidden layers of the trained HM-LSTM model. We applied the trained model to all 284 sentences in our stimulus text, generating a set of 284 × 2048-dimensional vectors. Next, we performed Principal Component Analysis (PCA) on the 2048 dimensions and extracted the first 100 principal components (PCs), resulting in 284 × 100-dimensional vectors for each regressor. These 284 × 100 matrices were then flattened into 28,400-dimensional vectors. Subsequently, we computed the correlation matrix for the z-transformed 28,400-dimensional vectors of our five linguistic regressors.

EEG recording and preprocessing

EEG was recorded using a standard 64-channel actiCAP mounted according to the international 10-20 system against a nose reference (Brain Vision Recorder, Brain Products). The ground electrode was set at the forehead. EEG signals were registered between 0.016 and 80 Hz with a sampling rate of 500 Hz. The impedances were kept below 20 kΩ. The EEG recordings were band-pass filtered between 0.1 Hz and 45 Hz using a linear-phase finite impulse response (FIR) filter. Independent component analysis (ICA) was then applied to remove eye blink artifacts. The EEG data were then segmented into epochs spanning 500 ms pre-stimulus onset to 10 minutes post-stimulus onset and were subsequently downsampled to 100 Hz. We further decomposed the epoched EEG time series for each section into six classic frequency bands components (delta 1–3 Hz, theta 4–7 Hz, alpha 8–12 Hz, beta 12–20 Hz, gamma 30–45 Hz) by convolving the data with complex Morlet wavelets as implemented in MNE-Python (version 0.24.0). The number of cycles in the Morlet wavelets was set

to frequency/4 for each frequency bin. The power values for each time point and frequency bin were obtained by taking the square root of the resulting time-frequency coefficients. These power values were normalized to reflect relative changes (expressed in dB) with respect to the 500 ms pre-stimulus baseline. This yielded a power value for each time point and frequency bin for each section.

Ridge regression at different time latencies

For each subject, we modeled the EEG responses at each sensor from the single-talker and mixed-talker conditions with our acoustic and linguistic features using ridge regression with a default regularization coefficient of 1 (see Figure 1B). For each sentence in the speech stimuli, we extracted the 2048-dimensional hidden layer activity from the HM-LSTM model to represent features at the phoneme, syllable, word, phrase and sentence levels. We then employed Principal Component Analysis (PCA) to reduce the 2048-dimensional vectors to the first 150 principal components. The 150-dimensional vectors for the 5 linguistic levels were then fit to the EEG signals time-locked to the offset of each sentence in the stimuli using ridge regression. The regressors are aligned to sentence offsets because our goal is to identify neural correlates for each model-derived feature of a whole sentence. If we align model activity with EEG data time-locked to sentence onsets, we would be finding neural responses to linguistic levels (from phoneme to sentence) of the whole sentence at the time when participants have not processed the sentence yet. Although this limits our analysis to a subset of the data (143 sentences × 2 sections × 400 ms windows), it targets the exact moment when full-sentence representations emerge against background speech, allowing us to examine each model-derived feature onto its neural signature. We understand that phonemes, syllables, words, phrases, and sentences differ in their durations. However, the five hidden activity vectors extracted from the model are designed to capture the representations of these five linguistic levels across *the entire sentence*. Specifically, for a sentence pair such as “It can fly <sep> This is an airplane,” the first 2048-dimensional vector represents all the phonemes in the two sentences (“t a_1 n əŋ_2 f ei_1 <sep> zh ə_4 sh iii_4 f ei_1 j ii_1”), the second vector captures all the syllables (“ta_1 nəŋ_2 fei_1 <sep> zhə_4 shiii_4 fei_1jii_1”), the third vector represents all the words, the fourth vector captures the phrases, and the fifth vector represents the sentence-level meaning. In our dataset, input pairs consist of adjacent sentences from the stimuli (e.g., Sentence 1 and Sentence 2, Sentence 2 and Sentence 3, and so on), and for each pair, the model generates five 2048-dimensional vectors, each corresponding to a specific linguistic level. To identify the neural correlates of these model-derived features—each intended to represent the full linguistic level across a complete sentence—we focused on the EEG signal surrounding the completion of the second sentence rather than on incremental processing. Accordingly, we extracted epochs from −100 ms to +300 ms relative to the offset of the second sentence and performed ridge regression analyses using the five model features (reduced to 150 dimensions via PCA) at every 50 ms across the epoch. To assess the temporal progression of the regression outcomes, we conducted the analysis at nine sequential time points, ranging from 100 milliseconds before to 300 milliseconds after the sentence offset, with a 50-millisecond interval. We chose this time window as lexical or phrasal processing typically occurs 200 ms after stimulus offsets (Bemis & Pylkkanen, 2011; Li et al., 2024; Li & Pylkkänen, 2021). Additionally, we included the −100 to 200 ms time period in our analysis to examine phoneme and syllable level processing (e.g., Gwilliams et al., 2022). Using the entire sentence duration was not feasible, as the sentences in the stimuli vary in length, making statistical analysis challenging. Additionally, extending the window (e.g. to 2 seconds) would risk overlapping adjacent sentences. This would introduce ambiguity as to whether the EEG responses correspond to the current or the adjacent sentences. Additionally, our model activity represents a “condensed final representation” at the five linguistic levels for the whole sentence, rather than incrementally during the sentence. We think the −100 to 300 ms time window relative to each sentence offset targets the exact moment when full-sentence representations are comprehended and a “condensed final representation” for the whole sentence across five linguistic level have been formed in the brain. The same regression procedure was applied to the 130-dimensional acoustic features (see the Supplementary video for a more detailed description of the ridge regression methods).

Note that we did not use Pearson's r as in some prior studies using ridge regression as brain encoding models (e.g., (Goldstein et al., 2022)) because our analysis did not involve a train-test split. Specifically, Goldstein et al. (2022) divided their data into training and testing sets, trained a ridge regression model on the training set, and then used the trained model to predict neural responses on the test set. They calculated Pearson's r to assess the correlation between the predicted and observed neural responses, making the correlation coefficient (r) their primary measure of model performance. In contrast, our analysis focused on computing the model fitting performance (R^2) of the ridge regression model for each sensor and time point for each subject. At the group level, we conducted one-sample t-tests with spatiotemporal cluster-based correction on the R^2 values to identify sensors and time windows where R^2 values were significantly greater than baseline. We established the baseline by normalizing the R^2 values using Fisher z-transformation across sensors within each subject. The ridge regression was performed using customary python codes, making heavy use of the scikit-learn (1.2.2) package.

Spatiotemporal clustering analysis

The timecourses of the z-transformed coefficient of determination (R^2) from the regression results at the nine time points for each sensor, corresponding to the linguistic and acoustic regressor for each subject, underwent spatiotemporal cluster permutation test to determine their statistical significance at the group level. For instance, to assess whether words from the attended stimuli better predict EEG signals during the mixed speech compared to words from the unattended stimuli, we used the 150-dimensional vectors corresponding to the word layer from our LSTM model for the attended and unattended stimuli as regressors. We then fit these regressors to the EEG signals at 9 time points (spanning −100 ms to 300 ms around the sentence offsets, with 50 ms intervals). We then conducted one-tailed two-sample t-tests to determine whether the differences in the contrasts of the z-transformed R^2 timecourses were statistically significant. We repeated these procedures 10,000 times, replacing the observed t-values with shuffled t-values for each participant to generate a null distribution of t-values for each sensor. Sensors whose t-values were in the top 5th percentile of the null distribution were deemed significant (sensor-wise significance $p < 0.05$). The same method was applied to analyze the contrasts between attended and unattended speech during mixed speech conditions, both within and between groups. All our analyses were performed using custom python codes, making heavy use of the mne (v.1.6.1), torch (v2.2.0), scipy (v1.12.0) and scikit-learn (1.2.2) packages.

mTRF analysis

To validate the linguistic features from our HM-LSTM models, we also examined baseline models for the linguistic features using mTRF analysis. Our predictors include phoneme, syllable, word, phrase, and sentence (i.e., marking a value of 1 at each unit boundary offset). Our EEG data spans the entire 10-minute duration for each condition, sampled at 10-ms intervals. Since our speech stimuli were computer-synthesized, the phonem and syllable boundaries were automatically generated. The word boundaries were manually annotated by a native Mandarin as in Li et al. (2022) . The phrase boundaries were automatically annotated by the Stanford parser (Levy & Manning, 2003) and manually checked by a native Mandarin speaker. These rate models are represented as five distinct binary time series, each aligned with the timing of the corresponding linguistic unit, making them well-suited for mTRF analysis. Note that the rate regressors only encode the timing of linguistic unit boundaries, while the model-derived features encode the representational content of the linguistic input. We also extracted the envelope at 10 ms intervals for both attended and unattended speech. We computed mTRFs independently for each subject and sensor using a basis of 50 ms Hamming windows spanning –100 ms to 300 ms relative to all linguistic boundaries and envelope onset. The mTRF analysis was conducted usin the Python package Eelbrain (v.0.39.8).

Data and code availability

All data and codes are available at *https://osf.io/fjv5n/* .

Supplementary

Same-vowel 4-syllable sentence			
nǎi	nai	mǎi	mài
ʃǎn	ʃən	ʃǎn	ʃàn
gūŋ	guŋ	tʃūŋ	dùŋ
mèi	mei	méi	lèi
tɕiàu	tɕiəu	tɕʰiáu	tɕiàu
jí	ji	jí	jì
dì	ɕí	ɕī	ɕì
tài	tai	bǎi	pāi
wá	wa	wā	wā
ʃū	fù	sù	dú
gū	fu	kǔ	dú
bà	ba	dá	kǎ
mā	ma	mà	mǎ
buɔ́	buɔ	buɔ̄	guɔ̌
ʃū	ʃu	ʃǔ	ʃù
gū	gu	bǔ	bù
lǎʊ	laʊ	náʊ	māʊ
puɔ́	puɔ́	duɔ̌	guɔ̄
tɕiě	tɕie	tɕié	tɕiē
dì	di	ɕǐ	dì

Table S1. **All four-syllable Chinese sentences with same vowels.**

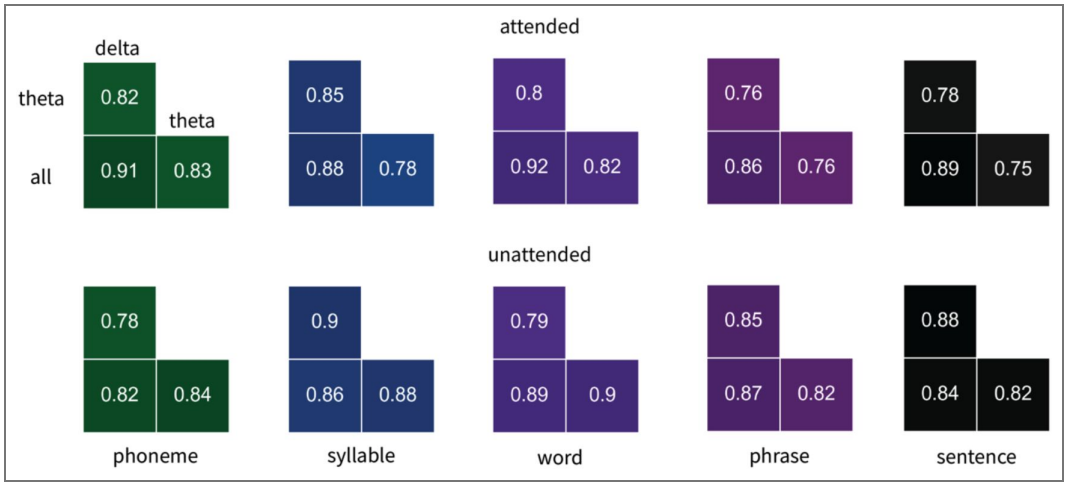


Figure S1. Correlation matrices of regression outcomes for the five linguistic predictors between the EEG data from delta, theta and all frequency bands.

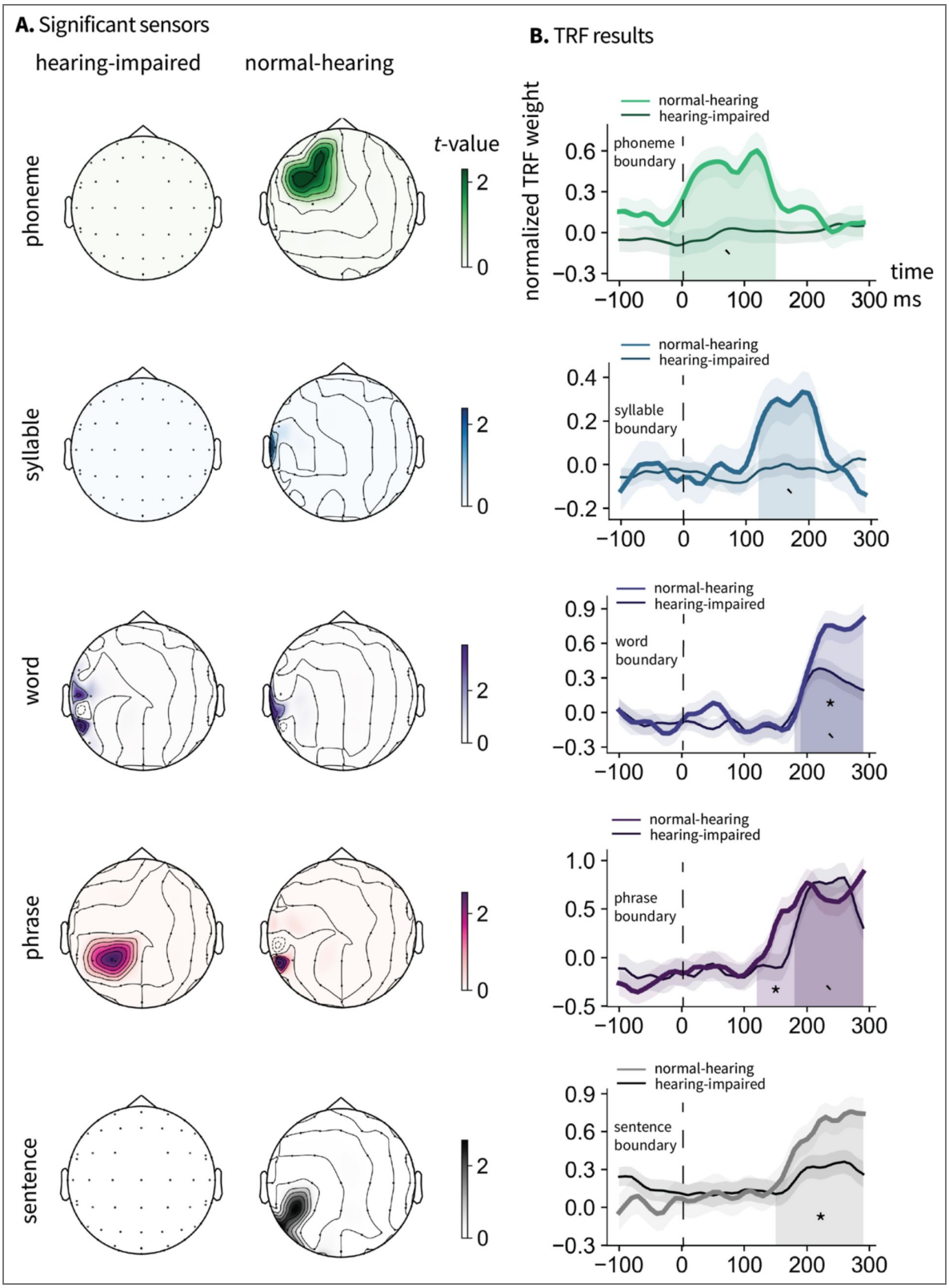


Figure S2. Contrast of TRF weights to the EEG data of attended and unattended speech for the five linguistic predictors.

A. Significant sensors showing higher model fit for the attended speech compared to the unattended speech at the acoustic and linguistic levels for the two listener groups and their contrast. Shaded regions indicate significant time windows. * denotes p<0.05, ** denotes p<0.01 and *** denotes p<0.001.

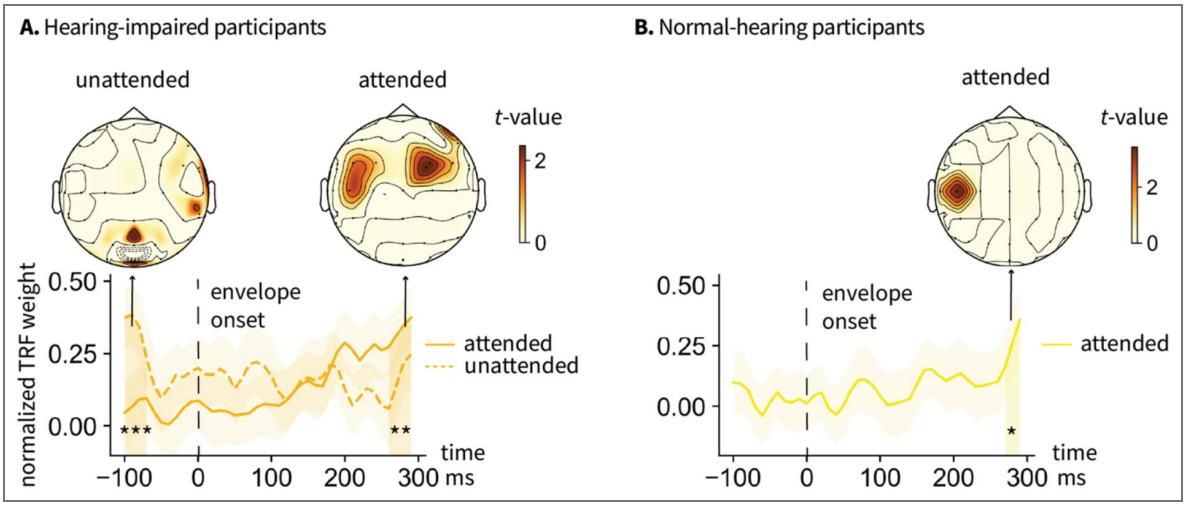


Figure S3. TRF weights to the EEG data of attended and unattended speech envelope.

Shaded regions indicate significant time windows. * denotes p<0.05, ** denotes p<0.01 and *** denotes p<0.001.

Additional information

Author ORCID iDs

Jixing Li: https://orcid.org/0000-0002-5210-6224

Author notes

Author contributions: Jixing Li designed the research, analyzed data, and wrote the paper; Qixuan Wang and Zhiwu Huang designed the research; Qian Zhou and Lu Yang collected data; Yutong Shen and Shujian Huang implemented the model; Shaonan Wang provided speech stimuli and helped write the paper; Liina Pylkkänen helped write the paper.

Funding: This work was supported by the CityU Start-up Grant 7020086 and CityU Strategic Research Grant 7200747 (Li, J.) the National Natural Science Foundation of China 82201273 (Wang Q.), Shanghai Science and Technology Commission Grant 22Y11902000 (Huang Z.) and award G1001 from NYUAD Institute, New York University Abu Dhabi (Pylkkänen, L.).

Competing interests: No competing interests declared

References

Apoux F., Bacon S. P (2004) Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *The Journal of the Acoustical Society of America* **116**:1671-1680

Badri R., Siegel J. H., Wright B. A (2011) Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms. *The Journal of the Acoustical Society of America* **129**:852-863

Bemis D. K., Pylkkanen L (2011) Simple composition: An MEG investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience* **31**:2801-2814

Bharadwaj H. M., Mai A. R., Simpson J. M., Choi I., Heinz M. G., Shinn-Cunningham B. G (2019) Non-invasive assays of cochlear synaptopathy – Candidates and considerations. *Neuroscience* **407**:53-66

Blank I. A., Fedorenko E (2020) No evidence for differences among language regions in their temporal receptive windows. *NeuroImage* **219**:116925

Brodbeck C., Hong L. E., Simon J. Z (2018) Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology* **28**:3976-3983.e5

Brungart D. S (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* **109**:1101-1109

Caucheteux C., King J.-R (2022) Brains and algorithms partially converge in natural language processing. *Communications Biology* **5**:134

Chang C. H. C., Nastase S. A., Hasson U (2022) Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences* **119**:e2209307119

Cherry E. C (1953) Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* **25**:975-979

Chung J., Ahn S., Bengio Y (2017) Hierarchical multiscale recurrent neural networks. *Iclr* **17**

Collins M. J., Cullen J. K., Berlin C. I (1981) Auditory signal processing in a hearing-impaired subject with residual ultra-audiometric hearing. *Audiology: Official Organ of the International Society of Audiology* **20**:347-361

Ding N., Mellon L., Zhang H., Tian X., Poeppel D (2015) Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*

Ding N., Simon J. Z (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* **109**:11854-11859

Du Y., Buchsbaum B. R., Grady C. L., Alain C (2016) Increased activity in frontal motor cortex compensates impaired speech perception in older adults. *Nature Communications* **7**:12241

Fuglsang S. A., Märcher-Rørsted J., Dau T., Hjortkjær J (2020) Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience* **40**:2562-2572

Gao C., Li J., Chen J., Huang S (2024) Measuring meaning composition in the human brain with composition scores from large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. pp. 11295-11308

Gazzaley A., Clapp W., Kelley J., McEvoy K., Knight R. T., D'Esposito M (2008) Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proceedings of the National Academy of Sciences of the United States of America* **105**:13122-13126

Gazzaley A., Cooney J. W., Rissman J., D'Esposito M (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience* **8**:1298-1300

Goldstein A., Zada Z., Buchnik E., Schain M., Price A., Aubrey B., Nastase S. A., Feder A., Emanuel D., Cohen A., *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* **25**

Goossens T., Vercammen C., Wouters J., van Wieringen A. (2018) Neural envelope encoding predicts speech perception performance for normal-hearing and hearing-impaired adults. *Hearing Research* **370**:189-200

Goossens T., Vercammen C., Wouters J., Wieringen A. van. (2016) Aging affects neural synchronization to speech-related acoustic modulations. *Frontiers in Aging Neuroscience* **8**

Gwilliams L., King J.-R., Marantz A., Poeppel D (2022) Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications* **13**

Hasson U., Yang E., Vallines I., Heeger D. J., Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience* **28**:2539-2550

Honey C. J., Thesen T., Donner T. H., Silbert L. J., Carlson C. E., Devinsky O., Doyle W. K., Rubin N., Heeger D. J., Hasson U (2012) Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**:423-434

Hubel D. H., Wiesel T. N (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* **160**:106-154

Lerner Y., Honey C. J., Silbert L. J., Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience* **31**:2906-2915

Levy R., Manning C. D (2003) Is it harder to parse Chinese, or the Chinese treebank?. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. pp. 439-446

Levy S. C., Freed D. J., Nilsson M., Moore B. C. J., Puria S (2015) Extended high-frequency bandwidth improves speech reception in the presence of spatially separated masking speech. *Ear and Hearing* **36**:e214-224

Li J., Bhattasali S., Zhang S., Franzluebbers B., Luh W.-M., Spreng R. N., Brennan J. R., Yang Y., Pallier C., Hale J (2022) Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data* **9**

Li J., Lai M., Pylkkänen L (2024) Semantic composition in experimental and naturalistic paradigms. *Imaging Neuroscience* **2**:1-17

Li J., Pylkkänen L (2021) Disentangling semantic composition and semantic association in the left temporal lobe. *Journal of Neuroscience* **41**:6526-6538

Luo H., Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54**:1001-1010

McDermott J. H (2009) The cocktail party problem. *Current Biology* **19**:R1024-R1027

Mesgarani N., Chang E. F (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**:233-236

Millman R. E., Mattys S. L., Gouws A. D., Prendergast G. (2017) Magnified neural envelope coding predicts deficits in speech perception in noise. *The Journal of Neuroscience* **37**:7727-7736

Motlagh Zadeh L., Silbert N. H., Sternasty K., Swanepoel D. W., Hunter L. L., Moore D. R. (2019) Extended high-frequency hearing enhances speech perception in noise. *Proceedings of the National Academy of Sciences* **116**:23753-23759

Murray J. D., Bernacchia A., Freedman D. J., Romo R., Wallis J. D., Cai X., Padoa-Schioppa C., Pasternak T., Seo H., Lee D., *et al.* (2014) A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience* **17**:1661-1663

Nelson M. J., Karoui I. E., Giber K., Yang X., Cohen L., Koopman H., Cash S. S., Naccache L., Hale J. T., Pallier C., *et al.* (2017) Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences* **114**:E3669-E3678

O'Sullivan J. A., Power A. J., Mesgarani N., Rajaram S., Foxe J. J., Shinn-Cunningham B. G., Slaney M., Shamma S. A., Lalor E. C (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex* **25**:1697-1706

Parthasarathy A., Herrmann B., Bartlett E. L (2019) Aging alters envelope representations of speech-like sounds in the inferior colliculus. *Neurobiology of Aging* **73**:30-40

Peelle J. E., Chandrasekaran K., Powers J., Smith E. E., Grossman M (2013) Age-related vulnerability in the neural systems supporting semantic processing. *Frontiers in Aging Neuroscience* **5**:46

Presacco A., Simon J. Z., Anderson S (2016) Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology* **116**:2346-2355

Schmitt L.-M., Erb J., Tune S., Rysop A. U., Hartwigsen G., Obleser J (2021) Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances* **7**:eabi6070

Schrimpf M., Blank I. A., Tuckute G., Kauf C., Hosseini E. A., Kanwisher N., Tenenbaum J. B., Fedorenko E (2021) The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**:e2105646118

Shinn-Cunningham B. G (2008) Object-based auditory and visual attention. *Trends in Cognitive Sciences* **12**:182-186

Sugimoto Y., Yoshida R., Jeong H., Koizumi M., Brennan J. R., Oseki Y (2024) Localizing syntactic composition with left-corner recurrent neural network grammars. *Neurobiology of Language* **5**:201-224

Wang Q., Yang L., Qian M., Hong Y., Wang X., Huang Z., Wu H (2021) Acute recreational noise-induced cochlear synaptic dysfunction in humans with normal hearing: A prospective cohort study. *Frontiers in Neuroscience* **15**

Zhang B., Lv H., Guo P., Shao Q., Yang C., Xie L., Xu X., Bu H., Chen X., Zeng C., *et al.* (2021) WenetSpeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. *arXiv*

Zion Golumbic E. M., Ding N., Bickel S., Lakatos P., Schevon C. A., McKhann G. M., Goodman R. R., Emerson R., Mehta A. D., Simon J. Z., *et al.* (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party.". *Neuron* **77**:980-991

Peer reviews

Reviewer #1 (Public review):

The authors relate a language model developed to predict whether a given sentence correctly followed another given sentence to EEG recordings in a novel way, showing receptive fields related to widely used TRFs. In these responses (or "regression results"), differences between representational levels are found, as well as differences between attended and unattended

speech stimuli, and whether there is hearing loss. These differences are found per EEG channel.

In addition to these novel regression results, which are apparently captured from the EEG specifically around the sentence stimulus offsets, the authors also perform a more standard mTRF analysis using a software package (Eelbrain) and TRF regressors that will be more familiar to researchers adjacent to these topics, which was highly appreciated for its comparative value. Comparing these TRFs with the authors' original regression results, several similarities can be seen. Specifically, response contrasts for attended versus unattended speaker during mixed speech, for the phoneme, syllable, and sentence regressors, are greater for normal-hearing participants than hearing-impaired participants for both analyses, and the temporal and spatial extents of the significant differences are roughly comparable (left-front and 0 - 200 ms for phoneme and syllable, and left and 200 - 300 ms for sentence).

The inclusion of the mTRF analysis is helpful also because some aspects of the authors' original regression results, between the EEG data and the HM-LSTM linguistic model, are less than clear. The authors state specifically that their regression analysis is only calculated in the -100 - 300 ms window around stimulus/sentence offsets. They clarify that this means that most of the EEG data acquired while the participants are listening to the sentences is not analyzed, because their HM-LSTM model implementation represents all acoustic and linguistic features in a condensed way, around the end of the sentence. Thus the regression between data and model only occurs where the model predictions exist, which is the end of the sentences. This is in contrast to the mTRF analysis, which seems to have been done in a typical way, regressing over the entire stimulus time, because those regressors (phoneme onset, word onset, etc.) exist over the entire sentence time. If my reading of their description of the HM-LSTM regression is correct, it is surprising that the regression weights are similar between the HM-LSTM model and the mTRF model.

However, the code that the authors uploaded to OSF seems to clarify this issue. In the file ridge_lstm.py, the authors construct the main regressor matrices called X1 and X2 which are passed to sklearn to do the ridge regression. This ridge regression step is calculated on the continuous 10-minute bouts of EEG and stimuli, and it is calculated in a loop over lag times, from -100 ms to 300 ms lag. These regressor matrices are initialized as zeros, and are then filled in two steps: the HM_LSTM model unit weights are read from numpy files and written to the matrices at one timepoint per sentence (as the authors describe in the text), and the traditional phoneme, syllable, etc. annotations are ALSO read in (from csv files) and written to the matrices, putting 1s at every timepoint of those corresponding onsets/offsets. Thus the actual model regressor matrix for the authors' main EEG results includes BOTH the HM_LSTM model weights for each sentence AND the feature/annotation times, for whichever of the 5 features is being analyzed (phonemes, syllables, words, phrases, or sentences).

So for instance, for the syllable HM_LSTM regression results, the regressor matrix contains: 1) the HM_LSTM model weights corresponding to syllables (a static representation, placed once per sentence offset time), AND 2) the syllable onsets themselves, placed as a row of 1s at every syllable onset time. And as another example, for the word HM_LSTM regression results, the regressor matrix contains: 1) the HM_LSTM model weights corresponding to words (a static representation, placed once per sentence offset time), AND 2) the word onsets themselves, placed as a row of 1s at every word onset time.

If my reading of the code is correct, there are two main points of clarification for interpreting these methods:

First, the authors' window of analysis of the EEG is not "limited" to 400 ms as they say; rather the time dimension of both their ridge regression results and their traditional mTRF analysis is simply lags (400 ms-worth), and the responses/receptive fields are calculated over the

entire 10-minute trials. This is the normal way of calculating receptive fields in a continuous paradigm. The authors seem to be focusing on the peri-sentence offset time points because that is where the HM_LSTM model weights are placed in the regressor matrix. Also because of this issue, it is not really correct when the authors say that some significant effect occurred at some latency "after sentence offset". The lag times of the regression results should have the traditional interpretation of lag/latency in receptive field analyses.

Second, as both the traditional linguistic feature annotations and the HM_LSTM model weights are part of the regression for the main ridge regression results here, it is not known what the contribution specifically of the HM_LSTM portion of the regression was. Because the more traditional mTRF analysis showed many similar results to the main ridge regression results here, it seems probable that the simple feature annotations themselves, rather than the HM_LSTM model weights, are responsible for the main EEG results. A further analysis separating these two sets of regressors would shed light on this question.

https://doi.org/10.7554/eLife.100056.3.sa2

Reviewer #3 (Public review):

Summary:

The authors aimed to investigate how the brain processes different linguistic units (from phonemes to sentences) in challenging listening conditions, such as multi-talker environments, and how this processing differs between individuals with normal hearing and those with hearing impairments. Using a hierarchical language model and EEG data, they sought to understand the neural underpinnings of speech comprehension at various temporal scales and identify specific challenges that hearing-impaired listeners face in noisy settings.

Strengths:

Overall, the combination of computational modeling, detailed EEG analysis, and comprehensive experimental design thoroughly investigates the neural mechanisms underlying speech comprehension in complex auditory environments.

The use of a hierarchical language model (HM-LSTM) offers a data-driven approach to dissect and analyze linguistic information at multiple temporal scales (phoneme, syllable, word, phrase, and sentence). This model allows for a comprehensive neural encoding examination of how different levels of linguistic processing are represented in the brain.

The study includes both single-talker and multi-talker conditions, as well as participants with normal hearing and those with hearing impairments. This design provides a robust framework for comparing neural processing across different listening scenarios and groups.

Weaknesses:

The study tests only a single deep neural network model for extracting linguistic features, which limits the robustness of the conclusions. A lower model fit does not necessarily indicate that a given type of information is absent from the neural signal-it may simply reflect that the model's representation was not optimal for capturing it. That said, this limitation is a common concern for data-driven, correlation-based approaches, and should be viewed as an inherent caveat rather than a critical flaw of the present work.

https://doi.org/10.7554/eLife.100056.3.sa1

Author response:

The following is the authors' response to the previous reviews

> ***eLife Assessment***
>
> *This valuable study combines a computational language model, i.e., HM-LSTM, and temporal response function (TRF) modeling to quantify the neural encoding of hierarchical linguistic information in speech, and addresses how hearing impairment affects neural encoding of speech. The analysis has been significantly improved during the revision but remain somewhat incomplete - The TRF analysis should be more clearly described and controlled. The study is of potential interest to audiologists and researchers who are interested in the neural encoding of speech.*

We thank the editors for the updated assessment. In the revised manuscript, we have added a more detailed description of the TRF analysis on p. of the revised manuscript. We have also updated Figure 1 to better visualize the analyses pipeline. Additionally, we have included a supplementary video to illustrate the architecture of the HM-LSTM model, the ridge regression methods using the model-derived features, and mTRF analysis using the acoustic envelop and the binary rate models.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *About R squared in the plots:*
>
> *The authors have used a z-scored R squared in the main ridge regression plots. While this may be interpretable, it seems non-standard and overly complicated. The authors could use a simple Pearson r to be most direct and informative (and in line with similar work, including Goldstein et al. 2022 which they mentioned). This way the sign of the relationships is preserved.*

We did not use Pearson's r as in Goldstein et al. (2022) because our analysis did not involve a train-test split, which was a key aspect of their approach. Specifically, Goldstein et al. (2022) divided their data into training and testing sets, trained a ridge regression model on the training set, and then used the trained model to predict neural responses on the test set. They calculated Pearson's r to assess the correlation between the predicted and observed neural responses, making the correlation coefficient (r) their primary measure of model performance. In contrast, our analysis focused on computing the model fitting performance (R^2) of the ridge regression model for each sensor and time point for each subject. At the group level, we conducted one-sample t-tests with spatiotemporal cluster-based correction on the R^2 values to identify sensors and time windows where R^2 values were significantly greater than baseline. We established the baseline by normalizing the R^2 values using Fisher z-transformation across sensors within each subject. We have added this explanation on p.13 of the revised manuscript.

> *About the new TRF analysis:*
>
> *The new TRF analysis is a necessary addition and much appreciated. However, it is missing the results for the acoustic regressors, which should be there analogous to the HM-LSTM ridge analysis. The authors should also specify which software they have utilized to conduct the new TRF analysis. It also seems that the linguistic predictors/regressors have been newly constructed in a way more consistent with previous literature (instead of using the HM-LSTM features); these specifics should also be included in the manuscript (did it come from Montreal Forced Aligner, etc.?). Now that*

> *the original HM-LSTM can be compared to a more standard TRF analysis, it is apparent that the results are similar.*

We used the Python package Eelbrain (*https://eelbrain.readthedocs.io/en/r0.39/auto_examples/temporal-response-functions/trf_intro.html*) to conduct the multivariate temporal response function (mTRF) analyses. As we previously explained in our response to R3, we did not apply mTRF to the acoustic features due to the high dimensionality of the input. Specifically, our acoustic representation consists of a 130-dimensional vector sampled every 10 ms throughout the speech stimuli (comprising a 129-dimensional spectrogram and a 1dimensional amplitude envelope). This led to interpreting the 130-dimensional TRF estimation difficult to interpret. A similar constraint applied to the hidden-layer activations from our HMLSTM model for the five linguistic features. After dimensionality reduction via PCA, each still resulted in 150-dimensional vectors. To address this, we instead used binary predictors marking the offset of each linguistic unit (phoneme, syllable, word, phrase, sentence). Since our speech stimuli were computer-synthesized, the phoneme and syllable boundaries were automatically generated. The word boundaries were manually annotated by a native Mandarin as in Li et al. (2022). The phrase boundaries were automatically annotated by the Stanford parser and manually checked by a native Mandarin speaker. These rate models are represented as five distinct binary time series, each aligned with the timing of the corresponding linguistic unit, making them well-suited for mTRF analysis. Although the TRF results from the 1-dimensional rate predictors and the ridge regression results from the high-dimensional HM-LSTM-derived features are similar, they encode different things: The rate regressors only encode the timing of linguistic unit boundaries, while the model-derived features encode the representational content of the linguistic input. Therefore, we do not consider the mTRF analyses to be analogous to the ridge regression analyses. Rather, these results complement each other and both provide informative results into the neural tracking of linguistic structures at different levels for the attended and unattended speech.

Since the TRF result for the continuous acoustic features also concerns R2, we have added an mTRF analysis where we fitted the one-dimensional speech envelope to the EEG. We extracted the envelope at 10 ms intervals for both attended and unattended speech and computed mTRFs independently for each subject and sensor using a basis of 50 ms Hamming windows spanning –100 ms to 300 ms relative to envelope onset. The results showed that in hearing-impaired participants, attended speech elicited a significant cluster in the bilateral temporal regions from 270 to 300 ms post-onset ($t = 2.40$, $p = 0.01$, Cohen's $d = 0.63$). Unattended speech elicited an early cluster in right temporal and occipital regions from –100 ms to –80 ms ($t = 3.07$, $p = 0.001$, $d = 0.83$). Normal-hearing participants showed significant envelope tracking in the left temporal region at 280–300 ms after envelope onset ($t = 2.37$, $p = 0.037$, $d = 0.48$), with no significant cluster for unattended speech. These results further suggest that hearing-impaired listeners may have difficulty suppressing unattended streams. We have added the new TRF results for envelope to Figure S3 and the "mTRF results for attended and unattended speech" on p.7 and the "mTRF analysis" in Material and Methods of the revised manuscript.

> *The authors' wording about this suggests that these new regressors have a nonzero sample at each linguistic event's offset, not onset. This should also be clarified. As the authors know, the onset would be more standard, and using the offset has implications for understanding the timing of the TRFs, as a phoneme has a different duration than a word, which has a different duration from a sentence, etc.*

In our rate-model mTRF analyses, we initially labelled linguistic boundaries as "offsets" because our ridge-regression with HM-LSTM features was aligned to sentence offsets rather than onsets. However, since each offset coincides with the next unit's onset—and our

regressors simply mark these transition points as 1—the "offset" and "onset" models yield identical mTRFs. To avoid confusion, we have relabeled "offset" as "boundary" in Figure S2.

As discussed in our prior responses, this design was based on the structure of our input to the HM-LSTM model, where each input consists of a pair of sentences encoded in phonemes, such as "t a_1 n əŋ_2 f ei_1 zh ə_4 sh iii_4 f ei_1 j ii_1" ("It can fly This is an airplane"). The two sentences are separated by a special token, and the model's objective is to determine whether the second sentence follows the first, similar to a next-sentence prediction task. Since the model processes both sentences in full before making a prediction, the neural activations of interest should correspond to the point at which the entire sentence has been processed by humans. To enable a fair comparison between the model's internal representations and brain responses, we aligned our neural analyses with the sentence offsets, capturing the time window after the sentence has been fully perceived by the participant. Thus, we extracted epochs from -100 to +300 ms relative to each sentence offset, consistent with our model-informed design.

We understand that phonemes, syllables, words, phrases, and sentences differ in their durations. However, the five hidden activity vectors extracted from the model are designed to capture the representations of these five linguistic levels across the entire sentence. Specifically, for a sentence pair such as "It can fly This is an airplane," the first 2048-dimensional vector represents all the phonemes in the two sentences ("t a_1 n əŋ_2 f ei_1 zh ə_4 sh iii_4 f ei_1 j ii_1"), the second vector captures all the syllables ("ta_1 nəŋ_2 fei_1 zhə_4 shiii_4 fei_1jii_1"), the third vector represents all the words, the fourth vector captures the phrases, and the fifth vector represents the sentence-level meaning. In our dataset, input pairs consist of adjacent sentences from the stimuli (e.g., Sentence 1 and Sentence 2, Sentence 2 and Sentence 3, and so on), and for each pair, the model generates five 2048-dimensional vectors, each corresponding to a specific linguistic level. To identify the neural correlates of these model-derived features—each intended to represent the full linguistic level across a complete sentence—we focused on the EEG signal surrounding the completion of the second sentence rather than on incremental processing. Accordingly, we extracted epochs from -100 ms to +300 ms relative to the offset of the second sentence and performed ridge regression analyses using the five model features (reduced to 150 dimensions via PCA) at every 50 ms across the epoch. We have added this clarification on p.12 of the revised manuscript.

> *About offsets:*
>
> *TRFs can still be interpretable using the offset timings though; however, the main original analysis seems to be utilizing the offset times in a different, more confusing way. The authors still seem to be saying that only the peri-offset time of the EEG was analyzed at all, meaning the vast majority of the EEG trial durations do not factor into the main HM-LSTM response results whatsoever. The way the authors describe this does not seem to be present in any other literature, including the papers that they cite. Therefore, much more clarification on this issue is needed. If the authors mean that the regressors are simply time-locked to the EEG by aligning their offsets (rather than their onsets, because they have varying onsets or some such experimental design complexity), then this would be fine. But it does not seem to be what the authors want to say. This may be a miscommunication about the methods, or the authors may have actually only analyzed a small portion of the data. Either way, this should be clarified to be able to be interpretable.*

We hope that our response in RE4, along with the supplementary video, has helped clarify this issue. We acknowledge that prior studies have not used EEG data surrounding sentence offsets to examine neural responses at the phoneme or syllable levels. However, this is largely due to a lack of model that represent all linguistic levels across an entire sentence. There is abundant work comparing model predictors with neural data time-locked to offsets because they mark the point at which participants has already processed the relevant

information (Brennan, 2016; Brennan et al., 2016; Gwilliams et al., 2024, 2025). Similarly, in our model– brain alignment study, our goal is to identify neural correlates for each model-derived feature. If we correlate model activity with EEG data aligned to sentence onsets, we would be examining linguistic representations at all levels (from phoneme to sentence) of the whole sentence at the time when participants have not heard the sentence yet. Although this limits our analysis to a subset of the data (143 sentences × 400 ms windows × 4 conditions), it targets the exact moment when full-sentence representations emerge against background speech, allowing us to examine each model-derived feature onto its neural signature. We have added this clarification on p.12 of the revised manuscript.

> ***Reviewer #2 (Public review):***
>
> *This study presents a valuable finding on the neural encoding of speech in listeners with normal hearing and hearing impairment, uncovering marked differences in how attention to different levels of speech information is allocated, especially when having to selectively attend to one speaker while ignoring an irrelevant speaker. The results overall support the claims of the authors, although a more explicit behavioural task to demonstrate successful attention allocation would have strengthened the study. Importantly, the use of more "temporally continuous" analysis frameworks could have provided a better methodology to assess the entire time course of neural activity during speech listening. Despite these limitations, this interesting work will be useful to the hearing impairment and speech processing research community. The study compares speech-in-quiet vs. multi-talker scenarios, allowing to assess within-participant the impact that the addition of a competing talker has on the neural tracking of speech. Moreover, the inclusion of a population with hearing loss is useful to disentangle the effects of attention orienting and hearing ability. The diagnosis of high-frequency hearing loss was done as part of the experimental procedure by professional audiologists, leading to a high control of the main contrast of interest for the experiment. Sample size was big, allowing to draw meaningful comparisons between the two populations.*

We thank you very much for your appreciation of our research and we have now added a more description of the mTRF analyses on p.13-14 of the revised manuscript.

> *An HM-LSTM model was employed to jointly extract speech features spanning from the stimulus acoustics to word-level and phrase-level information, represented by embeddings extracted at successive layers of the model. The model was specifically expanded to include lower level acoustic and phonetic information, reaching a good representation of all intermediate levels of speech. Despite conveniently extracting all features jointly, the HMLSTM model processes linguistic input sentence-by-sentence, and therefore only allows to assess the corresponding EEG data at sentence offset. If I understood correctly, while the sentence information extracted with the HM-LSTM reflects the entire sentence - in terms of its acoustic, phonetic and more abstract linguistic features - it only gives a condensed final representation of the sentence. As such, feature extraction with the HM-LSTM is not compatible with a continuous temporal mapping on the EEG signal, and this is the main reason behind the authors' decision to fit a regression at nine separate time points surrounding sentence offsets.*

Yes, you are correct. As explained in RE4, the model generates five hidden-layer activity vectors, each intended to represent all the phonemes, syllables, words, phrases within the entire sentence (“a condensed final representation”). This is the primary reason we extract EEG data surrounding the sentence offsets—this time point reflects when the full sentence has been processed by the human brain. We assume that even at this stage, residual neural responses corresponding to each linguistic level are still present and can be meaningfully analyzed.

> *While valid and previously used in the literature, this methodology, in the particular context of this experiment, might be obscuring important attentional effects impacted by hearing-loss. By fitting a regression only around sentence-final speech representations, the method might be overlooking the more "online" speech processing dynamics, and only assessing the permanence of information at different speech levels at sentence offset. In other words, the acoustic attentional bias between Attended and Unattended speech might exist even in hearing-impaired participants but, due to a lower encoding or permanence of acoustic information in this population, it might only emerge when using methodologies with a higher temporal resolution, such as Temporal Response Functions (TRFs). If a univariate TRF fit simply on the continuous speech envelope did not show any attentional bias (different trial lengths should not be a problem for fitting TRFs), I would be entirely convinced of the result. For now, I am unsure on how to interpret this finding.*

We agree and we have added the mTRF results using the rate models for the 5 linguistic levels in the prior revision. The rate model aligns with the boundaries of each linguistic unit at each level. As explained in RE3, the rate regressors encode the timing of linguistic unit boundaries, while the model-derived features encode the representational content of the linguistic input. The mTRF results showed similar patterns to those observed using features from our HM-LSTM model with ridge regression (see Figure S2). These results complement each other and both provide informative results into the neural tracking of linguistic structures at different levels for the attended and unattended speech.

We have also added TRF results fitting the envelope of attended and unattended speech at every 10 ms to the whole 10-minute EEG data at every 10 ms. Our results showed that in hearing-impaired participants, attended speech elicited a significant cluster in the bilateral temporal regions from 270 to 300 ms post-onset ($t = 2.40$, $p = 0.01$, Cohen's $d = 0.63$). Unattended speech elicited an early cluster in right temporal and occipital regions from –100 ms to –80 ms ($t = 3.07$, $p = 0.001$, $d = 0.83$). Normal-hearing participants showed significant envelope tracking in the left temporal region at 280–300 ms after envelope onset ($t = 2.37$, $p = 0.037$, $d = 0.48$), with no significant cluster for unattended speech. These results further suggest that hearing-impaired listeners may have difficulty suppressing unattended streams. We have added the new TRF results for envelope to Figure S3 and the "mTRF results for attended and unattended speech" on p.7 and the "mTRF analysis" in Material and Methods of the revised manuscript.

> *Despite my doubts on the appropriateness of condensed speech representations and singlepoint regression for acoustic features in particular, the current methodology allows the authors to explore their research questions, and the results support their conclusions. This work presents an interesting finding on the limits of attentional bias in a cocktail-party scenario, suggesting that fundamentally different neural attentional filters are employed by listeners with highfrequency hearing loss, even in terms of the tracking of speech acoustics. Moreover, the rich dataset collected by the authors is a great contribution to open science and will offer opportunities for re-analysis.*

We sincerely thank you again for your encouraging comments regarding the impact of our study.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *The authors aimed to investigate how the brain processes different linguistic units (from phonemes to sentences) in challenging listening conditions, such as multi-talker environments, and how this processing differs between individuals with normal hearing and those with hearing impairments. Using a hierarchical language model and EEG data, they sought to understand the neural underpinnings of speech comprehension at*

various temporal scales and identify specific challenges that hearing-impaired listeners face in noisy settings.

Strengths:

Overall, the combination of computational modeling, detailed EEG analysis, and comprehensive experimental design thoroughly investigates the neural mechanisms underlying speech comprehension in complex auditory environments. The use of a hierarchical language model (HM-LSTM) offers a data-driven approach to dissect and analyze linguistic information at multiple temporal scales (phoneme, syllable, word, phrase, and sentence). This model allows for a comprehensive neural encoding examination of how different levels of linguistic processing are represented in the brain. The study includes both single-talker and multi-talker conditions, as well as participants with normal hearing and those with hearing impairments. This design provides a robust framework for comparing neural processing across different listening scenarios and groups.

Weaknesses:

The analyses heavily rely on one specific computational model, which limits the robustness of the findings. The use of a single DNN-based hierarchical model to represent linguistic information, while innovative, may not capture the full range of neural coding present in different populations. A low-accuracy regression model-fit does not necessarily indicate the absence of neural coding for a specific type of information. The DNN model represents information in a manner constrained by its architecture and training objectives, which might fit one population better than another without proving the non-existence of such information in the other group. It is also not entirely clear if the DNN model used in this study effectively serves the authors' goal of capturing different linguistic information at various layers. More quantitative metrics on acoustic/linguistic-related downstream tasks, such as speaker identification and phoneme/syllable/word recognition based on these intermediate layers, can better characterize the capacity of the DNN model.

We agree that, before aligning model representations with neural data, it is essential to confirm that the model encodes linguistic information at multiple hierarchical levels. This is the purpose of our validation analysis: We evaluated the model's representations across five layers using a test set of 20 four-syllable sentences in which every syllable shares the same vowel—e.g., "mā ma mà mǎ" (mother scolds horse), "shū shu shǔ shù" (uncle counts numbers; see Table S1). We hypothesized that the activity in the phoneme and syllable layer would be more similar than other layers for same-vowel sentences. The results confirmed our hypothesis: Hidden-layer activity for same-vowel sentences exhibited much more similar distributions at the phoneme and syllable levels compared to those at the word, phrase and sentence levels Figure 3C displays the scatter plot of the model activity at the five linguistic levels for each of the 20 4-syllable sentences, post dimension reduction using multidimensional scaling (MDS). We used color-coding to represent the activity of five hidden layers after dimensionality reduction. Each dot on the plot corresponds to one test sentence. Only phonemes are labeled because each syllable in our test sentences contains the same vowels (see Table S1).The plot reveals that model representations at the phoneme and syllable levels are more dispersed for each sentence, while representations at the higher linguistic levels—word, phrase, and sentence—are more centralized. Additionally, similar phonemes tend to cluster together across the phoneme and syllable layers, indicating that the model captures a greater amount of information at these levels when the phonemes within the sentences are similar.

Apart from the DNN model, we also included the rate models which simply mark 1 at each unit boundaries across the 5 levels. We performed mTRF analyses with these rate models and

found similar patterns to our ridge-regression results with the DNN: (see Figure S2). This provides further evidence that the model reliably captures information across all five hierarchical levels.

> *Since EEG measures underlying neural activity in near real-time, it is expected that lower-level acoustic information, which is relatively transient, such as phonemes and syllables, would be distributed throughout the time course of the entire sentence. It is not evident if this limited time window effectively captures the neural responses to the entire sentence, especially for lower-level linguistic features. A more comprehensive analysis covering the entire time course of the sentence, or at least a longer temporal window, would provide a clearer understanding of how different linguistic units are processed over time.*

We agree that lower-level linguistic features may be distributed throughout the whole sentence, however, using the entire sentence duration was not feasible, as the sentences in the stimuli vary in length, making statistical analysis challenging. Additionally, since the stimuli consist of continuous speech, extending the time window would risk including linguistic units from subsequent sentences. This would introduce ambiguity as to whether the EEG responses correspond to the current or the following sentence. Additionally, our model activity represents a “condensed final representation” at the five linguistic levels for the whole sentence, rather than incrementally during the sentence. We think the -100 to 300 ms time window relative to each sentence offset targets the exact moment when full-sentence representations are comprehended and a “condensed final representation” for the whole sentence across five linguistic level have been formed in the brain. We have added this clarification on p.13 of the revised manuscript.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *Here are some specifics and clarifications of my public review:*
>
> *Initially I was interpreting the R squared as a continuous measure of predicted EEG relative to actual EEG, based on an encoding model, but this does not appear to be correct. Thank you for pointing out that the y axis is z-scored R squared in your main ridge regression plots. However, I am not sure why/how you chose to represent this that way. It seems to me that a simple Pearson r would be most informative here (and in line with similar work, including Goldstein et al. 2022 that you mentioned). That way you preserve the sign of the relationships between the regressors and the EEG. With R squared, we have a different interpretation, which is maybe also ok, but I also don't see the point of z-scoring R squared. Another possibility is that when you say "z-transformed" you are referring to the Fisher transformation; is that the case? In the plots you say "normalized", so that sounds like a z-score, but this needs to be clarified; as I say, a simple Pearson r would probably be best.*

We did not use Pearson’s r, as in Goldstein et al. (2022), because our analysis did not involve a train-test split, which was central to their approach. In their study, the data were divided into training and testing sets, and a ridge regression model was trained on the training set. They then used the trained model to predict neural responses on the held-out test set, and calculated Pearson’s r to assess the correlation between the predicted and observed neural responses. As a result, their final metric of model performance was the correlation coefficient (r). In contrast, our analysis is more aligned with standard temporal response function (TRF) approaches. We did not perform a train-test split; instead, we computed the model fitting performance (R^2) of the ridge regression model at each sensor and time point for each subject. At the group level, we conducted one-sample t-tests with spatiotemporal cluster-based correction on the R^2 values to determine which sensors and time windows showed significantly greater R^2 values than baseline. To establish a baseline, we z-scored the R^2

values across sensors and time points, effectively centering the distribution around zero. This normalization allowed us to interpret deviations from the mean R^2 as meaningful increases in model performance and provided a suitable baseline for the statistical tests. We have added this clarification on p.13 of the revised manuscript.

> *Thank you for doing the TRF analysis, but where are the acoustic TRFs, analogous to the acoustic results for your HM-LSTM ridge analyses? And what tools did you use to do the TRF analysis? If it is something like the mTRF MATLAB toolbox, then it is also using ridge regression, as you have already done in your original analysis, correct? If so, then it is pretty much the same as your original analysis, just with more dense timepoints, correct? This is what I meant by referring to TRFs originally, because what you have basically done originally was to make a 9-point TRF (and then the plots and analyses are contrasts of pairs of those), with lags between -100 and 300 ms relative to the temporal alignment between the regressors and the EEG, I think (more on this below).*
>
> *Also with the new TRF analysis, you say that the regressors/predictors had "a value of 1 at each unit boundary offset". So this means you re-made these predictors to be discrete as I and reviewer 3 were mentioning before (rather than using the HM-LSTM model layer(s)), and also, that you put each phoneme/word/etc. marker at its offset, rather than its onset? I'm also confused as to why you would do this rather than the onset, but I suppose it doesn't change the interpretation very much, just that the TRFs are slid over by a small amount.*

We used the Python package Eelbrain (*https://eelbrain.readthedocs.io/en/r0.39/auto_examples/temporal-response-functions/trf_intro.html*) to conduct the multivariate temporal response function (mTRF) analyses. As we previously explained in our response to Reviewer 3, we did not apply mTRF to the acoustic features due to the high dimensionality of the input. Specifically, our acoustic representation consists of a 130-dimensional vector sampled every 10 ms throughout the speech stimuli (comprising a 129-dimensional spectrogram and a 1-dimensional amplitude envelope). This renders the 130 TRF weights to the acoustic features uninterpretable. However, we have now added TRF results from the 1- dimension envelope to the attended and unattended speech at every 10 ms.

A similar constraint applied to the hidden-layer activations from our HM-LSTM model for the five linguistic features. After dimensionality reduction via PCA, each still resulted in 150-dimensional vectors, further preventing their use in mTRF analyses. To address this, we instead used binary predictors marking the offset of each linguistic unit (phoneme, syllable, word, phrase, sentence). These rate models are represented as five distinct binary time series, each aligned with the timing of the corresponding linguistic unit, making them well-suited for mTRF analysis. It is important to note that these rate predictors differ from the HM-LSTMderived features: They encode only the timing of linguistic unit boundaries, not the content or representational structure of the linguistic input. Therefore, we do not consider the mTRF analyses to be equivalent to the ridge regression analyses based on HM-LSTM features

For onset vs. offset, as explained RE4, we labelled them “offsets” because our ridge-regression with HM-LSTM features was aligned to sentence offsets rather than onsets (see RE4 and RE15 below for the rationale of using sentence offset). However, since each unit offset coincides with the next unit’s onset—and the rate model simply mark these transition points as 1—the “offset” and “onset” models yield identical mTRFs. To avoid confusion, we have relabeled “offset” as “boundary” in Figure S2.

> *I'm still confused about offsets generally. Does this maybe mean that the EEG, and each predictor, are all aligned by aligning their endpoints, which are usually/always the ends of sentences? So e.g. all the phoneme activity in the phoneme regressor actually*

> *corresponds to those phonemes of the stimuli in the EEG time, but those regressors and EEG do not have a common starting time (one trial to the next maybe?), so they have to be aligned with their ends instead?*

We chose to use sentence offsets rather than onsets based on the structure of our input to the HM-LSTM model, where each input consists of a pair of sentences encoded in phonemes, such as “t a_1 n əŋ_2 f ei_1 zh ə_4 sh iii_4 f ei_1 j ii_1” (“It can fly This is an airplane”). The two sentences are separated by a special token, and the model’s objective is to determine whether the second sentence follows the first, similar to a next-sentence prediction task. Since the model processes both sentences in full before making a prediction, the neural activations of interest should correspond to the point at which the entire sentence has been processed. To enable a fair comparison between the model’s internal representations and brain responses, we aligned our neural analyses with the sentence offsets, capturing the time window after the sentence has been fully perceived by the participant. Thus, we extracted epochs from -100 to +300 ms relative to each sentence offset, consistent with our modelinformed design. If we align model activity with EEG data aligned to sentence onsets, we would be examining linguistic representations at all levels (from phoneme to sentence) of the whole sentence at the time when participants have not heard the sentence yet. By contrast, aligning to sentence offsets ensures that participants have constructed a full-sentence representation.

We understand that it is a bit confusing why the regressor of each level is not aligned to their own offsets in the data. The hidden-layer activations of the HM-LSTM model corresponding to the five linguistic levels (phoneme, syllable, word, phrase, sentence) are consistently 150-dimensional vectors after PCA reduction. As a result, for each input sentence pair, the model produces five distinct hidden-layer activations, each capturing the representational content associated with one linguistic level for the whole sentence. We believe our -100 to 300 ms time window relative to sentence offset reflects a meaningful period during which the brain integrates and comprehends information across multiple linguistic levels.

> *Being "time-locked to the offset of each sentence at nine latencies" is not something I can really find in any of the references that you mentioned, regarding the offset aspect of this method. Can you point me more specifically to what you are trying to reference with that, or further explain? You said that "predicting EEG signals around the offset of each sentence" is "a method commonly employed in the literature", but the example you gave of Goldstein 2022 is using onsets of words, which is indeed much more in line with what I would expect (not offsets of sentences).*

You are correct that Goldstein (2022) aligned model predictions to onsets rather than offsets; however, many studies in the literature also align model predictions with unit offsets. typically because they mark the point at which participants has already processed the relevant information (Brennan, 2016; Brennan et al., 2016; Gwilliams et al., 2024, 2025). Similarly, in our study, we aim to identify neural correlates for each model-derived feature. If we correlate model activity with EEG data aligned to sentence onsets, we would be examining linguistic representations at all levels (from phoneme to sentence) of the whole sentence at the time when participants have not heard the sentence yet. By contrast, aligning to sentence offsets ensures that participants have constructed a full-sentence representation. Although this limits our analysis to a subset of the data (143 sentences × 400 ms windows × 4 conditions), it targets the exact moment when full-sentence representations emerge against background speech, allowing us to examine each model-derived feature onto its neural signature. We have added this clarification on p.12 of the revised manuscript.

> *This new sentence does not make sense to me: "The regressors are aligned to sentence offsets because all our regressors are taken from the hidden layer of our HM-LSTM model, which generates vector representations corresponding to the five linguistic levels of the entire sentence".*

Thank you for the suggestion. We hope our responses in RE4, 15 and 16, along with our supplementary video have now clarified the issue. We have deleted the sentence and provided a more detailed explanation on p.12 of the revised manuscript: The regressors are aligned to sentence offsets because our goal is to identify neural correlates for each model-derived feature of a whole sentence. If we align model activity with EEG data time-locked to sentence onsets, we would be finding neural responses to linguistic levels (from phoneme to sentence) of the whole sentence at the time when participants have not processed the sentence yet. By contrast, aligning to sentence offsets ensures that participants have constructed a full-sentence representation. Although this limits our analysis to a subset of the data (143 sentences × 2 sections × 400 ms windows), it targets the exact moment when full-sentence representations emerge against background speech, allowing us to examine each model-derived feature onto its neural signature. We understand that phonemes, syllables, words, phrases, and sentences differ in their durations. However, the five hidden activity vectors extracted from the model are designed to capture the representations of these five linguistic levels across the entire sentence Specifically, for a sentence pair such as “It can fly This is an airplane,” the first 2048dimensional vector represents all the phonemes in the two sentences (“t a_1 n əŋ_2 f ei_1 zh ə_4 sh iii_4 f ei_1 j ii_1”), the second vector captures all the syllables (“ta_1 nəŋ_2 fei_1 zhə_4 shiii_4 fei_1jii_1”), the third vector represents all the words, the fourth vector captures the phrases, and the fifth vector represents the sentence-level meaning. In our dataset, input pairs consist of adjacent sentences from the stimuli (e.g., Sentence 1 and Sentence 2, Sentence 2 and Sentence 3, and so on), and for each pair, the model generates five 2048dimensional vectors, each corresponding to a specific linguistic level. To identify the neural correlates of these model-derived features—each intended to represent the full linguistic level across a complete sentence—we focused on the EEG signal surrounding the completion of the second sentence rather than on incremental processing. Accordingly, we extracted epochs from -100 ms to +300 ms relative to the offset of the second sentence and performed ridge regression analyses using the five model features (reduced to 150 dimensions via PCA) at every 50 ms across the epoch.

> *More on the issue of sentence offsets: In response to reviewer 3's question about -100 - 300 ms around sentence offset, you said "Using the entire sentence duration was not feasible, as the sentences in the stimuli vary in length, making statistical analysis challenging. Additionally, since the stimuli consist of continuous speech, extending the time window would risk including linguistic units from subsequent sentence." This does not make sense to me, so can you elaborate? It sounds like you are actually saying that you only analyzed 400 ms of each trial, but that cannot be what you mean.*

Yes, we analyzed only the 400 ms window surrounding each sentence offset. Although this represents just a subset of our data (143 sentences × 400 ms × 4 conditions), it precisely captures when full-sentence representations emerge against background speech. Because our model produces a single, condensed representation for each linguistic level over the entire sentence—rather than incrementally—we think it is more appropriate to align to the period surrounding sentence offsets. Additionally, extending the window (e.g. to 2 seconds) would risk overlapping adjacent sentences, since sentence lengths vary. Our focus is on the exact period when integrated, level-specific information for each sentence has formed in the brain, and our results already demonstrate different response patterns to different linguistic levels for the two listener groups within this interval. We have added this clarification on p.13 of the revised manuscript.

> *In your mTRF analysis, you are now saying that the discrete predictors have "a value of 1" at each of the "boundary offsets", and those TRFs look very similar to your original plots. It sounds to me like you should not be referring to time zero in your original ridge analysis as "sentence offset". If what you mean is that sentence offset time is merely how you aligned the regressors and EEG in time, then your time zero still has a standard,*

> *typical TRF interpretation. It is just the point in time, or lag, at which the regressor(s) and EEG are aligned. So activity before zero is "predictive" and activity after zero is "reactive", to think of it crudely. So also in the text, when you say things like "50-150 ms after the sentence offsets", I think this is not really what you mean. I think you are referring to the lags of 50 - 150 ms, relative to the alignment of the regressor and the EEG.*

Thank you very much for the explanation. We agree that, in our ridge-regression time course, pre zero lags index “predictive” processing and post-zero lags index “reactive” processing. Unlike TRF analysis, we applied ridge regression to our high-dimensional model features at nine discrete lags around the sentence offset. At each lag, we tested whether the regression score exceeded a baseline defined as the mean regression score across all lags. For example, finding a significantly higher regression score between 50 and 150 ms suggests that our regressor reliably predicted EEG activity in that time window. So here time zero refers to the precise moment of the sentence offset—not the the alignment of the regressor and the EEG.

> *I look forward to discussing how much of my interpretation here makes sense or doesn't, both with the authors and reviewers.*

Thank you very much for these very constructive feedback and we hope that we have addressed all your questions.

https://doi.org/10.7554/eLife.100056.3.sa0