Leading and Following: Noise Differently Affects Semantic and Acoustic Processing during Naturalistic Speech Comprehension

Xinmiao Zhang[a, b], Jiawei Li[a, b], Zhuoran Li[a, b], Bo Hong[b c], Tongxiang Diao[d], Xin Ma[d], Guido Nolte[e], Andreas K. Engel[e], Dan Zhang[a, b, *]

a Department of Psychology, School of Social Sciences, Tsinghua University, Beijing 100084, China

b Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing 100084, China

c Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China

d Department of Otolaryngology, Head and Neck Surgery, People's Hospital, Peking University, Beijing 100044, China

e Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg Eppendorf, Hamburg 20246, Germany

* Corresponding author: Dan Zhang, dzhang@tsinghua.edu.cn

Abstract

While our daily verbal communication is challenging with unavoidable environmental noise, the neural mechanisms underlying speech-in-noise comprehension remain to be elucidated. The present study investigated the neural tracking of acoustic and semantic speech information during noisy naturalistic speech comprehension. Participants listened to narrative audios mixed with spectrally matched stationary noise at three signal-to-ratio (SNR) levels and 60-channel electroencephalography signals were recorded. A temporal response function (TRF) method was employed to derive event-related-like responses to the continuous speech stream at both the acoustic and the semantic levels. Whereas the amplitude envelope of the naturalistic speech was taken as the acoustic feature, word entropy and word surprisal were extracted via the natural language processing method as two semantic features. Theta-band frontocentral TRF responses to the acoustic feature were observed at around 400 ms following speech fluctuation onset over all three SNR levels, and the response latency was more delayed with increasing SNR. Delta-band frontal TRF responses to the semantic feature of word entropy were observed at around 200 to 600 ms leading to speech fluctuation onset over all three SNR levels, and the response latency became more leading with increasing SNR. While the following responses to speech acoustics were consistent with previous studies, the leading responses to speech semantics suggest our brain could be taking an active approach to deal with noisy speech comprehension.

Keywords: speech-in-noise comprehension, semantic processing, neural tracking, temporal response function, EEG

Highlight

1. With increasing noises, semantic prediction becomes increasingly advanced whereas acoustic processing becomes increasingly delayed.
2. Brain regions involved in semantic prediction exhibit a left-lateralized pattern in no- and low-noise conditions, and a shift in dominance in high-noise condition.
3. Delta and theta band neural tracking is related to acoustic and semantic processing, respectively.

1 Introduction

Noise is an inevitable accompaniment of real-life verbal communication, such as car horns on the streets, background music at parties, etc. Even though noise impairs auditory information, individuals with normal hearing may comprehend speech effectively. Reliable speech comprehension in a noisy environment plays a crucial role in our daily interactions. Understanding the neural mechanisms of noisy speech comprehension has attracted increasing interest during the past decades.

The neurophysiological studies using the conventional event-related technique have revealed important insights into how noise affects speech processing. The cortical auditory evoked potential (CAEP), such as the P1-N1-P2 complex in frontal-central electrodes, and MMN in central electrodes elicited by auditory and speech stimuli, have been found to show delayed latency and reduced amplitude under adverse conditions, which reflected impaired efficiency of neural processing (Bidelman & Howell, 2016; Billings et al., 2009, 2011; Kaplan-Neeman et al., 2006; Koerner et al., 2016; Koerner & Zhang, 2015; B. A. Martin & Stapells, 2005; Parbery-Clark et al., 2011; Whiting et al., 1998). In addition to auditory processing, semantic processing also plays a vital role in speech-in-noise comprehension and has been paid substantial emphasis. Semantic processing is closely related to several high-level cognitive capacities, such as knowledge representation (Bi, 2021; Patterson et al., 2007) and working memory (Loaiza & Camos, 2018; Shivde & Anderson, 2011), thus, contributes to context-based prediction and integration (DeLong et al., 2014; Federmeier, 2007; Hutchison, 2003; Kutas & Federmeier, 2011; R. C. Martin, 2021; Otten & Van Berkum, 2009). Furthermore, numerous studies have shown that semantic-based contextual processing is crucial for reliable comprehension in noise (Davis et al., 2011; Mattys et al., 2012; Miller et al., 1951; Obleser, 2014; Obleser & Kotz, 2010, 2011; Peelle & Wingfield, 2016; Pickering & Gambi, 2018; Zekveld et al., 2011). For example, Miller (1951) found that words in coherent sentences had higher intelligibility compared with the same words in unrelated word lists during speech-in-noise comprehension. Regarding the influence of noise on semantic processing such as the N400 component, existing studies have reported that as the noise level increased, the latency of N400 was delayed (Connolly et al., 1992; Daltrozzo et al., 2012), while the amplitude of N400 was increased (Romei et al., 2011; Zendel et al., 2015), decreased (Aydelott et al., 2006; Obleser & Kotz, 2011; Strauß et al., 2013), or exhibited a nonlinear trend (Jamison et al., 2016). These mixed studies provided valuable information on the complex relationship between noise

and cortical speech processing. However, the event-related design has restricted the speech stimuli in the form of highly-controlled and short-duration speech units, such as pure tones (e.g., Bennett et al., 2012; Billings et al., 2009), syllables (e.g., Bidelman & Howell, 2016; Koerner & Zhang, 2015) or simple sentences (e.g., Strauß et al., 2013). These materials contain limited semantic/contextual information, resulting in inadequate exploration of semantic processing under noise.

The recent rise of the naturalistic speech paradigm is expected to expand our knowledge of the neural mechanisms of the speech-in-noise comprehension (Alday, 2019; Z. Li & Zhang, 2023; Sonkusare et al., 2019; Wöstmann et al., 2017). Compared to the highly-controlled and short-duration speech units, the naturalistic speech stimuli usually have a longer duration and are more flexible in content, both of which provide a better resemblance to our daily communications (Alday, 2019; Alexandrou et al., 2020; Hartley & Poeppel, 2020; Sonkusare et al., 2019; Wöstmann et al., 2017). Most of all, the continuous naturalistic speech stimuli provide rich semantic-based contextual information (Alday, 2019; Alexandrou et al., 2020; Hamilton & Huth, 2020; Sonkusare et al., 2019), which is indispensable for reliable speech comprehension in chaotic daily environment (Peelle & Wingfield, 2016) but have not been sufficiently addressed in the conventional event-related studies. Mounting studies have recently embraced this promising approach. To analyze the neural activities related to the continuous naturalistic speech stimuli, researchers have mainly focused on the neural tracking (entrainment) of the speech stimuli by neural oscillations (Keitel et al., 2017; Zion Golumbic et al., 2013). The oscillatory neural activities at below 8 Hz have been frequently reported to be functionally important for the naturalistic speech processing (Brodbeck & Simon, 2020; Ding & Simon, 2014; Giraud & Poeppel, 2012; Obleser & Kayser, 2019). To further describe the spatiotemporal dynamics for naturalistic speech processing, system identification methods such as the temporal response function (TRF) method (Crosse et al., 2016, 2021) are adopted, and have been shown to be an effective tool to investigate neural correlates of speech processing in various conditions (e.g., Di Liberto et al., 2015; Fiedler et al., 2019; O'Sullivan et al., 2015). Existing TRF-based studies have provided valuable insights into naturalistic speech-in-noise comprehension. Compared to no-noise speech, noisy-speech-related TRF responses were found to have reduced amplitude and delayed latency (Ding & Simon, 2013; Gillis, Decruy, et al., 2022; Muncke et al., 2022; Zou et al., 2019), which was similar to previous event-related studies. However, while most of these studies focused on acoustic processing by relating the neural signal to

acoustic features such as amplitude envelope (e.g., Ding & Simon, 2013), the semantic processing of naturalistic speech during speech-in-noise comprehension has yet to be clarified. Although naturalistic stimuli provide rich semantic information, it also poses challenges to accurately quantifying semantic features in a continuous stream of information.

Recent advancements in natural language processing (NLP) methods have made computational linguistic models a powerful tool for understanding the processing of semantic-level speech information. With these models, semantic features in naturalistic speech can be precisely quantified, and neural tracking of the semantic information can be measured (Broderick et al., 2018, 2019, 2021; Gillis et al., 2021; Goldstein et al., 2022; Koskinen et al., 2020; Mesik et al., 2021; Weissbart et al., 2020; Willems et al., 2016). The state-of-art computational linguistic models can be used to extract a number of important semantic features. The two frequently used semantic features in speech-related neuroscience research are entropy and surprisal derived from information theory, which respectively measures the semantic uncertainty of the upcoming word and the unexpectedness of the current word (Willems et al., 2016). The entropy was found associated with neural activity within the left ventral premotor cortex, left middle frontal gyrus and right inferior frontal gyrus, etc. (Willems et al., 2016) and the neural activity evoked by entropy occurred both early at around 100 ms and late at around 500 ms after speech onset (Donhauser & Baillet, 2020; Weissbart et al., 2020). Furthermore, Goldstein et al., (2022) derived word entropy from deep language models (GPT-2) and correlated them with electrocorticography (ECoG) signals. The results indicated that entropy was related to neural activation in left-lateralized channels at several hundred milliseconds before the word onset. This leading/pre-onset response is consistent with the semantic prediction potential (SPP) in event-related studies, as a direct neural signature for semantic prediction (Grisoni et al., 2021; Pulvermüller & Grisoni, 2020) The surprisal was found associated with the superior temporal gyrus and inferior frontal sulcus, etc. (Willems et al., 2016), and is related to an N400-like neural response, i.e., negativity at around 400 ms within central-parietal electrodes (Broderick et al., 2021; Gillis et al., 2021; Goldstein et al., 2022, 2022; Heilbron et al., 2022; Mesik et al., 2021). These findings imply that the entropy- and surprisal-related neural activity may reflect both top-down and bottom-up semantic processing and proved the validity of these two semantic features. With the potent tool, NLP, the

investigation of how noise impacts semantic processing during naturalistic speech comprehension is anticipated to be advanced.

The current study aimed to investigate the neural mechanisms of speech-in-noise comprehension by simultaneously focusing on both the acoustic and semantic levels. A naturalistic speech comprehension paradigm was employed, as the naturalistic speech stimuli were expected to provide better ecological validity and contextual information (Alday, 2019; Sonkusare et al., 2019). 60-channel EEGs were recorded while the participants listened to spoken narratives at three signal-to-noise ratios (SNRs). The amplitude envelops of the speech stimuli were extracted as the acoustic feature, following previous studies (Di Liberto et al., 2015; O'Sullivan et al., 2015). Two typical semantic features, i.e., word entropy and word surprisal, were calculated by using a Chinese NLP model (Gillis et al., 2021; Weissbart et al., 2020; Willems et al., 2016; Koskinen et al., 2020; Mesik et al., 2021; Broderick et al., 2021). The neural responses to the acoustic and semantic features were estimated using the TRF method (Crosse et al., 2016), which yielded the spatiotemporal dynamics of how our brain tracked these features in naturalistic speeches. We hypothesize that the acoustic-level TRF could be related to delayed peak and reduced amplitude under noisy conditions compared with the no-noise condition to replicate previous studies (Ding & Simon, 2013; Zou et al., 2019). As for the semantic processing, we hypothesize that noise could affect the latency and amplitude of both the leading/pre-onset and the following/post-onset stages (Etard & Reichenbach, 2019; Goldstein et al., 2022). By simultaneously exploring both acoustic and semantic processing, as well as both the pre-onset and the post-onset stages, this study hopes to gain a more complete overview of the noise effect on neural processing during naturalistic speech comprehension.

2 Methods

2.1 Participants

Twenty college students (10 females, ages ranging from 19 to 28 years old) participated in the study. The sample size was determined to be sufficient by reference to previous TRF-based studies on the human speech processing (Broderick et al., 2018; J. Li et al., 2021). One male participant was excluded due to technical problems during data recording. The data of the remaining nineteen participants (age: mean ± SD = 21.80 ± 1.20) were included in the subsequent analyses. All participants were native Chinese speakers, right-handed, with normal hearing and normal or corrected-to-normal vision by self-report. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local Ethics Committee of Tsinghua University. Written informed consent was obtained from all participants.

2.2 Materials

Thirty narrative audios from our previous studies (Z. Li et al., 2021, 2022) were used as the stimuli. These audios were recorded from six native Chinese speakers with professional training in broadcasting. The participants were not familiar with the content of these narrative audios, which were about speakers' personal experiences on daily-life topics adapted from the National Mandarin Proficiency Test. Each narrative audio lasted for around 90 s and was recorded by a regular microphone at a sampling rate of 44,100 Hz in a sound-attenuated room.

These speech stimuli were further processed into three versions at three different SNR levels: no-noise (NN), low-noise (SNR = 3 dB), and high-noise (SNR = -3 dB) by adding spectrally matched stationary noise. The spectrally matched stationary noise was generated based on a 50th-order linear predictive coding (LPC) model estimated from the original speech recording (Broderick et al., 2018). The SNR levels were selected following previous studies (Ding & Simon, 2013; Z. Li et al., 2021), and were produced by varying the noise intensity while keeping the intensity of original speech (assessed by its root mean square) constant (Ding & Simon, 2013).

For each narrative audio, two four-choice questions were prepared by the experimenters for the assessment of one's speech comprehension performance. These questions and the corresponding choices were targeted at detailed narrative contents that would demand significant attentional efforts. For instance,

one question following a narrative audio about one's major was, "What is the speaker's most likely major as a graduate student? (说话人的研究生专业最可能是什么？)" and the four choices were 1) Social science, 2) International politics, 3) Pedagogy and 4) Psychology (1. 社会科学, 2. 国际政治, 3. 教育学 and 4. 心理学).

2.3 Procedure

Before the start of the experiment, the participants had one practice trial to get familiar with the procedure, with an additional speech audio stimulus not used in the formal experiment at the no-noise level. The formal experiment consisted of 30 trials, with 10 trials per noise level. In each trial, the participants listened to a narrative audio at one of the three noise levels. The participants were required to maintain visual fixation on a fixation cross displayed on the computer screen in front of them and to minimize eye blinks and all other motor activities during listening. The order of the narrative audios and their assigned noise levels was randomized for each participant.

After each trial, the participants were instructed to answer two four-choice questions about the content of the narrative audio using the computer keyboard. The averaged accuracies across all trials (separately for each noise level) were used to reflect the participants' comprehension performance. After completing these questions, the participants were instructed to rate the clarity and intelligibility of the narrative audio on 100-point rating scales and rested for at least 5 s before moving on to the next trial. No feedback was given to the participants about their performance during the experiment.

The experimental procedure was programmed in MATLAB using the Psychophysics Toolbox 3.0 (Brainard, 1997). The speech stimuli were delivered to listeners seated in a sound-attenuated room via an air-tube earphone (Etymotic ER2, Etymotic Research, Elk Grove Village, IL, USA) to avoid environmental

noise and equipment electromagnetic interference. All speech stimuli were presented at a comfortable level. The experimental procedure is illustrated in Fig. 1.

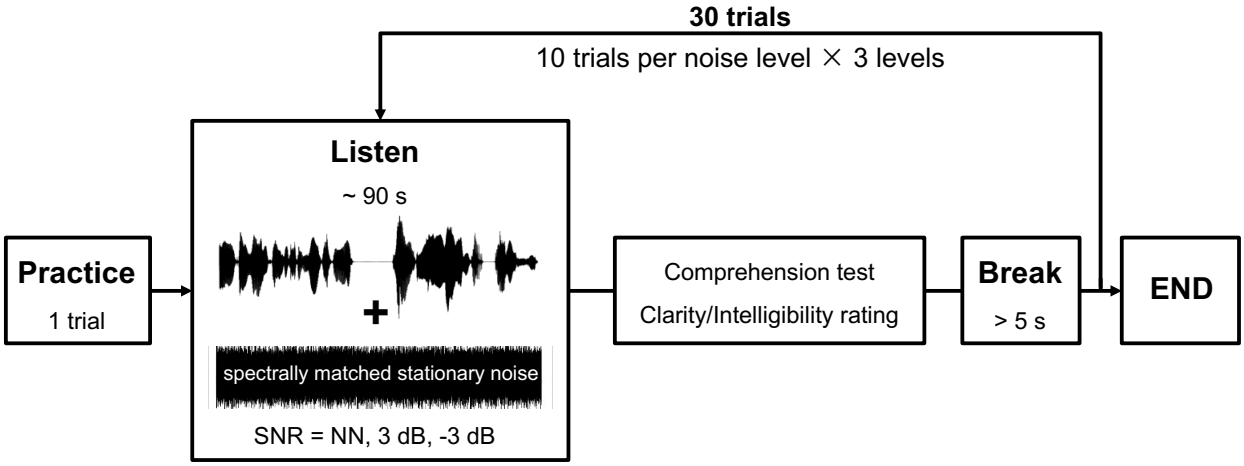

Fig. 1. The experimental procedure. The participants listened to 30 naturalistic narrative audios which each lasted around 90 s. These audios were mixed with three levels of spectrally matched stationary noise: no-noise (NN), low-noise (3 dB), and high-noise (-3 dB). The 60-channel EEG signals were recorded during listening. After listening to each audio, the participants were required to complete a comprehension test and reported the clarity and intelligibility rating. For the comprehension test, two four-choice questions per speech audio based on the narrative content were used.

2.4 EEG recording and preprocessing

EEG signals were recorded from 60 channels with a NeuroScan amplifier (SynAmp II, NeuroScan, Compumedics, USA) at a sampling rate of 1000 Hz. Electrodes were positioned according to the international 10–20 system, including FP1/2, FPZ, AF3/4, F7/8, F5/6, F3/4, F1/2, FZ, FT7/8, FC5/6, FC3/4, FC1/2, FCZ, T7/8, C5/6, C3/4, C1/2, CZ, TP7/8, CP5/6, CP3/4, CP1/2, CPZ, P7/8, P5/6, P3/4, P1/2, PZ, PO7/8, PO5/6, PO3/4, POZ, Oz, O1/2, referenced to an electrode between CZ and CPZ with a forehead ground at FZ. Electrode impedances were kept below 10 kOhm for all electrodes throughout the experiment.

The recorded EEG data were first notch filtered to remove the 50 Hz powerline noise. Independent Component Analysis (ICA) was performed to remove artifacts such as eye blinks and eye movements based on visual inspection. Around 4–12 independent components (ICs; mean = 6.6) were removed per participant. The remaining ICs were then back-projected onto the scalp EEG channels to reconstruct the artifact-free EEG signals. The EEG signals were then re-referenced to the average of all scalp channels and down-sampled to 128 Hz. Afterward, EEG signals were filtered into delta (1–4 Hz) and theta (4–8 Hz) bands, which have been previously reported to be important for speech processing (Ding et al., 2014; Ding &

Simon, 2014; Etard & Reichenbach, 2019; Fiedler et al., 2019; Keitel et al., 2017; Koskinen et al., 2020; Lu et al., 2022; Peelle et al., 2013). Causal FIR (Finite Impulse Response) filters were employed to ensure that the filtered EEG signals were determined only by the current and previous data samples, which was important for the present study focusing on the fine-grained time course, particularly considering the pre-onset neural responses (de Cheveigné & Nelken, 2019; Etard & Reichenbach, 2019).

These preprocessed EEG signals were segmented into 30 trials, from 5 to 90 s (duration = 85s) relative to the speech onsets of each trial to avoid possible onset and offset effects. All preprocessing was conducted offline using MATLAB and the Fieldtrip toolbox (Oostenveld et al., 2011).

2.5 Feature extraction

Three types of features were extracted to represent the acoustic (amplitude envelop) and semantic (word entropy, word surprisal) information for each narrative audio. An example of these speech features is illustrated in Fig. 2A.

Acoustic feature. The amplitude envelope for each narrative audio was calculated as the absolute values after a Hilbert transform and then down-sampled to the sampling rate of 128 Hz to match that of the EEG signals.

Semantic features. Before feature extraction, the narrative audios were converted to text by Iflyrec software (Iflytek Co., Ltd, Hefei, Anhui) and then segmented into words based on the THU Lexical Analyzer for Chinese (THULAC) toolbox (Sun et al., 2016).

Two semantic features, word entropy and word surprisal, were extracted. Word entropy measures the uncertainty of predicting the upcoming word based on the context so far and was calculated as equation (1):

$$Entropy(t) = -\sum_{w_{t+1}} P(w_{t+1}|w_1, \dots, w_t) logP(w_{t+1}|w_1, \dots, w_t) \quad (1)$$

Word surprisal measures how surprising the current word is given the previously encountered words and was calculated as equation (2):

$$Surprisal(t) = -logP(w_t|w_1, \dots, w_{t-1}) \quad (2)$$

Where $w_1, \dots, w_{t-1}$ are the existing word sequence and $P(w_t|w_1, \dots, w_{t-1})$ is the conditional probability of next word (Willems et al., 2016). These NLP calculations were conducted by ADAM, a widely accepted

Long-Short Term Memory (LSTM) Neural Network model (Kingma & Ba, 2015). The model was trained on the corpora corpus of the People's Daily with 534,246 Chinese words. See Supplementary Table S1 for more information of the model.

After extracting the semantic features of each word, the word onset timings were estimated via Iflyrec software (Iflytek Co., Ltd, Hefei, Anhui). Impulses at the word onset time were manipulated with corresponding semantic feature values to generate one-dimensional "semantic vectors" (Broderick et al., 2018; Gillis et al., 2021). The sampling rate of the semantic vectors was 128 Hz to match the EEG signals.

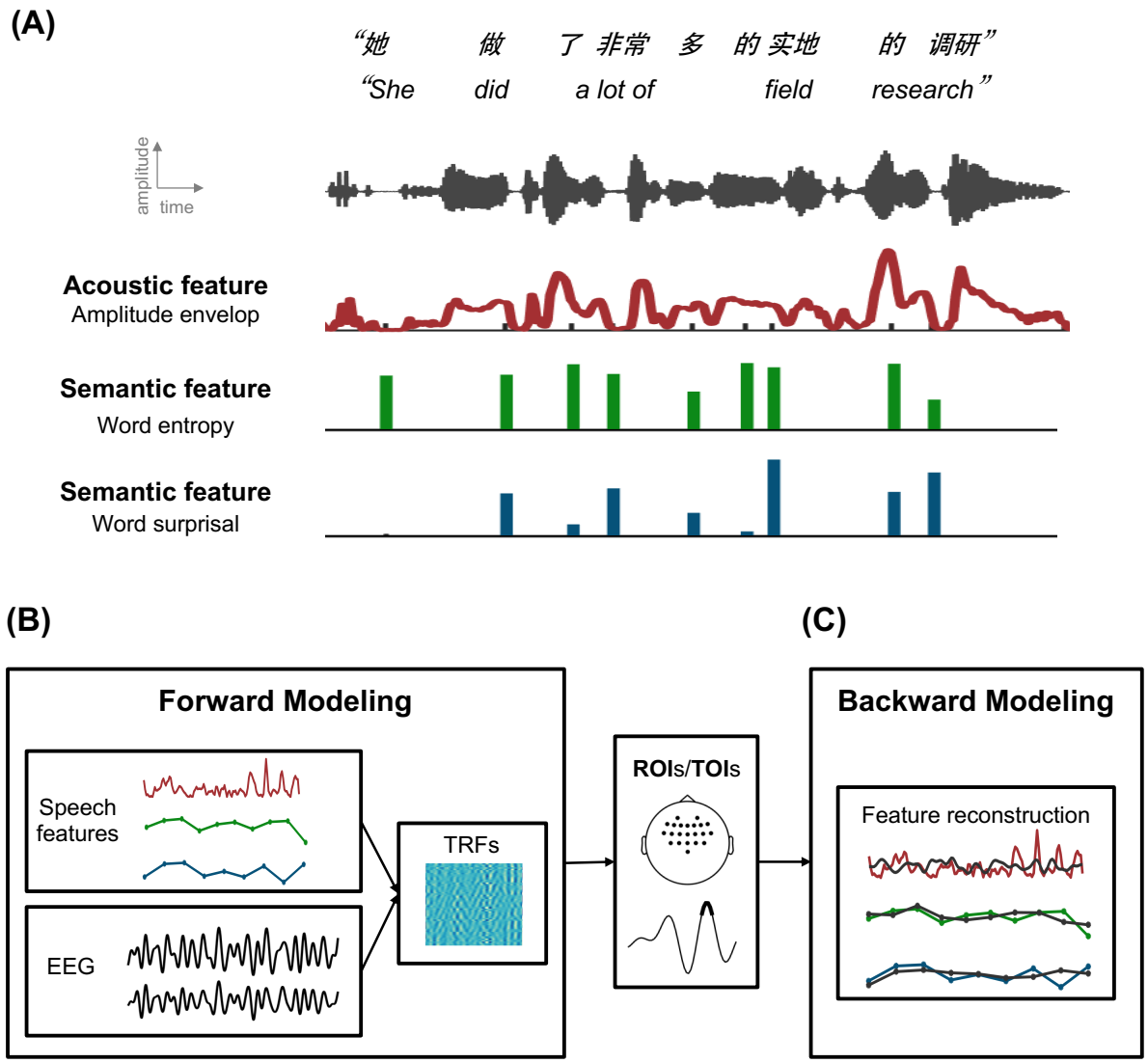

Fig. 2. Speech feature extraction and Temporal Response Function analysis. (A) Three types of speech features were extracted, including one acoustic feature (amplitude envelop) and two semantic features (word entropy and word surprisal). The two semantic features were derived from a computational linguistic model and one-dimensional vectors with impulses manipulated with semantic feature values at corresponding word onsets time were generated. (B) Forward modeling. TRFs were extracted by regressing

each of the three types of speech features against the EEG signals. The significance of these TRFs was estimated by comparing them with the corresponding null TRFs, which were modeled based on EEG signals and shuffled speech features. The resulting spatiotemporal ranges were identified as regions of interest (ROIs) and time of interest (TOIs). (C) Backward modeling. These three types of speech features were separately reconstructed through backward TRF models and the reconstruction accuracy (Pearson's r) depicted the strength of neural tracking. Null backward models were constructed with EEG signals and shuffled features.

2.6 Modeling of the stimulus-response relationship

The Temporal Response Function (TRF) based modeling method was adopted to explore the relationship between the neural activities and the three types of stimulus features (Crosse et al., 2016, 2021). Forward modeling was first used to illustrate the specific spatial-temporal response patterns and identify key electrodes and time windows in TRF responses of the corresponding speech feature, and then backward modeling was adopted to verify the possible contribution of these identified neural correlates (Broderick et al., 2019; Etard & Reichenbach, 2019). The overall procedure of the modeling analysis is shown in Fig. 2B and 2C.

Forward modeling. With a forward modeling approach, we described neural response patterns to different speech features by linear spatiotemporal filters called TRFs, which measure how the neural signals from different regions are modulated by stimulus features at different time lags (Crosse et al., 2016). The estimated TRF together with the corresponding speech feature was used to predict the EEG response from each electrode. The prediction accuracy measured as Pearson's correlation between the actual and predicted EEG signals represented the performance of the forward model. The TRF, w, can be expressed by equation (3):

$$w = (S^T S + \lambda I)^{-1} S^T r \tag{3}$$

Where S is the lagged time series of the stimulus features and I is the identity matrix. The time lags for forward modeling were chosen to cover a relatively broad time range, from -1000 to 1000 ms (J. Li et al., 2021, 2022). The λ is the ridge parameter used to prevent overfitting and ranged between 0.1 and 1000 in steps of the powers of 10 empirically (Gillis, Van Canneyt, et al., 2022). The cross-validation procedure was implemented in a leave-one-trial-out manner within each participant: each time the model was trained based on 9 trials and tested on the left-out trial, which was repeated for each of the 10 trials at three noise

levels separately. The λ value that produced the highest prediction accuracy averaged across trials after cross-validation was selected as the regularization parameter for all trials at a certain noise level per participant. TRF amplitude was further transformed into *z*-scores before the statistical analysis (Di Liberto et al., 2015; Ding et al., 2014; Gillis et al., 2021; J. Li et al., 2021).

The statistical significance of the estimated TRFs was estimated by constructing null TRF models (Gillis et al., 2021; Weissbart et al., 2020). We built the chance-level null models by constructing TRF models using shuffled stimulus features and the EEG recordings in the same way as for the computation of the actual TRFs. The shuffled amplitude envelope was constructed by randomly shuffling the feature value within a trial while keeping the timing of the quiet fragments. The shuffled word entropy and word surprisal were constructed by randomly shuffling the feature value within a trial while keeping the timing of impulses. Therefore the speech features that described acoustic and linguistic word onsets were not altered and thus left no impact on TRFs' significance (Weissbart et al., 2020). The shuffling was repeated 1,000 times and resulted in 1,000 null TRFs for a corresponding actual TRF.

Two-tailed one-sample *t*-tests were used to examine whether the amplitude on an electrode-time bin for the actual and null TRFs was significantly different from 0 (Gillis et al., 2021). We got the *p*-value of the actual TRF amplitude on a certain electrode-time bin based on the actual *t*-value and the distribution of the 1,000 corresponding null *t*-values. The same procedure was conducted on each electrode-time bin and a cluster-based analysis was employed for the multi-comparison correction (Maris & Oostenveld, 2007). In this procedure, neighboring channel-time bins with an uncorrected *p*-value below 0.01 were clustered together. Cluster-level statistics were calculated by summing the *t*-values in each cluster separately. The corrected *p*-values for each cluster were calculated on the permutation distribution of the maximum cluster-level statistic by the Monte Carlo method. We implemented the same statistical analysis procedure for each of the 18 TRFs (3 stimulus features × 2 frequency bands × 3 noise levels). All the significant clusters were then corrected together via the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995).

Backward modeling. With a backward modeling approach, we took neural signals from several electrodes simultaneously to reconstruct stimulus features with a decoder. The reconstruction accuracy was measured as Pearson's correlation between the actual and reconstructed stimulus features. The decoder, *g,* can be expressed by equation (4) where *R* is the lagged time series of the EEG data. The reconstructed

feature, $\hat{s}(t)$, be calculated by equation (5) where n is the EEG electrodes, and τ is the time lags (Broderick et al., 2019).

$$g = (R^T R + \lambda I)^{-1} R^T s \tag{4}$$

$$\hat{s}(t) = \sum_n \sum_\tau r_n(t+\tau) g_n(\tau) \tag{5}$$

The EEG electrodes and time windows that participated in backward modeling were chosen based on significant clusters found in the forward modeling approach. The same regularization and cross-validation procedure as in the forward modeling approach was conducted. We also estimated the null decoders through the same shuffling methods in forward modeling. The reconstruction accuracy from the 1,000 null decoders was averaged and compared with the actual model was compared with the actual decoder via a one-tailed paired *t*-test and the *p* values for all clusters were corrected via the FDR method.

In sum, the analysis of neural tracking followed two main steps. (1) We calculated the actual forward model and null forward models and identified ROIs/TOIs according to clusters with significant differences between them. (2) We estimated the reconstruction accuracy based on the ROIs/TOIs in neural tracking. This procedure resulted in (1) the specific spatiotemporal TRF response pattern and (2) the strength of neural tracking (reconstruction accuracy) for analysis.

To avoid missing any results, significant clusters in acoustic- and semantic-level TRF responses at each of the three noise levels were detected separately. We classified these clusters into two types based on their spatiotemporal dynamics: those with consistent patterns across all noise levels, which represent a reliable response at all noise levels, and those with unique patterns at a certain noise level, which may signify distinct processing mechanisms under certain circumstances. The consistent clusters were compared to explore how the commonly shared neural signature was affected by noise while the unique clusters received less attention in subsequent analysis.

Before modeling, the three types of stimulus features across all trials and EEG signals across all channels were z-scored as recommended to ensure consistent scaling and optimize the ridge parameter in the cross-validation (Crosse et al., 2016, 2021). Modeling and analysis for different stimulus features were conducted independently. Considering we cared about the neural tracking of underlying hierarchical information in speech rather than physical stimulus, we adopted the same stimulus features of no-noise speech for noise levels (Ding & Simon, 2013; Fuglsang et al., 2017). The forward and backward modeling

was conducted in MATLAB using the Multivariate Temporal Response Function (mTRF) toolbox (Crosse et al., 2016). The cluster-based correction was conducted in the FieldTrip toolbox (Oostenveld et al., 2011). Other statistical analysis was conducted via MATLAB functions and the software JASP (JASP Team, 2022).

3 Results

3.1 Behavioral performance

The speech comprehension performance was measured as the averaged response accuracy of the four-choice questions and was found significantly different among the three noise levels (rmANOVA, $F(2, 36) = 6.74$, $p = .003$). The speech comprehension performances were 95.26 ± 1.05%, 90.79 ± 2.10%, and 85.26 ± 2.63% (mean ± SE) at the no-noise level, low-noise level (3 dB), and high-noise level (-3 dB), respectively. The comprehension performance at the high-noise level was significantly lower than that at the no-noise level (post-hoc *t*-test, $p = .002$, Bonferroni corrected). In addition, it should be noted that the comprehension performance was still well above chance level even at the high-noise level (one-tailed *t*-test, $t(18) = 32.36$, $p < 0.001$).

The subjective ratings of clarity and intelligibility showed a similar pattern with significant differences among the noise levels (rmANOVA, $F(2, 36) = 144.84$ and 35.90, *p*s < .001). The clarity rating scores were 93.10 ± 1.53, 64.88 ± 3.95, and 34.76 ± 4.25 (mean ± SE), and the intelligibility rating scores were 92.12 ± 1.35, 87.01 ± 1.93, and 72.54 ± 3.51 (mean ± SE) at the no-noise, low-noise and high-noise level, respectively. Post hoc *t*-tests revealed significant pairwise differences for all possible comparisons (*p*s < .001, Bonferroni corrected) except for the intelligibility ratings between the no-noise and low-noise levels. The behavioral performances are illustrated in Fig. 3. These results suggested that the effect of noise on speech comprehension and perception was effectively manipulated.

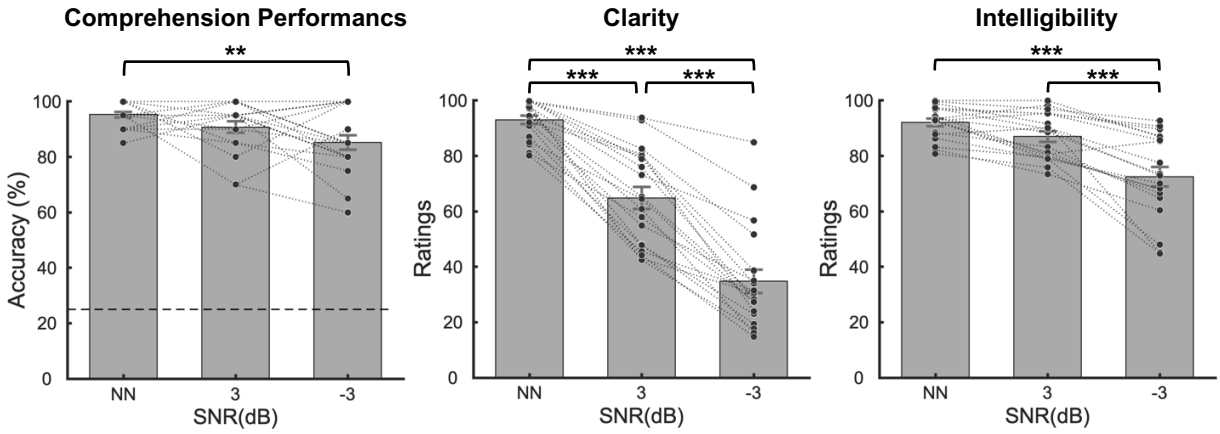

Fig. 3. Behavioral results. Black dots indicate data from each participant. Error bars denote the standard error. n.s. means not significant, † means $p < .1$, * means $p < .05$, ** means $p < .01$, ***means $p < .001$.

3.2 All significant clusters of acoustic- and semantic-level TRF responses

All significant clusters of acoustic- and semantic-level TRF responses are summarized in Fig. 4. Acoustic-level TRF responses occurred in the theta and delta bands whereas semantic-level (word entropy) TRF neural responses were dominated by the delta band. No significant TRF responses to word surprisal were found. Both the acoustic- and semantic-level TRF responses had several clusters with unique spatiotemporal dynamics. They all appeared in the no-noise condition and were dominated by the delta band. Theta-band acoustic-level TRF response within the occipital electrodes, i.e., A3 and A5, occurred in the low- and high-noise conditions but not the no-noise condition. Analysis of the peak latency and amplitude of the TRFs from the occipital ROIs indicated no significant difference between the low and high-noise conditions (paired-samples *t*-test, $t(18) = 0.51, 1.52$, $p = .617, .146$). See Supplementary Fig. S1 for more information. Post-onset semantic-level TRF responses within the central electrodes, i.e., E3 and E7, occurred in the no- and low-noise levels but not the no-noise condition. See Supplementary Fig. S2 for more information. Clusters with largely consistent spatiotemporal dynamics across all three noise conditions were compared as below.

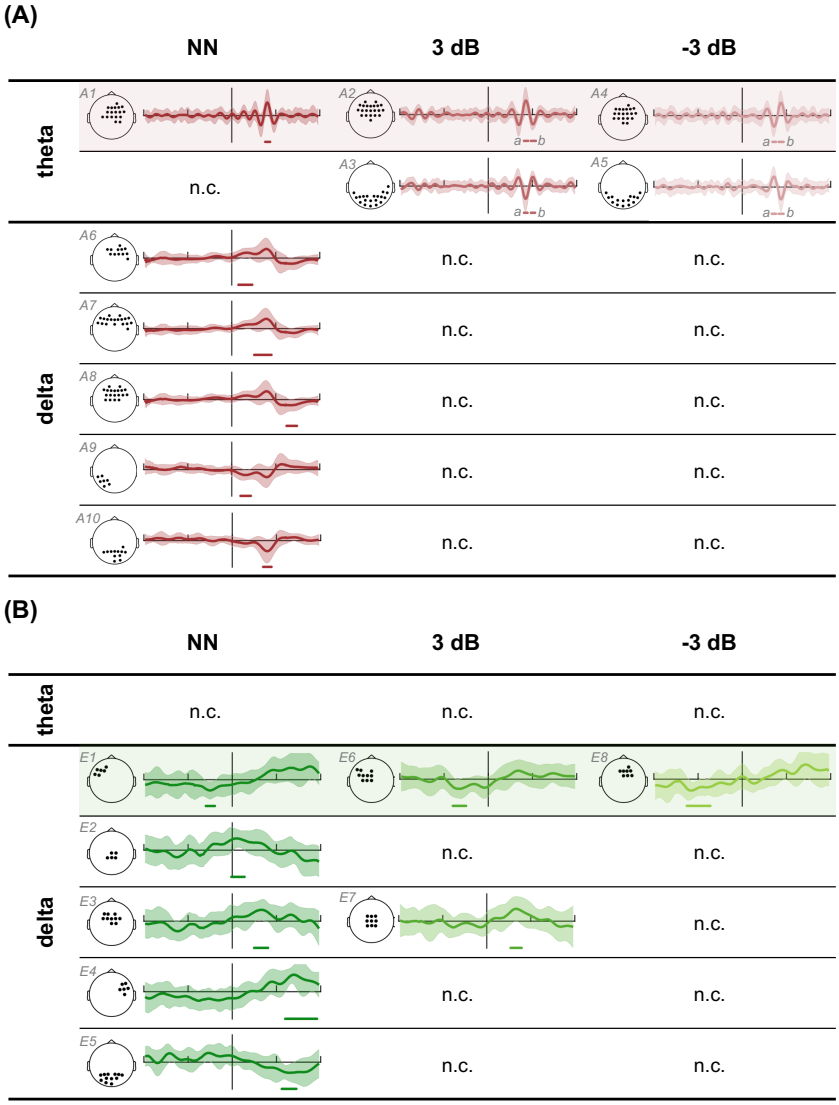

Fig. 4. Significant clusters in TRF responses for different noise levels and frequency bands to (A) amplitude envelop and (B) word entropy. Clusters with the similar spatiotemporal pattern are organized in the same row. Clusters shared by all noise levels are highlighted with a colored background. ROIs of clusters are shown as black dots in the corresponding topography. The colored curves are the mean of TRFs averaged among ROIs across participants. The shaded areas denote the standard error of TRFs. The colored horizontal line below the TRF curve indicates the TOIs of the cluster. n.c. means no significant cluster is found.

3.3 Acoustic-level TRF responses with delayed latencies as noise increases

Significant acoustic-level TRF responses in the theta band were found at all noise levels (Fig. 4A) and showed a similar positivity within central electrodes at around 300−500 ms demonstrated in Fig. 5A. At the no-noise level, the TRFs showed positivity in central electrodes with a latency of around 420 ms ($p < .01$, FDR corrected). At the low-noise and high-noise levels, the TRFs showed similar positivity in central electrodes with a latency of around 430 and 440 ms and extra negativity with a latency of around 500 ms or 350 ms ($ps < .01$, FDR corrected). The central positivity at around 420 ms latency consistently occurred at all three noise levels and thus was considered a prominent peak. We estimated the peak amplitude and latency for the prominent peak from TRF averaged among the intersection of the central ROIs in three conditions. The peak latency was significantly different among the three noise levels (rmANOVA, $F(2, 36) = 4.01$, $p = .027$) and post-hoc *t*-tests revealed significantly longer peak latency in high-noise condition compared with the no-noise condition ($p = .023$, Bonferroni corrected), as shown in Fig. 5C and 5D. No significant differences were found in the peak amplitude (rmANOVA, $F(2, 36) = 1.71$, $p = .194$).

The reconstruction accuracy from corresponding ROI/TOI was significantly higher in the actual decoder than in the null decoder ($ps < .05$, FDR corrected, Fig. 5B). Comparing the reconstruction accuracy levels revealed significant differences among the three noise levels (rmANOVA, $F(2, 36) = 6.17$, $p = .010$ with Greenhouse-Geisser correction) and post-hoc *t*-tests revealed significantly weakened neural tracking

in high-noise condition compared with the no-noise condition and low-noise condition (p = .007 and .029, Bonferroni corrected).

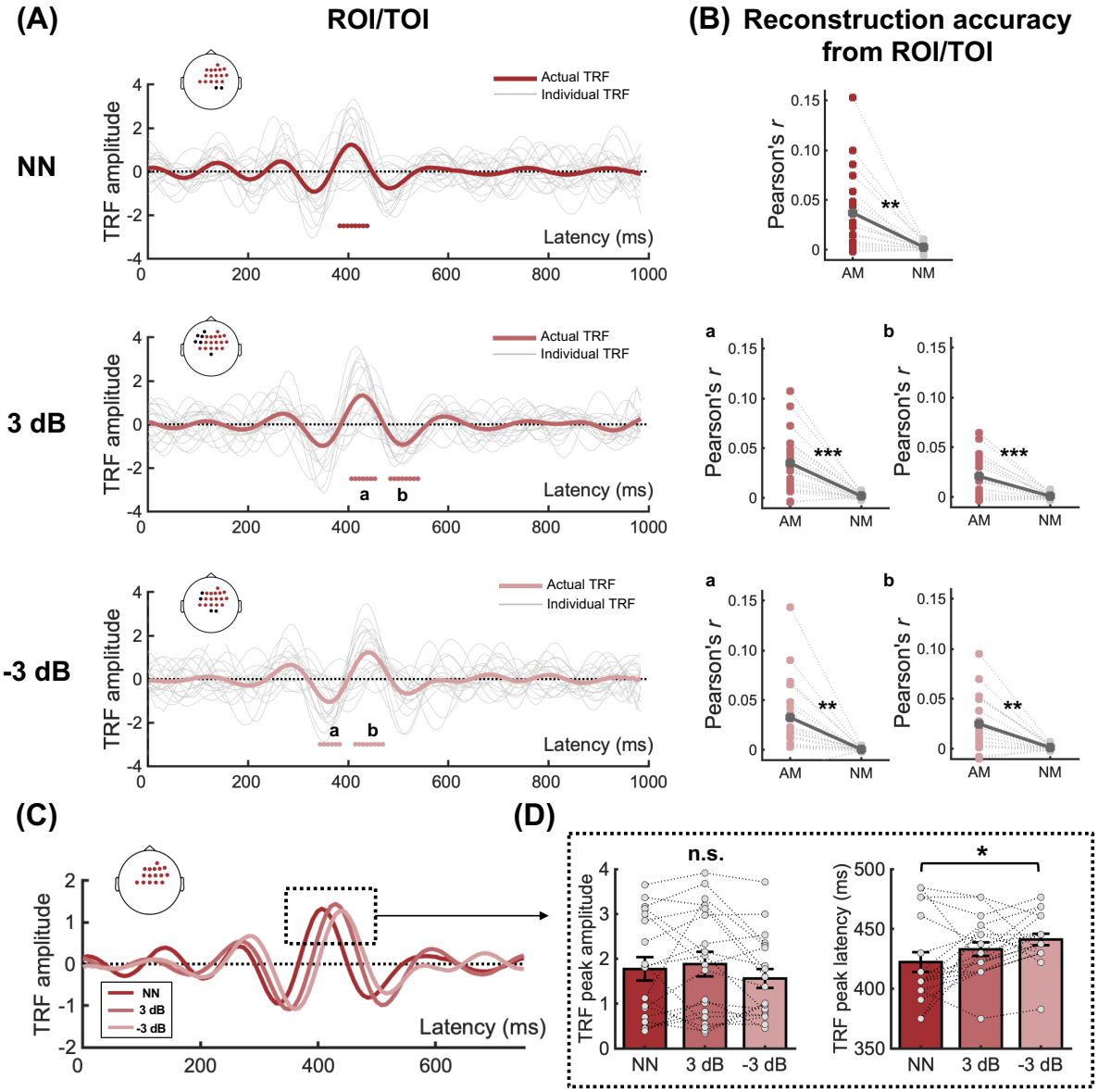

Fig. 5. Acoustic-level TRF response on the theta band under different noise conditions. (A) The bold curves in different shades of red are the mean of TRFs averaged among ROIs across participants in the three noise levels. The grey thin curves are TRFs averaged among the ROIs of each participant. The colored horizontal line at the bottom of each plot indicates TOIs over which TRFs differed significantly from the null models. Dots in the corresponding topographies depicted the ROIs and shared channels among different SNR conditions are highlighted in red. (B) Reconstruction accuracy calculated from the ROI/TOI in (A). (C) TRFs averaged across the shared ROIs of different SNR conditions. Amplitude and latency are extracted for the peak within the small dashed box. (D) Noise effect on peak amplitude and latency. Grey dots indicate data from each participant. Error bars denote the standard error. n.s. means not significant, † means $p < .1$, * means $p < .05$, ** means $p < .01$, ***means $p < .001$.

3.4 Semantic-level TRF responses with earlier latencies as noise increases

Significant semantic-level TRF responses to word entropy in the delta band were found at all noise levels (Fig. 4B) and showed similar negativity at around 200–600 ms leading to speech fluctuation onset, as demonstrated in Fig. 6A. The time window of pre-onset processing to word entropy showed a gradual advance trend as the increase of noise level. The time window was approximately from -300 ms to -180 ms at the no-noise level ($p < .05$, FDR corrected), and was from around -400 ms to -250 ms at the low-noise level ($p < .05$, FDR corrected), from -630 ms to -360 ms at the high-noise level ($p < .01$, FDR corrected). Confined to the high variability among participants' semantic-level TRFs (the average inter-subject correlation of TRF amplitude in semantic TRF responses is .05, which is significantly lower than that (.21) in acoustic TRF responses, $t(18) = 4.31$, $p < .001$), we were not able to extract individual pre-onset peak latency and conduct statistical tests. However, the gradual advancement in the time window was clear and noteworthy.

The pre-onset TRF responses to word entropy exhibited different spatiotemporal patterns in three noise conditions. At the no-noise level, the ROI included frontal-parietal electrodes and exhibited obvious left lateralization. At the low-noise level, the ROI showed similar left-lateralized topological distribution but included more electrodes while in the high-noise condition no obvious lateralization was observed in the frontal-parietal ROI.

The reconstruction accuracy from corresponding ROI/TOI was significantly higher in the actual decoder than in the null decoder at both low-noise level ($p < .05$, FDR corrected) and high-noise level ($p < .001$, FDR corrected), but only marginally significant at the no-noise level ($p = .0716$, FDR corrected),

as shown in Fig. 6B. Comparing the reconstruction accuracy among different noise levels revealed no significant difference (rmANOVA, $F(2, 36) = 1.643$, $p = .207$).

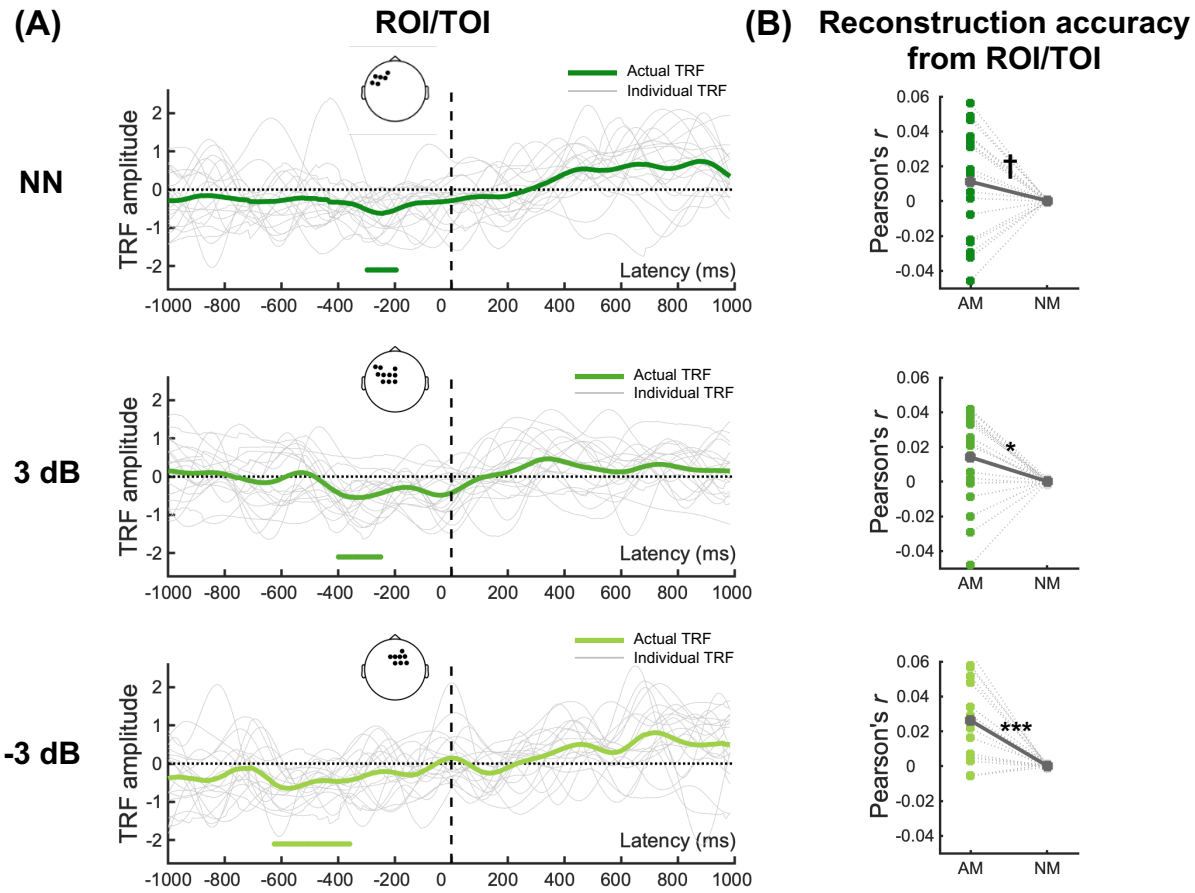

Fig. 6. Semantic-level TRF responses to word entropy on the delta band under different noise conditions. (A) The bold curves in different shades of green are the mean of TRFs averaged among ROIs across participants in the three noise levels. The grey thin curves are TRFs averaged among the ROIs of each participant. The colored horizontal line at the bottom of each plot indicates TOIs over which TRFs differed significantly from the null models. Dots in the corresponding topographies depicted the ROIs. (B) Reconstruction accuracy calculated from the ROI/TOI in (A). n.s. means not significant, † means $p < .1$, * means $p < .05$, ** means $p < .01$, ***means $p < .001$.

4 Discussion

The current study illustrated the neural tracking of hierarchical features of naturalistic speech in noisy situations using a TRF-based technique. Significant post-onset acoustic-level TRF responses were found within central electrodes at around 420 ms and the peak latency was prolonged as noise increased. Moreover, significant pre-onset semantic-level TRF responses were found within frontal electrodes at around -600–-200 ms and the pre-onset time lag showed a gradually advanced trend as noise increased. These findings reveal how the neural system adaptively processes hierarchical features of naturalistic speech in a noisy environment.

The delayed acoustic-level TRF response' peak latency with increasing noise level was in line with several previous studies, suggesting an impaired efficiency in demanding and challenging conditions with background noise (Ding & Simon, 2013; Gillis, Decruy, et al., 2022; Kong et al., 2014; Mirkovic et al., 2019; Zou et al., 2019). As the frontally and centrally distributed channels (corresponding to the primary auditory cortex, superior temporal gyrus, premotor cortex, etc.) have been frequently reported to be related to the processing of speech acoustics (Bidelman & Howell, 2016; Broderick et al., 2019; Hickok & Poeppel, 2007; Pérez et al., 2022; Zou et al., 2019) , the present TRF results would imply similar recruitment of these brain regions for acoustic-level processing for naturalistic speech under various noise levels. However, the post-onset 400-ms latency was prolonged as compared to the commonly reported latency of < 300 ms in previous studies (e.g., Broderick et al., 2019; Di Liberto et al., 2015). This discrepancy could be due to the causal filter used for EEG signal preprocessing in the present study, possibly resulting in a delayed TRF compared to previous studies using noncausal filters, as reported by Etard and Reichenbach (2019). Alternatively, it could be possible that the latency modulation started earlier but only reached significance later for the present dataset, as the observed TRF responses actually exhibited an oscillatory pattern starting much earlier than 400 ms (Fig. 5C). While an impaired processing efficiency has been associated with both amplitude and latency modulation by noise in previous studies (Muncke et al., 2022; Zion Golumbic et al., 2013; Zou et al., 2019), the present study together with a series of other studies reporting latency-only results would suggest latency as a more sensitive candidate for noisy speech processing at the acoustic level (Ding & Simon, 2013; Kaplan-Neeman et al., 2006; Whiting et al., 1998).

More importantly, increasingly advanced pre-onset TRF responses to word entropy with increasing noises suggest a distinct mechanism for speech-in-noise comprehension at the semantic level. While both the pre-onset latency and the left-lateralization spatial pattern of the TRF responses to word entropy under the no-noise condition was similar to recent studies on semantic prediction (Goldstein et al., 2022; Weissbart et al., 2020; Willems et al., 2016), our results further extend the present understanding of the predictive mechanism with the manipulation of the noise levels: The pre-onset latency became more advanced with increasing noises. One possible explanation for this phenomenon is that our brain may adjust the timing of predictive processing in response to adverse environments. As semantic prediction can facilitate speech comprehension (Mattys et al., 2012; Miller et al., 1951; Obleser & Kotz, 2010; Pickering & Gambi, 2018; Zekveld et al., 2011), the brain relies on it more heavily as the noise level increases to counteract distorted auditory input. Nevertheless, noise can increase the processing load and decrease the processing efficiency (Gillis, Decruy, et al., 2022; Kaplan-Neeman et al., 2006; Kong et al., 2014; Mirkovic et al., 2019). To compensate for the interference, the neural system initiates the pre-onset response earlier and extends it for a longer duration, giving our brain more time for the preparation of the upcoming speech information. This finding suggests that the brain has a temporally active adaptation mechanism for reliable speech comprehension in noisy environments. As such a pre-onset signature was not observed at the acoustic level, our results suggest the predictive mechanism might be mainly focused at the semantic level (Etard & Reichenbach, 2019; Goldstein et al., 2022; Grisoni et al., 2021), where the speech information is expected to be more abstract and more robust against noises (Yasmin et al., 2023).

Interestingly, the spatial patterns of TRF responses to word entropy showed a shift from the left to the right hemisphere with increasing noises. As the left-lateralization has been frequently reported in studies on speech-in-noise comprehension (Z. Li et al., 2021; Vander Ghinst et al., 2016, 2019), and was found sensitive to linguistic content (Peelle et al., 2013), word entropy (Willems et al., 2016) and semantic expectancy (Golestani et al., 2013), the left-lateralization results under the no-noise and low-noise conditions would support the left-lateralized brain regions for predictive speech processing at the semantic level (entropy in the present study). In the meanwhile, the right-lateralized regions such as the right inferior frontal gyrus have been shown to be sensitive to semantic features such as entropy (Willems et al., 2016), and the weakening of left-lateralization and a shift in dominance to the right hemisphere under degraded

conditions have also been reported in previous studies and were hypothesized as the possible recruitment of additional regions for compensation (Bidelman & Howell, 2016; Shtyrov et al., 1998, 1999). Accordingly, our results would suggest that the hemispheric dominance shifting in adverse environments might reflect a semantic-related compensation mechanism.

Our results regarding the TRF-based representation of speech acoustics and semantics in different frequency bands provide further evidence for the frequency-specificity of speech processing. In this study, acoustic-level TRF response was primarily associated with the theta band whereas semantic-level TRF neural response was dominated by the delta band (Fig. 4). These findings correspond to previous studies indicating that theta-band neural tracking was related to acoustic processing and delta-band neural tracking was associated with the semantic and syntactic processing (Dai et al., 2022; Ding et al., 2014; Etard & Reichenbach, 2019; Keshavarzi et al., 2021; J. Li et al., 2022; see a review, Kösem & van Wassenhove, 2017). And our study provides straightforward evidence for this functional differentiation of frequency bands with an NLP-based semantic quantification method. Additionally, the present study paid attention to the pre-onset stage of TRFs, and discovered that the semantic predictive responses also occurred on the delta band even in noisy conditions. This finding further suggests that delta-band neural tracking is involved in predictive processes, which echos the previous finding that temporal predictions were encoded in delta oscillations (Herbst & Obleser, 2019; Stefanics et al., 2010), extending our comprehension of the functional role of delta-band speech tracking.

This study has some limitations to be noted. First, there were several significant TRF responses not included in the above analysis and discussions, which were primarily observed in the no-noise condition. As the focus of the present study was speech-in-noise comprehension, these responses were not further discussed. Nevertheless, they also reflected speech information processing that would deserve investigations in future studies. Second, while the present study only adopted word entropy and word surprisal due to their relatively solid research foundation (Gillis et al., 2021; Goldstein et al., 2022; Heilbron et al., 2022; Weissbart et al., 2020; Willems et al., 2016), advancements in NLP methods present us with a wider range of options (Broderick et al., 2018, 2019, 2021; Mesik et al., 2021; Yasmin et al., 2023). Thus, future studies could employ additional indexes to fully demonstrate the semantic-level adaption mechanism. Third, despite the advantage of high temporal resolution of EEG in exploring temporal dynamics, the

relatively poor spatial resolution limits the ability to investigate brain regions involved in predictive mechanisms. A finer spatial-temporal dynamics of semantic prediction would require techniques such as fMRI, ECoG, or multimodal approaches.

In summary, the current study investigated how noise affected acoustic and semantic processing during naturalistic speech comprehension. With increasing noises, acoustic processing became increasingly delayed whereas semantic processing became increasingly advanced. Our results suggest that, while the efficiency of brain processing of speech information is indeed impaired by noise, the brain can compensate for the associated effects through active participation at the semantic level. Overall, the present findings are expected to contribute to the growing body of research on the neural mechanisms of naturalistic speech comprehension.

Declaration of Competing Interest

All the co-authors declare that they have no conflict of interest.

Acknowledgment

The authors would like to thank Prof. Xiaoqin Wang for providing the shielded room for the experiment. The authors would like to thank Prof. Zhiyuan Liu and members of his lab for computing the NLP models. The authors would like to thank Lingyi Tao and Xuelin Wang for their help with data collection.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in project Crossmodal Learning (NSFC 62061136001/DFG TRR-169/C1, B1), and the National Natural Science Foundation of China (61977041).

Data availability statement

Data in the study are available for reviewers if required.

References

Alday, P. M. (2019). M/EEG analysis of naturalistic stories: A review from speech to language processing. *Language, Cognition and Neuroscience*, *34*(4), 457–473. https://doi.org/10.1080/23273798.2018.1546882

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2020). Cortical entrainment: What we can learn from studying naturalistic speech perception. *Language, Cognition and Neuroscience*, *35*(6), 681–693. https://doi.org/10.1080/23273798.2018.1518534

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology*, *43*(5), 454–464. https://doi.org/10.1111/j.1469-8986.2006.00448.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bennett, K. O., Billings, C. J., Molis, M. R., & Leek, M. R. (2012). Neural encoding and perception of speech signals in informational masking. *Ear and Hearing*, *33*(2), 231–238. https://doi.org/10.1097/AUD.0b013e31823173fd

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883–895. https://doi.org/10.1016/j.tics.2021.07.006

Bidelman, G. M., & Howell, M. (2016). Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception. *NeuroImage*, *124*(Pt A), 581–590. https://doi.org/10.1016/j.neuroimage.2015.09.020

Billings, C. J., Bennett, K. O., Molis, M. R., & Leek, M. R. (2011). Cortical encoding of signals in noise: Effects of stimulus type and recording paradigm. *Ear and Hearing*, *32*(1), 53–60. https://doi.org/10.1097/AUD.0b013e3181ec5c46

Billings, C. J., Tremblay, K. L., Stecker, G. C., & Tolin, W. M. (2009). Human evoked cortical activity to signal-to-noise ratio and absolute signal level. *Hearing Research*, *254*(1), 15–24. https://doi.org/10.1016/j.heares.2009.04.002

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25–31. https://doi.org/10.1016/j.cophys.2020.07.014

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *Journal of Neuroscience*, *39*(38), 7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*(1), 4963. https://doi.org/10.1038/s41598-021-84597-9

Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language*, *43*(1), 1–18. https://doi.org/10.1016/0093-934X(92)90018-A

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. *Frontiers in Neuroscience*, *15*, 1350. https://doi.org/10.3389/fnins.2021.705621

Dai, B., McQueen, J. M., Terporten, R., Hagoort, P., & Kösem, A. (2022). Distracting Linguistic Information Impairs Neural Tracking of Attended Speech. *Current Research in Neurobiology*, 100043. https://doi.org/10.1016/j.crneur.2022.100043

Daltrozzo, J., Wioland, N., & Kotchoubey, B. (2012). The N400 and Late Positive Complex (LPC) Effects Reflect Controlled Rather than Automatic Mechanisms of Sentence Processing. *Brain Sciences*, *2*(3), Article 3. https://doi.org/10.3390/brainsci2030267

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does Semantic Context Benefit Speech Understanding through "Top–Down" Processes? Evidence from Time-resolved Sparse fMRI. *Journal of Cognitive Neuroscience*, *23*(12), 3914–3932. https://doi.org/10.1162/jocn_a_00084

de Cheveigné, A., & Nelken, I. (2019). Filters: When, Why, and How (Not) to Use Them. *Neuron*, *102*(2), 280–293. https://doi.org/10.1016/j.neuron.2019.02.039

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*(12), 631–645. https://doi.org/10.1111/lnc3.12093

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311.

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393.

Etard, O., & Reichenbach, T. (2019). Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. *Journal of Neuroscience*, *39*(29), 5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

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

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

Gillis, M., Decruy, L., Vanthornhout, J., & Francart, T. (2022). Hearing loss is associated with delayed neural responses to continuous speech. *European Journal of Neuroscience*, *55*(6), 1671–1690. https://doi.org/10.1111/ejn.15644

Gillis, M., Van Canneyt, J., Francart, T., & Vanthornhout, J. (2022). Neural tracking as a diagnostic tool to assess the auditory pathway. *Hearing Research*, 108607. https://doi.org/10.1016/j.heares.2022.108607

Gillis, M., Vanthornhout, J., Simon, J. Z., Tom Francart, & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0812-21.2021

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Golestani, N., Hervais-Adelman, A., Obleser, J., & Scott, S. K. (2013). Semantic versus perceptual interactions in neural processing of speech-in-noise. *NeuroImage*, *79*, 52–61. https://doi.org/10.1016/j.neuroimage.2013.04.049

Grisoni, L., Tomasello, R., & Pulvermüller, F. (2021). Correlated Brain Indexes of Semantic Prediction and Prediction Error: Brain Localization and Category Specificity. *Cerebral Cortex*, *31*(3), 1553–1568. https://doi.org/10.1093/cercor/bhaa308

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Hartley, C. A., & Poeppel, D. (2020). Beyond the Stimulus: A Neurohumanities Approach to Language, Music, and Emotion. *Neuron*, *108*(4), 597–599. https://doi.org/10.1016/j.neuron.2020.10.021

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Herbst, S. K., & Obleser, J. (2019). Implicit temporal predictability enhances pitch discrimination sensitivity and biases the phase of delta oscillations in auditory cortex. *NeuroImage*, *203*, 116198. https://doi.org/10.1016/j.neuroimage.2019.116198

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. https://doi.org/10.3758/BF03196544

Jamison, C., Aiken, S. J., Kiefte, M., Newman, A. J., Bance, M., & Sculthorpe, -Petley Lauren. (2016). Preliminary Investigation of the Passively Evoked N400 as a Tool for Estimating Speech-in-Noise Thresholds. *American Journal of Audiology*, *25*(4), 344–358. https://doi.org/10.1044/2016_AJA-15-0080

JASP Team. (2022). *JASP (version 0.16.3)[Computer software]*. https://jasp-stats.org/

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., & Muchnik, C. (2006). Identification of syllables in noise: Electrophysiological and behavioral correlates. *The Journal of the Acoustical Society of America*, *120*(2), 926–933. https://doi.org/10.1121/1.2217567

Keitel, A., Ince, R. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage*, *147*, 32–42.

Keshavarzi, M., Varano, E., & Reichenbach, T. (2021). Cortical tracking of a background speaker modulates the comprehension of a foreground speech signal. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3200-20.2021

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization3rd international conference on learning representations, ICLR 2015. *International Conference on Learning Representations, ICLR*.

Koerner, T. K., & Zhang, Y. (2015). Effects of background noise on inter-trial phase coherence and auditory N1–P2 responses to speech stimuli. *Hearing Research*, *328*, 113–119. https://doi.org/10.1016/j.heares.2015.08.002

Koerner, T. K., Zhang, Y., Nelson, P. B., Wang, B., & Zou, H. (2016). Neural indices of phonemic discrimination and sentence-level speech intelligibility in quiet and noise: A mismatch negativity study. *Hearing Research*, *339*, 40–49. https://doi.org/10.1016/j.heares.2016.06.001

Kong, Y.-Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research*, *316*, 73–81.

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544. https://doi.org/10.1080/23273798.2016.1238495

Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A., & Hari, R. (2020). Brain activity reflects the predictability of word sequences in listened continuous speech. *NeuroImage*, *219*, 116936. https://doi.org/10.1016/j.neuroimage.2020.116936

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Li, J., Hong, B., Nolte, G., Engel, A. K., & Zhang, D. (2021). Preparatory delta phase response is correlated with naturalistic speech comprehension performance. *Cognitive Neurodynamics*. https://doi.org/10.1007/s11571-021-09711-z

Li, J., Hong, B., Nolte, G., Engel, A. K., & Zhang, D. (2022). *An integrative view on attentional modulation in naturalistic speech* (p. 2022.10.02.510499). bioRxiv. https://doi.org/10.1101/2022.10.02.510499

Li, Z., Hong, B., Wang, D., Nolte, G., Engel, A. K., & Zhang, D. (2022). Speaker–listener neural coupling reveals a right-lateralized mechanism for non-native speech-in-noise comprehension. *Cerebral Cortex*, bhac302. https://doi.org/10.1093/cercor/bhac302

Li, Z., Li, J., Hong, B., Nolte, G., Engel, A. K., & Zhang, D. (2021). Speaker–Listener Neural Coupling Reveals an Adaptive Mechanism for Speech Comprehension in a Noisy Environment. *Cerebral Cortex*, *bhab118*. https://doi.org/10.1093/cercor/bhab118

Li, Z., & Zhang, D. (2023). How does the human brain process noisy speech in real life? Insights from the second-person neuroscience perspective. *Cognitive Neurodynamics*. https://doi.org/10.1007/s11571-022-09924-w

Loaiza, V. M., & Camos, V. (2018). The role of semantic representations in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 863–881. https://doi.org/10.1037/xlm0000475

Lu, Y., Jin, P., Pan, X., & Ding, N. (2022). Delta-band neural activity primarily tracks sentences instead of semantic properties of words. *NeuroImage*, 118979. https://doi.org/10.1016/j.neuroimage.2022.118979

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martin, B. A., & Stapells, D. R. (2005). Effects of low-pass noise masking on auditory event-related potentials to speech. *Ear and Hearing*, *26*(2), 195–213. https://doi.org/10.1097/00003446-200504000-00007

Martin, R. C. (2021). The Critical Role of Semantic Working Memory in Language Comprehension and Production. *Current Directions in Psychological Science*, *30*(4), 283–291. https://doi.org/10.1177/0963721421995178

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mesik, J., Ray, L., & Wojtczak, M. (2021). Effects of Age on Cortical Tracking of Word-Level Features of Continuous Competing Speech. *Frontiers in Neuroscience*, *15*. https://doi.org/10.3389/fnins.2021.635126

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329.

Mirkovic, B., Debener, S., Schmidt, J., Jaeger, M., & Neher, T. (2019). Effects of directional sound processing and listener's motivation on EEG responses to continuous noisy speech: Do normal-hearing and aided hearing-impaired listeners differ? *Hearing Research*, *377*, 260–270. https://doi.org/10.1016/j.heares.2019.04.005

Muncke, J., Kuruvila, I., & Hoppe, U. (2022). Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise. *Frontiers in Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fnins.2022.876421

Obleser, J. (2014). Putting the Listening Brain in Context. *Language and Linguistics Compass*, *8*(12), 646–658. https://doi.org/10.1111/lnc3.12098

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences*, *23*(11), 913–926.

Obleser, J., & Kotz, S. A. (2010). Expectancy Constraints in Degraded Speech Modulate the Language Comprehension Network. *Cerebral Cortex*, *20*(3), 633–640. https://doi.org/10.1093/cercor/bhp128

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

O'Sullivan, 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*(7), 1697–1706.

Otten, M., & Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research*, *1291*, 92–101. https://doi.org/10.1016/j.brainres.2009.07.042

Parbery-Clark, A., Marmel, F., Bair, J., & Kraus, N. (2011). What subcortical-cortical relationships tell us about processing speech in noise: Neural correlates of speech in noise. *European Journal of Neuroscience*, *33*(3), 549–557. https://doi.org/10.1111/j.1460-9568.2010.07546.x

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), Article 12. https://doi.org/10.1038/nrn2277

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Peelle, J. E., & Wingfield, A. (2016). The Neural Consequences of Age-Related Hearing Loss. *Trends in Neurosciences*, *39*(7), 486–497. https://doi.org/10.1016/j.tins.2016.05.001

Pérez, A., Davis, M. H., Ince, R. A. A., Zhang, H., Fu, Z., Lamarca, M., Lambon Ralph, M. A., & Monahan, P. J. (2022). Timing of brain entrainment to the speech envelope during speaking, listening and self-listening. *Cognition*, *224*, 105051. https://doi.org/10.1016/j.cognition.2022.105051

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

Pulvermüller, F., & Grisoni, L. (2020). Semantic Prediction in Brain and Mind. *Trends in Cognitive Sciences*, *24*(10), 781–784. https://doi.org/10.1016/j.tics.2020.07.002

Romei, L., Wambacq, I. J. A., Besing, J., Koehnke, J., & Jerger, J. (2011). Neural indices of spoken word processing in background multi-talker babble. *International Journal of Audiology*, *50*(5), 321–333. https://doi.org/10.3109/14992027.2010.547875

Shivde, G., & Anderson, M. C. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1342–1370. https://doi.org/10.1037/a0024832

Shtyrov, Y., Kujala, T., Ahveninen, J., Tervaniemi, M., Alku, P., Ilmoniemi, R. J., & Näätänen, R. (1998). Background acoustic noise and the hemispheric lateralization of speech processing in the human brain: Magnetic mismatch negativity study. *Neuroscience Letters*, *251*(2), 141–144. https://doi.org/10.1016/S0304-3940(98)00529-1

Shtyrov, Y., Kujala, T., Ilmoniemi, R. J., & Näätänen, R. (1999). Noise affects speech-signal processing differently in the cerebral hemispheres. *NeuroReport*, *10*(10), 2189–2192. Scopus. https://doi.org/10.1097/00001756-199907130-00034

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Stefanics, G., Hangya, B., Hernádi, I., Winkler, I., Lakatos, P., & Ulbert, I. (2010). Phase Entrainment of Human Delta Oscillations Can Mediate the Effects of Expectation on Reaction Speed. *Journal of Neuroscience*, *30*(41), 13578–13585. https://doi.org/10.1523/JNEUROSCI.0703-10.2010

Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed Expectancies under Degraded Speech: Revisiting the N400. *Journal of Cognitive Neuroscience*, *25*(8), 1383–1395. https://doi.org/10.1162/jocn_a_00389

Sun, M., Chen, X., Zhang, K., Guo, Z., & Liu, Z. (2016). *THULAC: An Efficient Lexical Analyzer for Chinese*.

Vander Ghinst, M., Bourguignon, M., Niesen, M., Wens, V., Hassid, S., Choufani, G., Jousmäki, V., Hari, R., Goldman, S., & Tiège, X. D. (2019). Cortical Tracking of Speech-in-Noise Develops from Childhood to Adulthood. *Journal of Neuroscience*, *39*(15), 2938–2950. https://doi.org/10.1523/JNEUROSCI.1732-18.2019

Vander Ghinst, M., Bourguignon, M., Op de Beeck, M., Wens, V., Marty, B., Hassid, S., Choufani, G., Jousmäki, V., Hari, R., Van Bogaert, P., Goldman, S., & De Tiège, X. (2016). Left Superior Temporal Gyrus Is Coupled to Attended Speech in a Cocktail-Party Auditory Scene. *Journal of Neuroscience*, *36*(5), 1596–1606. https://doi.org/10.1523/JNEUROSCI.1730-15.2016

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

Whiting, K. A., Martin, B. A., & Stapells, D. R. (1998). The Effects of Broadband Noise Masking on Cortical Event-Related Potentials to Speech Sounds /ba/ and /da/. *Ear and Hearing*, *19*(3), 218–231. https://journals.lww.com/ear-hearing/Abstract/1998/06000/The_Effects_of_Broadband_Noise_Masking_on_Cortical.5.aspx

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. *Language, Cognition and Neuroscience*, *32*(7), 855–869.

Yasmin, S., Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2023). *The Effects of Speech Masking on Neural Tracking of Acoustic and Semantic Features of Natural Speech* (p. 2023.02.10.527537). bioRxiv. https://doi.org/10.1101/2023.02.10.527537

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Festen, J. M., van Beek, J. H. M., & Rönnberg, J. (2011). The Influence of Semantically Related and Unrelated Text Cues on the Intelligibility of Sentences in Noise. *Ear and Hearing*, *32*(6), e16. https://doi.org/10.1097/AUD.0b013e318228036a

Zendel, B. R., Tremblay, C.-D., Belleville, S., & Peretz, I. (2015). The Impact of Musicianship on the Cortical Mechanisms Related to Separating Speech from Background Noise. *Journal of Cognitive Neuroscience*, *27*(5), 1044–1059. https://doi.org/10.1162/jocn_a_00758

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. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. http://dx.doi.org/10.1016/j.neuron.2012.12.037

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., Zheng, J., & Ding, N. (2019). Auditory and language contributions to neural encoding of speech features in noisy environments. *NeuroImage*, *192*, 66–75. https://doi.org/10.1016/j.neuroimage.2019.02.047