Cognitive Neurodynamics (2022) 16:337–352
https://doi.org/10.1007/s11571-021-09711-z

RESEARCH ARTICLE

Preparatory delta phase response is correlated with naturalistic speech comprehension performance

Jiawei Li[1,3] · Bo Hong[2,3] · Guido Nolte[4] · Andreas K. Engel[4] · Dan Zhang[1,3]

Received: 30 October 2020 / Revised: 9 July 2021 / Accepted: 12 August 2021 / Published online: 31 August 2021

Abstract

While human speech comprehension is thought to be an active process that involves top-down predictions, it remains unclear how predictive information is used to prepare for the processing of upcoming speech information. We aimed to identify the neural signatures of the preparatory processing of upcoming speech. Participants selectively attended to one of two competing naturalistic, narrative speech streams, and a temporal response function (TRF) method was applied to derive event-related-like neural responses from electroencephalographic data. The phase responses to the attended speech at the delta band (1–4 Hz) were correlated with the comprehension performance of individual participants, with a latency of − 200–0 ms relative to the onset of speech amplitude envelope fluctuations over the fronto-central and left-lateralized parietal electrodes. The phase responses to the attended speech at the alpha band also correlated with comprehension performance but with a latency of 650–980 ms post-onset over the fronto-central electrodes. Distinct neural signatures were found for the attentional modulation, taking the form of TRF-based amplitude responses at a latency of 240–320 ms post-onset over the left-lateralized fronto-central and occipital electrodes. Our findings reveal how the brain gets prepared to process an upcoming speech in a continuous, naturalistic speech context.

Keywords Preparatory processing · Attention · Speech comprehension · Electroencephalogram · Temporal response function

Introduction

Humans can effectively comprehend complex and rapidly changing speech in challenging conditions, e.g., in a cocktail party scenario with multiple competing speech streams and high background noise. To achieve such a capacity, the human brain is equipped with an efficient neural architecture that is dedicated to bottom-up processing of perceived speech information, from the low-level acoustics, to the phoneme, syllable, and sentence levels (Pisoni and Luce 1987; DeWitt and Rauschecker 2012; Friederici 2012; Hickok 2012; Verhulst et al. 2018). In recent years, increasing evidence has also suggested that human speech comprehension is an active process that involves top-down predictions (Rao and Ballard 1999; Federmeier 2007; Arnal et al. 2011; Hickok et al. 2011; Kutas and Federmeier 2011; Fries 2015; Tian et al. 2018). For instance, in the cocktail party scenario, it is believed that a listener should continuously predict what their attended speaker is going to say next in order to efficiently understand the corresponding speech (Cherry 1953; Ding and Simon 2012a; Zion Golumbic et al. 2013a; O'Sullivan et al. 2015; Bednar and Lalor 2020).

Although the idea of prediction in human speech comprehension is gaining popularity, it remains unclear how the brain gets prepared for the processing of upcoming speech information. The preparatory process could be an

✉ Dan Zhang
dzhang@tsinghua.edu.cn

1 Department of Psychology, School of Social Sciences, Tsinghua University, Room 334, Mingzhai Building, Beijing, China

2 Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, China

3 Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

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

important part of the prediction mechanism, reflecting how the predictive information could guide the neural system to fine-tune its state to process the upcoming speech efficiently. The available findings on prediction in speech, however, are not sufficient to determine the neural mechanisms underlying preparation. For instance, the classic studies of active speech prediction have mainly focused on neural activity in response to prediction errors. Event-related potential (ERP) components such as the N400 and P600 are frequently reported when the perceived word violates semantic and syntactic congruency of the preceding speech context, respectively (Lau et al. 2008; Kutas and Federmeier 2011; Van Petten and Luka 2012; Wang et al. 2018). These ERP components normally occur>400 ms after the presentation of the perceived speech and therefore provide only indirect support for the preparatory process.

Recent studies have reported evidence of the brain's pre-activation before the onset of the upcoming speech. Some researchers have focused on preparatory attentional orientation to specific acoustic features (e.g., spatial location, pitch, etc.) in speech or general auditory tasks (Hill and Miller 2010; Lee et al. 2013; Holmes et al. 2016, 2018; ElShafei et al. 2018; Nolden et al. 2019). For instance, the visual cues prior to the onset of an auditory stimulus could elicit distinct neural responses over the left superior temporal sulcus (STS), depending on its associated task instructions about the to-be-attended pitch feature of the upcoming auditory stimulus (Lee et al. 2013). In the meanwhile, researchers have reported pre-activations that are more specific for speech processing (DeLong et al. 2005; Dikker and Pylkkänen 2013; Söderström et al. 2016, 2018): event-related neural responses to a preceding speech unit (e.g., words) were found to be informative about possible upcoming speech units in the continuous speech materials (e.g., sentences). While these speech-related pre-activations could reflect the brain's preparation for processing the upcoming speech, they were represented by event-related responses that were associated with either the attention-related cues or the preceding speech units. In other words, these results were constrained by the processing of the preceding events and therefore the preparatory process might be only indirectly expressed. Ideally, the most direct preparatory processing should be linked to the to-be-processed speech but occur before its onset.

While this direct evidence has not been investigated in the speech domain, studies on general sensory processing have provided support for the possible existence of such a preparatory process. In the visual domain, the amplitude and phase of pre-stimulus oscillatory activities, especially in the alpha band, have been reported to have a significant impact on subsequent perceptual consequences (Van Dijk et al. 2008; Kok et al. 2017; Harris et al. 2018; Galindo-Leon et al. 2019; Rassi et al. 2019), such as threshold-level perception, attention orientation, visual search performance, etc. Auditory information processing has also been shown to be affected by pre-stimulus oscillations, mainly at lower frequency bands such as theta and delta (Ng et al. 2012; Kayser et al. 2016; Zoefel et al. 2018). However, these studies have mainly employed simple and abstract sensory stimuli such as pure tone, visual shapes as preparatory cues, that do not resemble real-world speech scenarios, therefore were limited in explaining possible neural mechanisms underlying the preparatory processing of human speech. Recently, there is an emergence of naturalistic stimuli in auditory studies (Sonkusare et al. 2019). Researchers adopted naturalistic, continuous audios (e.g., poetry, long sentences) as the stimulus (Etard and Reichenbach 2019; Teng et al. 2020; Donhauser and Baillet 2020). The naturalistic speech presents listeners with a variety and multitude of different linguistic contents (Alexandrou et al. 2018), catering for different levels of preparation. Thus, our study utilized naturalistic and continuous materials to study the preparatory processing.

One crucial issue that needs to be considered is the possible dependence of the preparatory process on top-down selective attention. As attention regulates the processing of the input sensory information, it can be expected to affect prediction and consequently preparation (Schröger et al. 2015a, b). A number of recent studies have indeed reported attention-dependent neural responses to prediction errors (Kok et al. 2012; Auksztulewicz and Friston 2015; Hisagi et al. 2015; Marzecová et al. 2017; Smout et al. 2019), with ongoing debates on the direction of the interplay and the involved sensory processing stages. Most of these studies have been conducted within the visual domain, with limited exploration in the auditory domain, let alone speech processing.

The present study aimed to identify neural signatures that directly reflect the preparatory processing of human speech. Naturalistic, narrative speech materials were present to the participants with a 60-channel electroencephalogram (EEG) recording; this procedure is believed to be of high ecological validity, thus providing necessary contextual information for the engagement of top-down prediction and therefore preparation (Rao and Ballard 1999; Friston 2005; Federmeier 2007; Jehee and Ballard 2009). A cocktail party paradigm was used to introduce a complex perceptual environment that imposed further demands on preparation as compared to a noise-free environment (Cherry 1953; Ding and Simon 2012a; Zion Golumbic et al. 2013a; O'Sullivan et al. 2015; Broderick et al. 2018). To characterize the neural responses to the continuous, naturalistic speech streams, a temporal response function (TRF) method was used to derive event-related-like neural responses from the EEG signal, based on the speech amplitude envelope of both the attended and

the unattended speech streams (Lalor et al. 2006; Crosse et al. 2016). We employed the amplitude envelope for the following considerations: (1) speech amplitude envelope has been successfully used in previous studies on naturalistic speech processing to investigate attention, clarity, and comprehension performance (Di Liberto et al. 2015; Etard and Reichenbach 2019); (2) speech amplitude envelope could include sufficient information at all levels, from acoustics to semantics (Di Liberto et al. 2015; Daube et al. 2019). The TRF-based responses were expected to reveal the temporal dynamics of neural activities underlying human speech processing, and the responses with latencies earlier than the onset of speech amplitude envelope fluctuations are regarded to represent the preparatory phase. Specifically, the TRF-based responses were further decomposed into amplitude and phase responses, as amplitude and phase have been proposed to play unique roles in networks underlying human cognition (Bonnefond and Jensen 2012; Engel et al. 2013; Fries 2015). Following the studies on the perceptual influence of pre-stimulus neural activities (Smith et al. 2006; Iemi et al. 2019; Rassi et al. 2019; Avramiea et al. 2020), we were interested in whether the TRF-based responses at the preparatory stage could be correlated to speech comprehension performance, as measured by speech-content-related questionnaires, and how amplitude and phase responses contributed to speech preparation. With the employment of the cocktail party paradigm, we also addressed the issue of the attention-dependency of the to-be-explored performance-related preparatory activities. Our study is expected to reveal the neural mechanisms underlying how the brain gets prepared to process an upcoming speech in a continuous, naturalistic speech context.

Materials and methods

Ethics statement

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.

Experimental model and participant details

Twenty college students (10 females; mean age: 24.7 years; range: 20–43 years) from Tsinghua University participated in the study as paid volunteers. All participants were native Chinese speakers, and reported having normal hearing and normal or corrected-to-normal vision.

Note that we did not perform a power analysis for sample size due to the lack of previous speech-related studies with a similar analysis framework (i.e., a cluster-based permutation of correlation values, see below). Instead, the sample size (N=20) was decided empirically following previous TRF-based studies on human speech processing (Di Liberto et al. 2015; Mirkovic et al. 2015; Broderick et al. 2018).

Stimuli

The speech stimuli were recorded from two male speakers using the microphone of an iPad2 mini (Apple Inc., Cupertino, CA) at a sampling rate of 44,100 Hz. The speakers were college students from Tsinghua University who had more than four years of professional training in broadcasting. Both speakers were required to tell twenty-eight 1-min narrative stories in Mandarin Chinese. Half of these stories were about daily-life topics recommended by the experimenter, and the speakers improvised on their own (14 stories); and the other half were selected from the National Mandarin Proficiency Test (14 stories). The recommended topic or story materials were presented to the speakers on the computer screen. They were allowed to prepare for as long as required before telling the story (usually $\sim$ 3 min). When they were ready, the speakers pressed the SPACE key on the computer keyboard, and the recording began with the presentation of three consecutive pure-tone beep sounds at 1000 Hz (duration: 1000 ms; inter-beep interval: 1500 ms). The beep sounds served as the event markers to synchronize the speech streams to be presented simultaneously in the main experiment. The speakers were asked to start speaking as soon as the third beep had ended (within around 3 s). The speakers were allowed to start the recording again if the audio did not meet the requirements of either the experimenter or the speakers themselves (mainly concerned with the coherence of speech). The actual speaking time of each story ranged from 51 to 76 s.

Two four-choice questions per story (two for the attended story and two for the unattended story) were then prepared by the experimenter and two college students who were familiar with comprehension performance assessment. These questions and the corresponding choices concerned story details that required significant attentional efforts. For instance, one question following a story about one's hometown was, "What is the most dissatisfying thing about the speaker's hometown? (推测讲述人对于家乡最不满意的地方在于?)", and the four choices were (A) There is no heating in winter; (B) There are no hot springs in summer; (C) There is no fruit in autumn; (D) There are no flowers in spring (A. 冬天没暖气; B. 夏天没温泉; C. 秋天没水果; D. 春天没鲜花).

Experimental procedure

The main experiment consisted of four blocks, each containing seven trials. In each trial, two narrative stories were presented simultaneously, one to the left ear and the other to the right ear. The participants were instructed to attend to one spatial side. The two speech streams within each trial were from the two different male speakers to facilitate selective attention. Considering the possible duration difference between the two audio streams, the trial ended after the longer speech audio had ended. Each trial began when participants pressed the SPACE key on the computer keyboard. Participants were instructed which side to attend to by plain text ("Please pay attention to the [LEFT/RIGHT]") displayed on the computer screen. A white fixation cross was also displayed throughout the trial. The speech stimuli were played immediately after the keypress and were preceded by the three beep sounds to allow participants to prepare.

At the end of each trial, four questions (two for the attended story and the other two for the unattended story) were presented sequentially in random order on the computer screen, and the participants made their choices using the computer keyboard. The listeners were not explicitly informed about the correspondence between the questions and the stories. The single-trial comprehension accuracy could be 0% (two wrong answers), 50% (one correct answer), or 100% (two correct answers) for both the attended and the unattended stories. While the goal of such a design was to measure the listener's comprehension performance in feasible experimental time, using only two questions per story might not provide a sufficient test for all the story content. Therefore, the averaged accuracies across all trials (separately for the attended and the unattended stories) were computed to have a reliable estimation of the listener's overall comprehension performance.

After completing these questions, participants scored their attention level of the attended stream, the experienced difficulty of performing the attention task, and the familiarity with the attended material using three 10-point Likert scales. No feedback was given to the participants about their performance during the experiment. Throughout the trial, participants were required to maintain visual fixation on the fixation cross while listening to the speech and to minimize eye blinks and all other motor activity. The participants were recommended to take a short break (around 1 min) after every trial within one block and a long break (no longer than 10 min) between blocks.

The to-be-attended side was fixed within each block (two blocks for attending to the left side and two for attending to the right side). Within each block, the speaker identity remained unchanged for the left and right sides. The to-be-attended spatial side and the corresponding speaker identity were balanced within the participant, with seven trials per side for both speakers. The assignment of the stories to the four blocks was randomized across the participants.

The experiment was carried out in a sound-attenuated, dimly lit, and electrically shielded room. The participants were seated in a comfortable chair in front of a 19.7-inch LCD monitor (Lenovo LT2013s). The viewing distance was approximately 60 cm. The experimental procedure was programmed in MATLAB using the Psychophysics Toolbox 3.0 extensions (Brainard and Brainard 1997). The speech stimuli were delivered binaurally via an air-tube earphone (Etymotic ER2, Etymotic Research, Elk Grove Village, IL, USA) to avoid possible electromagnetic interferences from auditory devices. The volume of the audio stimuli was adjusted to be at a comfortable level that was well above the auditory threshold. Furthermore, the speech stimuli driving the earphone were used as an analog input to the EEG amplifier through one of its bipolar inputs together with the EEG recordings. In this way, the audio and the EEG recordings were precisely synchronized, with a maximal delay of 1 ms (at a sampling rate of 1000 Hz).

Data acquisition and pre-processing

EEG was recorded from 60 electrodes (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, and O1/2), which were referenced to an electrode between Cz and CPz, with a forehead ground at Fz. A NeuroScan amplifier (SynAmp II, NeuroScan, Compumedics, USA) was used to record EEG at a sampling rate of 1000 Hz. Electrode impedances were kept below 10 kOhm for all electrodes.

The recorded EEG data were first notch filtered to remove the 50 Hz powerline noise and then subjected to an artifact rejection procedure using independent component analysis. Independent components (ICs) with large weights over the frontal or temporal areas, together with a corresponding temporal course showing eye movement or muscle movement activities, were removed. The remaining ICs were then back-projected onto the scalp EEG channels, reconstructing the artifact-free EEG signals. While the relatively long duration of the speech trials in the present study ($\sim$ 1 min per story, see Experimental procedure) has made it more difficult for the participants to avoid inducing movement-related artifacts as compared to the classical ERP-based studies, a temporally continuous, non-interrupted EEG segment per trial was preferred for the employment of the TRF method. Therefore, any ICs with artifact-like EEG activities for more than 20% of the trial

time (i.e., ~ 12 s) were rejected, leading to around 4–11 ICs rejected per participant. The cleaned EEG data were used for the TRF analysis without any further artifact rejection procedures. Then the EEG signals were re-referenced to a common average reference, following previous speech-related studies using TRFs (e.g., O'Sullivan et al. 2015; Bednar and Lalor 2020).

Next, the EEG data were segmented into 28 trials according to the markers representing speech onsets. The analysis window for each trial extended from 10 to 55 s (duration: 45 s) to avoid the onset and the offset of the stories (Crosse et al. 2016).

Temporal response function modeling

The analysis workflow for the analysis related to the attended speech stream is shown in Fig. 1. The neural responses to the speech stimuli were characterized using a temporal response function (TRF)-based modeling method. The TRF response describes the impulse response to fluctuations of an input signal and is based on system identification theories (Lalor et al. 2006; Crosse et al. 2016). We used the amplitude envelope of the speech signal as the input signal required by TRF, which has been demonstrated to be a valid index to extract speech-related neural responses (Ding and Simon 2012a; Mirkovic et al. 2015; O'Sullivan et al. 2015; Huang et al. 2018; Broderick et al. 2018; Bednar and Lalor 2018). In the present study, the TRF-based responses were expected to characterize how the human brain responded to the fluctuations of the input speech amplitude envelope signals. The TRF-based responses provided the opportunity to inspect the dynamics of the speech-related neural activities: while the TRF-based responses following the speech amplitude envelope fluctuations are considered to represent post-processing of the speech stream, the pre-onset responses can the pre-onset responses can express pre-activation or preparation for processing incoming speech information (Etard and Reichenbach 2019). While the temporal dynamics of speech-related neural activities could also be investigated using simpler methods without the employment of modelling methods (e.g., performing simple correlation or

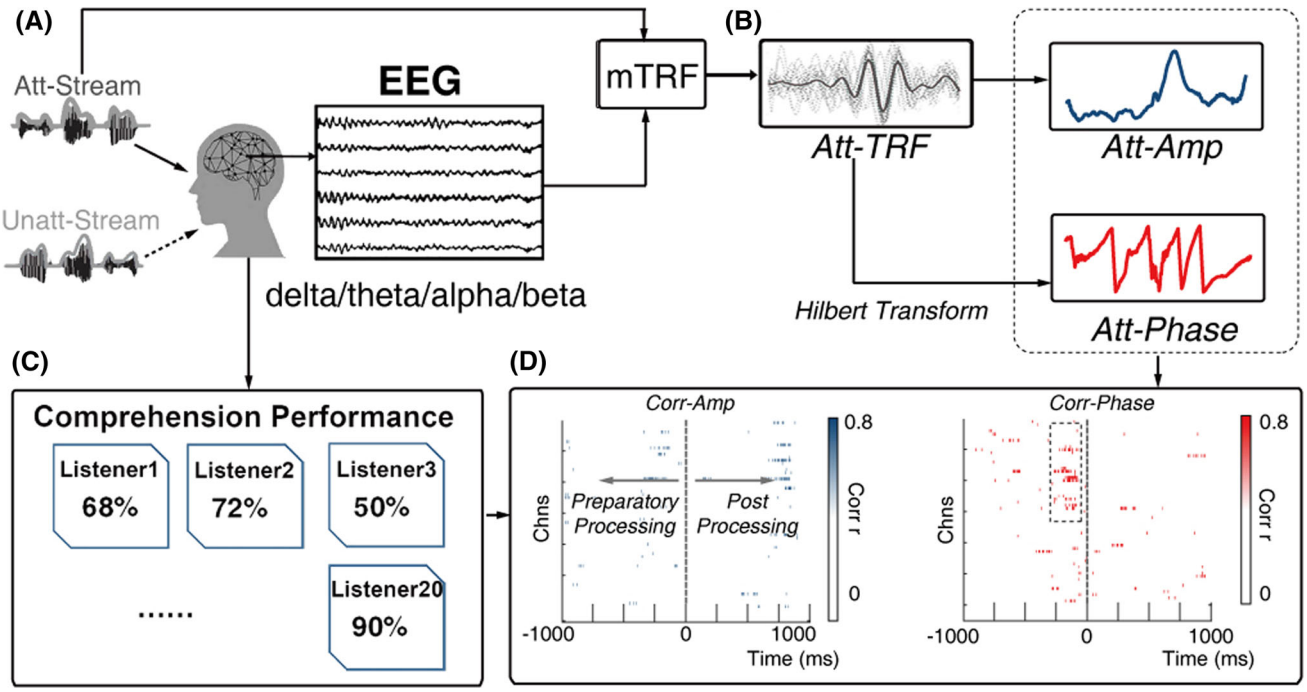


Fig. 1 The analysis workflow. **a** The experimental paradigm. Participants attended to one of two simultaneously presented naturalistic, narrative speech streams while 60-channel EEG was recorded. **b** EEG data analysis. Neural responses were characterized using a TRF-based modeling method. The TRF-based neural responses were decomposed into the amplitude and the phase responses using the Hilbert transform. This procedure was conducted separately for attended (Att-) and unattended (Unatt-, not shown) speech streams, and separately for EEG data filtered at delta, theta, alpha and beta bands. **c** Comprehension performance. The participants completed a comprehension task after each speech comprehension trial. The average response accuracy over all trials per participant was taken as his/her comprehension performance. **d** Correlation analysis for comprehension performance-related neural responses. We calculated the correlation between either amplitude or phase responses and comprehension performance for each channel-latency bin. We defined neural activity before 0 ms as preparatory-processing and activity after 0 ms as post-processing. The results of the delta band were illustrated here. The colored channel-latency bins showed uncorrected significant correlation with comprehension performance. The dashed box in the 'Corr-Phase' plot indicates a significant channel-latency cluster by a cluster-based permutation test

cross-correlation analyses) (Ringach and Shapley 2004; Kong et al. 2014; Müller et al. 2019), calculating the TRF-based responses could facilitate the extraction of neural activities that were more related to the input speech signals (Crosse et al. 2016).

Prior to the modeling, the preprocessed EEG signals were re-referenced to the average of all scalp channels and then downsampled to 128 Hz. Then, the EEG data were filtered in delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz) and beta (12–30 Hz) bands (filter order: 64, one-pass forward filter). The use of a causal FIR filter ensured that filtered EEG signals were decided only by the current and previous data samples (de Cheveigné and Nelken 2019), which is important for the present research aim of preparatory speech processing. The filter order of 64 was chosen to keep a balance of temporal resolution and filter performance: the filtered EEG signals were therefore calculated based on the preceding 500 ms data (64 at 128 Hz).

The amplitude envelopes of the speech signals were obtained using a Hilbert transform and then downsampled to the same sampling rate of 128 Hz. The EEG data were z-scored before TRF computation as recommended (Crosse et al. 2016). When denoting $R^j_{i,k}(t)$ as the downsampled EEG signals from channel i, trial k filtered at one specific frequency band j (representing the four frequency bands) and $S_k(t)$ as the input speech amplitude envelope corresponding to trial k, the corresponding neural response $TRF^j_{i,k}(t)$ can be formulated as follows:

$$R^j_{i,k}(t) = TRF^j_{i,k}(t) * S_k(t) \tag{1}$$

where * represents the convolution operator. The latency in the neural response models $TRF^j_{i,k}(t)$ was set to vary from − 1000 ms to 1000 ms post-stimulus. Note that the model latency parameters were set to vary to a wider time range as compared to previous studies (normally − 100 to 400 ms, e.g., Di Liberto et al. 2015; O'Sullivan et al. 2015; Broderick et al. 2018). The time range was decided to allow a reliable exploration of the temporal dynamics at latencies beyond previous studies, especially on the pre-onset end, while avoiding possible regression artifacts at the extremes of the model latencies (Crosse et al. 2016).

To control for overfitting, we varied the lambda from 10^{-1} to 10^3 (lambda=10^{-1}, 10^0,…, 10^3) in the ridge regression (Di Liberto et al. 2015; Crosse et al. 2016; Broderick et al. 2018). The lambda value corresponding to the backward decoder that produced the highest cross-validated speech amplitude envelope reconstruction accuracy (as the correlation between the reconstructed and original envelopes), averaged across trials, was selected as the regularization parameter for all trials per participant (Broderick et al. 2019). The cross-validation procedure was implemented as a leave-one-trial-out manner: each time the TRFs were trained on the basis of data from 27 trials and tested on the left-out trial. All TRF-based responses were further transformed into z-scores within the − 1000 ms to 1000 ms time window for each channel separately per participant to account for across participants and across session differences (Pasley et al. 2012; Kleen et al. 2016). These z-scores were then used for the following analyses.

The above-mentioned TRF calculation procedure was performed for the EEG signals from each EEG channel filtered at the four frequency bands, with either the attended or the unattended speech amplitude envelopes as the reference signals. In other words, we obtained the TRF-based responses to both the attended and the unattended speech streams, for all the EEG channels and the four frequency bands.

Amplitude and phase were calculated using the Hilbert transform of the TRF-based neural responses at the single-trial level. Hereby, the instantaneous amplitude $A^j_{i,k}(t)$ and the instantaneous phase $\phi^j_{i,k}(t)$ can be computed as

$$A^j_{i,k}(t) = \sqrt{\left(TRF^j_{i,k}(t)\right)^2 + h\left(TRF^j_{i,k}(t)\right)^2} \tag{2}$$

$$\phi^j_{i,k}(t) = \tan^{-1}\left(\frac{h\left(TRF^j_{i,k}(t)\right)}{TRF^j_{i,k}(t)}\right) \tag{3}$$

where $h(TRF^j_{i,k}(t))$ represents its Hilbert transform.

These single-trial instantaneous amplitude and phase were then averaged across all trials per participant to reflect one's overall EEG responses to the naturalistic speech streams. Specifically, the single-trial amplitudes were averaged by taking their arithmetic mean value, whereas the single-trial phase responses were averaged by computing their circular mean (i.e., the mean phase angle). The circular-linear correlation between the mean phase $\varnothing_{i,j}(t)$ and the linear variable of comprehension performance x was defined as follow:

$$r_{i,k}(t) = \sqrt{\frac{r^2_{cx} + r^2_{sx} - 2r_{cx}r_{sx}r_{cs}}{1 - r^2_{cs}}} \tag{4}$$

where r_{sx} means the Pearson correlation between $\sin(\varnothing_{i,j}(t))$ and x, r_{cx} means the Pearson correlation between $\cos(\varnothing_{i,j}(t))$ and x, r_{cs} means the Pearson correlation between $\sin(\varnothing_{i,j}(t))$ and$\cos(\varnothing_{i,j}(t))$. $r_{i,k}(t)$ varies from − 1 to 1, where+1 means the two variables have a perfect positive relationship, − 1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.

In addition, the inter-trial phase locking (ITPL) was also calculated to evaluate the phase consistency across trials within each participant. Given the number of trials denoted by N, the TRF-based ITPL were calculated as follows:

$$ITPL(t) = \left|\sum_{k=1}^{N} \exp(\varnothing_{i,k}^{j}(t)\right|/N \quad (5)$$

The phase-related ITPL value varies between 0 and 1; 0 refers to a situation in which the phase responses of different trials are uniformly distributed between 0 and 2π, and 1 means the phase responses from all trials are entirely locked to a fixed phase angle.

The TRF analysis was conducted in MATLAB using the Multivariate Temporal Response Function (mTRF) toolbox (Crosse et al. 2016). All the other EEG processing procedures, as well as the statistical analyses, were conducted using the FieldTrip toolbox (Oostenveld et al. 2011).

Quantification and statistical analysis

The extracted TRF-based amplitude and phase responses were used to correlate with the speech comprehension performance of the attended speech at the participant level. The Spearman's correlation was calculated between the amplitude response and the comprehension performance at each EEG channel and each individual latency across the participants. The correlations between the phase response and the comprehension performance were evaluated by computing the circular linear correlation using the CircStat toolbox (Berens 2009). Both the TRFs to the attended and unattended speech at the four frequency bands were included for this analysis.

Statistical analysis was performed to examine the significance of correlations over all channel-latency bins by computing the correlation r-values. To account for multiple comparisons, a nonparametric cluster-based permutation analysis was applied (Maris and Oostenveld 2007). In this procedure, neighboring channel-latency bins with an uncorrected correlational p-value below 0.01 were combined into clusters, for which the sum of the correlational t-statistics corresponding to the correlation r-values were obtained. A null-distribution was created through permutations of data across participants (n=1,000 permutations), which defined the maximum cluster-level test statistics and corrected p-values for each cluster.

In addition, we investigated the attention modulation of the TRF-based responses. The purpose of this analysis was to test whether the well-established attention effect (Ding and Simon 2012b; O'Sullivan et al. 2015) could be replicated on the present dataset. To this end, paired t-tests were performed, contrasting the TRFs to the attended speech versus the unattended speech. Both amplitude and phase were included in the analysis. The phase difference was calculated as the phase angle difference by the CircStat toolbox as well. A similar cluster-based permutation was used to control for the multiple comparison problem ($p<$ 0.01 as the threshold, n=1, 000 permutations).

The above statistical analysis followed the standard cluster-based permutation procedure as employed in classical ERP and related studies (Arnal et al. 2011; Henry and Obleser 2012; Zhang et al. 2012). Note that the reported p-values were only corrected for the tests performed within each frequency band by using the cluster-based permutation tests. No multiple comparison correction was employed across different frequency bands.

Results

Behavioral results

The average comprehension performance was significantly better for the 28 attended stories than for the 28 unattended stories (67.0±2.5% (standard error) vs. 36.0±1.6%, $t(19)$ =10.948, $p<0.001$; the four-choice chance level: 25%). The participants reported a moderate level of attention (8.146±0.343 on a 10-point Likert scale) and attention difficulties (2.039±0.530 on a 10-point Likert scale). The accuracy for the attended story was significantly correlated with both the self-reported attention level (r=0.476, p= 0.043) and attention difficulty (r=− 0.677, p=0.001). The self-reported story familiarity level was low for all the participants (0.860±0.220 on a 10-point Likert scale) and was not correlated with comprehension performance (r= − 0.224, p=0.342). These results suggest that participants' selective attention was effectively manipulated and the measurement of comprehension performance was reliable. Most importantly, there was large inter-individual variability in the participant-wise average comprehension performance for the attended stories; the response accuracy varied from 48.2% to 91.1%, which supports the feasibility of using these accuracy values as a behavioral indicator of comprehension-relevant neural signatures. In the meanwhile, the response accuracy varied from 25.0% to 51.8% for the unattended stories.

Speech comprehension performance related TRF-based responses

Figure 2 depicts the distribution of the average reconstruction accuracy for each listener in each frequency band. The average reconstruction accuracies (as the correlation between the reconstructed and original envelopes) for the attended speech streams were 0.144±0.014 (standard error), 0.170±0.016, 0.147±0.014, and 0.082±0.009 for the delta, theta, alpha, and beta bands, respectively. The average reconstruction accuracies were significantly lower for the unattended speech streams (ps<0.05 with Bonferroni correction), with 0.082±0.010, 0.114±0.012, 0.096± 0.010, and 0.054±0.004 for the delta, theta, alpha, and

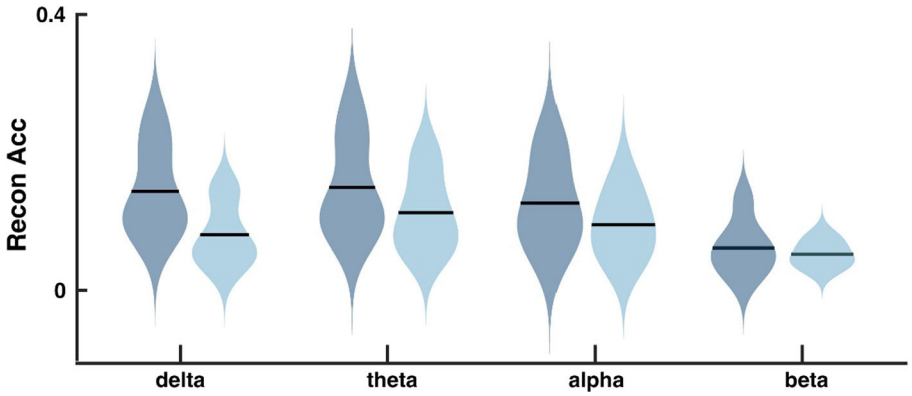


Fig. 2 The violin plots depict the distribution of the average reconstruction accuracy (as the correlation between the reconstructed and original envelopes) for each listener in each frequency band for the attended (the dark blue plots) and the unattended speech amplitude envelopes (the light blue plots). Black horizontal bars indicate the means

beta bands, respectively. These results were comparable with previous studies (O'Sullivan et al. 2015), constituting the basis for our follow-up analysis.

The nonparametric cluster-based permutation analysis revealed a significant correlation between the multi-channel amplitude and phase representation of the TRFs and individual speech comprehension performance of the attended speech. This corresponded to two clusters in the observed data (cluster-based permutation $p<0.05$). Both the two significant clusters are reflected by TRF-based phase responses to the attended speech, with one at the delta band and the other at the alpha band (Fig. 3a, b). No significant correlations were found for the TRFs to the unattended speech.

As shown in Fig. 3c.I and II, the delta cluster was represented by TRF-based phase responses to the attended speech at − 200–0 ms before the onset of speech amplitude envelope fluctuations over the fronto-central and left-lateralized parietal electrodes (cluster-based permutation $p=0.012$, mean circular linear correlation $r=0.787$). The participants with their comprehension performance ranking intop, middle, and bottom tertiles were associated with different delta phase angles during this pre-speech-onset time window. As shown in Fig. 3c.I, the top-performing participants showed a negative peak within the time window of the cluster, whereas the bottom-performing participants showed a positive peak during this time period. Figure 3c.III provides more quantitative information regarding this phase effect, and average phase angles for the participants in the three tertiles were $\Phi_{top}=73.411°$, $\Phi_{middle}=80.175°$, and $\Phi_{bottom}=-134.767°$, respectively. However, the participants' ITPL did not significantly correlate with their comprehension performance ($r=0.127$, $p=.593$, Fig. 3c.IV).

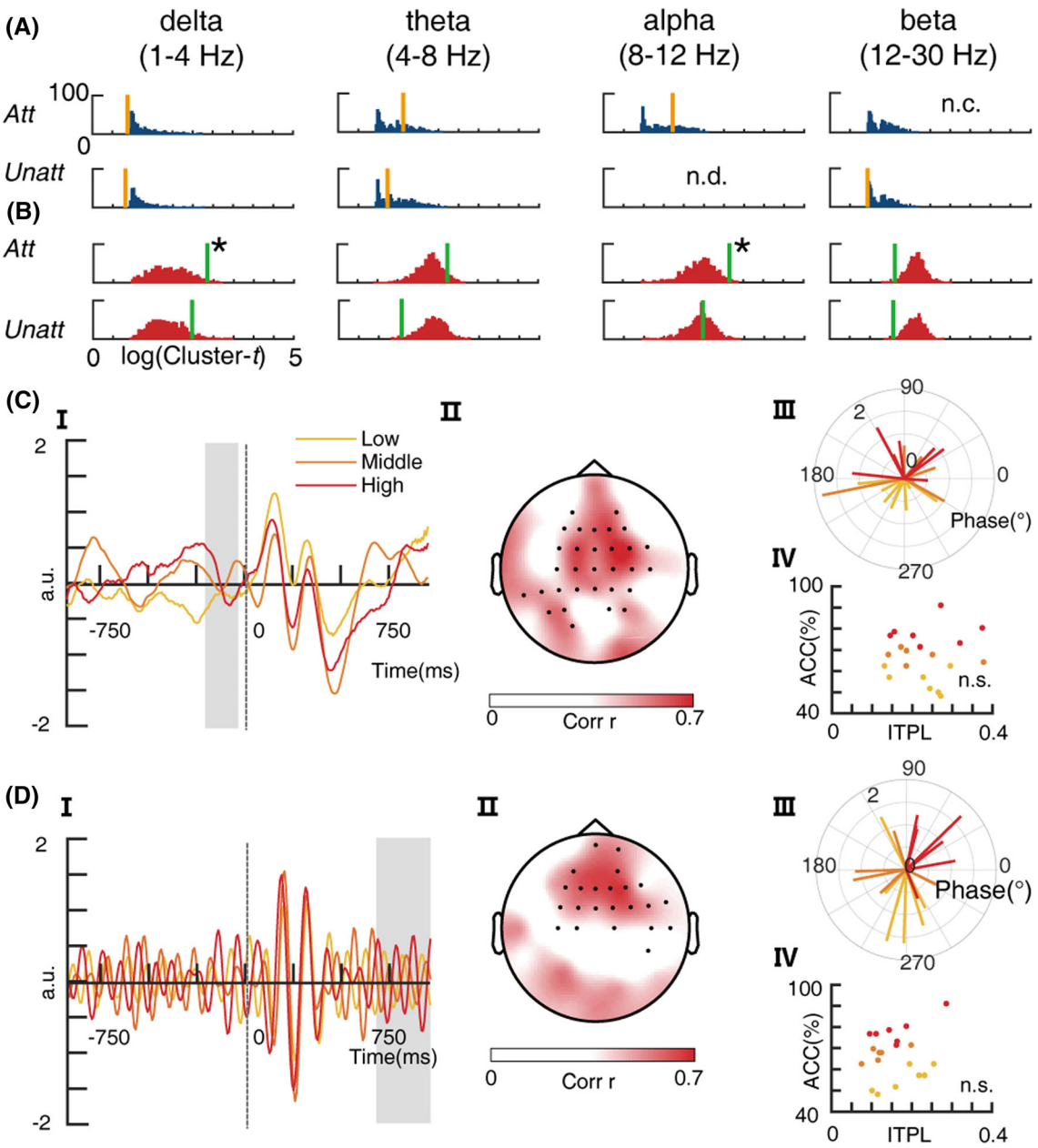
The alpha cluster was represented by TRF-based phase responses to the attended speech at 650–980 ms post speech onset, with a fronto-central distribution (cluster-based permutation $p=0.031$, mean $r=0.784$, see Fig. 3d.I and II). The individual difference in the phase angle of their responses was large, with average $\Phi_{top}=41.904°$ for the top-performing participants and $\Phi_{middle}=-135.461°$, $\Phi_{bottom}=-115.916°$ for the other two groups (Fig. 3d.III). Again, there was no significant correlation between the ITPL and the comprehension performance ($r=0.267$, $p=0.255$, Fig. 3d.IV).

Attention related TRF-based responses

The nonparametric cluster-based permutation analysis revealed a significant difference between the TRFs to the attended and unattended speech, as shown in Fig. 4. The difference was manifested as a cluster involving a set of channels covering the left-lateralized fronto-central and occipital electrodes (cluster-based permutation $p<0.001$) at a latency of 240–320 ms post the onset of speech amplitude envelope fluctuations (Fig. 4c). This attentional difference is reflected on TRF-based amplitude response at the theta band. No significant differences were observed in amplitude responses at other frequency bands and phase responses at all frequency bands (Fig. 4a, b).

Discussion

The present study aimed to identify neural signatures that directly reflect the preparatory processing of upcoming speech. We used naturalistic narrative speech materials in a selective attention paradigm and a TRF-based approach for modeling the neural activity and observed preparatory neural activities before the onset of speech amplitude envelope fluctuations. We found a significant correlation between the comprehension performance of individual participants and the phase responses to the attended speech at the delta band (1–4 Hz), with a latency of − 200–0 ms relative to the fluctuation onset over the fronto-central and left-lateralized parietal electrodes. The comprehension performance was also correlated with the phase responses to the attended speech at the alpha band, but with a latency


(A)
delta (1-4 Hz)
theta (4-8 Hz)
alpha (8-12 Hz)
beta (12-30 Hz)
Att
Unatt
100
0
n.c.
n.d.
(B)
*
0 log(Cluster-t) 5
(C)
I
II
III
IV
Low
Middle
High
2
-2
a.u.
-750
0
750
Time(ms)
0 Corr r 0.7
90
180
0
270
Phase(°)
100
40
ACC(%)
n.s.
0 ITPL 0.4
(D)

◀**Fig. 3** Speech comprehension performance related TRF-based responses to naturalistic speech. **a** Histograms showing the distributions of the cluster-level correlational *t*-statistics (log-transformation) from the 1000 permutated calculations of the correlation between the comprehension performance and the TRF-based amplitude response at the four frequency bands. The y-axis shows the histogram counts of clusters from the permutated calculation at the corresponding log (Cluster-*t*) value and the asterisk shows the statistically significant cluster from the real data. The upper and lower panel for the attended and unattended responses, respectively. The vertical orange or green lines indicate the *t*-statistics of the clusters from the real data. N.C. means no cluster was formed and N.D. means the permutated distribution could not be generated (i.e., no cluster formed during the permutation calculation). **b** Distributions of the cluster-level correlational *t*-statistics from the 1000 permutated calculations of the correlation between the comprehension performance and the TRF-based phase response at the four frequency bands. The upper and lower panel for the attended and unattended responses, respectively. The two vertical lines with asterisks indicate statistically significant clusters from the real data. **c** Illustration of the significant phase-response cluster at the delta band. I. The time course of the TRFs at one representative channel (FC1). The three waveforms represent the average responses over the participants with comprehension performance of the attended speech ranking in the top (red), middle (orange), and bottom (yellow) tertiles (7, 6, and 7 participants, respectively). The three waveforms represent the average TRF responses over the participants with comprehension performance of the attended speech ranking in the top (red), middle (orange), and bottom (yellow) tertiles. The shaded area depicts the time window in which the phase has a significant circular correlation with the comprehension performance, which is further demonstrated in III. II. The topography of the average correlation *r*-values in the time window of interest. Black dots indicate the channels of interest in the cluster. III. The polar plot is showing the average phase angles at channel FC1 per participant. The vector lengths indicate the response amplitude and the vector directions indicate the phase angle. The three different colors indicate the participants' comprehension performance rankings as in I. IV. The scatter plot showing the participants' comprehension performance (accuracy in percentage) versus their inter-trial phase-locking (ITPL) values. The color of the dots indicates the comprehension performance rankings as in I and III. 'n.s.' means no significant correlation. **d** Illustration of the significant phase-response cluster at the alpha band. The plots in I and III are from channel FC1. The explanations of the sub-plots follow (C)

of 650–980 ms post onset over fronto-central electrodes. As Distinct neural signatures were found for the attentional modulation, taking the form of TRF-based amplitude responses at a latency of 240–320 ms post onset over the left-lateralized fronto-central and occipital electrodes. Our results provide direct neural evidence for how the brain prepares for the processing of upcoming speech.

Before detailed discussions, it is necessary to state that our assumption for a preparatory process is based on the observation that the TRF-based neural activities before the onset of speech amplitude envelope fluctuations during the continuous speech stream were significantly correlated with comprehension performance. Recent TRF-based studies using naturalistic stimuli have reported reasonable latencies that resembled their ERP counterparts for describing selective auditory attention (∼200 ms) (Mirkovic et al. 2015; O'Sullivan et al. 2015), semantic violation processing (∼400 ms) (Broderick et al. 2018), and visual working memory (200–400 ms) (Huang et al. 2018). Although our findings have mainly focused on the window of<0 ms, these studies support the rationale of using the TRF-based responses to reflect the time course of information processing in general. In addition, the present study also replicated the timing of previously reported attentional modulation of TRF-based responses to speech stimuli (Mirkovic et al. 2015; O'Sullivan et al. 2015). Therefore, the pre-onset latencies observed in the present study can be considered to represent a preparatory state that precedes speech processing. Moreover, the TRF method has enabled us to investigate the neural dynamics in an event-related-like manner but with naturalistic speech materials that are expected to better resemble real-world speech comprehension tasks.

Our results highlight in particular that the delta band phase at − 200–0 ms before the speech onset determines the zcomprehension accuracy of the listeners, which serves for the preparation of up-comping speech information. As the analysis was performed on the neural responses to the to-be-processed speech at the pre-onset stage rather than those related to the preceding speech-related information (i.e., attentional cue, preceding words), our study has focused on a distinct processing phase other than previous studies that have mainly explored either the preparatory attention orientation and the speech-specific pre-activations by the preceding speech unit (DeLong et al. 2005; Söderström et al. 2016, 2018). Notably, the employment of the naturalistic speech materials has ensured sufficient variations of the interval between the preceding and the to-be-processed speech unit. Therefore, the present results are not likely to reflect the neural responses to the preceding speech unit. Although there was one study that has reported a sustained difference in the neural activities at around 400 ms prior to the onset of the upcoming speech (Lee et al. 2013), their 'pre-onset' activities mainly reflected the preparatory attention orientation by a visually-presented abstract attentional cue. In contrast, the results of the present study are expected to reflect how the human brain makes use of the rich contextual information in the naturalistic speech materials to infer and prepare for the upcoming speech information.

The timing and the spatial distribution of the reported pre-onset preparatory activity extended previous studies on pre-stimulus oscillatory activities to the speech domain. Visual pre-stimulus studies have generally reported perceptual-relevant neural activities with timings of 100–400 ms prior to stimulus onset (Harris et al. 2018; Wöstmann et al. 2019). In our research, we have found a similar time window for those studies. Our results, therefore, suggest that approximately 0–200 ms prior to stimulus

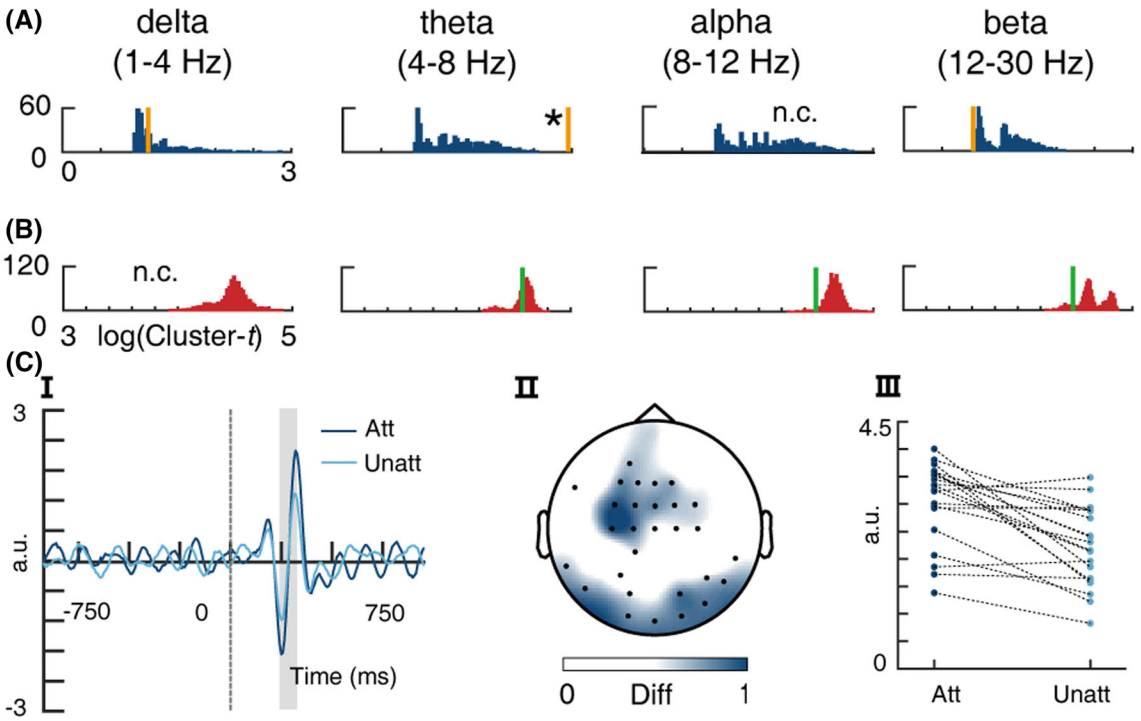


Fig. 4 Attention related TRF-based responses to naturalistic speech. **a** Distributions of the cluster-level *t*-statistics (log-transformation) from the 1000 permutated calculations of the paired comparisons between the TRF-based attended and unattended amplitude responses at the four frequency bands. The y-axis shows the histogram counts of clusters from the permutated calculation at the corresponding log (Cluster-*t*) value. Vertical orange or green lines indicate the max cluster-level *t*-statistics in real data and the asterisk shows the statistically significant cluster from the real data. N.C. means no cluster was formed. **b** Distributions of the cluster-level *t*-statistics from the 1, 000 permutated calculations of the paired comparisons between the TRF-based attended and unattended phase responses at the four frequency bands. **c** Illustration of the significant amplitude-response cluster at the theta band. I. The time course of the TRFs at one representative channel (FC1). The two waveforms represent the average responses to the attended (dark blue) and unattended (light blue) speech. The shaded region depicts the time window of interest with significant differences between the two waveforms. II. The topography of the average amplitude response differences in the time window of interest. Black dots indicate the channels of interest in the cluster. III. The amplitude response difference per participant

onset might have a general implication for the preparation of sensory information processing. As the present study employed naturalistic (speech) materials that are expected to provide much richer contextual information as compared to the simple and abstract stimuli in most of the previous studies, our results might provide more reliable support for the observed timing on preparatory sensory processing. The result is also in line with a recent study, which indicated that the delta band entrainment at 100 ms before the stimulus is related to the noise-induced comprehension difference (Etard and Reichenbach 2019). Meanwhile, the electrodes in central-parietal regions found in our study involved in the preparatory processing are also consistent with the previous auditory pre-stimulus studies (Stefanics et al. 2010; Kayser et al. 2016). The central-parietal responses could be related to the predictive processing of speech meaning and could recruit a mechanism that is similar to that underlying the classical central-parietal N400 response (Federmeier 2007; Lau et al. 2008; Szewczyk and Schriefers 2018) or possibly preparatory attentional orientation as well (Holmes et al. 2016). The left-lateralized parietal electrodes could also be linked to the speech-specific processing, e.g., the Wernicke's area (Hickok and Poeppel 2007).

Our study expanded on findings from the studies about the delta band's functional role in speech preparatory processing. The delta band could help the segmentation or identification of intonation phrases, which is essential for the preparation and prediction of upcoming speech (Giraud and Poeppel 2012; Ding et al. 2015; Kösem et al. 2018; Meyer 2018; Morillon et al. 2019). In particular, the phase of the delta band before the stimulus is related to the hit rate afterward (Ng et al. 2012), or the behavioral consequence (i.e., the reaction time) in the auditory studies using simple, isolated stimulus (Lakatos et al. 2008; Stefanics et al. 2010; Henry and Obleser 2012). It is in contrast to visual studies in which the alpha band is most pronounced in the pre-stimulus stage (Busch et al. 2009; Mathewson et al. 2011; Milton and Pleydell-Pearce 2016), suggesting active auditory perception is dominated by lower-

frequency (Ng et al. 2012; VanRullen 2016). In addition, the brain activity has been shown to be capable of dynamically tracks speech streams using the delta phase (Zion Golumbic et al. 2013a), and temporal predictions have been reported to be encoded by delta neural oscillations (Morillon and Baillet 2017; Auksztulewicz et al. 2018). It should be further noted that although we did not find any significant correlation between the ITPL and the comprehension performance, the observed ITPL values within the found clusters (mostly within the range of 0.1–0.4, as depicted in Fig. 3c.IV, d.IV) implied relatively consistent phases across the trials as referenced to previous studies (Sorati and Behne 2019; Teng and Poeppel 2020), in support of a reliable estimation of the phase angles. Taken together, our findings are in line with the previous studies about the delta band phase's role in preparatory processing and unambiguously show that the comprehension performance could be elevated when the delta is better prepared at a particular phase angle.

The neural mechanisms of the preparatory process were investigated using correlation analysis of the TRF-based neural activity. In line with recent TRF-based studies, we observed attention-related neural responses (Mirkovic et al. 2015, 2016; O'Sullivan et al. 2015), with the peak attention effect represented by theta amplitude activities at 250–320 ms post-stimulus onset over the central and occipital electrodes. In contrast, the comprehension-related post-onset neural signatures were in the alpha band at 650–980 ms. The result could be interpreted for a functional role of alpha-band for a top-down control mechanism to achieve the preparatory process, as the phase of alpha oscillations has an active role in attentional temporal predictions (Händel et al. 2011; Bonnefond and Jensen 2012; Samaha et al. 2015). Meanwhile, in auditory studies, the alpha frequency band has also been suggested to be associated with working memory (Bonnefond and Jensen 2012; Meyer 2018), capable of storing semantics in sentences (Haarmann and Cameron 2005) and syntax information (Bonhage et al. 2017) and lexical decision (Strauß et al. 2015), which are also closely related to speech comprehension. It should be admitted that the observed alpha signature took a phase-based form at a relatively late latency that was different from most of the previous studies. While the above-mentioned studies could provide some hints for its possible functional role, further studies are necessary to refine the interpretation. Nevertheless, the temporal dissociation of the neural signatures at delta and alpha frequency bands further supports the functional specificity of the delta band for preparatory speech processing.

The neural mechanisms of the preparatory process were further explored by inspecting their relationship with attention. While our results are in line with previous research that has reported low-frequency phase track the envelope of attended speech (Ding and Simon 2012a; Zion Golumbic et al. 2013b), we provide further evidence on how such interactions could affect behavior (i.e., comprehension). Indeed, we only found the correlation between the attended TRF-based neural activities and the comprehension performance in the pre-onset stage, suggesting possible attentional facilitation of preparation of speech processing.

This study has some limitations that should be noted. While the use of naturalistic speech materials is expected to better resemble the real-world speech comprehension scenarios, the paradigm could be further improved by presenting both speech streams to both ears (Bidet-Caulet et al. 2007; Oberfeld and Klöckner-Nowotny 2016; Bednar and Lalor 2020). Still, the use of naturalistic speech materials has made it difficult to infer which specific types of information (e.g., timing, phoneme, etc.) were the main contributor for the observed preparatory activities. The distributed brain regions involved in the preparatory process may provide a guidance for designing further experiments to have an in-depth exploration. The present study used the speech amplitude envelope as the reference signal from which the TRF models were derived, which could reflect the speech information at all linguistic levels due to the highly redundant information shared across levels (Di Liberto et al. 2015; Daube et al. 2019). While such an operation has the advantage of providing a general overview about preparatory processing, further investigations are necessary to differentiate possible contributions at different linguistic levels (Di Liberto et al. 2015; Broderick et al. 2018). Meanwhile, caution must be taken when interpreting the timing of the preparatory activities. While the preparatory activity as early as $\sim$ 200 ms before speech onset could be the result of an optimized utilization of the rich contextual information provided by the naturalistic speech materials, such timings may be dependent upon the materials per se. Alternatively, the regularity of the speech materials could potentially lead to preparatory-like responses. Further studies are necessary to extensively investigate the possible material dependence of these timings, for instance, by employed an extended amount of speech materials. Also, denser MEG recordings together with source localization methods are expected to more precisely identify the brain regions for preparatory speech processing (Mazaheri et al. 2009; Nolte and Müller 2010). Besides, due to design limitations, we analyzed the correspondence between trial-averaged comprehension and trial-averaged phase/amplitude TRFs at an inter-individual level. While the average comprehension questionnaire accuracies across all stories within each participant were employed to provide a more reliable estimation of the speech comprehension performance than the single-trial

accuracies, our results do not necessarily imply that the observed neural signatures reflect the participants' trait-like, stable speech processing style. Alternatively, it could be more plausible to consider these neural signatures to reflect a more or less efficient speech processing state. Further studies are needed to clarify the underlying mechanisms, for instance, by using speech audio trials of longer duration together with a more comprehensive test per trial to allow for single-trial analyses.

Abbreviations EEG: Electroencephalogram; ERP: Event-related potential; IC: Independent component; TRF: Temporal response function

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in project Crossmodal Learning (Grant No.: NSFC 62061136001/DFG TRR-169/C1, B1), the National Natural Science Foundation of China (grant number: 61977041 and U1736220), and Tsinghua University Initiative Scientific Research Program (Grant No.: 20197010006). The authors would like to thank Prof. Dr. Xiaoqin Wang and Dr. Yue Ding for providing the shielded room for the experiment as well as necessary technical support.

Author contributions J.L. drafted the paper, conducted the experiment and data analysis; D.Z. designed the experiment and drafted the paper; B.H., G.D., and A.K.E. edited the manuscript.

Data availability Dataset generated in our study has been uploaded to OSF (https://osf.io/87srv/?view_only=13f01e1f1f7b4cf98555ffacd878a53b). We have also provided the experimental materials, including the speech audios and the comprehension questions. The data analysis codes are available upon request.

Declarations

Conflict of interests The authors declare no competing interests.

References

Alexandrou AM, Saarinen T, Kujala J, Salmelin R (2018) Cortical entrainment: what we can learn from studying naturalistic speech perception. Lang Cogn Neurosci 35:681–693. https://doi.org/10.1080/23273798.2018.1518534

Arnal LH, Wyart V, Giraud AL (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14:797–801. https://doi.org/10.1038/nn.2810

Auksztulewicz R, Friston K (2015) Attentional enhancement of auditory mismatch responses: a DCM/MEG study. Cereb Cortex 25:4273–4283. https://doi.org/10.1093/cercor/bhu323

Auksztulewicz R, Schwiedrzik CM, Thesen T et al (2018) Not all predictions are equal: "what" and "when" predictions modulate activity in auditory cortex through different mechanisms. J Neurosci 38:8680–8693. https://doi.org/10.1523/JNEUROSCI.0369-18.2018

Avramiea AE, Hardstone R, Lueckmann JM et al (2020) Pre-stimulus phase and amplitude regulation of phase-locked responses is maximized in the critical state. Elife 9:1–17. https://doi.org/10.7554/eLife.53016

Bednar A, Lalor EC (2020) Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG. Neuroimage 205:116283. https://doi.org/10.1016/j.neuroimage.2019.116283

Bednar A, Lalor EC (2018) Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG. Neuroimage 181:683–691. https://doi.org/10.1016/j.neuroimage.2018.07.054

Berens P (2009) CircStat: a MATLAB toolbox for circular statistics. J Stat Softw 31:293–295

Bidet-Caulet A, Fischer C, Besle J et al (2007) Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J Neurosci 27:9252–9261. https://doi.org/10.1523/JNEUROSCI.1402-07.2007

Bonhage CE, Meyer L, Gruber T et al (2017) Oscillatory EEG dynamics underlying automatic chunking during sentence processing. Neuroimage 152:647–657. https://doi.org/10.1016/j.neuroimage.2017.03.018

Bonnefond M, Jensen O (2012) Alpha oscillations serve to protect working memory maintenance against anticipated distracters. Curr Biol 22:1969–1974. https://doi.org/10.1016/j.cub.2012.08.029

Brainard DH, Brainard DH (1997) The psychophysics toolbox. In: Spatial vision, pp 433–436

Broderick MP, Anderson AJ, Di Liberto GM et al (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Broderick MP, Anderson AJ, Lalor EC (2019) Semantic context enhances the early auditory encoding of natural speech. J Neurosci 39:7564–7575. https://doi.org/10.1523/JNEUROSCI.0584-19.2019

Busch NA, Dubois J, VanRullen R (2009) The phase of ongoing EEG oscillations predicts visual perception. J Neurosci 29:7869–7876. https://doi.org/10.1523/JNEUROSCI.0113-09.2009

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604. https://doi.org/10.3389/fnhum.2016.00604

Daube C, Ince RAA, Gross J (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr Biol 29:1924-1937.e9. https://doi.org/10.1016/j.cub.2019.04.067

de Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. Neuron 102:280–293

DeLong KA, Urbach TP, Kutas M (2005) Probabilistic word preactivation during language comprehension inferred from electrical brain activity. Nat Neurosci 8:1117–1121. https://doi.org/10.1038/nn1504

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci USA 109:505–514. https://doi.org/10.1073/pnas.1113427109

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Dikker S, Pylkkänen L (2013) Predicting language: MEG evidence for lexical preactivation. Brain Lang 127:55–64. https://doi.org/10.1016/j.bandl.2012.08.004

Ding N, Melloni L, Zhang H et al (2015) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164. https://doi.org/10.1038/nn.4186

Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad

Sci U S A 109:11854–11859. https://doi.org/10.1073/pnas.1205381109
Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89. https://doi.org/10.1152/jn.00297.2011
Donhauser PW, Baillet S (2020) Two distinct neural timescales for predictive speech processing. Neuron 105:385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019
ElShafei HA, Bouet R, Bertrand O, Bidet-Caulet A (2018) Two sides of the same coin: distinct sub-bands in the α rhythm reflect facilitation and suppression mechanisms during auditory anticipatory attention. eNeuro 5:1–14. https://doi.org/10.1523/ENEURO.0141-18.2018
Engel AK, Gerloff C, Hilgetag CC, Nolte G (2013) Intrinsic coupling modes: multiscale interactions in ongoing brain activity. Neuron 80:867–886. https://doi.org/10.1016/j.neuron.2013.09.038
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. J Neurosci 39:5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019
Federmeier KD (2007) Thinking ahead: the role and roots of prediction in language comprehension. Psychophysiology 44:491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x.Thinking
Friederici AD (2012) The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn Sci 16:262–268. https://doi.org/10.1016/j.tics.2012.04.001
Fries P (2015) Rhythms for cognition: communication through coherence. Neuron 88:220–235. https://doi.org/10.1016/j.neuron.2015.09.034
Friston K (2005) A theory of cortical responses. Philos Trans R Soc B Biol Sci 360:815–836. https://doi.org/10.1098/rstb.2005.1622
Galindo-Leon EE, Stitt I, Pieper F et al (2019) Context-specific modulation of intrinsic coupling modes shapes multisensory processing. Sci Adv 5:1–13. https://doi.org/10.1126/sciadv.aar7633
Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: Emerging computational principles and operations. Nat Neurosci 15:511–517. https://doi.org/10.1038/nn.3063
Haarmann H, Cameron K (2005) Active maintenance of sentence meaning in working memory: evidence from EEG coherences. Int J Psychophysiol 57:115–128. https://doi.org/10.1016/j.ijpsycho.2005.03.017
Händel BF, Haarmeier T, Jensen O (2011) Alpha oscillations correlate with the successful inhibition of unattended stimuli. J Cogn Neurosci 23:2494–2502. https://doi.org/10.1162/jocn.2010.21557
Harris AM, Dux PE, Mattingley JB (2018) Detecting unattended stimuli depends on the phase of prestimulus neural oscillations. J Neurosci 38:3092–3101. https://doi.org/10.1523/jneurosci.3006-17.2018
Henry MJ, Obleser J (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc Natl Acad Sci 109:20095–20100. https://doi.org/10.1073/pnas.1213390109
Hickok G (2012) Computational neuroanatomy of speech production. Nat Rev Neurosci 13:135–145. https://doi.org/10.1038/nrn3158
Hickok G, Houde J, Rong F (2011) Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69:407–422. https://doi.org/10.1016/j.neuron.2011.01.019
Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402. https://doi.org/10.1038/nrn2113
Hill KT, Miller LM (2010) Auditory attentional control and selection during cocktail party listening. Cereb Cortex 20:583–590. https://doi.org/10.1093/cercor/bhp124
Hisagi M, Shafer VL, Strange W, Sussman ES (2015) Neural measures of a Japanese consonant length discrimination by Japanese and American English listeners: effects of attention. Brain Res 1626:218–231. https://doi.org/10.1016/j.brainres.2015.06.001
Holmes E, Kitterick PT, Summerfield AQ (2018) Cueing listeners to attend to a target talker progressively improves word report as the duration of the cue-target interval lengthens to 2,000 ms. Attent Percept Psychophys 80:1520–1538. https://doi.org/10.3758/s13414-018-1531-x
Holmes E, Kitterick PT, Summerfield AQ (2016) EEG activity evoked in preparation for multi-talker listening by adults and children. Hear Res 336:83–100. https://doi.org/10.1016/j.heares.2016.04.007
Huang Q, Jia J, Han Q, Luo H (2018) Fast-backward replay of sequentially memorized items in humans. Elife 7:e35164. https://doi.org/10.7554/eLife.35164.001
Iemi L, Busch NA, Laudini A et al (2019) Multiple mechanisms link prestimulus neural oscillations to sensory responses. Elife 8:e43620. https://doi.org/10.1101/461558
Jehee JFM, Ballard DH (2009) Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. PLoS Comput Biol 5:e1000373. https://doi.org/10.1371/journal.pcbi.1000373
Kayser SJ, McNair SW, Kayser C (2016) Prestimulus influences on auditory perception from sensory representations and decision processes. Proc Natl Acad Sci 113:4842–4847. https://doi.org/10.1073/pnas.1524087113
Kleen JK, Testorf ME, Roberts DW et al (2016) Oscillation phase locking and late ERP components of intracranial hippocampal recordings correlate to patient performance in a working memory task. Front Hum Neurosci 10:1–14. https://doi.org/10.3389/fnhum.2016.00287
Kok P, Jehee JFM, de Lange FP (2012) Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75:265–270. https://doi.org/10.1016/j.neuron.2012.04.034
Kok P, Mostert P, De Lange FP (2017) Prior expectations induce prestimulus sensory templates. Proc Natl Acad Sci USA 114:10473–10478. https://doi.org/10.1073/pnas.1705652114
Kong YY, Mullangi A, Ding N (2014) Differential modulation of auditory responses to attended and unattended speech in different listening conditions. Hear Res 316:73–81. https://doi.org/10.1016/j.heares.2014.07.009
Kösem A, Bosker HR, Takashima A et al (2018) Neural entrainment determines the words we hear. Curr Biol 28:2867-2875.e3. https://doi.org/10.1016/j.cub.2018.07.023
Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol 62:621–647. https://doi.org/10.1146/annurev.psych.093008.131123
Lakatos P, Karmos G, Mehta AD et al (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320:110–113. https://doi.org/10.1126/science.1154735
Lalor EC, Pearlmutter BA, Reilly RB et al (2006) The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage 32:1549–1561. https://doi.org/10.1016/j.neuroimage.2006.05.054
Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci 9:920–933. https://doi.org/10.1038/Nrn2532
Lee AKC, Rajaram S, Xia J et al (2013) Auditory selective attention reveals preparatory activity in different cortical regions for

selection based on source location and source pitch. Front Neurosci 6:1–9. https://doi.org/10.3389/fnins.2012.00190
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
Marzecová A, Widmann A, SanMiguel I et al (2017) Interrelation of attention and prediction in visual processing: Effects of task-relevance and stimulus probability. Biol Psychol 125:76–90. https://doi.org/10.1016/j.biopsycho.2017.02.009
Mathewson KE, Lleras A, Beck DM et al (2011) Pulsed out of awareness: EEG alpha oscillations represent a pulsed-inhibition of ongoing cortical processing. Front Psychol 2:1–15. https://doi.org/10.3389/fpsyg.2011.00099
Mazaheri A, Nieuwenhuis ILC, Van Dijk H, Jensen O (2009) Prestimulus alpha and mu activity predicts failure to inhibit motor responses. Hum Brain Mapp 30:1791–1800. https://doi.org/10.1002/hbm.20763
Meyer L (2018) The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. Eur J Neurosci 48:2609–2621. https://doi.org/10.1111/ejn.13748
Milton A, Pleydell-Pearce CW (2016) The phase of pre-stimulus alpha oscillations influences the visual perception of stimulus timing. Neuroimage 133:53–61. https://doi.org/10.1016/j.neuroimage.2016.02.065
Mirkovic B, Bleichner MG, De Vos M, Debener S (2016) Target speaker detection with concealed EEG around the ear. Front Neurosci 10:1–11. https://doi.org/10.3389/fnins.2016.00349
Mirkovic B, Debener S, Jaeger M, De VM (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng 12:046007. https://doi.org/10.1088/1741-2560/12/4/046007
Morillon B, Arnal LH, Schroeder CE, Keitel A (2019) Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. Neurosci Biobehav Rev 107:136–142. https://doi.org/10.1016/j.neubiorev.2019.09.012
Morillon B, Baillet S (2017) Motor origin of temporal predictions in auditory attention. Proc Natl Acad Sci 114:E8913–E8921. https://doi.org/10.1073/pnas.1705373114
Müller JA, Wendt D, Kollmeier B et al (2019) Effect of Speech Rate on Neural Tracking of Speech. Front Psychol 10:1–15. https://doi.org/10.3389/fpsyg.2019.00449
Ng BSW, Schroeder T, Kayser C (2012) A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J Neurosci 32:12268–12276. https://doi.org/10.1523/jneurosci.1877-12.2012
Nolden S, Ibrahim CN, Koch I (2019) Cognitive control in the cocktail party: Preparing selective attention to dichotically presented voices supports distractor suppression. Attent Percept Psychophys 81:727–737. https://doi.org/10.3758/s13414-018-1620-x
Nolte G, Müller KR (2010) Localizing and estimating causal relations of interacting brain rhythms. Front Hum Neurosci 4:1–5. https://doi.org/10.3389/fnhum.2010.00209
O'Sullivan JA, Power AJ, Mesgarani N et al (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706. https://doi.org/10.1093/cercor/bht355
Oberfeld D, Klöckner-Nowotny F (2016) Individual differences in selective attention predict speech identification at a cocktail party. Elife 5:1–24. https://doi.org/10.7554/eLife.16747
Oostenveld R, Fries P, Maris E, Schoffelen J (2011) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput Intell Neurosci 2011:1–9. https://doi.org/10.1155/2011/156869
Pasley BN, David SV, Mesgarani N et al (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:e1001251. https://doi.org/10.1371/journal.pbio.1001251
Pisoni DB, Luce PA (1987) Acoustic-phonetic representations in word recognition. Cognition 25:21–52. https://doi.org/10.1016/0010-0277(87)90003-5
Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87. https://doi.org/10.1038/4580
Rassi E, Wutz A, Müller-Voggel N, Weisz N (2019) Prestimulus feedback connectivity biases the content of visual experiences. Proc Natl Acad Sci 116:16056–16061. https://doi.org/10.1073/pnas.1817317116
Ringach D, Shapley R (2004) Reverse correlation in neurophysiology. Cogn Sci 28:147–166. https://doi.org/10.1016/j.cogsci.2003.11.003
Samaha J, Bauer P, Cimaroli S, Postle BR (2015) Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. Proc Natl Acad Sci 112:8439–8444. https://doi.org/10.1073/pnas.1520473112
Schröger E, Kotz SA, SanMiguel I (2015a) Bridging prediction and attention in current research on perception and action. Brain Res 1626:1–13. https://doi.org/10.1016/j.brainres.2015.08.037
Schröger E, Marzecová A, Sanmiguel I (2015b) Attention and prediction in human audition: a lesson from cognitive psychophysiology. Eur J Neurosci 41:641–664. https://doi.org/10.1111/ejn.12816
Smith JL, Johnstone SJ, Barry RJ (2006) Effects of pre-stimulus processing on subsequent events in a warned Go/NoGo paradigm: Response preparation, execution and inhibition. Int J Psychophysiol 61:121–133. https://doi.org/10.1016/j.ijpsycho.2005.07.013
Smout CA, Tang MF, Garrido MI, Mattingley JB (2019) Attention promotes the neural encoding of prediction errors. PLoS Biol 17:1–22. https://doi.org/10.1371/journal.pbio.2006812
Söderström P, Horne M, Frid J, Roll M (2016) Pre-activation negativity (PrAN) in brain potentials to unfolding words. Front Hum Neurosci 10:1–11. https://doi.org/10.3389/fnhum.2016.00512
Söderström P, Horne M, Mannfolk P et al (2018) Rapid syntactic pre-activation in Broca's area: concurrent electrophysiological and haemodynamic recordings. Brain Res 1697:76–82. https://doi.org/10.1016/j.brainres.2018.06.004
Sonkusare S, Breakspear M, Guo C (2019) Naturalistic stimuli in neuroscience: critically acclaimed. Trends Cogn Sci 23:699–714. https://doi.org/10.1016/j.tics.2019.05.004
Sorati M, Behne DM (2019) Musical expertise affects audiovisual speech perception: findings from event-related potentials and inter-trial phase coherence. Front Psychol 10:1–19. https://doi.org/10.3389/fpsyg.2019.02562
Stefanics G, Hangya B, Hernadi I et al (2010) Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. J Neurosci 30:13578–13585. https://doi.org/10.1523/JNEUROSCI.0703-10.2010
Strauß XA, Henry MJ, Scharinger XM, Obleser XJ (2015) Alpha Phase Determines Successful Lexical Decision in Noise 35:3256–3262. https://doi.org/10.1523/JNEUROSCI.3357-14.2015
Szewczyk JM, Schriefers H (2018) The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. Lang Cogn Neurosci 33:665–686. https://doi.org/10.1080/23273798.2017.1401101
Teng X, Ma M, Yang J (2020) Constrained structure of ancient chinese poetry facilitates speech content grouping. Curr Biol 30:1–7. https://doi.org/10.1016/j.cub.2020.01.059

Teng X, Poeppel D (2020) Theta and gamma bands encode acoustic dynamics over wide-ranging timescales. Cereb Cortex 30:2600–2614. https://doi.org/10.1093/cercor/bhz263

Tian X, Ding N, Teng X et al (2018) Imagined speech influences perceived loudness of sound. Nat Hum Behav 2:225–234. https://doi.org/10.1038/s41562-018-0305-8

Van Dijk H, Schoffelen JM, Oostenveld R, Jensen O (2008) Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. J Neurosci 28:1816–1823. https://doi.org/10.1523/JNEUROSCI.1853-07.2008

Van Petten C, Luka BJ (2012) Prediction during language comprehension: benefits, costs, and ERP components. Int J Psychophysiol 83:176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

VanRullen R (2016) Perceptual cycles. Trends Cogn Sci 20:723–735. https://doi.org/10.1016/j.tics.2016.07.006

Verhulst S, Altoè A, Vasilkov V (2018) Computational modeling of the human auditory periphery: auditory-nerve responses, evoked potentials and hearing loss. Hear Res 360:55–75. https://doi.org/10.1016/j.heares.2017.12.018

Wang L, Kuperberg G, Jensen O (2018) Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. Elife 7:1–24. https://doi.org/10.7554/eLife.39061

Wöstmann M, Waschke L, Obleser J (2019) Prestimulus neural alpha power predicts confidence in discriminating identical auditory stimuli. Eur J Neurosci 49:94–105. https://doi.org/10.1111/ejn.14226

Zhang ZG, Hu L, Hung YS et al (2012) Gamma-band oscillations in the primary somatosensory cortex-a direct and obligatory correlate of subjective pain intensity. J Neurosci 32:7429–7438. https://doi.org/10.1523/JNEUROSCI.5877-11.2012

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." J Neurosci 33:1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zion Golumbic EM, Ding N, Bickel S et al (2013b) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." Neuron 77:980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zoefel B, Archer-Boyd A, Davis MH (2018) Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. Curr Biol 28:401-408.e5. https://doi.org/10.1016/j.cub.2017.11.071