Attention-dependent preparatory processing of naturalistic narratives is correlated with speech comprehension

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

[1]Department of Psychology, School of Social Sciences, Tsinghua University, 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

[*]**Correspondence**:

Dan Zhang, Ph.D.

Room 334, Mingzhai Building, Tsinghua University, Beijing 100084, China

E-mail: dzhang@tsinghua.edu.cn; Tel: +86-10-62796737

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 preparatory processing of upcoming speech. Participants selectively attended to one of two competing naturalistic, narrative speech streams, and a temporal response function method was applied to derive event-related-like neural responses from electroencephalographic data. Regression analysis revealed that neural signatures with latencies as early as -450 ms prior to speech onset were significantly correlated with speech comprehension performance. The preparatory process involved a distributed network. These preparatory signatures were attention dependent; activity prior to the attended speech was negatively correlated with comprehension performance, whereas the opposite was found for unattended speech. Our findings suggest that attention plays an important role in the preparation to process upcoming speech.

Keywords

preparatory processing, attention, speech comprehension, electroencephalogram, temporal response function

Introduction

Humans are a powerful speech recognition system that can comprehend complex and rapidly changing human 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 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 (DeWitt & Rauschecker, 2012; Friederici, 2012; Hickok, 2012a; Pisoni & Luce, 1987; Verhulst, Altoè, & Vasilkov, 2018). In recent years, increasing evidence has also suggested that human speech comprehension is an active process that involves top-down predictions (Arnal, Wyart, & Giraud, 2011; Federmeier, 2007; Fries, 2015; Hickok, Houde, & Rong, 2011; Kutas & Federmeier, 2011; Rao & Ballard, 1999; Tian, Ding, Teng, Bai, & Poeppel, 2018). In the cocktail party scenario, it is believed that a listener should continuously predict what their attended speaker is going to say next to efficiently understand the corresponding speech (Cherry, 1953; O'Sullivan et al., 2015; Zion Golumbic, Cogan, Schroeder, & Poeppel, 2013). These predictions supposedly inform the brain about the 'what' and 'when' of upcoming speech information (Arnal & Giraud, 2012; Auksztulewicz et al., 2018), which allows a listener to prepare for follow-up processing.

Although the idea of top-down prediction in human speech comprehension is gaining popularity, it remains unclear how the brain uses predictive information to prepare for the processing of upcoming speech information. Understanding the preparatory process is essential because it reflects the influence of prediction on subsequent information

processing. Moreover, the available findings on prediction in speech are not sufficient to determine the neural mechanisms underlying preparation. For instance, the classic studies of active speech prediction have mainly focused on the 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 (Kutas & Federmeier, 2011; Lau, Phillips, & Poeppel, 2008; Van Petten & Luka, 2012). These ERP components normally occur >400 ms after the presentation of the perceived speech, and so provide only indirect support for the preparatory process. Recent studies have also reported evidence of the brain's pre-activation before the onset of the upcoming speech (DeLong, Urbach, & Kutas, 2005; Dikker & Pylkkänen, 2013; Söderström, Horne, Frid, & Roll, 2016; Söderström, Horne, Mannfolk, van Westen, & Roll, 2018); these pre-activations have been interpreted as 'predictive' because they have been found to be correlated with the relative likelihoods of the upcoming speech unit (e.g. words) in the continuous speech materials (e.g. sentences). However, this is still only indirect evidence for preparation, as these pre-activations have been represented by event-related neural responses to the preceding speech unit that are informative about possible upcoming speech units. Direct neural evidence for the preparatory response should be derived from neural activity that is directly related to the processing of the upcoming speech information, and which occurs immediately before speech onset.

While this direct evidence has not been investigated in the speech domain, several studies on general sensory processing have provided ample support for the existence of

such a preparatory process. For instance, pre-stimulus oscillatory activity has been reported to have a significant impact on subsequent perceptual consequences (Cao, Thut, & Gross, 2017; Galindo-Leon et al., 2019; Harris, Dux, & Mattingley, 2018; Kok, Mostert, & De Lange, 2017; Rassi, Wutz, Müller-Voggel, & Weisz, 2019). Moreover, synchronization within neural populations responsible for the specific sensory processing has been proposed to underlie preparation (Engel, Fries, & Singer, 2001; Galindo-Leon et al., 2019; Lakatos et al., 2009). Following on from this work, the present study investigated neural activity prior to the onset of upcoming speech information to identify possible neural signatures of the preparatory process.

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. Indeed, recent studies have demonstrated the interplay between attention and prediction (Schröger, Kotz, & SanMiguel, 2015; Schröger, Marzecová, & Sanmiguel, 2015). Specifically, the magnitude of the prediction error-related neural response has been shown to be magnified or reversed, depending on the attentional state (Auksztulewicz & Friston, 2015; Hisagi, Shafer, Strange, & Sussman, 2015; Kok, Jehee, & de Lange, 2012; Marzecová, Widmann, SanMiguel, Kotz, & Schröger, 2017; Smout, Tang, Garrido, & Mattingley, 2019). Most of these studies have been conducted within the visual domain, with limited exploration in the auditory domain, let alone speech processing. Compared to vision, the fast temporal dynamics of auditory stimuli and speech signals require neuroimaging tools such as EEG that can

track online changes in neural activity with a high temporal resolution.

The present study aimed to identify neural signatures that directly reflect the preparatory processing of human speech. A 60-channel electroencephalogram (EEG) was recorded from participants while they listened to naturalistic narratives; this procedure is believed to be of high ecological validity and thus to provide necessary contextual information for the engagement of top-down prediction and therefore preparation (Federmeier, 2007; Friston, 2005; Jehee & Ballard, 2009; Rao & Ballard, 1999). A cocktail party paradigm was used, whereby we introduced a complex perceptual environment that imposed further demands on prediction and preparation (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018). To obtain the neural responses to continuous, naturalistic speech, we used a temporal response function (TRF) method, to derive event-related-like neural responses from EEG data, for both the attended and unattended speech streams in the cocktail party scenario (Crosse, Di Liberto, Bednar, & Lalor, 2016; Lalor, Pearlmutter, Reilly, McDarby, & Foxe, 2006). Following studies on the perceptual influence of pre-stimulus neural activities (Iemi et al., 2019; Rassi et al., 2019; Smith, Johnstone, & Barry, 2006), these TRF-based responses were analyzed for neural signatures related to speech comprehension performance, as measured by speech-content-related questionnaires. We considered that performance-relevant TRF-based responses before speech onset would be direct neural evidence for the preparatory process, as the comparison between the predicted and the actual perceived sensorial information cannot be performed during this period. Furthermore, our experimental design allows the investigation on the attention

dependence of preparatory speech processing. Specifically, regression analyses were employed with the TRF-based neural responses to attended and unattended speech streams as the independent variables, and speech comprehension performance as the dependent variable. Results revealed that neural signatures with latencies as early as -450 ms prior to speech onset were significantly correlated with speech comprehension performance. A distributed network was involved in the preparatory process of speech comprehension. The preparatory activity to the attended speech was found to be negatively correlated with comprehension performance, whereas the opposite was found for unattended speech. Our findings suggest that attention plays an important role in the preparation to process upcoming speech.

Results

Twenty participants took part in 28 'cocktail party' trials. In each trial, two narrative stories were presented simultaneously to both the left and the right ears and the participants were instructed to attend to one spatial side. Comprehension performance was evaluated by questionnaires about the story contents, which were implemented at the end of each story. There were two four-choice questions for the two simultaneously heard stories, respectively. The 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%; the four-choice chance level: 25%; $t(19) = 10.95$, $p < .001$). The participants reported a moderate level of attention (8.15±0.34 on a 10-point Likert scale) and attention difficulties (2.04±0.53 on a 10-point Likert scale). The accuracy for the attended story was significantly correlated with both the self-reported attention level (*r*

= .476, p = .043) and attention difficulty (r = -.677, p = .001). The self-reported story familiarity level was low for all the participants (0.86±0.22 on a 10-point Likert scale) and was not correlated with comprehension performance (r = -.224, p = .342). These results suggest that participants' selective attention was effectively manipulated, as well as good reliability of the measured comprehension performance. Most importantly, there was a large inter-individual difference 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.

The analysis workflow is shown in Figure 1. TRF-derived neural responses to the attended and unattended speech were calculated separately, at latencies of -500 ms to 500 ms relative to speech onset. Responses within the -500–0 ms latency window are considered to represent preparatory activity, whereas responses within the 0–500 ms latency window reflect post-processing of the speech stream. These TRF responses also underwent time-frequency analysis, and the average single-trial amplitudes and inter-trial phase-locking (ITPL) values were calculated. We hypothesized that this decomposition into amplitude and phase responses would yield more detailed insights into the underlying neural mechanisms of speech processing, as amplitude and phase have been proposed to play unique roles in networks underlying human cognition (Engel, Gerloff, Hilgetag, & Nolte, 2013; Fries, 2015; Klimesch, 2012). To achieve this, we established linear regression models with either amplitude or phase responses from both the attended and unattended TRFs as the independent variables, and

comprehension accuracy of the attended speech as the dependent variable. The TRF-based amplitude and phase responses at different channels, latencies, and frequencies were used in separate regression models. We used regression analysis to take a full consideration of possible joint contributions from the attended and unattended TRFs, by deriving different regression coefficients respectively. The regression analyses were performed by treating each participant's comprehension accuracy averaged over all the 28 attended stories as the dependent variable. This provides a robust estimation of the comprehension performance, as only two four-choice questions were asked per attended story. In addition, this design could allow more flexibility in neural data analysis, e.g. exploring the inter-story variability in phase responses.

Regression models were built separately for the neural responses at each channel-latency-frequency bin. The regression *R*-values were obtained to reveal how well the regression models correlated with individual comprehension performance. Statistical analyses were performed based on these regression *R*-values using a nonparametric cluster-based permutation method (Maris & Oostenveld, 2007). Any significant results at a latency <0 ms were taken to indicate the neural correlates of preparation for upcoming speech information. We also calculated the mean of the regression coefficient, and drew the distribution for every cluster.

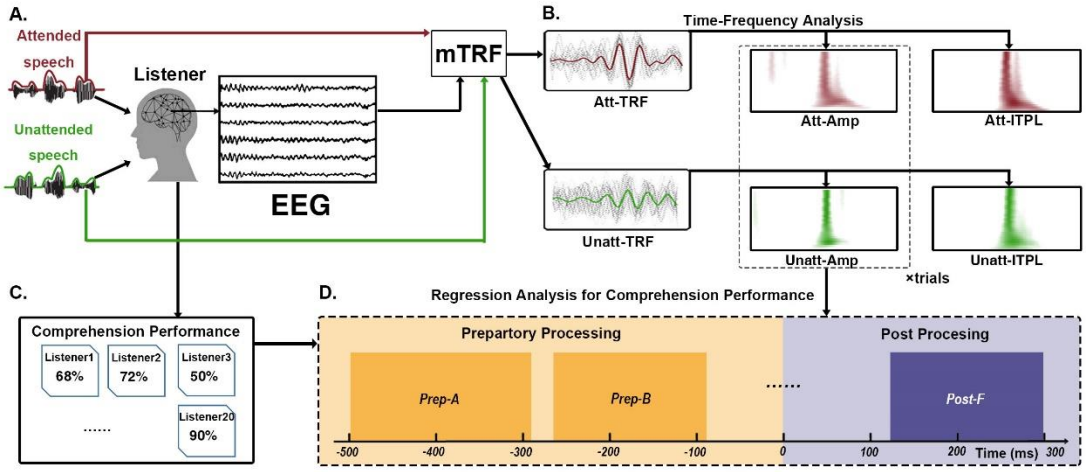

Figure 1. The analysis workflow

(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 then further subjected to a time-frequency analysis at the single-trial level and decomposed into single-trial amplitude and phase (by inter-trial phase locking, ITPL) responses based on the time-frequency representations (denoted by 'Amp' and 'ITPL'). This procedure was conducted for attended (Att-) and unattended (Unatt-) speech streams separately. (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 comprehension performance. (D) Regression analysis for comprehension performance-related neural responses. We established linear regression models with either amplitude or phase responses from both the attended and unattended TRFs as the independent variables, and comprehension accuracy of the attended speech as the dependent variable, for each channel-latency-frequency bin. The neural responses with significant regression model fitting are reported. We defined neural activity before 0 ms as preparatory processing and activity after 0 ms as post processing.

Preparatory neural activities were correlated with speech comprehension performance

The nonparametric cluster-based permutation analysis revealed a significant correlation between the multi-channel time-frequency representation of the TRFs and individual speech comprehension performance of the attended speech. This corresponded to six

clusters in the observed data (all, $p < .05$), as shown in Figure 2.

There were four clusters with latencies prior to 0 ms, which suggested that the brain actively prepares for upcoming speech information. The earliest cluster (*Prep-A*) extended from around -450 ms to -300 ms, and spread from the theta to low beta range (6–18 Hz) over the right parietal region (permutation $p = .005$). The average prediction model within the cluster revealed a significant correlation between the predicted and the actual comprehension performance ($r = .678$, $p = .001$). The following cluster (*Prep-B*) extended from around -400 ms to -350 ms, and spread from the alpha to beta range (12–25 Hz) over the left frontal region (permutation $p = .030$; model prediction $r = .625$, $p = .003$). There were two clusters with latencies around 0 ms. While *Prep-C* was based on amplitude responses, much like *Prep-A* and *Prep-B*, *Prep-D* was based on the ITPL. The amplitude cluster (*Prep-C*) extended from -50 ms to 0 ms, and spread from the alpha to low beta range (9–19 Hz) over the central-parietal region (permutation $p = .015$; model prediction $r = .725$, $p < .001$). The ITPL cluster (*Prep-D*) extended from -30 ms to 20 ms, and spread from the alpha to low beta range (10–17 Hz) over the right central region (permutation $p = .002$; model prediction $r = .746$, $p < .001$).

There were also two clusters with latencies >0 ms. The first cluster (*Post-E*) occurred at 80–200 ms, and spread within the alpha range (7–9 Hz) over the right temporal region (permutation $p = .035$; model prediction $r = .774$, $p < .001$). The other cluster (*Post-F*) occurred at 180–300 ms, and spread from the theta to alpha range (4–11 Hz) over the central parietal region (permutation $p = .008$; model prediction $r = .673$, $p = .001$). These clusters were speech-following responses that likely reflect post processing of

the speech information for comprehension.

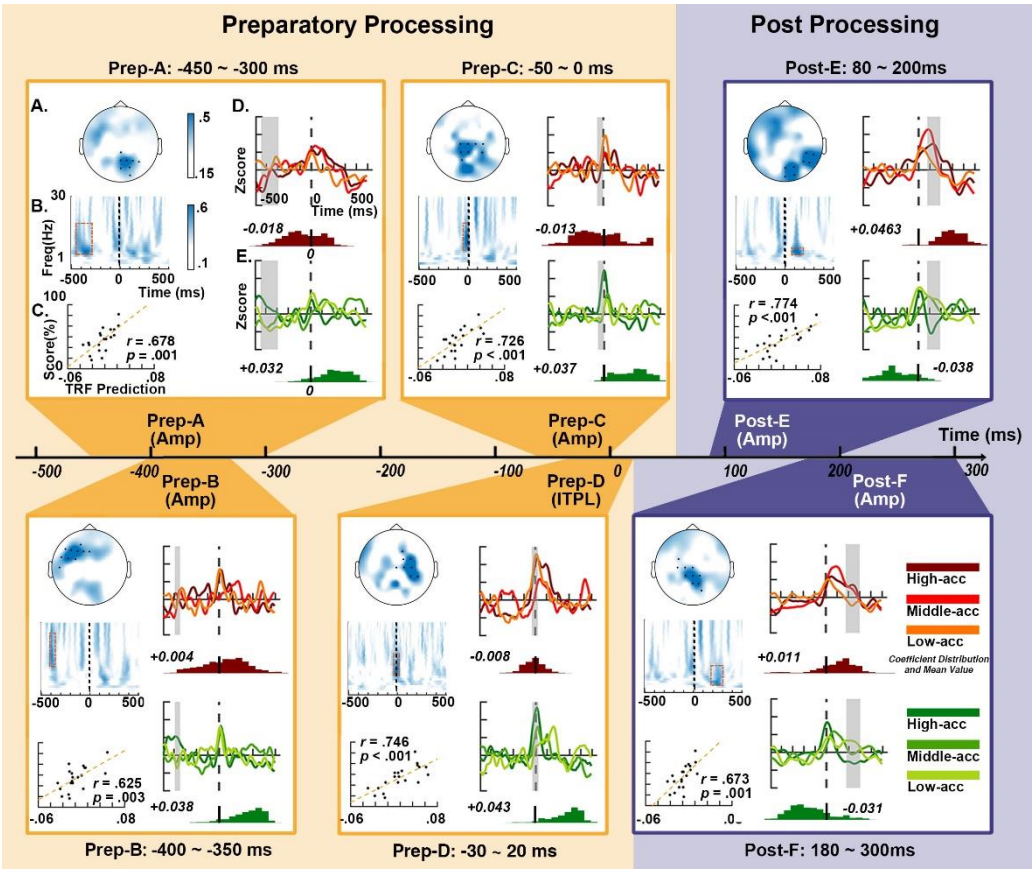

Figure 2. Neural responses that were correlated with speech comprehension performance

There are six clusters showing significant correlations with the comprehension performance of the attended speech, shown as six sub-plots. Each sub-plot is divided into five panels (A~E, labeled only in the upper left sub-plot titled 'Prep-A'), as explained below.

(A) Grand average topography of the regression *R*-values in the time windows of the significant clusters (depicted by the gray shadowed area in (D) and (E)). Black dots indicate the channels of interest included in these clusters.

(B) Time-frequency profile of the *R*-values averaged over the channels of interest. The dashed red rectangles indicate the time and frequency of interest included in the clusters.

(C) Scatter plots of the comprehension performance (y-axis) and the predicted values by the regression model averaged over all the channel-latency-frequency bins within the corresponding clusters (x-axis). Each dot represents an individual participant.

(D) and (E) Response time courses of attended-TRF(D) and unattended-TRF(E). The three red (D) and green (E) lines of different darkness represent the averaged responses over the participants with comprehension performance of the attended speech ranking in the top, middle, and bottom tertiles (7, 6, and 7 participants, respectively). The

histograms illustrate the distribution of coefficients of the selected channel-latency-frequency bins. The number displayed beside the histogram is the mean regression coefficient within the selected channel-latency-frequency bins.

Preparatory activities were attention dependent

The regression models for both attended and unattended responses revealed a joint contribution of the preparatory activities related to the attended and the unattended speech streams for the speech comprehension performance. For all six clusters, the mean regression coefficients within the selected channel-latency-frequency bins were significantly different from zero. For example, the mean of the coefficients for the attended and unattended activities of *Prep-A* were -0.018 and 0.032, respectively (the 99% bootstrap confidence intervals for the attended activity: [-0.019, -0.016]; for unattended activity: [+0.030, +0.033]; for all means and distributions, see Fig. 2D and 2E). Thus, both the attended and unattended activities significantly contributed to individual comprehension performance. By plotting the average TRF responses within the channel-latency-frequency of interest in these clusters for the participants with comprehension performance ranking in the top, middle, and bottom tertiles, we observed primarily reversed trends for the attended and unattended responses. Interestingly, reduced preparatory activities to the attended speech were seen for the top-performance tertile (corresponding to the negative mean coefficients), whereas the opposite effect was seen for the unattended speech (except for *Prep-B*, in which both mean coefficients were positive). The post-processing activities showed the reverse pattern, whereby participants with a better performance exhibited enhanced responses to the attended speech (as reflected by positive mean coefficients) and reduced

responses to the unattended speech (as reflected by negative mean coefficients). The 99% bootstrap confidence intervals for the means of all the coefficients are provided in Table S1.

We further computed partial correlations between these preparatory neural activities and the comprehension performance, while controlling for the post-processing neural activities (i.e., *Post-E* and *Post-F*). *Prep-B* and *Prep-C* had significant partial correlations with the comprehension performance (r = .610 and .560, p = .007 and .016, respectively), which is suggestive of a unique functional contribution to comprehension performance of the two preparatory neural responses. The partial correlations of *Prep-A* and *Prep-D* with the comprehension performance failed to reach significance (r = .32 and .44, p = .189 and .068, respectively). Nevertheless, as these preparatory activities occurred substantially earlier than the post-processing activities, the non-significant results do not necessarily undermine the importance of these responses, but rather indicate there to be shared neural mechanisms for preparation and post-processing. Indeed, both the spectral and spatial signatures of *Prep-A* and *Post-F* largely overlapped, and *Prep-D* and *Post-E* shared similar spatial patterns. In support of these observations, significant correlations were found between *Prep-A* and *Post-F* (r =. 648, p = .002) and between *Prep-D* and *Post-E* (r = .672, p = .001). The pairwise correlations between all six clusters are shown in Table S2, and the partial correlation results are shown in Table S3.

Attention modulation was reflected by post-onset processing: replication of previous TRF-based studies

Several previous studies have employed a similar paradigm to investigate the attentional modulation of speech processing (Broderick et al., 2018; Mirkovic, Debener, Jaeger, & Vos, 2015; O'Sullivan et al., 2015). In these studies, the attentional modulation effect was operationalized as the difference between TRF responses to attended and unattended speech, whereas the present study focused on neural signatures that were associated with speech comprehension performance. To replicate their findings, we computed the attention-related neural activities according to these previous studies.

To this end, a cluster-based permutation analysis was performed to search for differences between the TRF-based neural responses to the attended and unattended speech, based on the multichannel time-frequency responses of either amplitude or phase responses. Selective attention resulted in significant differences the neural activities related to the attended and unattended speech streams, as represented by an amplitude-based cluster and a phase-based cluster. Both clusters had latencies well after 0 ms. Compared to the unattended speech, the attended speech was associated with a larger amplitude response at 50–180 ms at 5–13 Hz over the parietal region (p = .018) and a stronger ITPL at 100–260 ms at 4–8 Hz over the left-central parietal region (p = .028). None of the clusters, however, were significantly correlated with speech comprehension performance (r = -.110 and -.187, p = .646 and .431).

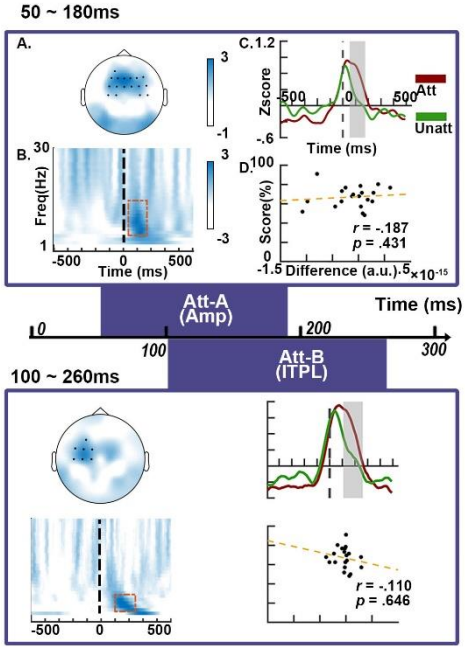

Figure 3. Attention-related neural responses were not correlated with speech comprehension performance

There are two clusters showing significant difference between attended and unattended speech streams. Each sub-plot is divided into four panels (A~D, labeled only in the upper sub-plot), as explained below.

(A) Grand average topography of *t*-values on the difference between the attended and unattended responses in the time windows of the significant clusters (depicted by the gray shadowed area in (C)). Black dots represent the channels of interest.

(B) Time-frequency profile of the *t*-values averaged over the channels of interest. The dashed red rectangles indicate the time and frequency of interest included in the clusters.

(C) Grand average time courses of the attended (red) and unattended (green) TRF responses averaged over the channels and frequencies of interest. Gray shaded areas indicate the time windows of interest.

(D) Scatter plots of the comprehension performance (y-axis) and the response differences averaged over all channel-latency-frequency bins within the corresponding clusters (x-axis). Each dot represents one participant.

Discussion

The present study aimed to identify neural signatures that directly reflect preparatory processing of upcoming speech. We used naturalistic narrative speech materials in a

selective attention paradigm using a TRF-based approach for modeling the neural activity, and observed preparatory neural activities before the onset of speech power envelope fluctuations. These preparatory activities were correlated with the comprehension performance of individual participants, with latencies as early as -450 ms. The preparatory process involved spatially distributed brain areas, taking the form of an amplitude response rather than phase synchronization, with the most relevant frequencies within the alpha and beta ranges. There was also an interplay between attention and preparation, whereby preparatory activities to the attended and the unattended speech contributed to comprehension performance, but with opposite mechanisms. 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 prior to speech onset 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, Jia, Han, & Luo, 2018). Although our findings have mainly focused 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. Therefore, the pre-onset latencies observed in the present study can be considered to represent a preparatory state that precedes

363 speech processing.

364 **Preparatory activities involve a distributed neural network**

365 To prepare for the processing of upcoming speech information, multiple neural
366 signatures with different time, space, and frequency characteristics were identified,
367 which is indicative of the engagement of multi-center neural networks for active speech
368 perception. We did not base our analysis on preselected regions of interest, and so our
369 results provide a complete overview of all activities for preparation. Notably, one
370 preparatory cluster was found to be located over the left frontal region (*Prep-B*), which
371 supports the popular notion of a left-lateralized frontal network for top-down speech
372 prediction (Federmeier, 2007; Hickok, 2012b). Previous studies that have reported
373 involvement of the left frontal region in prediction have focused on post-processing of
374 either violations of linguistic congruency (e.g. the MMN and N400 responses; (Kutas
375 & Hillyard, 1984; Lau et al., 2008; Szewczyk & Schriefers, 2018) or contextual speech
376 cues (Dikker & Pylkkänen, 2013; Söderström et al., 2016). However, our results
377 indicate that this region plays an active role in preparation of speech processing, with
378 latencies of ~400 ms before speech onset.

379 Furthermore, the preparatory process was broadly distributed beyond the left frontal
380 region, including the parietal (*Prep-A*), central-parietal (*Prep-C*), and right central
381 (*Prep-D*) regions. The central-parietal responses could be related to the predictive
382 processing of speech meaning, and could recruit a mechanism that is similar to that
383 underlying the classical central-parietal N400 response (Federmeier, 2007; Lau et al.,

2008; Szewczyk & Schriefers, 2018). The right-lateralized finding (i.e., *Prep-D*), however, may indicate a possible functional contribution of the right hemisphere to prediction. Some studies have suggested that the right hemisphere is engaged in language processing, primarily during complex narratives (Brownell HH, Michel D, Powelson J, & Gardner H, 1983; George, Kutas, Martinez, & Sereno, 1999; Robertson et al., 2000). As naturalistic speech materials used in the present study were likely to engage speech processing at all levels, our results demonstrate the involvement of a distributed neural network for the preparation of naturalistic speech processing.

Higher frequency neural activity for preparation and lower frequency activity for post-processing

The neural mechanisms of the preparatory process were investigated using a time-frequency analysis of the TRF-based neural activity. Similarly to recent TRF-based studies, we observed attention-related neural responses (Mirkovic, Bleichner, De Vos, & Debener, 2016; Mirkovic et al., 2015; O'Sullivan et al., 2015), with the peak attention effect represented by theta and alpha oscillatory activities at 100–200 ms post-stimulus onset over the frontal regions. Similarly, the comprehension-related post-onset processing was mainly reflected in lower frequency bands from theta to low alpha bands, as has been frequently reported in previous speech literatures (Ding & Simon, 2014; Giraud & Poeppel, 2012; Luo & Poeppel, 2007). In contrast, the comprehension-related pre-onset neural signatures were in a higher frequency range, mainly within the alpha and beta bands. This observation is in accordance with recent studies on pre-stimulus ERPs in sensory perception, in which pre-stimulus alpha-band activity was

found to be significantly correlated with post-stimulus perception, especially in the visual modalities (Bauer, Stenner, Friston, & Dolan, 2014; Milton & Pleydell-Pearce, 2016; Rohenkohl & Nobre, 2011; van Ede, Jensen, & Maris, 2010). These results have been interpreted for a functional role of alpha band for a top-down inhibitory mechanism to achieve the preparatory process. Meanwhile, several studies have suggested that beta-band power reflects updating the content of a prediction (Bauer et al., 2014; Sedley et al., 2016), as well as maintenance of ongoing cognitive context (Engel & Fries, 2010). Accordingly, the pre-onset alpha and beta activity in our study may reflect inhibitory of unattended speech and maintain the expectation in speech preparatory processing (Kayser, Ince, Gross, & Kayser, 2015; Keitel, Gross, & Kayser, 2018). Taken together, these results suggest possibly of distinct functional roles of neural activity at different frequency bands for speech processing, with alpha- or higher-band activity reflecting top-down speech preparation, and the lower-frequency activity reflecting post-stimulus processing.

In addition, three out of the four preparatory activities took the form of an amplitude response rather than ITPL, including the earliest activities (*Prep-A* and *Prep-B*). At a first glance, our observation may seem to be inconsistent with the popular view on the functional roles of amplitude and phase responses, as phase synchronization is frequently suggested to reflect the coordination of long-distance neuron communication and therefore more likely to reflect top-down regulation of sensory information processing (Engel et al., 2001, 2013; Galindo-Leon et al., 2019; Klimesch, 2012; Lakatos et al., 2009; Salinas & Sejnowski, 2001; Sauseng et al., 2007; Schyns, Thut, &

Gross, 2011; Zhang, Hong, Gao, & Röder, 2017). Nevertheless, as the preparation process reflects the usage of top-down predictive information for facilitated speech processing rather than prediction *per se*, it is likely that these preparatory activities mainly exhibited the actual implementation of prediction in speech-processing-specific brain regions. In support of such a hypothesis, the distributed neural network indeed covered the typical speech processing regions. Therefore, the amplitude-based preparatory activities could be the result of localized speech-related information processing as proposed in previous studies on general sensory processing (Engel et al., 2001; Klimesch, Sauseng, & Hanslmayr, 2007; Mathewson et al., 2011; Zhang et al., 2017). These activities could be related to the processing of the contextual information relevant for the upcoming speech and thus provide the basis for the enhanced comprehension performance. As limited previous studies have addressed the dissociation between amplitude and phase responses, further work is necessary to elucidate this issue.

Attention dependence of the preparatory neural signatures

The neural mechanisms of the preparatory process were further explored by inspecting their relationship with attention. We further decomposed neural activity into amplitude and phase responses. Specifically, our findings of the earliest preparation-related amplitude modulation (*Prep-A*) in the alpha and beta bands support the recent reports of beta power reduction during temporal prediction (Arnal & Giraud, 2012; Nobre & Van Ede, 2018); in high-performing participants, comprehension performance was associated with reduced amplitudes in response to attended speech and enhanced

amplitudes in response to unattended speech (i.e., the positive and negative regression coefficients as displayed in Fig. 2D and 2E). Similar patterns were also seen for *Prep-C* and *Prep-D* (ITPL in this case). The reduced amplitude responses and ITPL could reflect a well-prepared state for processing upcoming attended speech (Bauer et al., 2014; Chao, Takaura, Wang, Fujii, & Dehaene, 2018; Jensen & Mazaheri, 2010). The neural activities related to the unattended speech were also correlated with speech performance, but with opposite effects. Thus, the different modulation effects could contribute towards an enlarged activity difference between the neural activities to the attended and the unattended speech, for an efficient processing and thus comprehension of the attended speech information.

Interestingly, although the neural activities related to both the attended and the unattended speech also jointly contributed to comprehension performance at the post-processing stage (*Post-E* and *Post-F*), a reversed pattern was observed as compared to the preparatory stage. In participants with better performances, performance was associated with enhanced responses to the attended speech and reduced responses to the unattended speech. This reversed pattern is in accordance with the classical view on attention modulation, and reflects enhanced processing to attended information and suppressed processing to unattended information (Carrasco, 2011; Luck, Woodman, & Vogel, 2000). The sharp contrast between the preparatory and the post-processing processing stages supports the idea that there is an interplay between preparation and attention. While our results are in line with previous research that has reported there to be an interaction between attention and prediction (Friston, 2009; Kok, Rahnev, Jehee,

Lau, & de Lange, 2012; Smout et al., 2019), we provide further evidence on how such interactions could affect behavior (i.e., comprehension). Namely, given that the preparatory process supposedly reflects how prediction is implemented to facilitate information processing, our results imply that attended speech may be favored by the predictive or preparatory mechanism. Indeed, the neural activity associated with the attended speech was inhibited during the preparatory stage, but enhanced during the post processing stage, both mechanisms have been linked to more efficient processing by previous studies (Rohenkohl & Nobre, 2011; Rommers, Dickson, Norton, Wlotko, & Federmeier, 2017; Smith et al., 2006).

This study has some limitations that should be noted. The present study used the speech power 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 (Daube, Ince, & Gross, 2019; Di Liberto, O'Sullivan, & Lalor, 2015). 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 (Broderick et al., 2018; Di Liberto et al., 2015). Meanwhile, caution must be taken when interpreting the timing of the preparatory activities. While the preparatory activity as early as 450 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. Further studies are necessary to investigate the possible material dependence of these timings, for instance, by employed an extended

amount of speech materials. In addition, an inter-individual level regression analysis method was chosen, as the average comprehension questionnaire accuracies across all stories within each participant was believed to provide a more reliable estimation the speech comprehension performance than the single-trial accuracies. Thus, 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. More theoretical and empirical research is needed to clarify the underlying mechanisms.

Summary

We found that individual participants' comprehension performance was significantly correlated with neural responses as early as -450 ms relative to speech onset. A widely distributed brain network was involved in the preparatory process. Higher-frequency activity in the alpha and beta bands were more closely related to top-down processing, while lower-frequency activity was more closely associated with post processing. Neural activities related to both the attended and the unattended speech contributed to the comprehension performance, but with distinct mechanisms. Attended speech was more efficiently processed when neural activity was inhibited in the preparatory stage and enhanced during post processing, whereas the opposite effects were observed for unattended speech. Our study provides a mechanistic description of how the brain prepares to process upcoming speech information.

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 female; 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, reported having normal hearing, and had normal or corrected-to-normal vision.

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 a common average, 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, bandpass filtered to 0.5–40 Hz 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. Around 4–9 ICs were rejected per participant.

Next, the EEG data were segmented into 28 trials according to the markers representing speech onsets. The analysis window for each trial extended from 5 to 55 s (duration: 50 s) to avoid the onset and the offset of the stories.

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 28 1-min narrative stories in Mandarin Chinese; the stories were either those about daily-life topics recommended by the experimenter and told by the speaker improvising on their own (14 stories), or those selected from the National Mandarin Proficiency Test (14 stories). The speakers were presented with the recommended topic or story materials on the computer screen. They were allowed to prepare for as long as required before telling the story (usually ~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 marker to synchronize the speech audios in the main experiment, in which two speech streams were presented simultaneously. The speakers were asked to start speaking as soon as the third beep had ended (within around 3 sec). 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 (which mainly concerned speech coherence). The actual speaking time per story ranged from 51 to 76 sec.

Two four-choice questions per 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). Both the speech audio and corresponding questions are available for downloads.

Experimental procedure

The main experiment consisted of 4 blocks of 7 trials. During each trial, two speech streams were played simultaneously to the left and right ears. The two speech streams within each trial were from the two different 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 each story) were presented sequentially in a random order on the computer screen, and the participants made their choices using the computer keyboard. 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. 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. We recommended that participants take a short break (of 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. In this way, 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 Lenovo LT2013s Wide LCD monitor. The viewing distance was approximately 60 cm. The experimental procedure was programmed in MATLAB using the Psychophysics Toolbox 3.0 extensions (Brainard & 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 1ms (at a sampling rate of 1000 Hz).

Temporal response function modeling

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 (Crosse et al., 2016; Lalor et al., 2006). We used the power envelope of the speech signal as the input signal required by TRF, which has been demonstrated to be a valid index by which to extract speech-related neural responses (Bednar & Lalor, 2018; Broderick et al., 2018; Ding & Simon, 2012; Huang et al., 2018; Mirkovic et al., 2015; O'Sullivan et al., 2015).

Prior to the modeling, the preprocessed EEG signals were re-referenced to the average of all scalp channels and then downsampled to 128 Hz. Likewise, the power envelopes of the speech signals were obtained using a Hilbert transform and then downsampled to the same sampling rate of 128 Hz. When denoting the downsampled EEG signals from channel *i*, trial *k* as *R(i,t)* and the input speech power envelope as *S(t)*, the corresponding neural response $TRF_{i,k}$ can be formulated as follows:

$$R(i,t) = TRF_{i,k} * S(t) \quad (1)$$

Where * represents the convolution operator. The latency in the neural response models ($TRF_{i,k}$) was set to vary from -1000 ms to 1000 ms post-stimulus to provide sufficient data for the planned latency of -500 ms to 500 ms in the following time-frequency analysis.

The TRF modeling analysis was performed on each EEG channel for each trial per participant. TRF models were calculated for attended and unattended speech processing separately using the corresponding speech streams as the input signal. It should be noted that we did not consider the input lateralization for the TRF models, as the observed behaviorally related findings were insensitive to the physical origin of the speech audios, but rather likely to reflect the lateralization of the human speech network. Figure S3 provides the topographical information of our main results; it is similar to Figure 2, but all results were calculated separately for speech stimuli from the left and right sides. The topographies were comparable, even for the highly lateralized responses (e.g., *Prep-B and Post-E*).

The TRF-based neural responses were then further subjected to a time-frequency analysis at the single-trial level. The TRF temporal profile was transformed using the Hanning taper (2-cycle time window; for example, FWHM = 2 sec for 1 Hz wavelet) at each time sample from -500 ms to 500 ms, with frequencies ranging from 1 to 30 Hz at 1-Hz increments. Both single-trial amplitude and phase were recorded and denoted as $\mathrm{TRF}^{A}_{i,k}(\tau, f)$ and $\mathrm{TRF}^{P}_{i,k}(\tau, f)$, where τ represents TRF latency relative to the onset of speech power envelope fluctuations, and *f* represents the TRF frequency.

Given the number of trials denoted by *N*, the TRF-based single-trial amplitude and phase responses were calculated as follows:

$$\mathrm{A}_i(\tau, f) = \sum_{k=1}^{N} TRF^{A}_{i,k}(\tau, f) \tag{2}$$

$$\mathrm{ITPL}_i(\tau, f) = \left| \sum_{k=1}^{N} \exp\left(j \cdot TRF^{P}_{i,k}(\tau, f)\right) \right| / N \tag{3}$$

A large-amplitude value indicates a neural response of high magnitude across all trials, and 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.

These phase and amplitude responses were further transformed into z-scores within the -500 ms to 500 ms time window for each channel separately per participant. These z-scores were then used for the cross-participant statistical analyses.

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, Fries, Maris, & Schoffelen, 2011).

Quantification and statistical analysis

The extracted TRF-based amplitude and phase responses were used as independent variables into a regression model to predict the speech comprehension performance of the attended speech at the participant level (denoted by *CompreScore*). Given that we aimed to explore the neural correlates of speech comprehension, we built different regression models for amplitude and phase responses, for each EEG channel (i) at individual latency (τ) and frequency (f) separately. Nevertheless, the responses to the attended and unattended speech streams were incorporated into the same regression model, as follows:

$$CompreScore = \alpha_1 \cdot \mathrm{A}_i^{Attended}(\tau, f) + \alpha_2 \cdot \mathrm{A}_i^{Unattended}(\tau, f) \quad (4)$$

$$CompreScore = \alpha_1 \cdot ITPL_i^{Attended}(\tau, f) + \alpha_2 \cdot ITPL_i^{Unattended}(\tau, f) \quad (5)$$

Statistical analysis was performed to examine the significance of these regression model predictions over all channel-latency-frequency bins by computing the regression R-values. Nonparametric cluster-based permutation analysis was applied to control for multiple comparisons. In this procedure, neighboring channel-latency-frequency bins with an uncorrected p-value below 0.01 were combined into clusters, for which the sum of correlational t-statistics corresponding to the regression R-values were obtained. A null-distribution was created through permutations of data across participants ($n = 1000$

permutations), which defines the maximum cluster-level test statistics and corrected *p*-values for each cluster.

We also examined the coefficients of the regression. We calculated the distribution and the mean of every selected channel-latency-frequency bin. We also calculated 99% regression coefficient confidence intervals using the bootstrap method for every cluster.

To investigate the attention modulation effect, we performed paired t-tests on the TRF-based neural activities related to the attended speech versus the unattended speech. Both amplitude and ITPL were included in the analysis. A similar cluster-based permutation was used to control for the multiple comparison problem ($p < .01$ as the threshold, $n = 1000$ permutations).

Supporting information

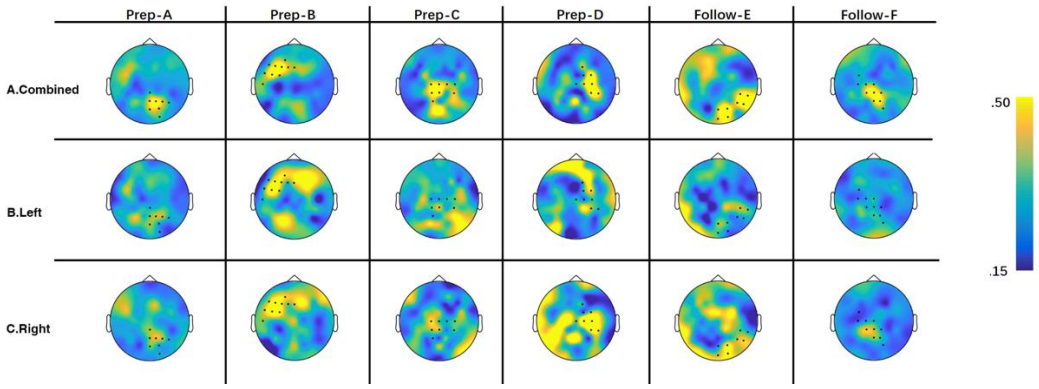

Fig S1. Topographies of the six responses as shown in Figure 2 (A) and calculated separately for the speech stimuli delivered to the left side (B) and the right side (C) only.

Table S1: CIs of all Clusters

	Att	*Unatt*
Prep-A	[-0.019, -0.016]	[+0.030, +0.033]
Prep-B	[+0.002,+0.007]	[+0.036, +0.040]
Prep-C	[-0.016, -0.009]	[+0.035, +0.040]
Prep-D	[-0.010, -0.006]	[+0.040, +0.045]
Post-E	[+0.044, +0.049]	[-0.041, -0.035]
Post-F	[+0.009, +0.013]	[-0.033, -0.029]

693

Table S2: Pairwise Correlation between all clusters

	Prep-A	*Prep-B*	*Prep-C*	*Prep-D*	*Post-E*
Prep-B	.589**				
Prep-C	.421	.439			
Prep-D	.480*	.408	.486*		
Post-E	.596**	.361	.474*	.672**	
Post-F	.648**	.378	.483*	.562**	.646**

** p<.01

* p<.05

694

Table S3: Partial correlation

	Partial-r	*p*
Prep-A	.323	.189
Prep-B	.560	.016*
Prep-C	.610	.007*
Prep-D	.439	.068

* p<.05

695

696

Acknowledgments

This work was supported by the National Science Foundation of China (NSFC) and the German Research Foundation (DFG) in project Crossmodal Learning (grant number: NSFC 61621136008/DFG TRR-169/C1, B1), the National Key Research and Development Plan (grant number: 2016YFB1001200), the National Natural Science Foundation of China (grant number: 61977041 and U1736220), and the National Social Science Foundation of China (grant number: 17ZDA323).

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. conducted the experiments and data analysis; D.Z. designed the experiments and wrote the paper; B.H., G.D., and A.K.E. edited the manuscript.

Declaration of interests

The authors declare no competing interests.

References

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

Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations

reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, *14*(6), 797–801. https://doi.org/10.1038/nn.2810

Auksztulewicz, R., & Friston, K. (2015). Attentional enhancement of auditory mismatch responses: A DCM/MEG study. *Cerebral Cortex*, *25*(11), 4273–4283. https://doi.org/10.1093/cercor/bhu323

Auksztulewicz, R., Schwiedrzik, C. M., Thesen, T., Doyle, W., Devinsky, O., Nobre, A. C., … Melloni, L. (2018). Not all predictions are equal: "what" and "when" predictions modulate activity in auditory cortex through different mechanisms. *Journal of Neuroscience*, *38*(40), 8680–8693. https://doi.org/10.1523/JNEUROSCI.0369-18.2018

Bauer, M., Stenner, M. P., Friston, K. J., & Dolan, R. J. (2014). Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes. *Journal of Neuroscience*, *34*(48), 16117–16125. https://doi.org/10.1523/JNEUROSCI.3474-13.2014

Bednar, A., & Lalor, E. C. (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

Brainard, D. H., & Brainard, D. H. (1997). The Psychophysics Toolbox. In *Spatial vision* (pp. 433–436).

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the

738 Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803-
739 809.e3. https://doi.org/10.1016/j.cub.2018.01.080

740 Brownell HH, Michel D, Powelson J, & Gardner H. (1983). Surprise but not
741 coherence: sensitivity to verbal humor in right-hemisphere patients. *Brain and*
742 *Language*, *18*(1), 20–27.

743 Cao, L., Thut, G., & Gross, J. (2017). The role of brain oscillations in predicting self-
744 generated sounds. *NeuroImage*, *147*, 895–903.
745 https://doi.org/10.1016/j.neuroimage.2016.11.001

746 Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13),
747 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

748 Chao, Z. C., Takaura, K., Wang, L., Fujii, N., & Dehaene, S. (2018). Large-Scale
749 Cortical Networks for Hierarchical Prediction and Prediction Error in the
750 Primate Brain. *Neuron*, *100*(5), 1252-1266.e3.
751 https://doi.org/10.1016/j.neuron.2018.10.004

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

754 Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate
755 Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for
756 Relating Neural Signals to Continuous Stimuli. *Frontiers in Human*
757 *Neuroscience*, *10*(November), 604. https://doi.org/10.3389/fnhum.2016.00604

758 Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple Acoustic Features Can Explain
759 Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*,
760 *29*(12), 1924-1937.e9. https://doi.org/10.1016/j.cub.2019.04.067

761 DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation
762 during language comprehension inferred from electrical brain activity. *Nature*
763 *Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

764 DeWitt, I., & Rauschecker, J. P. (2012). Phoneme and word recognition in the
765 auditory ventral stream. *Proceedings of the National Academy of Sciences of the*
766 *United States of America*, *109*(8), 505–514.
767 https://doi.org/10.1073/pnas.1113427109

768 Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical
769 entrainment to speech reflects phoneme-level processing. *Current Biology*,
770 *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

771 Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical
772 preactivation. *Brain and Language*, *127*(1), 55–64.
773 https://doi.org/10.1016/j.bandl.2012.08.004

774 Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects
775 while listening to competing speakers. *Proceedings of the National Academy of*
776 *Sciences of the United States of America*, *109*(29), 11854–11859.
777 https://doi.org/10.1073/pnas.1205381109

778 Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech:

functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*(May), 1–7. https://doi.org/10.3389/fnhum.2014.00311

Engel, A. K., & Fries, P. (2010). Beta-band oscillations--signalling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, *2*(10), 704–716. https://doi.org/10.1038/35094565

Engel, A. K., Gerloff, C., Hilgetag, C. C., & Nolte, G. (2013). Intrinsic Coupling Modes: Multiscale Interactions in Ongoing Brain Activity. *Neuron*, *80*(4), 867–886. https://doi.org/10.1016/j.neuron.2013.09.038

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.Thinking

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268. https://doi.org/10.1016/j.tics.2012.04.001

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, *88*(1), 220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the*

Royal Society B: Biological Sciences, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Galindo-Leon, E. E., Stitt, I., Pieper, F., Stieglitz, T., Engler, G., & Engel, A. K. (2019). Context-specific modulation of intrinsic coupling modes shapes multisensory processing. *Science Advances*, *5*(4), 1–13. https://doi.org/10.1126/sciadv.aar7633

George, M. S., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic engagement in reading. *Brain*, *122*, 1317–1325.

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

Harris, A. M., Dux, P. E., & Mattingley, J. B. (2018). Detecting Unattended Stimuli Depends on the Phase of Prestimulus Neural Oscillations. *The Journal of Neuroscience*, *38*(12), 3092–3101. https://doi.org/10.1523/jneurosci.3006-17.2018

Hickok, G. (2012a). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(january), 135–145. https://doi.org/10.1038/nrn3158

Hickok, G. (2012b). The cortical organization of speech processing: Feedback control

and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, *45*(6), 393–402. https://doi.org/10.1016/j.jcomdis.2012.06.004

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2011.01.019

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2015). Neural measures of a Japanese consonant length discrimination by Japanese and American English listeners: Effects of attention. *Brain Research*, *1626*, 218–231. https://doi.org/10.1016/j.brainres.2015.06.001

Huang, Q., Jia, J., Han, Q., & Luo, H. (2018). Fast-backward replay of sequentially memorized items in humans. *Elife*, *7*(e35164), 376202. https://doi.org/10.7554/eLife.35164.001

Iemi, L., Busch, N. A., Laudini, A., Haegens, S., Samaha, J., Villringer, A., & Nikulin, V. V. (2019). Multiple mechanisms link prestimulus neural oscillations to sensory responses. *Elife*, *8*(e43620), 461558. https://doi.org/10.1101/461558

Jehee, J. F. M., & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology*, *5*(5). https://doi.org/10.1371/journal.pcbi.1000373

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*,

4(November), 1–8. https://doi.org/10.3389/fnhum.2010.00186

Kayser, S. J., Ince, R. A. A., Gross, J., & Kayser, C. (2015). Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *Journal of Neuroscience*, *35*(44), 14691–14701. https://doi.org/10.1523/JNEUROSCI.2243-15.2015

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, *16*(3), 1–19. https://doi.org/10.1371/journal.pbio.2004473

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, *16*(12), 606–617. https://doi.org/10.1016/j.tics.2012.10.007

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Research Reviews*, *53*(1), 63–88. https://doi.org/10.1016/j.brainresrev.2006.06.003

Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, *75*(2), 265–270. https://doi.org/10.1016/j.neuron.2012.04.034

Kok, P., Mostert, P., & De Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(39), 10473–10478. https://doi.org/10.1073/pnas.1705652114

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., & de Lange, F. P. (2012). Attention Reverses the Effect of Prediction in Silencing Sensory Signals. *Cerebral Cortex*, *22*(9), 2197–2206. https://doi.org/10.1093/cercor/bhr310

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

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Lakatos, P., O'Connell, M. N., Barczak, A., Mills, A., Javitt, D. C., & Schroeder, C. E. (2009). The Leading Sense: Supramodal Control of Neurophysiological Context by Attention. *Neuron*, *64*(3), 419–430. https://doi.org/10.1016/j.neuron.2009.10.014

Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., McDarby, G., & Foxe, J. J. (2006). The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage*, *32*(4), 1549–1561. https://doi.org/10.1016/j.neuroimage.2006.05.054

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/Doi 10.1038/Nrn2532

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies

of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440. https://doi.org/10.1016/S1364-6613(00)01545-X

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

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

Marzecová, A., Widmann, A., SanMiguel, I., Kotz, S. A., & Schröger, E. (2017). Interrelation of attention and prediction in visual processing: Effects of task-relevance and stimulus probability. *Biological Psychology*, *125*, 76–90. https://doi.org/10.1016/j.biopsycho.2017.02.009

Mathewson, K. E., Lleras, A., Beck, D. M., Fabiani, M., Ro, T., & Gratton, G. (2011). Pulsed out of awareness: EEG alpha oscillations represent a pulsed-inhibition of ongoing cortical processing. *Frontiers in Psychology*, *2*(MAY), 1–15. https://doi.org/10.3389/fpsyg.2011.00099

Milton, A., & Pleydell-Pearce, C. W. (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, M. G., De Vos, M., & Debener, S. (2016). Target speaker detection with concealed EEG around the ear. *Frontiers in Neuroscience*,

10(JUL), 1–11. https://doi.org/10.3389/fnins.2016.00349

Mirkovic, B., Debener, S., Jaeger, M., & Vos, M. De. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, *12*(4), 046007. https://doi.org/10.1088/1741-2560/12/4/046007

Nobre, A. C., & Van Ede, F. (2018). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience*, *19*(1), 34–48. https://doi.org/10.1038/nrn.2017.141

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

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

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1–2), 21–52. https://doi.org/10.1016/0010-0277(87)90003-5

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature*

Neuroscience, *2*(1), 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. *Proceedings of the National Academy of Sciences*, *116*(32), 16056–16061. https://doi.org/10.1073/pnas.1817317116

Robertson, D. A., Gernsbacher, M. A., Guidotti, S. J., Robertson, R. R. W., Irwin, W., Mock, B. J., & Campana, M. E. (2000). Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. *Psychological Science*, *11*(3), 255–260. https://doi.org/10.1111/1467-9280.00251

Rohenkohl, G., & Nobre, A. C. (2011). Alpha Oscillations Related To Anticipatory Attention Follow Temporal Expectations. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(40), 14076–14084. https://doi.org/10.1523/JNEUROSCI.3387-11.2011

Rommers, J., Dickson, D. S., Norton, J. J. S., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, *32*(5), 576–589. https://doi.org/10.1080/23273798.2016.1183799

Salinas, E., & Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience*, *2*(8), 539–550. https://doi.org/10.1038/35086012

Sauseng, P., Klimesch, W., Gruber, W. R., Hanslmayr, S., Freunberger, R., &

945 Doppelmayr, M. (2007). Are event-related potential components generated by
946 phase resetting of brain oscillations? A critical discussion. *Neuroscience*, *146*(4),
947 1435–1444. https://doi.org/10.1016/j.neuroscience.2007.03.014

948 Schröger, E., Kotz, S. A., & SanMiguel, I. (2015). Bridging prediction and attention
949 in current research on perception and action. *Brain Research*, *1626*, 1–13.
950 https://doi.org/10.1016/j.brainres.2015.08.037

951 Schröger, E., Marzecová, A., & Sanmiguel, I. (2015). Attention and prediction in
952 human audition: A lesson from cognitive psychophysiology. *European Journal*
953 *of Neuroscience*, *41*(5), 641–664. https://doi.org/10.1111/ejn.12816

954 Schyns, P. G., Thut, G., & Gross, J. (2011). Cracking the code of oscillatory activity.
955 *PLoS Biology*, *9*(5), 1001063. https://doi.org/10.1371/journal.pbio.1001064

956 Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., …
957 Griffiths, T. D. (2016). Neural signatures of perceptual inference. *ELife*,
958 *5*(MARCH2016), 1–13. https://doi.org/10.7554/eLife.11476

959 Smith, J. L., Johnstone, S. J., & Barry, R. J. (2006). Effects of pre-stimulus processing
960 on subsequent events in a warned Go/NoGo paradigm: Response preparation,
961 execution and inhibition. *International Journal of Psychophysiology*, *61*(2), 121–
962 133. https://doi.org/10.1016/j.ijpsycho.2005.07.013

963 Smout, C. A., Tang, M. F., Garrido, M. I., & Mattingley, J. B. (2019). Attention
964 promotes the neural encoding of prediction errors. *PLoS Biology*, *17*(2), 1–22.
965 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. *Frontiers in Human Neuroscience*, *10*(October), 1–11. https://doi.org/10.3389/fnhum.2016.00512

Söderström, P., Horne, M., Mannfolk, P., van Westen, D., & Roll, M. (2018). Rapid syntactic pre-activation in Broca's area: Concurrent electrophysiological and haemodynamic recordings. *Brain Research*, *1697*, 76–82. https://doi.org/10.1016/j.brainres.2018.06.004

Szewczyk, J. M., & Schriefers, H. (2018). The N400 as an index of lexical preactivation and its implications for prediction in language comprehension. *Language, Cognition and Neuroscience*, *33*(6), 665–686. https://doi.org/10.1080/23273798.2017.1401101

Tian, X., Ding, N., Teng, X., Bai, F., & Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. *Nature Human Behaviour*, *2*(3), 225–234. https://doi.org/10.1038/s41562-018-0305-8

van Ede, F., Jensen, O., & Maris, E. (2010). Tactile expectation modulates pre-stimulus A-band oscillations in human sensorimotor cortex. *NeuroImage*, *51*(2), 867–876. https://doi.org/10.1016/j.neuroimage.2010.02.053

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

987 Verhulst, S., Altoè, A., & Vasilkov, V. (2018). Computational modeling of the human
988 auditory periphery: Auditory-nerve responses, evoked potentials and hearing
989 loss. *Hearing Research*, *360*, 55–75.
990 https://doi.org/10.1016/j.heares.2017.12.018

991 Zhang, D., Hong, B., Gao, S., & Röder, B. (2017). Exploring the temporal dynamics
992 of sustained and transient spatial attention using steady-state visual evoked
993 potentials. *Experimental Brain Research*, *235*(5), 1575–1591.
994 https://doi.org/10.1007/s00221-017-4907-6

995 Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual
996 input enhances selective speech envelope tracking in auditory cortex at a
997 "cocktail party". *The Journal of Neuroscience*, *33*(4), 1417–1426.
998 https://doi.org/10.1523/JNEUROSCI.3675-12.2013

999