Strike different *neural notes* at distinct stages:

An integrative process of attention modulation in naturalistic speech

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

When we are following someone when there are multiple speakers, are we just paying attention to the sound the speaker makes, the word the speaker uses, or are we aligning our minds to the attended speaker? In this process, selective attention plays a critical role. Previous studies have mainly focused on attention modulation on the single types of features from the speaker. However, attending to a speaker needs an integrative process: the sound waves, which represent acoustic features, should be heard; the meaning of the text, which represents semantic features, needs to be understood; listeners and speakers need to form a common grounding during the process, which represent the inter-brain feature. We still do not know how our brain entrains different types of features from the speaker in an integrative way. In this study, naturalistic speech was adopted to explore attention modulation. The natural language processing models and the flourishing inter-brain recording methods were adopted to qualify the different types of information from the speaker.

Our study revealed that attention modulated the acoustic, semantic, and speaker's neural activity in an integrative way. When the listener pays attention to a speaker, the sound is the first to be processed: The neural activity in the theta band entrains the acoustic feature with a latency of 200–350 ms. Then, the meaning of the attended speech is parsed: the delta band modulates the attention with a latency of 200–600 ms. The listener entrains the speaker's neural activity 5 seconds before the speech onset. The entrainment of the acoustic feature and the semantic feature is correlated with each other. Only the entrainment of the attended speaker has a significant negative correlation with the comprehension score. Together, these results illustrate an integrative view of how our brain is selectively entrained to different types of information from the speaker.

Keywords: attention, entrainment, electroencephalogram, temporal response function, speech, inter-brain

Introduction

"Men do not understand one another by actually exchanging signs for things…they do it by striking the same note on their mental instruments."

Wilhelm von Humboldt

We don't always hear toasts in the daily life. There are many speakers in a "cocktail party" situation. It is not a difficult task for most people to follow the person that they pay attention to and ignore others. However, the neural process behind this "cocktail party" situation (Cherry, 1953; McDermott, 2009; Middlebrooks et al., 2017) is complex: the listener needs to transform the attended sound wave to the meaning (Hasson et al., 2012; Heilbron et al., 2022). Moreover, the listener needs to actively perceive the speech and form a common grounding with the speaker (Friston, 2009; Jiang et al., 2012, 2021; Pulvermüller & Fadiga, 2010; Stolk et al., 2016; Yeshurun et al., 2021). We still do not know when we pay attention to someone how our brain integrates different types of information from the speaker.

Previous studies mainly focused on the modulation mechanism of the speech itself. The auditory scene analysis studies mainly focused on the *acoustic feature* of the speech information in the cocktail party problem (Bregman, 1990; Brodbeck & Simon, 2022; Ding & Simon, 2012a, 2012b; Shamma et al., 2011; Shinn-Cunningham, 2008; Teoh et al., 2022; Yuanye Wang et al., 2019). These studies demonstrated that the attention

modulation of the acoustic feature mainly occurs 100–250 ms after the speech onset on the low-frequency bands (e.g., 2–8 Hz) (Broderick et al., 2018, 2021; Mesik et al., 2021; Weissbart et al., 2019). Some other studies revealed that *semantic or linguistic features* could be modulated by the attentional process, for the features in the unattended stream could also be processed (Broderick et al., 2018; Connolly et al., 1990; Dai et al., 2022; Har-shai Yahav & Zion Golumbic, 2021; Heil et al., 2004). However, when we pay attention to someone, the speaker is also very important. As famous German philosopher Gadamer (Gadamer, 1975) once said, "there can be no speech that does not bind the speaker and the person spoken to." The inter-brain studies provided another angle to the traditional attention studies and included the speaker in the scene: the listeners' neural activity could entrain not only the sound wave but also coupled with the speaker's neural activity(Pérez et al., 2017; Stephens et al., 2010; Yeshurun et al., 2021). One study reported that interpersonal neural synchronization (INS) between the listener and the attended speaker was selectively enhanced using fNIRS recording (Dai et al., 2018). It indicates the listener could actively "understand" the speaker not only through the speech itself but rely on the "beyond the stimulus" grounding (Bashivan et al., 2019; Hartley & Poeppel, 2020; Hasson et al., 2012; Jiang et al., 2021; Redcay & Schilbach, 2019; Stolk et al., 2016). While attention modulation of different features was explored separately, to our knowledge, there hasn't been a single study to explore

the attentional modulation to multiple types of information from the speaker at the same time.

An integrative view enables us to investigate the attention modulations to the different features simultaneously, and it may answer some critical questions about attentional modulation. The *first* question concerns the temporal dynamics of attention modulation of different features. Does the attention modulate all the features at the same time? Or does the attention modulate them in sequential order? It was hard for the previous studies to answer because they mostly focused on one or two types of speech features (Broderick et al., 2019, 2021; Mesik et al., 2021), and they didn't consider the speaker. The time range of attentional modulations varied in the previous studies. Most single brain studies chose the time range within ~1 second after speech onset (Ding & Simon, 2012a, 2012b; Lalor & Foxe, 2010; Power et al., 2012; Teoh et al., 2022; E. M. Zion Golumbic et al., 2013). However, the inter-brain studies revealed that the listener could selectively tune in to the attended speaker up to 5 seconds before the speech onset (Dai et al., 2018; Kuhlen et al., 2012). This phenomenon was rarely observed in the high temporal resolution EEG studies (Kuhlen et al., 2012), due to the different selection of the time windows. We still do not know which speech processing stages attention could modulate. The *second* question centers on the relationship between different types of attention modulations. Previous studies revealed that the entrainment of acoustic features and semantic features could interact with each other during the

monologue condition (Anderson et al., 2019; Gillis et al., 2021; Heilbron et al., 2022). The semantic feature was reported to enhance the entrainment of the acoustic feature (Anderson et al., 2019; Gillis et al., 2021; Heilbron et al., 2022). There were few studies providing evidence about how these three features interact with each other in a "cocktail party situation"(Dai et al., 2018; Pérez et al., 2017, 2019). The correlation between acoustic features, semantic features, and inter-brain features remains elusive.

The present study aimed to reveal the neural mechanism of attention modulation in an integrative view, which means the attention modulation of the three types of information from the speaker would be explored simultaneously. The naturalistic speech was used as stimuli, which contains much richer information than either the sound sequence or the single word in the previous studies (Broderick et al., 2018; Hamilton & Huth, 2018; Hartley & Poeppel, 2020; Nastase et al., 2021; Sonkusare et al., 2019; Willems et al., 2020). The sound, the meaning, and the speaker could appear at the same time in an ecological situation and enables us to investigate different types of information from the speaker simultaneously (Hasson et al., 2012; Willems et al., 2020). As Figure 1(A) illustrates, the amplitude envelope was the representation of the *acoustic feature*. A natural language processing model (an LSTM model) was applied to exact the semantic feature in the text. We chose surprisal as the representation of the *semantic feature* in the previous studies, which is one of the most important semantic features that have received sufficient investigation in previous neuroscience studies

(Brodbeck et al., 2022; Frank & Willems, 2017; Willems et al., 2016). The early event-related potentials studies also used the congruent or incongruent words to explore the neural mechanism of semantic processing, which is an early version of surprisal (Kutas & Hillyard, 1984, 1989; Lau et al., 2008). A sequential dual-brain approach was used, and the electroencephalogram (EEG) of both speaker and listener was recorded. The speaker's neural activity represented the *inter-brain feature* (Hasson et al., 2012; Jiang et al., 2021; Leong et al., 2017; Pérez et al., 2017, 2019; Stolk et al., 2016). A temporal response function (TRF) method was used to measure the difference between the entrainment to the attended features and the unattended features. Based on the previous studies, we hypothesized that attention modulates the different features in distinct ways, which means the modulation to different features would happen on different frequency bands and time ranges. For the frequency bands, we hypothesize that the delta band reflects attention modulation of the semantic feature (Dai et al., 2022; Teoh et al., 2022; Yu et al., 2022), and the theta band represents the acoustic feature (Ding et al., 2014; Etard & Reichenbach, 2019). The entrainment of the speaker's neural feature needs further exploration. For the time course, we assume that the attention modulation of the acoustic feature occurs at first (Ding & Simon, 2012a, 2012b; Hillyard et al., 1973; O'Sullivan et al., 2015; Power et al., 2011; Teoh et al., 2022), and the attention effect of the semantic feature lasts longer (Broderick et al., 2018; Dai et al., 2022). The listener's neural activity is aligned with the speaker's neural activity in a broader time

range (Dai et al., 2018; Kuhlen et al., 2012). We also hypothesize that the attention modulation between the acoustic feature and the semantic feature correlates to each other (Broderick et al., 2019; Heilbron et al., 2022), but the inter-brain feature has a distinct pattern, which is independent of the acoustic and semantic feature. Together, our study adopted an integrative view to investigate three different types of features when attending to a speaker, which would further our understanding of the attention modulation from the speech to the speaker.

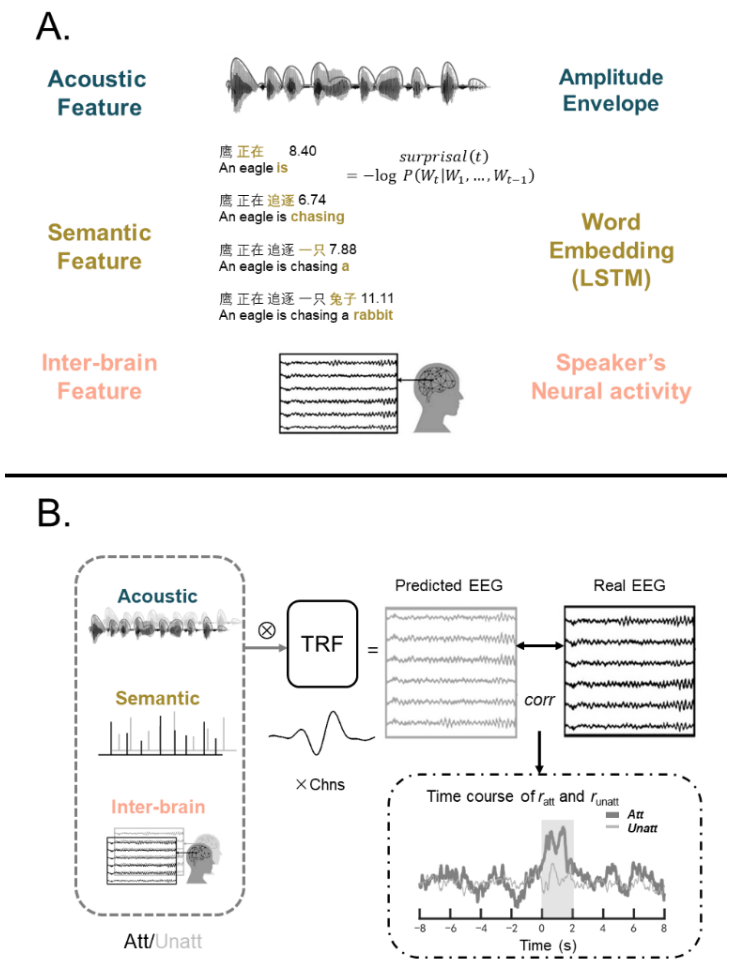

Figure 1 The stimuli, the experimental paradigm and the analysis process

(A)The demonstration of acoustic, semantic, and inter-brain features. The envelope of the speech represents the acoustic feature. The surprisal index, which was calculated by an LSTM model, was used as the semantic feature. The speaker's neural activity is the inter-brain feature. (B) A "cocktail party" selective attention paradigm was used, in which the listener was asked to pay attention to one side of the speech stream and ignore the other. While listening to the speech stream, the listener's EEG signals were recorded. The Encoding *r* for the attended feature and the unattended feature was calculated time point by time point by applying the TRF method.

Results

Behavioral Performance of the Listeners

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%, $t(19) = 10.948$, $p < .001$; 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 ($r = .476$, $p = .043$) and attention difficulty ($r = -.677$, $p = .001$). 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 ($r = -.224$, $p = .342$). These results suggest that participants' selective attention was effectively manipulated, and the measurement of comprehension performance was reliable. The response accuracy varied from 25.0 % to 51.8% for unattended stories.

The theta band and the delta band reflect distinct attentional to different features

The delta and theta bands have different functional roles in attentional modulation, as Figure 2 illustrates. A cluster-based permutation (Maris & Oostenveld, 2007) was conducted to reveal the difference between the encoding of the attended and the unattended features and control the multiple comparisons.

The theta band only modulates the acoustic feature. There are two significant acoustic clusters that appear in the theta band. They were designated as *Acoustic_Early* (cluster-based $p < .001$) and *Acoustic_Late* (cluster-based $p = .005$) depending on the time they occurred.

In contrast to the acoustic feature, semantic clusters were found in the delta band. Two clusters illustrated the difference in entrainment to the attended and unattended semantic features. They were labeled *Semantic_early* (cluster-based $p = .002$) and *Semantic_late* (cluster-based $p < .001$). One inter-brain cluster was also found in the delta band (cluster-based $p < .001$).

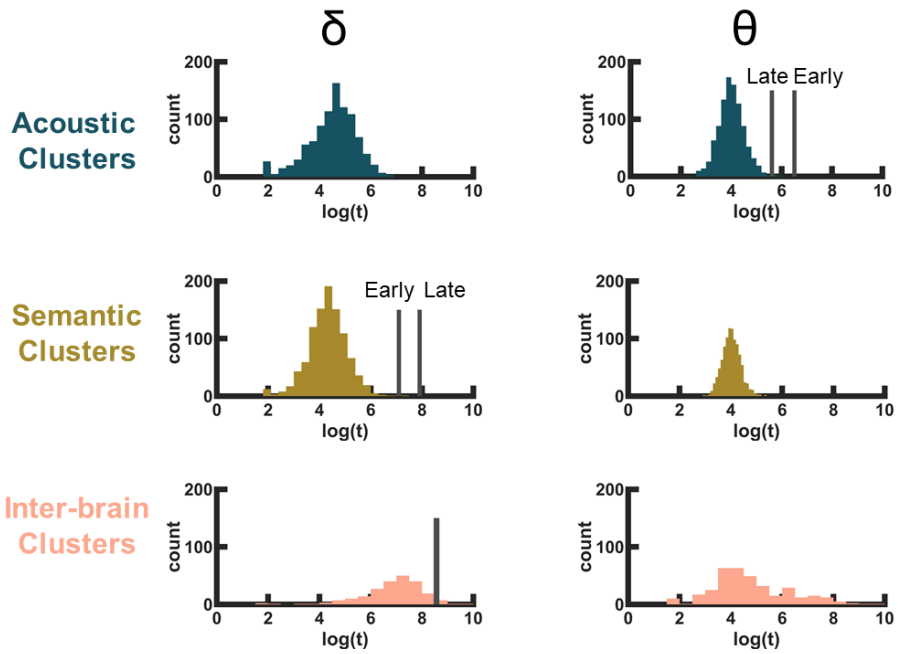

Figure 2 Different attentional modulation roles of different bands. The null distribution of the *t*-statistics of every feature in the delta band(left) and theta band(right). The grey lines indicate significant clusters.

The attention modulation of different features unfolds in a distinct time range

As shown in Figure 3, the *Acoustic_Early* cluster involved the left-lateralized fronto-central and occipital electrodes (cluster-based permutation $p < .001$) at a latency of 0.219–0.359 s after the onset of speech. The *Acoustic_Late* cluster had a later latency of 1.508–1.602 s with the electrodes in the right-frontal regions (cluster-based $p = .005$).

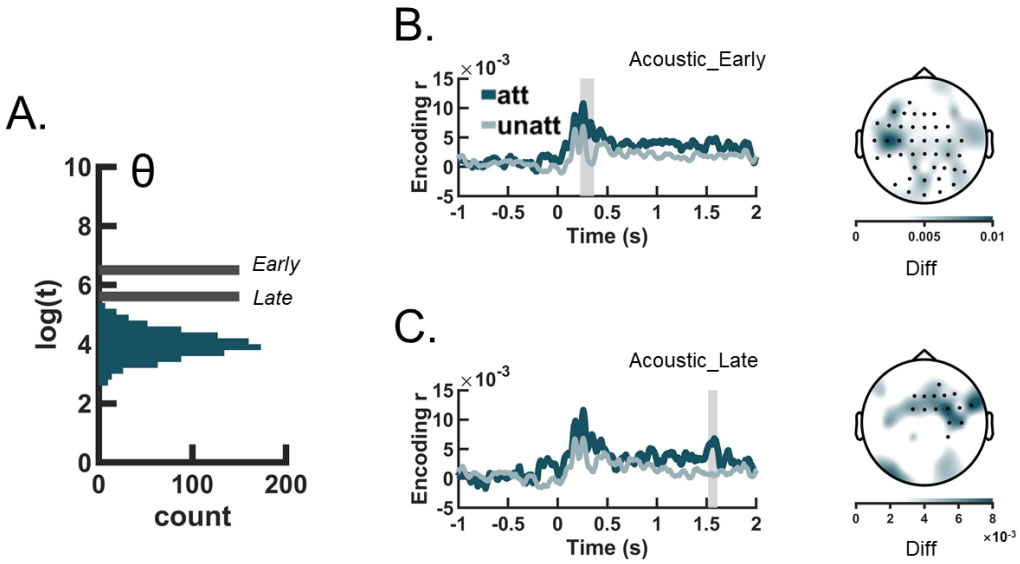

Figure 3 The attention modulation for the acoustic feature. (A) The null distribution of the cluster-based *t*-statistics of the acoustic feature. (B) and (C)The time course and the topo-plot of the significant acoustic cluster. The dark blue line represents the Encoding r_{att}, and the light line represents the Encoding r_{unatt}. The shaded region depicts a significant difference in the time window. The topo-plot of the average difference between in Encoding r_{att} and Encoding r_{unatt} cluster. The black dots indicate the channels in the cluster.

As figure 4 indicates, the *Semantic_Early* cluster occurred at 0.227–0.621 s covering the electrodes in frontal and central regions. The *Semantic_Late* cluster was found at 1.073–1.516 s involving the wide distribution of the electrodes.

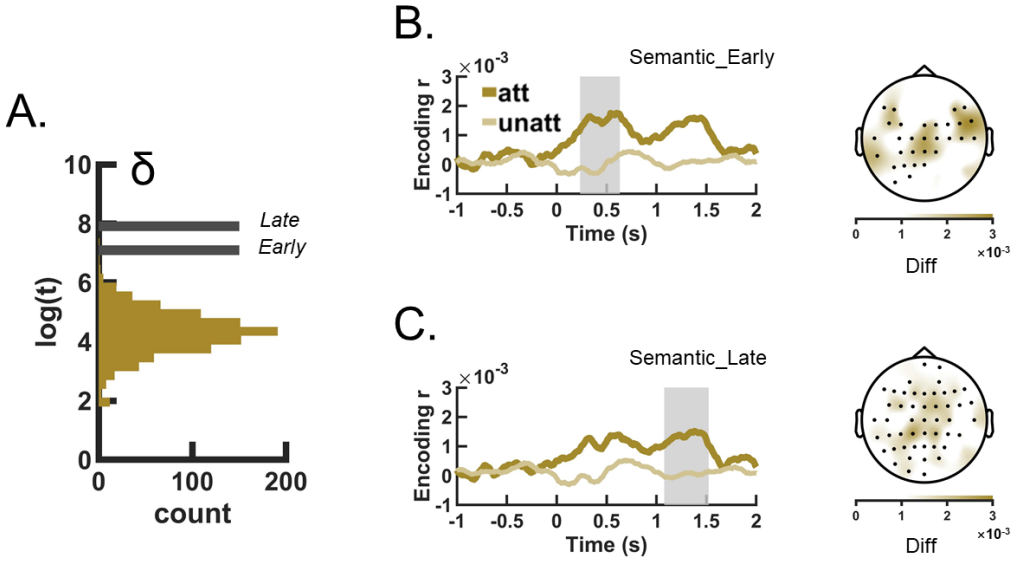

Figure 4 The attention modulation for the semantic feature. (A) The null distribution of the cluster-based *t*-statistics of the semantic feature. (B) and (C)The time course and the topo-plot of the significant semantic cluster. The dark brown line represents the Encoding r_{att}, and the light brown represents the Encoding r_{unatt}. The shaded region depicts a significant difference in the time window. The topo-plot of the average difference between in Encoding r_{att} and Encoding r_{unatt} cluster. The black dots indicate the channels in the cluster.

There was only one cluster inter-brain cluster. Unlike the acoustic clusters and semantic clusters, the inter-brain cluster had a wide time range of -4.836 to -0.539 s with the electrodes in the left frontal region, as shown in Figure 5.

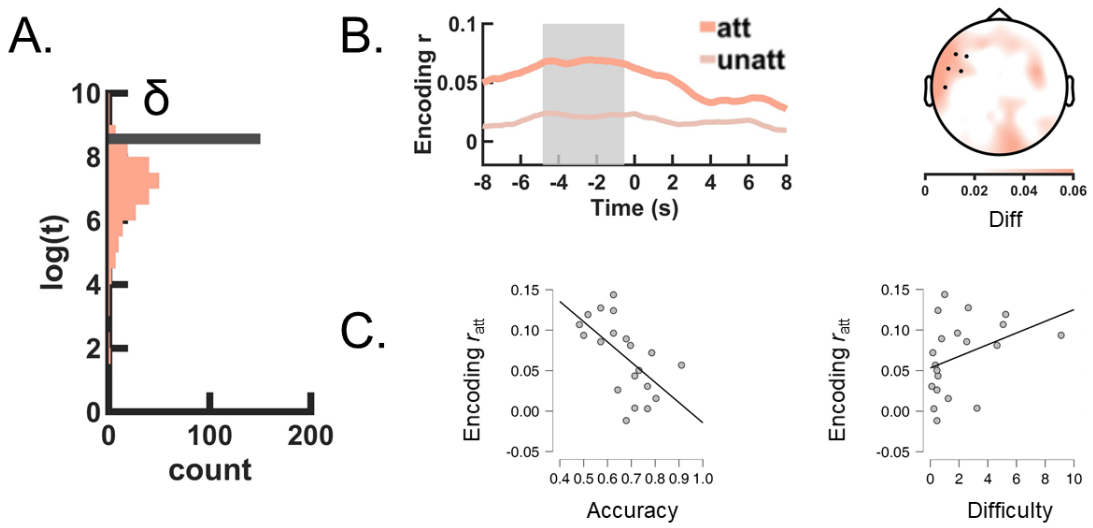

Figure 5 The attention modulation for the inter-brain feature. (A) The null distribution of the cluster-based *t*-statistics of the inter-brain feature. (B) The time course and the topo-plot of the significant inter-brain cluster. The light pink line represents the Encoding r_{att}, and the dark pink represents the Encoding r_{unatt}. The shaded region depicts a significant difference in the time window. The topo-plot of the average difference between in Encoding r_{att} and Encoding r_{unatt} cluster. The black dots indicate the channels in the cluster. (C) The correlation between the Encoding r_{att} and the behavioral results.

The entrainment to the inter-brain feature is independent of the acoustic and semantic features

As Figure 6 indicates, the average Encoding r_{att} in *Semantic_Early* cluster and the average Encoding r_{att} in *Acoustic_Early* clusters were highly correlated ($r(18) = .786$, $p < .001$, FDR_corrected). The average Encoding r_{att} in *Semantic_late* cluster and the average Encoding r_{att} in *Acoustic_Early* cluster were also highly correlated ($r(18) = .565$, $p = .045$, FDR_corrected). There were no other significant correlations between other clusters (*p*s > .05).

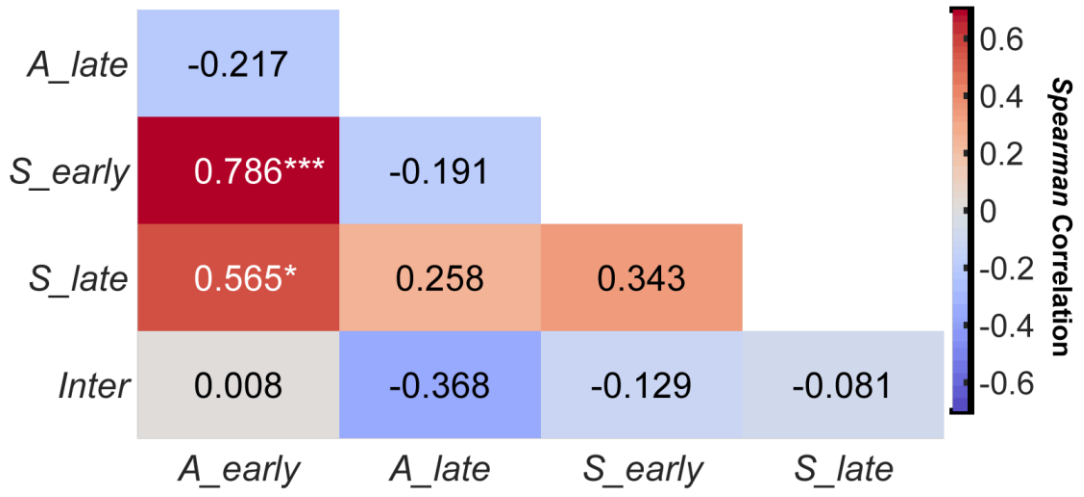

Figure 6 The correlations between Encoding r_{att} in different clusters. A_early stands for Acoustic_Early, A_late stands for Acoustic_Late, S_early is short for Semantic_early, S_late stands for Semantic_Late, and Inter is short for Inter-brain.

The entrainment to the inter-brain feature was correlated to the comprehension performance

The partial correlation between the behavioral performance and the coefficients in was calculated to reveal the unique contribution of a certain feature to the comprehension performance. Specifically, the correlations between the mean Encoding r_{att} in the cluster and the accuracy of the questions were calculated while controlling other features.

As Table 1 illustrated, only the inter-brain cluster has a significate partial correlation with the behavioral performance, partial correlation r (18) = -.769, p = .002 (FDR corrected). All the other clusters didn't reveal significant correlations. We further

analyzed the Encoding r_{att} in the inter-brain cluster and difficulty. We found significant positive correlation, $r(18) = .499$, $p = .025$ as shown in Figure 5(c).

Table 1 The partial correlation between the entrainment coefficients and behavioral performance

Clusters	*rho*	*p* values (FDR-corrected)
Acoustic-early	.512	.089
Acoustic-late	-.306	.291
Semantic-early	-.412	.170
Semantic-late	.098	.710
Inter-brain	-.769 **	.002

*$p < .05$, ** $p < .01$ (FDR corrected)

Discussion

Our study provided an integrative view of how our brain pays attention to the different types of information from the speaker in the "cocktail party problem", as Figure 7 illustrates: the attention modulated the sound firstly in the theta band at 200–350 ms. The meaning of speech was modulated later, in the delta band at 200–600 ms. The listeners aligned to the speaker's neural activity 5 s before the speech onset. The entrainment to the acoustic feature and semantic feature were correlated, but the entrainment to the speaker's neural activity is independent of the speech stimuli. Noticeably, only the entrainment to the speaker's neural activity has a negative correlation with the comprehension score, which indicates the compensation role of the entrainment to the speaker's neural activity.

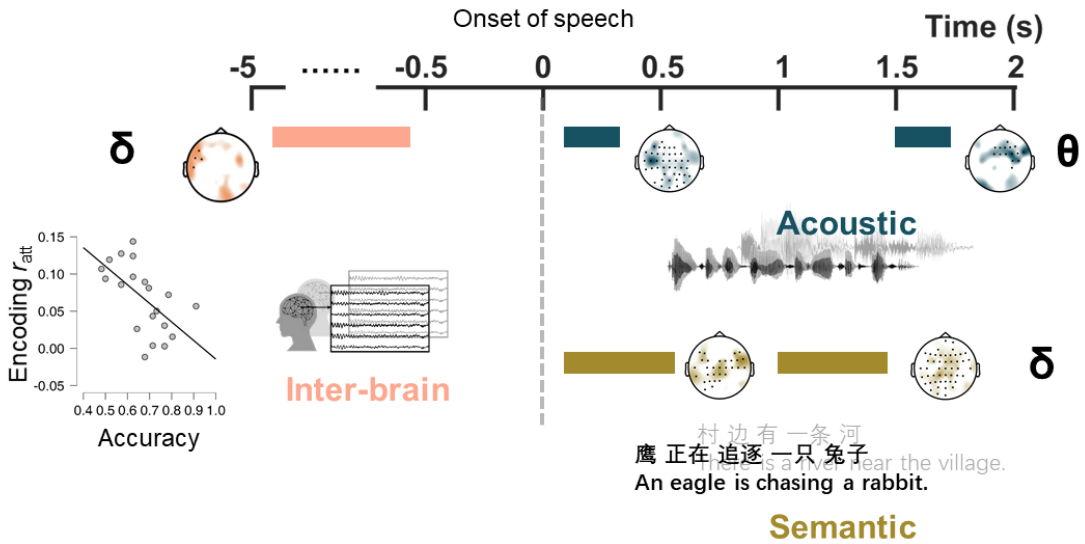

Figure 7 An integrative view of the attentional modulations of different features in speech.

Our study clearly illustrated the temporal dynamics of attention modulation of the different features in an integrative view. *The acoustic feature* was modulated occurred at first, with a latency of 200–350 ms. The time range was in line with the classical attention studies applying naturalistic speech as the stimulus (Ding & Simon, 2012a; Lalor & Foxe, 2010; O'Sullivan et al., 2015; Yadong Wang et al., 2012). The attention effect of the semantic feature lasted longer, from 200–600 ms. The N400 effect happened around 400 ms after the speech onset, and it was the most crucial neural signature of the semantic process (Kutas & Hillyard, 1989; Lau et al., 2008). Our result demonstrated that the attention modulation of the semantic feature lasted longer than the acoustic feature, and the time range was consistent with the classical N400 (Broderick et al., 2018). We also found two late clusters for the acoustic feature and the semantic feature, which have rarely been reported in previous studies. It may indicate

the attention modulation on the sentence level (e.g., the terminal of the sentence), as the early ERP studies suggested (Connolly et al., 1990; Connolly & Phillips, 1994; Sanders & Neville, 2003). The time range of the inter-brain feature revealed a different pattern. The listener aligned to the speaker's neural activity 5 s before the speech onset, which was much earlier than the other two features. This result replicated previous inter-brain studies using fMRI or fNIRS in the high temporal resolution EEG signals (Dai et al., 2018; Jiang et al., 2012; L. Liu, Zhang, et al., 2020; Stephens et al., 2010), which indicates that attention modulation may happen far before the previous attention modulation studies. Our result further proved that the entrainment to the speaker's neural activity was different from other entrainments, which emphasized the importance of the speaker (Dai et al., 2018; Pérez et al., 2017, 2019).

Our study reveals the distinct roles of delta and theta in attention modulation. The theta band modulates the acoustic feature, and the delta band modulates the semantic feature and inter-brain feature. This result extended our understanding of the functional roles of the two bands. It is consistent with the previous finding that the theta band processes the stimulus-link features, like the acoustic feature (Ding et al., 2014; Ding & Simon, 2012a, 2012b; Teng et al., 2017), and the delta band reflects the comprehension-related features, like the semantic feature (Ding et al., 2014; Etard & Reichenbach, 2019; J. Li et al., 2022; Lu et al., 2022). In particular, the modulation of

the inter-brain feature was also found in the delta band, which may indicate that the delta bands reflects the comprehension-related functions.

Our study extends to the attended target from the speech to the speaker by adopting the inter-brain feature. The entrainment of the speaker's neural activity was the only predictor of comprehension performance, and it didn't correlate with other features. Noticeably, the entrainment to the speaker's neural activity was a negative correlation with the comprehension performance. We further found that the encoding index was positively correlated with the perceived task difficulty reported by the listeners. We called that a "compensation" mechanism: when the listeners feel hard to complete the task, they start to guess what the speaker wants to say. However, guessing is not always correct. Therefore, the comprehension performance is decreasing. The spatial distribution of inter-brain clusters is also different from the distribution of the acoustic and semantic clusters. While the central electrodes are primarily involved in the acoustic and semantic modulations, only the left-frontal electrodes are recruited in the inter-brain modulation. The left frontal regions play a critical region in the language process (Har-shai Yahav & Zion Golumbic, 2021; Hickok & Poeppel, 2007) and are a 'high-order' area in attention selection (E. Zion Golumbic et al., 2013). The left-frontal electrodes may indicate a unique contribution of IFG when listeners are under adverse listening conditions in previous inter-brain studies (Dai et al., 2018; Z. Li et al., 2021; L. Liu, Yan, et al., 2020). Our study highlighted the critical role of the speaker in the

attention process. We proved speech itself only serves as a trigger and a signal, while the attended speaker and the listener aligned their mind "beyond the speech stimulus"(Hartley & Poeppel, 2020) even before the speech onset.

To our knowledge, our study is the first study to combine the NLP method and the inter-brain method to extract the different levels of features in speech and investigate their attention modulation of them, which enables us to extend the attentional modulation studies to an integrative and "beyond the stimulus" perspective (Hartley & Poeppel, 2020). In our study, the speech was separated into three levels features: the acoustic feature, the semantic feature, and the inter-brain feature. With the help of the NLP models (Armeni et al., 2019; Brodbeck et al., 2022; Broderick et al., 2018; Kingma & Ba, 2015), we could calculate semantic features in the text to which the listeners attended. Inspired by the inter-brain studies (Dai et al., 2018; Jiang et al., 2012, 2015; Leong et al., 2017; Stephens et al., 2010), how the listeners pay attention to the "hidden meaning" behind the text was also analyzed. Meanwhile, we applied the TRF method to describe the precise neural activity towards the attended and unattended features and simultaneously compared entrainment to the different features.

In conclusion, our study used the attention as a spotlight and revealed that the listener would strike *the different neural notes* at distinct stages in an integrative way: the *acoustic note* is struck on the theta bands at first, and the *semantic note* comes later and lasts longer on the delta band. The striking on the mental instrument, which is

achieved by the inter-brain coupling, appears even before the speech onset. Our study depicts the temporal dynamics of the attention modulation and the functional roles of different frequency bands, which contributes to the old “cocktail party” a new integrative perspective.

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.

Participants

Two participants (both male, aged 26 and 24 years) were recruited for this study as speakers. Both speakers were from the broadcasting station of Tsinghua University and had experience related to broadcasting and hosting.

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

The sample size (N = 20) was decided empirically following previous TRF-based studies on human speech processing (Broderick et al., 2018; Di Liberto et al., 2015; J. Li et al., 2022; Mirkovic et al., 2015).

Experimental procedure for the speakers

A sequential inter-brain approach was adopted by the present study (Redcay & Schilbach, 2019), in which the neural activities of the speakers were recorded prior to the listeners. The sequential design was more appropriate for this study than the real-time interactive design because the speakers' audio and neural activity remained consistent for all listeners (Leong et al., 2017; Y. Liu et al., 2017; Stephens et al., 2010).

In this experiment, each speaker participated in 30 trials, each of which was approximately 51–76 seconds in length, while the speakers' audio signals and EEG signals were recorded. The experimenter selected 28 trials for the listener's experiment, excluding the two most unqualified trials.

The speaker first read the relevant material on the screen. There was a wide variety of content to be covered, including one's hometown, a recent book, a fable, etc. The speaker could decide how long they wanted to spend on preparation and start talking when they were fully prepared (the length of preparation was usually 3 minutes). When the speaker was prepared, they would press the space bar on the computer keyboard, and the recording would begin. When the space bar was pressed, three 1000 Hz pure tone cues were triggered (duration: 1000 ms; cue interval: 1500 ms). The cues were presented as event markers, synchronized with the sound in the listener's experiment to ensure that the neural signals of the speaker and listener remain aligned with the sound stimuli. The speaker was asked to start speaking immediately after the end of the third beep (within approximately 3 s). A fixation and a countdown timer appeared on the

screen during the talking part. The speaker was asked to stare at the fixation and to complete the speaking as clearly, completely, and naturally as possible. During the recording process, the experimenter listened to the speaker's narration simultaneously and controlled the quality. The experimenter had the right to ask the speaker to retell the clip if there was a reason for the lack of fluency, length, etc., that might affect the listeners' perception. The materials of both speakers' content were varied. Between each trial, the speakers were allowed to rest on their own. During the experiment, the speakers were asked to control their head movements and facial muscles to obtain better quality EEG signals.

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.

Experimental procedure for the listeners

The experiment consisted of four blocks, each containing seven trials. Two speech streams were presented simultaneously in each trial, one to the left ear and the other to the right ear. Two speech streams of the same trial matched the volume, i.e., the root mean squared intensity of the amplitude of the speech streams in the same trial were the same. The participants were instructed to attend to one spatial side according to the hints on the screen ("Please pay attention to the [LEFT/RIGHT]"). Considering the possible duration difference between the two audio streams, the researchers set the end

of the trial after the longer speech audio had ended. Each trial began when participants pressed the SPACE key on the computer keyboard. 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.

At the end of each trial, four multiple-choice questions (two for the attended story and the other two for the unattended story) were presented sequentially in random order on the computer screen. Each question had four options, and participants entered the letter of the correct option as their answer. The listeners were not explicitly informed about the correspondence between the questions and the stories. 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). 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. No feedback on whether the questions were answered correctly or not.

After completing these questions, participants rated their concentration level of the attended stream, the experienced difficulty 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. Meanwhile, they were asked to minimize eye blinks and all other motor activities. 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.

In each block, the side being attended to was fixed (two blocks for attending to the left side and two for attending to the right side). Within each block, the identity of the speaker is kept constant on 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 conducted 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 & 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 interference from auditory devices. The volume of the audio stimuli was adjusted to be at a comfortable level (~70 dB SPL) that was well above the auditory threshold. The average presentation level was

measured with a B&K (Brüel & Kjær, Nærum, Denmark) Sound Level Meter (Type 2250 Investigator) with a 1-inch Free-field Microphone (Type 4144) and an Artificial Ear (Type 4152).

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 ten 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 (~1 minute 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 CCA method. Therefore, any ICs with artifact-like EEG activities for more than 20% of the trial time (i.e., ~12 sec) were rejected, leading to around 4–11 ICs rejected per participant. The cleaned EEG data were used for the mTRF analysis without any further artifact rejection procedures.

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

The pre-processed EEG signals were re-referenced to the average of all scalp channels and then downsampled to 128 Hz before the modeling. Then, the EEG data were filtered in delta (1–4 Hz) and theta (4–8 Hz) (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 & Nelken, 2019), which is essential for accurate time-course analysis. 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).

Speech Representations

- *Acoustic Features*

The amplitude envelope of the speech represented the acoustic features of the speech. It was obtained using a Hilbert transform and then down-sampled to the same sampling rate of 128 Hz.

- *Semantic Features*

The original audio recorded by the speaker during the EEG recording was converted to the text firstly automatedly by *Iflyrec* software (Iflytek Co., Ltd, Hefei, Anhui) and then doubled checked manually. The onset time of every word was extracted during this process.

The recent emergence of Natural Process Language (NLP) models has enabled the description of the semantic features in speech (Brodbeck et al., 2022; Broderick et al., 2018, 2021). Next word prediction is one of the fundamental NLP tasks using the semantic information in the texts (Schrimpf et al., 2021; Vaswani et al., 2017). The goal of the task is to predict the next word when given a sequence of words $W_1, W_2, \dots, W_{t-1}$, which was consistent with the human understanding process. The probability of the next word is $P(W_t|W_1, \dots, W_{t-1})$ and can be calculated by varied NLP models. The surprisal of the word was defined as follow, which reflected how surprised the next word (Willems et al., 2016):

$$surprisal(t) = -\log P(W_t|W_1, \dots, W_{t-1}) \quad (1)$$

The index was calculated based on ADAM, a widely accepted classical natural language process model (Bengio et al., 2003; Kingma & Ba, 2015). The model was trained on a couple of *People's Daily.* There were 534,246 words involved in the model training. 66,781 words were in the cross-validation set, and 66,781 words used as a test set. The details of the model are described in Table S1.

After calculating the surprisal index of every word, we generated a "semantic vector" at the same sampling rate as the EEG data (Broderick et al., 2018). The vectors contained the time-aligned impulses at the start of each word of the surprisal value for every audio clip.

- *Inter-brain Features*

The Inter-brain recording method enables us to study how the listeners are aligned with the attended speaker (Dai et al., 2018; Hasson et al., 2012; Jiang et al., 2021; Stephens et al., 2010). We used the speaker's neural activity as the representation of the inter-brain feature.

The speaker's EEG served as the inter-brain feature. It followed the same pre-processing procedure as the listener's EEG.

Temporal response function modeling

The analysis workflow for the analysis related to the attended speech stream is shown in Figure 1. The neural responses to the three different features were

characterized using a temporal response function (TRF)-based modeling method (Crosse et al., 2016, 2021).

Three different features mentioned above are the input signal required by TRF. The corresponding neural response $r(t,n)$ can be formulated as follows:

$$r(t,n) = \sum_{\tau} w(\tau,n)\, S(t-\tau) + \varepsilon(t,n) \tag{2}$$

where $r(t,n)$ is the actual EEG response at every channel n; t = 1 … T is the time point; $S(t-\tau)$ means the multivariate stimulus representation; $w(\tau, n)$ is the channel specific TRF at lag τ and $\varepsilon(t,n)$ is the residual.

The TRF is estimated by minimizing the mean square error between the actual neural response $r(t,n)$ and the neural response predicted by the model $\hat{r}(t,n)$. The Pearson's correlation between the actual neural response and predicted neural response was referred as Encoding *r*. The mTRF toolbox (Crosse et al., 2016) was used to estimate the TRF(w) as fellow:

$$w = (S^T S + \lambda I)^{-1} S^T r \tag{2}$$

where the lambda(λ) is the ridge regression parameter, *I* is the identity matrix, and the matrix *S* is the stimulus matrix. The lambda varied from 10^{-1} to 10^{8} (lambda = 10^{-1}, 10^{0}, …, 10^{8}) to make the model optimal (Crosse et al., 2021). The lambda value, which produces the highest encoding *r*, averaged across trials and channels, was selected as

the regularization parameter for all trials per participant (Broderick et al., 2019). The cross-validation procedure was implemented in a leave-one-trial-out manner: the TRFs were trained based on data from 27 trials and tested on the left-out trial each time.

The TRF was trained at individual time lags of -8 s to 8 s to investigate the specific interval of attentional modulation for each feature. At a sampling rate of 128 Hz, there are 2049 individual time-lag intervals of 7.625 ms.

The TRF calculation procedure was performed for the EEG signals from each EEG channel filtered at the four frequency bands. Only attended features are used as input to the model, and TRFs trained by the attended features and the neural response were applied to the tests of attended and unattended features, referring to Encoding r_{att} and Encoding r_{unatt}, respectively.

Quantification and statistical analysis

The paired *t*-tests were performed to investigate the attentional modulation for different features, contrasting the encoding *r* of the attended speech versus the unattended speech at each time lag. Encoding *r* was normalized using the Fisher-*z* transform (Corey et al., 1998).

A nonparametric cluster-based permutation analysis was applied to account for multiple comparisons (Maris & Oostenveld, 2007). In this procedure, neighboring channel-latency bins with uncorrected *t*-tests *p*-value below 0.05 were combined into

clusters, for which the sum of the correlational *t*-statistics corresponding to the *t*-tests was obtained. The combing process was initially automated by the toolbox and then manually double-checked. Two clusters were combined if they shared a similar spatial distribution or time lag. 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. Clusters with *p*-values below 0.01 based on clusters were selected for further analysis.

The above statistical analysis followed the standard cluster-based permutation procedure as employed in classical ERP and related studies (Arnal et al., 2011; Henry & 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 cluster-based permutation tests. No multiple comparison correction was employed across different frequency bands.

Correlation between clusters

The *Spearman* correlation of Encoding r_{att} in each cluster was calculated for each pair of clusters to analyze the correlation between them.

Partial correlation

The partial correlation between every cluster and the comprehension performance was calculated to investigate the unique contribution of each cluster to the behavioral performance.

Data availability statement

Dataset generated in our study has been uploaded to Mendeley Data (Li, Jiawei (2021), "Speaker-listener Cocktail Party", Mendeley Data, V3, doi: 10.17632/wkh75s27kn.3). We have also provided data of individual participant, including the pre-processed EEG data and behavioral performance data.

Where existing toolboxes were used for data analysis, citations have been provided.

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), and the National Natural Science Foundation of China (grant number: 61977041).

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.

The authors would like to acknowledge Prof. Zhiyuan Liu and members of his lab for computing the NLP models.

Supporting Information

Table S1 Details of Neural Language Processing Model

Type	Parameter
model type	LSTM
embedding size	200
hidden units per layer	200
number of layers	2
initial learning rate	3
gradient clipping	0.25
sequence length	35
drop out	0.2
epoch	50
batch size	3

Reference

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., & Lalor, E. C. (2019). An Integrated Neural Decoder of Linguistic and Experiential Meaning. *The Journal of Neuroscience*, *39*(45), 8969–8987. https://doi.org/10.1523/JNEUROSCI.2575-18.2019

Armeni, K., Willems, R. M., van den Bosch, A., & Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, *198*, 283–295. https://doi.org/10.1016/j.neuroimage.2019.04.083

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

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439). https://doi.org/10.1126/science.aav9436

Bengio, Y., Ducharme, R., & Vincent, P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

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

Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. In *Auditory scene analysis: The perceptual organization of sound.* The MIT Press.

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A. L., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *ELife*, *11*, 1–28. https://doi.org/10.7554/eLife.72056

Brodbeck, C., & Simon, J. Z. (2022). Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. *Frontiers in Neuroscience*, *16*(August), 1–11. https://doi.org/10.3389/fnins.2022.828546

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

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

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-84597-9

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience*, *6*(3), 256–266. https://doi.org/10.1162/jocn.1994.6.3.256

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language*, *39*(2), 302–318.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for

Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604

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

Dai, B., Chen, C., Long, Y., Zheng, L., Zhao, H., Bai, X., Liu, W., Zhang, Y., Liu, L., Guo, T., Ding, G., & Lu, C. (2018). Neural mechanisms for selectively tuning in to the target speaker in a naturalistic noisy situation. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-04819-z

Dai, B., McQueen, J. M., Terporten, R., Hagoort, P., & Kösem, A. (2022). Distracting linguistic information impairs neural tracking of attended speech. *Current Research in Neurobiology*, *3*, 100043. https://doi.org/10.1016/j.crneur.2022.100043

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

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

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

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

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

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. *The Journal of Neuroscience*, *39*(29), 5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019

Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, *32*(9), 1192–1203. https://doi.org/10.1080/23273798.2017.1323109

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

Gadamer, H. G. (1975). *Truth and Method*. Seabury Press. https://books.google.co.jp/books?id=zQnXAAAAMAAJ

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. *The Journal of Neuroscience*, *41*(50), 10316–10329. https://doi.org/10.1523/JNEUROSCI.0812-21.2021

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 1–10. https://doi.org/10.1080/23273798.2018.1499946

Har-shai Yahav, P., & Zion Golumbic, E. (2021). Linguistic processing of task-irrelevant speech at a cocktail party. *ELife*, *10*, 1–24. https://doi.org/10.7554/eLife.65096

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

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Heil, M., Rolke, B., Pecchinenda, A., Heil, M., Rolke, B., & Pecchinenda, A. (2004). Automatic Semantic Activation Is No Myth. *Psychological Science*, *15*(12), 852–857. https://doi.org/doi.org/10.1111

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), 2020.12.03.410399. https://doi.org/10.1073/pnas.2201968119

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences*, *109*(49), 20095–20100. https://doi.org/10.1073/pnas.1213390109

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180. https://doi.org/10.1126/science.182.4108.177

Jiang, J., Chen, C., Dai, B., Shi, G., Ding, G., Liu, L., & Lu, C. (2015). Leader emergence through interpersonal neural synchronization. *Proceedings of the National Academy of Sciences*, *112*(14), 4274–4279. https://doi.org/10.1073/pnas.1422930112

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience*, *32*(45), 16064–16069. https://doi.org/10.1523/JNEUROSCI.2926-12.2012

Jiang, J., Zheng, L., & Lu, C. (2021). A hierarchical model for interpersonal verbal communication. *Social Cognitive and Affective Neuroscience*, *16*(1–10), 246–255. https://doi.org/10.1093/scan/nsaa151

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Kuhlen, A. K., Allefeld, C., & Haynes, J.-D. (2012). Content-specific coordination of listeners' to speakers' EEG during communication. *Frontiers in Human Neuroscience*, *6*(October), 1–15. https://doi.org/10.3389/fnhum.2012.00266

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

Kutas, M., & Hillyard, S. A. (1989). An Electrophysiological Probe of Incidental Semantic Association. *Journal of Cognitive Neuroscience*, *1*(1), 38–49. https://doi.org/10.1162/jocn.1989.1.1.38

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

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

Leong, V., Byrne, E., Clackson, K., Georgieva, S., Lam, S., & Wass, S. (2017). Speaker gaze increases information coupling between infant and adult brains. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(50), 13290–13295. https://doi.org/10.1073/pnas.1702493114

Li, J., Hong, B., Nolte, G., Engel, A. K., & Zhang, D. (2022). Preparatory delta phase response is correlated with naturalistic speech comprehension performance. *Cognitive Neurodynamics*, *16*(2), 337–352. https://doi.org/10.1007/s11571-021-09711-z

Li, Z., Li, J., Hong, B., Nolte, G., Engel, A. K., & Zhang, D. (2021). Speaker–Listener Neural Coupling Reveals an Adaptive Mechanism for Speech Comprehension in a Noisy Environment. *Cerebral Cortex*, *31*(10), 4719–4729. https://doi.org/10.1093/cercor/bhab118

Liu, L., Yan, X., Li, H., Gao, D., & Ding, G. (2020). Identifying a supramodal language network in human brain with individual fingerprint. *NeuroImage*, *220*(July), 117131. https://doi.org/10.1016/j.neuroimage.2020.117131

Liu, L., Zhang, Y., Zhou, Q., Garrett, D. D., Lu, C., Chen, A., Qiu, J., & Ding, G. (2020). Auditory-Articulatory Neural Alignment between Listener and Speaker during Verbal Communication. *Cerebral Cortex*, *30*(3), 942–951. https://doi.org/10.1093/cercor/bhz138

Liu, Y., Piazza, E. A., Simony, E., Shewokis, P. A., Onaral, B., Hasson, U., & Ayaz, H. (2017). Measuring speaker–listener neural coupling with functional near infrared spectroscopy. *Scientific Reports*, *7*(1), 43293. https://doi.org/10.1038/srep43293

Lu, Y., Jin, P., Pan, X., & Ding, N. (2022). Delta-band neural activity primarily tracks sentences instead of semantic properties of words. *NeuroImage*, *251*(January), 118979. https://doi.org/10.1016/j.neuroimage.2022.118979

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

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), 1024–1027. https://doi.org/10.1016/j.cub.2009.09.005

Mesik, J., Ray, L., & Wojtczak, M. (2021). Effects of Age on Cortical Tracking of Word-Level Features of Continuous Competing Speech. *Frontiers in Neuroscience*, *15*(April), 1–21. https://doi.org/10.3389/fnins.2021.635126

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., & Fay, R. R. (2017). The Auditory System at the Cocktail Party. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *Springer International Publishing*. Springer International Publishing. https://doi.org/10.1007/978-3-319-51662-2

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (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

Nastase, S. A., Liu, Y. F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1), 1–22. https://doi.org/10.1038/s41597-021-01033-3

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

Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-To-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-04464-4

Pérez, A., Dumas, G., Karadag, M., & Duñabeitia, J. A. (2019). Differential brain-to-brain entrainment while speaking and listening in native and foreign languages. *Cortex*, *111*, 303–315. https://doi.org/10.1016/j.cortex.2018.11.026

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503. https://doi.org/10.1111/j.1460-9568.2012.08060.x

Power, A. J., Lalor, E. C., & Reilly, R. B. (2011). Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. *Cerebral Cortex*, *21*(6), 1223–1230. https://doi.org/10.1093/cercor/bhq233

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360. https://doi.org/10.1038/nrn2811

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Sanders, L. D., & Neville, H. J. (2003). An ERP study of continuous speech processing. *Cognitive Brain Research*, *15*(3), 228–240. https://doi.org/10.1016/s0926-6410(02)00195-7

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

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

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, *107*(32), 14425–14430. https://doi.org/10.1073/pnas.1008662107

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual Alignment: How Brains Achieve Mutual Understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007

Teng, X., Tian, X., Rowland, J., & Poeppel, D. (2017). Concurrent temporal channels for auditory processing: Oscillatory neural entrainment reveals segregation of function at different scales. *PLoS Biology*, *15*(11), 1–29. https://doi.org/10.1371/journal.pbio.2000812

Teoh, E. S., Ahmed, F., & Lalor, E. C. (2022). Attention Differentially Affects Acoustic and Phonetic Feature Encoding in a Multispeaker Environment. *The Journal of Neuroscience*, *42*(4), 682–691. https://doi.org/10.1523/JNEUROSCI.1455-20.2021

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Wang, Yadong, Ding, N., Ahma, N., Xiang, J., Poeppel, D., & Simon, J. Z. (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. *Journal of Neurophysiology*, *107*(8), 2033–2041. https://doi.org/10.1152/jn.00310.2011

Wang, Yuanye, Zhang, J., Zou, J., Luo, H., & Ding, N. (2019). Prior Knowledge Guides Speech Segregation in Human Auditory Cortex. *Cerebral Cortex*, *29*(4), 1561–1571. https://doi.org/10.1093/cercor/bhy052

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2019). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience*, *32*(1), 155–166. https://doi.org/10.1162/jocn_a_01467

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

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for Neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003

Yeshurun, Y., Nguyen, M., & Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, *22*(3), 181–192. https://doi.org/10.1038/s41583-020-00420-w

Yu, W., Sun, W., & Ding, N. (2022). Asymmetrical cross-modal influence on neural encoding of auditory and visual features in natural scenes. *NeuroImage*, *255*(March), 119182. https://doi.org/10.1016/j.neuroimage.2022.119182

Zhang, Z. G., Hu, L., Hung, Y. S., Mouraux, A., & Iannetti, G. D. (2012). Gamma-band oscillations in the primary somatosensory cortex-A direct and obligatory correlate of subjective pain intensity. *Journal of Neuroscience*, *32*(22), 7429–7438. https://doi.org/10.1523/JNEUROSCI.5877-11.2012

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

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037