frontiers | Frontiers in Computational Neuroscience

METHODS
published: 07 July 2022
doi: 10.3389/fncom.2022.919215

Detection of Brain Network Communities During Natural Speech Comprehension From Functionally Aligned EEG Sources

Di Zhou[1], *Gaoyan Zhang*[2]*, *Jianwu Dang*[1,2]*, *Masashi Unoki*[1] *and Xin Liu*[1]

[1] School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan, [2] College of Intelligence and Computing, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China

In recent years, electroencephalograph (EEG) studies on speech comprehension have been extended from a controlled paradigm to a natural paradigm. Under the hypothesis that the brain can be approximated as a linear time-invariant system, the neural response to natural speech has been investigated extensively using temporal response functions (TRFs). However, most studies have modeled TRFs in the electrode space, which is a mixture of brain sources and thus cannot fully reveal the functional mechanism underlying speech comprehension. In this paper, we propose methods for investigating the brain networks of natural speech comprehension using TRFs on the basis of EEG source reconstruction. We first propose a functional hyper-alignment method with an additive average method to reduce EEG noise. Then, we reconstruct neural sources within the brain based on the EEG signals to estimate TRFs from speech stimuli to source areas, and then investigate the brain networks in the neural source space on the basis of the community detection method. To evaluate TRF-based brain networks, EEG data were recorded in story listening tasks with normal speech and time-reversed speech. To obtain reliable structures of brain networks, we detected TRF-based communities from multiple scales. As a result, the proposed functional hyper-alignment method could effectively reduce the noise caused by individual settings in an EEG experiment and thus improve the accuracy of source reconstruction. The detected brain networks for normal speech comprehension were clearly distinctive from those for non-semantically driven (time-reversed speech) audio processing. Our result indicates that the proposed source TRFs can reflect the cognitive processing of spoken language and that the multi-scale community detection method is powerful for investigating brain networks.

Keywords: community detection, neural entrainment, temporal response function (TRF), source localization, electroencephalography

OPEN ACCESS

Edited by:
Xinting Ge,
Shandong Normal University, China

Reviewed by:
Vangelis P. Oikonomou,
Centre for Research and Technology Hellas (CERTH), Greece
German Castellanos-Dominguez,
National University of Colombia, Manizales, Colombia

***Correspondence:**
Gaoyan Zhang
zhanggaoyan@tju.edu.cn
Jianwu Dang
jdang@jaist.ac.jp

Received: *13 April 2022*
Accepted: *14 June 2022*
Published: *07 July 2022*

Citation:
Zhou D, Zhang G, Dang J, Unoki M and Liu X (2022) Detection of Brain Network Communities During Natural Speech Comprehension From Functionally Aligned EEG Sources. Front. Comput. Neurosci. 16:919215. doi: 10.3389/fncom.2022.919215

1. INTRODUCTION

Speech comprehension is the acquisition of communicative information from speech sounds, which links auditory stimuli and cognitive processes (Zhang et al., 2019). During speech comprehension, low-level uninterrupted acoustic features are transferred to high-level linguistic information, which in turn is integrated into meaningful sentences and also stories

(Gaskell and Mirkovic, 2016). One of the main objectives of auditory neuroscience is to investigate how the human brain comprehends perceived speech. In past decades, researchers tried to use isolated words or simple sentences to investigate speech comprehension in the human brain. In those studies, subjects were asked to participate in specific tasks such as identifying whether the perceived word is a real word or a pseudoword (Binder et al., 2009; Zhang et al., 2019), or assessing whether a word in a sentence is congruent or incongruent with the rest of the sentence (Chow and Phillips, 2013). With this kind of well-designed paradigm, researchers can use a statistical analysis method (such as t-tests, analysis-of-variance) to estimate the mechanism of speech processing by comparing neural behaviors between different conditions. For example, in an experiment designed to study the N400 amplitude and word expectancy, it was found that N400s to sensible and equally low-cloze-probability completions of strongly constraining sentences (e.g., The bill was due at the end of the hour) were much larger than those to high-cloze-probability endings (The bill was due at the end of the month) (Kutas and Federmeier, 2011). However, such a task is far away from human speech comprehension in daily life. In recent years, studies have extended the controlled experimental paradigm, from isolated words or simple sentences to a more natural experimental setting, such as listening to continuous speech with a complete storyline (Brennan, 2016). Under the natural language paradigm, previous studies treat the neural system as a linear time-invariant (LTI) system. Although it is too rough to treat the complicated neural system as an LTI system, the reversible properties of the LTI system are helpful for evaluating the performance of the linearized neural system. The impulse response of the neural system is referred to as a temporal response function (TRF) hereafter. The TRF can describe the neural response from a speech stimulus, which can be represented by a speech envelope, phoneme sequence, or a spectrogram, to a specific electrode of the EEG setting, which is also called neural entrainment to speech (Ding and Simon, 2012; Brodbeck et al., 2018b; Pan et al., 2019). Many studies have proved that TRFs can reflect high-level language processing to some extent, such as categorical representations of phonemes (Liberto et al., 2015), perception attention and speech comprehension processing (Broderick et al., 2018; Weissbart et al., 2020), and lexical processing (Brodbeck et al., 2018a). Because most language cognitive processes occur within hundreds of milliseconds, high temporal resolution electroencephalograph/magnetoencephalography (EEG/MEG) techniques are often used for the natural spoken language paradigm.

Although EEG/MEG is an effective, non-invasive technique for investigating the neural mechanism behind auditory processing, due to the low spatial resolution of EEG/MEG, most previous studies investigate natural spoken language only in the electrode/sensor space (Liberto et al., 2015; Broderick et al., 2018; Etard and Reichenbach, 2019). Because the signal of an EEG electrode is a mixture of many source components, previous research cannot explain the cortical origins of the underlying natural spoken language processes. With the development of EEG signal processing, it has been proved that with enough sensors or using an accurate individual head model with reasonable conductivity values, EEG source localization may be precise enough to reflect the cortical origins of language processing (Klamer et al., 2015). In addition, as generators of neural activity cannot unambiguously be interpreted from sensor EEG data, which cannot provide useful information to explore the underlying brain mechanism, study done in the source space is beneficial to explicitly exploring brain functions in response to continuous speech (Stropahl et al., 2018). Therefore, this study investigates the neural responses to continuous speech on the basis of sources reconstructed from EEG signals.

More recent studies got some exciting results for both speech production and speech perception based on EEG source localization techniques (Stropahl et al., 2018; Zhang et al., 2019; Janssen et al., 2020). However, most of them are based on event-related potential (ERP) paradigm (Handy, 2005). In particular, in the natural speech paradigm, stimuli are typically long segments from lectures or stories and presented to subjects only once to avoid a priming effect. There are two key problems that need to be solved for single-trial paradigm before source localization. First, It is acknowledged that the generated electrical fields are easily contaminated by external noise (e.g., eye movement, head movement) that occurs during the transmission from the neural population to the top layer of the scalp through the brain tissue and skull. If we reconstruct a source single from a single trial and then fit TRFs to cortical sources directly, the unexpected noise may affect the accuracy and interpretability of source-based TRFs. In addition, most of the recent source localization techniques are developed for the ERP paradigm based on the assumption of spatiotemporal sparsity (Grech et al., 2008; Pirondini et al., 2017; Mannepalli and Routray, 2018; Liu et al., 2019; Asadzadeh et al., 2020). However, the natural speech paradigm does not allow for additive averaging across repeated trials common in the source localization of responses evoked from EEG. To solve the problem, we propose using additive averaging across subjects to improve the accuracy of source localization. Assuming that the brain functions for speech processing are consistent across individuals, a similar neural response can be expected from different subjects for the same speech stimulus. In contrast, external noise, involuntary breathing, and attentiveness differ from individual to individual, and such noises can be suppressed by averaging the neural signals of the same stimuli for all subjects. However, this is difficult due to the lack of methods that account for subjects' differences in terms of the setup positions of the electrodes. Addressing this problem well before averaging neural activities across multiple subjects should result in more accurate source localization.

To do so, we first propose a functional hyper-alignment method to reduce the mismatching caused by individual experiment settings, and source reconstruction is then performed on the basis of additive averaging over all subjects for each trial. TRFs are then estimated independently for each localized brain source; they are then related to one another by using Pearson correlation to construct the whole brain functional network. Previous research proved that the brain network can be characterized by its community structure, and community detection for functional brain networks has facilitated the

understanding of the underlying brain organization and its related cognitive function (Forbes et al., 2018). Therefore, we use the community detection method to detect the community structure of the brain networks underlying continuous speech comprehension. We compared the community organization between the understanding of a naturally told story and a time-reversed story to study the brain mechanism related to semantic processing underlying continuous speech comprehension.

2. METHODS

2.1. Noise Reduction by Additive Averaging

There are many kinds of external noises caused by eye movement, heartbeat, electrical noise, and so on. They occur during the transmission from the neural population to the scalp through the brain tissue and skull and are mixed into EEG signals (Cohen, 2014). Supposing these noises are random, ERP analysis removes these kinds of noises by applying an addition operation to a number of trials for the same task, namely additive averaging. Our study adopts an idea similar to that used in the ERP technique to reduce noise, but we apply additive averaging to EEG signals for the same stimulus material over all subjects since subjects can listen to the same material only once. In our previous study, we proved that additive averaging across subjects can improve the accuracy of TRFs (Di Zhou et al., 2020).

To apply additive averaging across subjects, it is better to calibrate the individual differences in the setup positions of the electrodes. To tackle this problem, we propose a functional hyper-alignment method for soft calibration. It uses a well-designed spatial filter to align the setup positions of electrodes by minimizing the distance of the signal features among the subjects. Due to the lack of methods that account for subjects' differences in the setup stage of the electrodes, the position of an electrode n for subject i may not be the same as that of subject j. Thus, the additive average over the EEG data $x_i(t,n)(i = 1,2,...,I)$ cannot be used to perform denoising properly, where I is the subject number. For this reason, we propose using a functional hyper-alignment method for eliminating this effect. The main idea of the functional hyper-alignment is to rotate $x_i(t,n)$ and $x_j(t,n)$ $(i \neq j \in [1,2,...,I])$ to maximize their correlation among subjects. So far, several methods have been proposed for this purpose, such as group task-related component analysis (gTRCA) (Tanaka, 2020) and multi-set canonical correlation analysis (MCCA) (de Cheveigné et al., 2019). We choose MCCA to maximize the data correlation among subjects, which satisfies the requirement of our study.

The goal of MCCA is to find projection vectors ω that maximize the correlation between multiple data sets $X_i, i = 1,2...,I$. The correlation ρ of all data sets can be calculated as the ratio of the summations of the between-set covariance $V_{x_ix_j}$ over the within-set covariance $V_{x_ix_i}$,

$$p(\widetilde{X}_1,\widetilde{X}_2,...,\widetilde{X}_i,...,\widetilde{X}_I) = \frac{1}{N-1}\frac{\sum_{i=1}^{I}\sum_{j=1,i\neq j}^{I}\omega_i^T V_{x_ix_j}\omega_j}{\sum_{i=1}^{I}\omega_i^T V_{x_ix_i}\omega_i}, \quad (1)$$

where

$$V_{x_ix_j} = (X_i - \bar{X}_i)^T(X_j - \bar{X}_j), \quad (2)$$

$$V_{x_ix_i} = (X_i - \bar{X}_i)^T(X_i - \bar{X}_i). \quad (3)$$

$\bar{X}_i$, $\bar{X}_j$ are the means for set i and set j. $\frac{1}{N-1}$ ensures that the correlation ρ scales between 0 and 1. Altogether, the above equation can be summarized into a generalized eigenvalue problem,

$$B\omega = \lambda R\omega, \quad (4)$$

where

$$B = \begin{bmatrix} O & V_{x_1x_2} & \cdots & V_{x_1x_I} \\ V_{x_2x_1} & O & \cdots & V_{x_2x_I} \\ \vdots & \vdots & \ddots & \vdots \\ V_{x_Ix_1} & V_{x_Ix_2} & \cdots & O \end{bmatrix}, R = \begin{bmatrix} V_{x_1x_1} & O & \cdots & O \\ O & V_{x_2x_2} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & V_{x_Ix_I} \end{bmatrix}. \quad (5)$$

B is a matrix combining all between-set covariance $V_{x_ix_j}$, and R is a diagonal matrix that contains all within-set covariance $V_{x_ix_i}$. ω is a spatial vector set for an entire data set $\omega = [\omega_1^T, \omega_2^T, \ldots, \omega_I^T]$. Finally, the spatial filter for aligning the positions of the electrodes is reduced to solve the generalized eigenvalue problem.

2.2. Source Reconstruction Based on Denoising EEG Data

After the denoising, the EEG data are used to estimate their cortical source activations in the brain, namely source reconstruction. In this study, the forward and reverse models for source localization were calculated by the Brainstorm toolbox (Tadel et al., 2011). The finite element method (FEM) as implemented in DUNEuro was used to compute the forward head model using Brainstorms default parameters with an MNI MRI template (ICBM152) (Vorwerk et al., 2016; Schrader et al., 2021). The FEM models provide more accurate results than the spherical forward models and more realistic geometry and tissue properties than the boundary element method (BEM) methods (Gramfort et al., 2010). For source estimation, the number of potential sources (grid on the cortex surface) is set to 15,002. And the option of constrained dipole orientations was selected, which means dipoles are oriented perpendicular to the cortical surface (Tadel et al., 2011). We then apply the method of standardized low-resolution electromagnetic tomography (sLORETA) (Pascual-Marqui, 2002) to obtain plausible EEG source estimates. Although the spatial resolution of sLORETA is low, sLORETA can provide smooth and good localization with few localization errors (Asadzadeh et al., 2020). Finally, according to the Desikan-Killiany Atlas (DKA), the cortical surface is divided into 68 anatomical regions of interest (ROIs) (Desikan et al., 2006). The time series of each ROI is calculated from the average value of all dipoles in the respective region. As a result, we obtain a series of brain areas (sources) that are activated in speech comprehension.

2.3. Source-Based TRFs Estimation

To evaluate the brain functions explicitly, we calculate the TRFs from speech input to sources in the brain cortex, instead of the electrodes on the scalp. The mTRF toolbox (https://github.com/mickcrosse/mTRF-Toolbox) is used to linearly map the speech envelope and the neural response in the sources by approximating the brain as an LTI system (Crosse et al., 2016). Let $\hat{r}(\tau, ROI_n)$ be the TRF of a brain region ROI_n for an input speech envelope $s(t)$, the neural response signal $\hat{x}(t, ROI_n)$ of the source ROI_n can be described as follows.

$$\hat{x}(t, ROI_n) = \sum_{\tau} \hat{r}(\tau, ROI_n) s(t - \tau) \tag{6}$$

The range for τ is from 0 to 800 ms in our study, as most common ERP components in language research are within 800 ms (Beres, 2017). The broadband temporal speech envelopes of $s(t)$ are obtained from a gammatone filterbank followed by a power law (Biesmans et al., 2016; Peng et al., 2018, 2021). For the modeling process, the envelope is decimated to the same sampling rate as the source signal, enabling us to relate its dynamics to the source.

Under the theory for the LTI system, the backward approach can be modeled using a decoder $\hat{r}^{-1}(\tau, ROI_n)$, which is the inverse function of $\hat{r}(\tau, ROI_n)$. The optimal decoder $\hat{r}^{-1}(\tau, ROI_n)$ is acquired by minimizing the MSE between the original and predicted speech envelope, and n denotes the number of regions. Thus, the input speech stimulus $s(t)$ can be decoded from the source neural signal $\hat{x}(t, ROI_n)$ using the decoder function $\hat{r}^{-1}(\tau, ROI_n)$. This can be expressed as follows.

$$s(t) = \sum_{n} \sum_{\tau} \hat{r}^{-1}(\tau, ROI_n) \hat{x}(t - \tau, ROI_n) \tag{7}$$

The encoder $\hat{r}(\tau, ROI_n)$ and decoder $\hat{r}^{-1}(\tau, ROI_n)$ approach the optimal values when iterating the above calculation.

2.4. Brain Network Analysis Based on Community Detection Method

2.4.1. Construction of Preliminary Brain Network

The brain network can be characterized as a community structure. Therefore, community detection is often used in exploring the brain network during a given task (Jin et al., 2019). To do so, we first need to define the nodes of the brain and links of the network (Yu et al., 2018). In large-scale brain networks, nodes usually represent brain regions, and links represent anatomical, functional, or effective connections (Friston et al., 1994). The predefined spatial regions of interest (ROIs) assessed by anatomical atlases are one of the most popular methods for defining brain nodes (Smith, 2012). This study uses the 68 nodes (brain region) that were defined in the DKA, and it uses Pearson correlation to describe the functional link among the nodes (Smith et al., 2013). This would result in 2,278 ($= C_{68}^2$) edges if linking all pairwise nodes for each trial. Differing from the previous studies, the link weights (temporal correlations) here are calculated using the TRF of each node, but not the source neural signal. As a result, we obtain a preliminary brain network that consists of all of the brain regions and pairwise links with a weighted edge.

2.4.2. Community Detection in Functional Brain Networks

Community detection divides the nodes of the preliminary network into a number of non-overlapping clusters and then detects the communities in a functional network by maximizing the module quality metric Q (Newman and Girvan, 2004), which is also called modularity. A higher value of modularity represents that the detection approaches a more evident community structure. Therefore, this algorithm provides not only a community partition but also an index for evaluating whether a network community structure is evident. Recent studies have proposed that the negative correlation in the functional connection matrix also possesses some physiological significance, and correlated and anti-correlated brain activities may signify cooperative and competitive interactions between brain areas that subserve adaptive behaviors (Khambhati et al., 2018; Zhang and Liu, 2021). Therefore, we used an optimized community detection algorithm in the Brain Connectivity Toolbox (Anwar et al., 2016) on a preliminary connection matrix to detect the community structure of the functional connection matrix for different densities, which takes into account both positive and negative correlations in a network (Bolton et al., 2018; Zhang and Liu, 2021). Since the density as a threshold reflects how many edges are effective in a network, density selection also has a significant impact on studying brain function (Liu et al., 2011; Jrad et al., 2016; Jin et al., 2019). To determine the brain network more accurately, this study uses different scales to divide the brain network; thus, it can detect the optimal communities of a brain network using different densities. While running this algorithm, the default resolution parameter of γ is set to 1, yielding modularity scores Q_d and density D_t for $t = 0.01, 0.02, ..., 0.1, ..., 0.5$, where t is the variation of density.

2.4.3. Brain Network Selection

For community detection with different scales, it is important to find the optimal functional brain networks from the scales. The scale means the different thresholds which can sparse the brain networks in different densities. On the basis of a previous study (Jin et al., 2019), this study uses the variation of information (VI) to evaluate the similarity of community structures with different scales in functional brain networks. VI can compare two community structures by means of their information exchange loss and gain. VI can be described as

$$VI(X, Y) = H(X) + H(Y) - 2I(X; Y), \tag{8}$$

where X and Y are two different community structures of the brain network. $H(X)$ and $H(Y)$ are the entropy for X and Y, respectively. $I(X; Y)$ is the mutual information between X and Y.

When VI is equal to zero, it represents the most stable partition across densities (He et al., 2018). According to VI values, we can obtain the most reasonable community partition of functional brain networks. In addition to VI, this study also uses cluster analysis on the different scales to find the best discrimination for separating natural speech and time-reversed speech in brain networks. Finally, we evaluate the selected

brain network using the current novel results of brain research (Walenski et al., 2019).

3. EXPERIMENTS

3.1. Participants

Twenty-four healthy Mandarin Chinese speakers (mean ± standard deviation age, 22 ± 2.4 years; nine males; right-handed) were recruited from Tianjin University and Tianjin University of Finance and Economics. The experiments were conducted in accordance with the Declaration of Helsinki (World Medical Association, 2014) and were approved by the local ethics committee. The subjects signed informed consent forms before the experiment and were paid for their participation afterward. All the subjects reported no history of hearing impairment or neurological disorders.

3.2. Materials and Experimental Procedure

Subjects undertook 48 non-repetitive trials. Among them, 24 trials were short stories (around 60 s) with a complete storyline, recorded by a male Chinese announcer in a soundproof room. The other 24 trials were the same story segments but were time reversed. All stimuli were mono speech with a 44.1 kHz sampling rate, and the stimulus amplitudes were normalized to have the same root mean square (RMS) intensity. The 48 trials were randomly presented to the subjects. All speech segments were also modified to truncate silence gaps to <0.5 s (Brodbeck et al., 2018b).

The experiments were carried out in an electronically and magnetically shielded soundproof room. Speech sounds were presented to subjects through Etymotic Research ER-2 insert earphones (Etymotic Research, Elk Grove Village, IL, USA) at a suitable volume (around 65 dB). During each trial, subjects were instructed to focus on a crosshair mark in the center of the screen to minimize head movements and other bodily movements. There was a 5 s interval between each trial, and the subjects were given a 5 min break every 10 trials. After each story trial, subjects were asked immediately to answer multiple-choice questions about the content of the story to ensure that they focused on the auditory task. We embedded unique tones in some trials to draw more of the subjects' attention to the reversed stimuli. Subjects were requested to detect the tones and indicate how many times they appeared after the trial. The EEG data corresponding to the embedded tones was removed in further analysis.

3.3. EEG Data Acquisition and Pre-processing

Scalp EEG signals were recorded with a 128-channel Neuroscan SynAmps system (Neuroscan, USA) at a sampling rate of 1,000 Hz. Six of the channels were used for recording a vertical electrooculogram (VEOG), a horizontal electrooculogram (HEOG), and two mastoid signals. The impedance of each electrode was kept below 5 kΩ during data acquisition. Three subjects' data were discarded in further analysis because they did not give a proper answer to the multiple-choice questions or the electrodes detached during the EEG data recording. The raw EEG data were pre-processed using the EEGLAB toolbox (https://sccn.ucsd.edu/eeglab/index.php) in MATLAB (MathWorks) (Delorme and Makeig, 2004). This involved removing sinusoidal (i.e., line) noise and bad channels (i.e., low-frequency drifts, noisy channels, short-time bursts) and repairing the data segments (Perrin et al., 1989; Plechawska-Wojcik et al., 2018). Then, the EEG data was downsampled to 250 Hz, 1 Hz high-pass filtering was performed to remove linear drift, and 40 Hz low-pass filtering was performed to remove power frequency interference and high-frequency noise. Adaptive mixture independent component analysis (AMICA) (Palmer et al., 2012) and ICLabel (Pion-Tonachini et al., 2019) were used to automatically identify and remove artifact components.

3.4. Overview of Proposed Approach

As mentioned above, we reasonably assume that brain functions for processing the same speech material are consistent across individuals; thus, consistent neural responses were expected for all subjects to the same speech stimulus. In contrast, potential noise, eye movement, involuntary breathing, and attentiveness differ from individual to individual as well as the individual electrode setting. To calibrate the electrode positions across subjects, we used functional hyper-alignment and applied it to the EEG data first. After the calibration, we suppressed the random noises by additive averaging of the neural signals of the same stimulus over all subjects. Then, the denoised EEG singles were used to reconstruct the neural sources of EEG data in the brain, and the TRFs for the neural sources were estimated. Using the TRFs, we constructed a functional brain network of natural speech processes on different scales. Finally, an optimal functional brain network was decided on the basis of the VI of the communities with the multiple scales. A flowchart for the proposed approach is shown in **Figure 1**.

4. RESULTS

4.1. Accuracy for Source-Based TRFs

4.1.1. Evaluation With Results of Speech Envelope Prediction

The accuracy of the backward prediction of stimulus input from neural output is usually used for evaluating the performance of TRFs (O'Sullivan et al., 2015; Das et al., 2020). Here, we use the backward prediction approach to evaluate the performance of the proposed functional hyper-alignment approach through comparison with the traditional method with single-trial estimation. For training processes, we used a leave-one-out cross-validation procedure, where 23 trials were used for training, and the remaining one was used for testing in each fold. The prediction accuracy was described by the Pearson correlation coefficient between the predicted speech envelopes and the original ones. **Figure 2** shows comparisons of the prediction accuracies for the single-trial method without noise reduction (O'Sullivan et al., 2015), the additive average method over the subjects without functional hyper-alignment (Di Zhou et al., 2020), and the functional hyper-alignment method combined with the additive average method. For a fair comparison, the correlation coefficient was first transformed into a z-value by Fisher's z transformation to satisfy a normal distribution

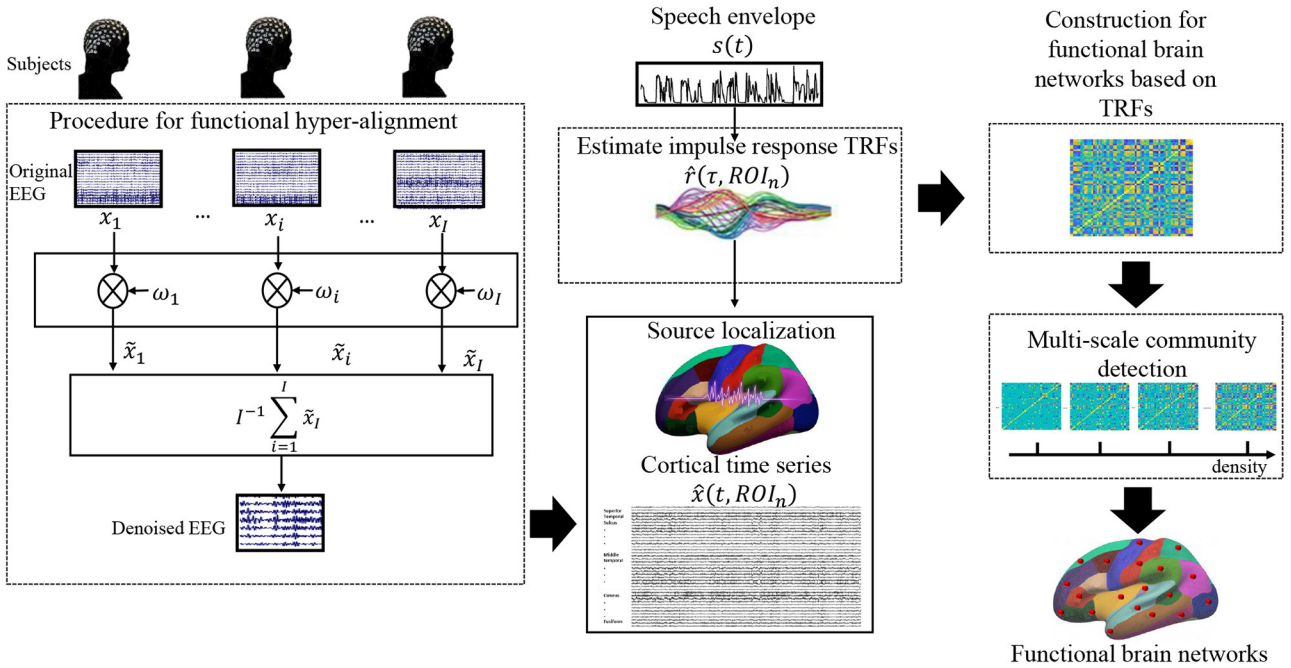

FIGURE 1 | Flowchart of the data processing approaches in this study.

(Corey et al., 1998). Then, an analysis-of-variance (ANOVA) of the z-values revealed a significant effect on the prediction methods ($F = 193.53, p < 0.001$). The results of the ANOVA demonstrated that the prediction accuracy of our proposed method, the functional hyper-alignment method combined with the additive average method, was the highest among the three methods. A post-hoc test for the ANOVA showed that the prediction accuracy was significantly improved by the functional hyper-alignment method, compared with the average without the functional hyper-alignment method ($F = 9.68, p < 0.005$) and single-trial method ($F = 400.78, p < 0.001$). Therefore, the functional hyper-alignment with the additive average method was used in the following TRF estimation.

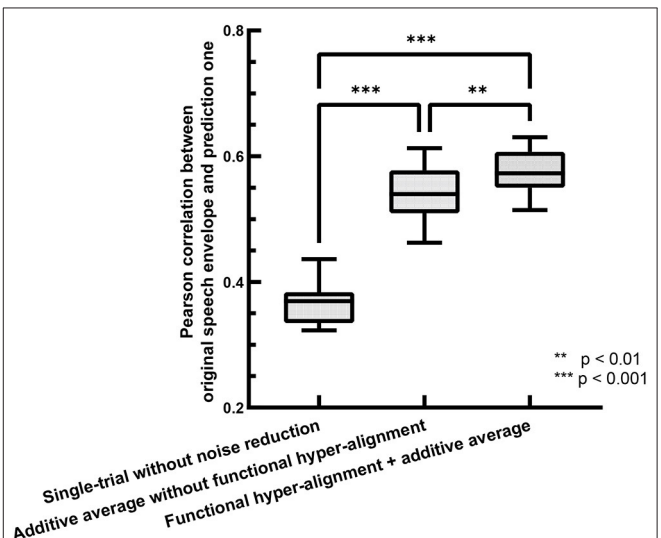

FIGURE 2 | Comparison of envelope prediction accuracies between functional hyper-alignment method and other two methods.

4.1.2. Evaluation by Comprehension Behaviors

To evaluate the behaviors of the different methods from a brain function point of view, we used a distinct degree of the brain networks in speech comprehension. We judged whether the methods could or could not distinguish the brain network for the normal speech listening task from that for the time-reversed speech listening task since the latter one tackles non-semantic audio sequences. The method with a higher distinct degree between normal play and reverse play would be the better method for our study. Therefore, we estimated source-based TRFs using the proposed method and the single-trial method, and we then constructed functional brain networks based on the TRFs. The link between two brain regions was expressed by a Pearson correlation ranging from −1 to 1. The higher the correlation, the higher the similarity between the two brain regions, and vice versa.

During the experiment, we asked the subjects to answer multiple-choice questions about the story presented in the listening task after each trial. The accuracy of the answers of the question was 88.25 ± 4.62% for the normal speech, which indicates that the subjects mostly comprehended it seriously. The

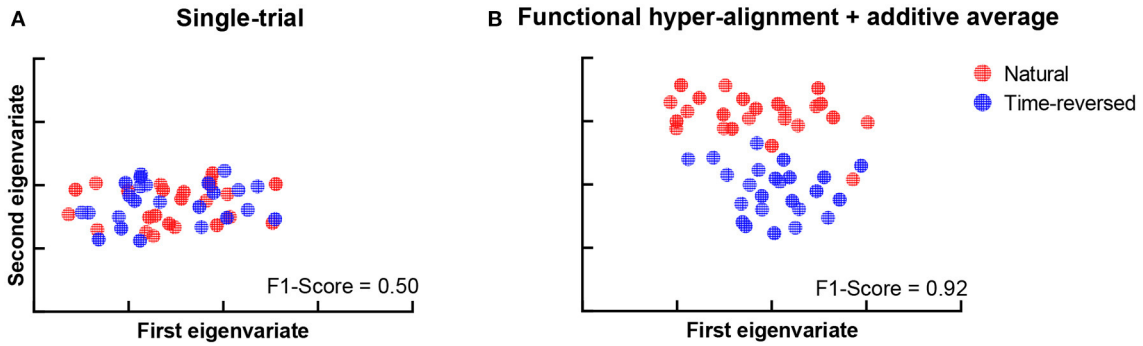

FIGURE 3 | K-means clustering of t-SNE embedded distributions obtained by single-trial method **(A)** and proposed method **(B)**.

speech intelligibility was also evaluated on a numerical rating scale from 1 to 5 by the subjects, where "very easy to understand" was scored 5, and "completely incomprehensible" was scored 1. The speech intelligibility was 4.74 ± 0.45 for the normally played natural story but was 1.46 ± 0.81 for the time-reversed one. This means that speech was not comprehended in the time-reversed case since there was little semantic information. For these reasons, functional brain networks are expected to be separated into two clusters. One cluster is for semantically driven brain activation, and the other is for non-semantically driven audio processing. To clarify our expectation, we used t-distributed stochastic neighbor embedding (t-SNE) (der Maaten and Hinton, 2008) to visualize the brain networks in a two-dimensional representation, and we verified whether or not semantically driven brain activation was separable from non-semantically driven activation. We first reshaped the connection matrix with a size of 68×68 to a vector with 2,278 dimensions to analyze the pairwise connections of all the brain nodes for each trial. For 48 trials, we had a matrix with a size of 48 × 2,278, and we applied t-SNE analysis to it. Finally, the results of the t-SNE analysis were clustered by using the k-means algorithm (Liberto et al., 2015). We set the cluster number to 2, and we performed 1,000 k-means repetitions with random initial states. A two-dimensional scatter of the semantically and non-semantically driven brain networks is shown in **Figure 3**, where the left panel is the result obtained with the single-trial method, and the right panel is for the proposed method. One can see that the functional connections based on our proposed method show distinct clusters for natural and time-reversed speech, while the single-trial method does not show any obvious clusters. The F1-scores of the actual groupings and the k-means clusters were calculated for all repetitions. The average F1-scores of 1,000 repetitions was 0.5 for the single-trial method (**Figure 3A**) and 0.93 (**Figure 3B**) for the proposed method. These results indicate that a brain network based on our proposed method can reveal high-level speech comprehension processing; however, that of the single-trial TRF-based brain network cannot.

To address the differences in the brain networks between the natural speech and time-reversed speech in detail, we investigated the difference in the estimated TRFs between the two cases. **Figure 4** shows examples of the TRFs of the left superior temporal sulcus (STS) and the left middle temporal gyrus (MTG) for natural speech and time-reversed speech, where STS and MTG are considered to correspond to speech comprehension (Price, 2012). One can see that the patterns of the peaks and troughs for STS (**Figure 4A**) show a significant difference at time lags between 300 and 450 ms of the TRF (paired t-test, $p = 5.2 \times 10^{-5}$; effect size d = 1). In **Figure 4B**, the TRF patterns for MTG show a significant difference between 150 and 450 ms (paired t-test, $p = 2.1 \times 10^{-14}$; effect size d = 2).

According to Beres (2017), the difference in intervals between 150 and 450 ms plausibly corresponds to semantic processing (N400) or syntactic processing (left-anterior negativity, LAN). To verify whether the differences are related to speech comprehension or not, we applied t-SNE and k-means again. To do this, we segment the TRFs into four periods: 0 ∼ 150, 150 ∼ 300, 300 ∼ 450, and 450 ∼ 600 ms, and we applied k-means to the amplitudes of the TRFs of the 68 brain regions for each period, respectively. **Figure 5** shows the clusters of each period: 0 ∼ 150 ms (**Figure 5A**), 150 ∼ 300 ms (**Figure 5B**), 300 ∼ 450 ms (**Figure 5C**) and 450 ∼ 600 ms (**Figure 5D**). According to the F1-score, one can see that the time between 150 ∼ 450 ms (**Figure 5C**) had the best clustering, which is consistent with the common knowledge that N400 is concerned with semantic processing.

4.2. Multi-Scale Community Detection

The investigation above was carried out with a brain network with full connection. As we knew, different speech functions have different brain network communities as speech planning, semantic processing networks. Therefore, some edges in the connection matrix may be invalid or noise for a specific speech process. It is thus necessary to determine the optimal connection matrix on the basis of the different thresholds of the connection.

As described in the section on our method, we analyze the functional connection matrix using different densities, and we investigate the functional brain network at multiple scales (Jin et al., 2019). **Figure 6** shows an illustration of the brain connection matrix for four different densities.

When the modularity Q ranges from 0.3 to 0.8, in general, the network contains community structures (Jin et al., 2019). The modularity score Q_d is around 0.64 at any density D_t for the

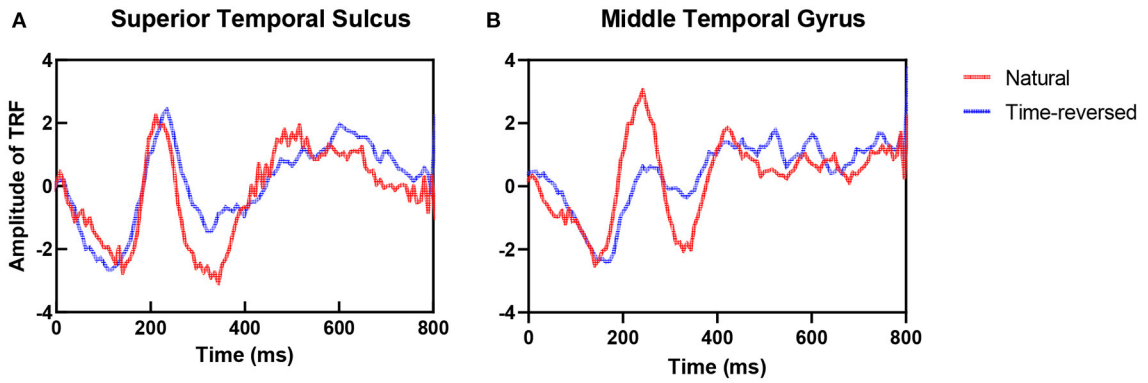

FIGURE 4 | TRFs for natural and time-reversed speech for STS **(A)** and MTG **(B)**.

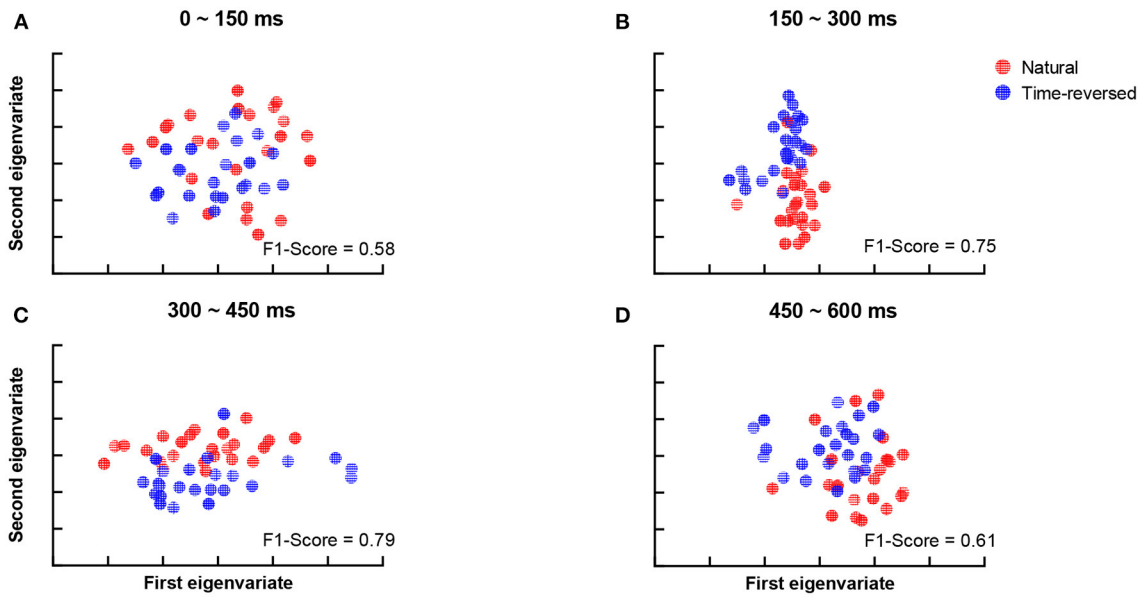

FIGURE 5 | Clustering results for TRF amplitude in different time intervals: 0 ~ 150 ms **(A)**, 150 ~ 300 ms **(B)**, 300 ~ 450 ms **(C)**, 450 ~ 600 ms **(D)**.

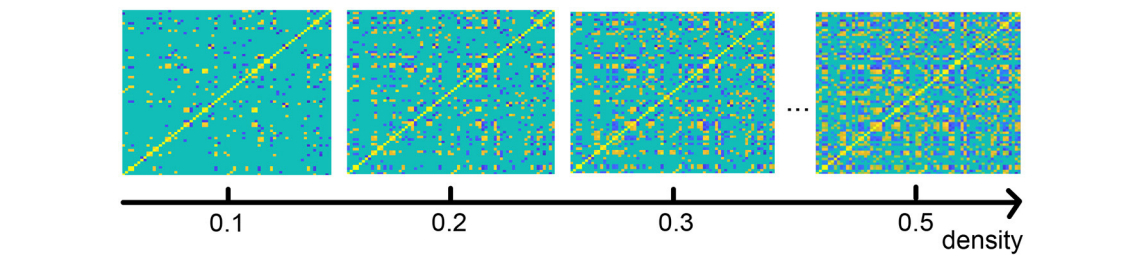

FIGURE 6 | Functional brain connection matrix with different densities. Yellow denotes edge value 1, blue −1.

functional brain network based on our proposed method. This implies that community detection is available for our data. In addition to the modularity values, we introduce VI as monitoring variables along with densities to explore the best community division. **Figure 7** shows the VI (**Figure 7A**) along with the density. When the community structure approached a stable situation, the VI value reached the minimum. At the same time, we refer to the F1-score (**Figure 7B**).

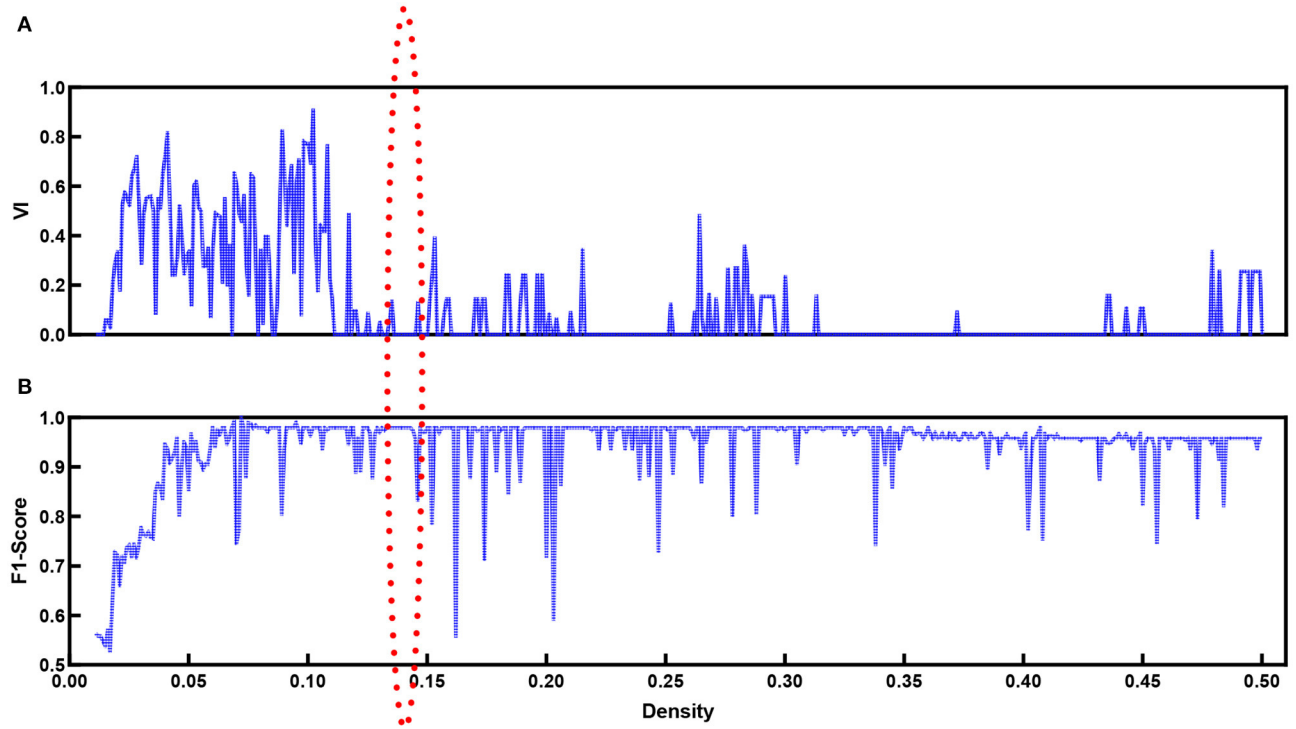

FIGURE 7 | VI **(A)** values and F1-score **(B)** for functional brain network with different density values.

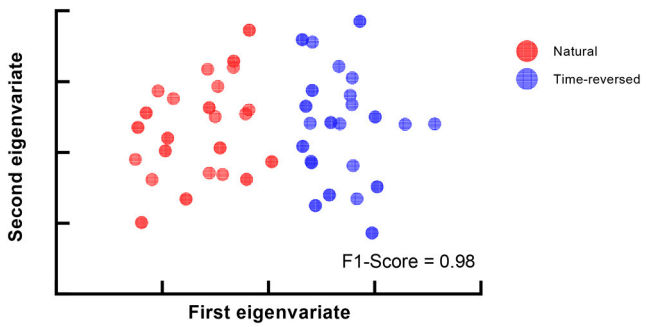

FIGURE 8 | Best clustering results for multi-scale density.

Figure 7B shows the F1-score for k-means used to separate natural and time-reversed speech at different densities of brain networks. Altogether, it seems that the three conditions could be satisfied when the density ranged from 0.136 to 0.143. In this range, we could also obtain the best separation of the brain networks for the natural speech and time-reversed speech, as shown in **Figure 8**.

In terms of the above investigation, we finally determined the optimal density value to be 0.14, and we obtained community structures using the community detection method. As a result, 16 communities were discovered in the functional brain network during natural speech comprehension, and 19 communities were discovered for the audio processing of the time-reversed speech. From the community detection results, we chose the two largest communities for natural speech and time-reversed speech, respectively, and show them in **Table 1**. Since the common processing in both cases was audio signal processing, the majority of the brain areas were the same for the natural and time-reversed speech. In the first community, 14 brain regions appeared in both cases simultaneously, including the transverse temporal cortex, superior temporal gyrus, temporal pole, and so on, which are related to auditory information processing, phonological encoding, and lexical selection, as well as syntactic and phrase level processing (Walenski et al., 2019). It is interesting to note that some of the subjects told us that the time-reversed speech sounded like a foreign language. This implies that the subjects possibly recruited brain resources to comprehend the audio sequences of the time-reversed speech even though it had no semantic information. For the normal speech, the subjects decoded the meaningful information and carried out high-level cognitive processing, whereas, for the time-reversed speech, brain areas such as the caudal middle frontal gyrus and precuneus cortex were not activated.

In previous research, the coupling of the auditory cortex and frontal areas was reported, and this coupling increases when speech has higher intelligibility (Park et al., 2015). They hypothesized that top-down signals from the frontal brain areas causally modulate the phases of brain oscillations in the auditory cortex. To verify this hypothesis, we checked the TRFs of the pars

TABLE 1 | The detected brain network communities under the conditions of natural and time-reversed speech.

Community	Natural speech	Time-reversed speech
1	L_Caudal anterior-cingulate cortex	L_Caudal anterior-cingulate cortex
	R_Caudal anterior-cingulate cortex	R_Caudal anterior-cingulate cortex
	R_Fusiform gyrus	R_Fusiform gyrus
	R_Insula cortex	R_Insula cortex
	L_Pars opercularis	L_Pars opercularis
	L_Pars triangularis	L_Pars triangularis
	L_Posterior-cingulate cortex	L_Posterior-cingulate cortex
	R_Posterior-cingulate cortex	R_Posterior-cingulate cortex
	L_Precentral gyrus	L_Precentral gyrus
	L_Superior temporal gyrus	L_Superior temporal gyrus
	L_Temporal pole	L_Temporal pole
	R_Temporal pole	R_Temporal pole
	L_Transverse temporal cortex	L_Transverse temporal cortex
	R_Transverse temporal cortex	R_Transverse temporal cortex
	L_Caudal middle frontal gyrus	L_Inferior temporal gyrus
	R_Caudal middle frontal gyrus	L_Medial orbital frontal cortex
	L_Cuneus cortex	R_Pars triangularis
	L_Insula cortex	L_Rostral anterior cingulate cortex
	L_Lateral occipital cortex	R_Rostral middle frontal gyrus
	L_Precuneus cortex	
	R_Precuneus cortex	
2	R_Superior temporal sulcus	R_Superior temporal sulcus
	L_Entorhinal cortex	L_Entorhinal cortex
	L_Fusiform gyrus	L_Fusiform gyrus
	R_Inferior parietal cortex	R_Inferior parietal cortex
	L_Middle temporal gyrus	L_Middle temporal gyrus
	R_Middle temporal gyrus	R_Middle temporal gyrus
	R_Parahippocampal gyrus	R_Parahippocampal gyrus
	L_Paracentral lobule	L_Paracentral lobule
	R_Pars opercularis	R_Pars opercularis
	L_Postcentral gyrus	L_Postcentral gyrus
	R_Postcentral gyrus	R_Postcentral gyrus
	R_Precentral gyrus	R_Precentral gyrus
	L_Superior frontal gyrus	L_Superior frontal gyrus
	L_Superior parietal cortex	L_Superior parietal cortex
	R_Superior parietal cortex	R_Superior parietal cortex
	L_Supramarginal gyrus	L_Supramarginal gyrus
	R_Supramarginal gyrus	R_Supramarginal gyrus
	R_Entorhinal cortex	R_Isthmus-cingulate cortex
	R_Inferior temporal gyrus	L_Lateral orbital frontal cortex
	L_Parahippocampal gyrus	R_Lateral orbital frontal cortex
	L_Superior temporal sulcus	R_Paracentral lobule
		L_Pars orbitalis
		R_Rostral anterior cingulate cortex
		L_Rostral middle frontal gyrus
		R_Superior frontal gyrus
		R_Superior temporal gyrus

The different brain regions in the communities were highlighted.

opercularis in the frontal area and the primary auditory cortex (transverse temporal cortex), and we show them in **Figure 9**. One can see that the coupling between the frontal area and primary auditory was stronger for the natural speech than the time-reversed speech, where the correlation coefficient was 0.65 for the natural speech but 0.23 for the time-reversed speech. In particular, the coupling rapidly decreased after 400 ms for the time-reversed speech. This may indicate that high-level language brain areas were not recruited in the audio processing for the time-reversed speech because it was not comprehended by the subjects.

Furthermore, we attempted to investigate how much the brain networks could be separated on the basis of community detection. The results are shown in **Figure 10**. The brain networks for the normal and time-reversed speech were clearly separated into two independent clusters for both communities 1 and 2.

5. DISCUSSION

5.1. Proposed Approach Extends Cognitive Understanding of TRFs to Brain Source Space

TRFs have been used to reflect functional roles for the cognitive processing of speech (Liberto et al., 2015; Broderick et al., 2018), but their estimation is predominantly limited to the electrode/sensor space. The cortical origins of the underlying speech processing are still not clear. Although recent electrocorticography (ECoG) studies provide some exciting results for the cortical origins of speech processing (Anumanchipalli et al., 2019; Zhang et al., 2021), such intracranial electrography is not friendly for healthy people, and it is also hard to investigate the whole-brain distribution of the sources due to its limited spatial range. As the accuracy of EEG source reconstruction has been improved, in this paper, we proposed an approach to extending TRFs from the electrode space to the source space by using a source reconstruction method. Source-based TRFs can reflect the cortical origins of the underlying natural spoken language processes. Then, we used the source-based TRFs to investigate the brain network during speech comprehension on the basis of community detection. The contributions of this paper are as follows. First, we proposed an approach to perform source localization for the single-trial natural speech paradigm. For accurate source localization results, we introduced a functional hyper-alignment method combined with additive averaging over all subjects. From the accuracy of speech envelope prediction, our proposed approach showed good performance. Second, from the clustering results of natural speech and time-reversed speech, the source TRFs based on our method can be used in revealing the cognitive mechanism for natural speech processing. Third, the findings obtained with our approach are highly consistent with previous meta-analysis of natural language processing (see Section More Regions Recruited in Brain Network Communities Under Natural Speech Paradigm). To the best of our knowledge, this is the first study that tries to use community detection to address the EEG-based brain network during natural speech comprehension task. As a result, our approach can be a reasonable way of performing community detection for complex natural speech processing tasks using EEG in the future.

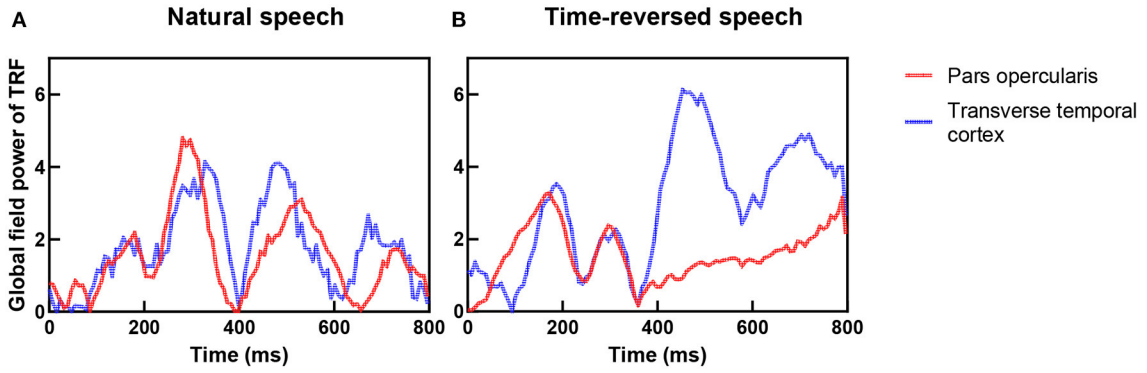

FIGURE 9 | Coupling between frontal area (red color) and auditory cortex (blue color) for natural speech **(A)** and time-reversed speech **(B)**.

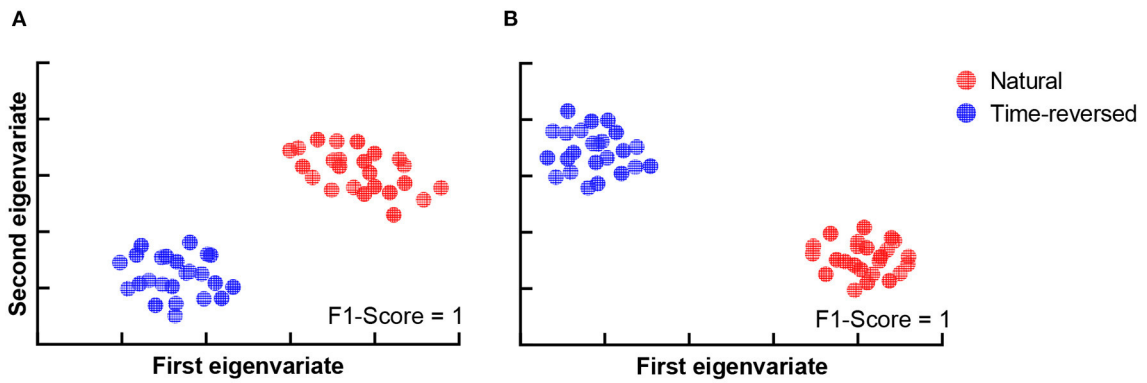

FIGURE 10 | Clustering results for two communities. **(A)** Community 1, **(B)** Community 2.

5.2. Neural Response to Speech Envelope Reflects High-Level Semantic Processing

In previous research, some contrary results regarding the neural entrainment to speech envelope were reported. For example, the speech envelope is usually considered to be related to low-level acoustic features such as the syllable boundary, while some studies reported that the neural entrainment to speech envelope was stronger when speech was easy to understand (Park et al., 2015; Vanthornhout et al., 2018). They considered a high-level top-down prediction mechanism driving the coupling between neural signal and speech envelope more strongly during intelligible speech perception than unintelligible speech perception. However, some other studies state that there was no difference in the neural entrainment to speech envelope between accessible and inaccessible speech (Howard and Poeppel, 2010; Millman et al., 2015; Zoefel and VanRullen, 2016). Here, we have enough evidence to speculate that the opposite may be caused by unexpected noises in the neural signal. The TRF-based brain network is a kind of supervised network structure. In this study, it is only related to the speech envelope. From the clustering results of **Figure 3**, the cognitive-level difference between an intelligible speech envelope and unintelligible speech envelope is affected by unexpected noises. However, after the noise reduction by the proposed method, this kind of semantic-driven speech envelope can be separated from non-semantic driven sound. From **Figure 4**, the TRFs for STS and MTG showed the typical semantic processing components around 400 ms in the natural speech condition, and from the community results, we found that the middle temporal gyri, left inferior frontal gyrus, and auditory cortex are in the same community, which may indicate a top-down prediction mechanism for improving the coupling between neural responses and speech envelopes.

5.3. More Regions Recruited in Brain Network Communities Under Natural Speech Paradigm

The natural speech results show less left-lateralization than seen under the traditional paradigm. For over a century, it has been thought that the frontal and temporoparietal regions of the left hemisphere are crucial for speech processes. However, in the natural language paradigm, it has been shown that daily speech comprehension involves bilateral networks, not only left-lateralization in traditional studies (Jung-Beeman, 2005; Huth et al., 2016; Tang et al., 2017; Hamilton et al., 2018). The natural language paradigm reveals more widespread responses to the speech comprehension process, not only to language-specific

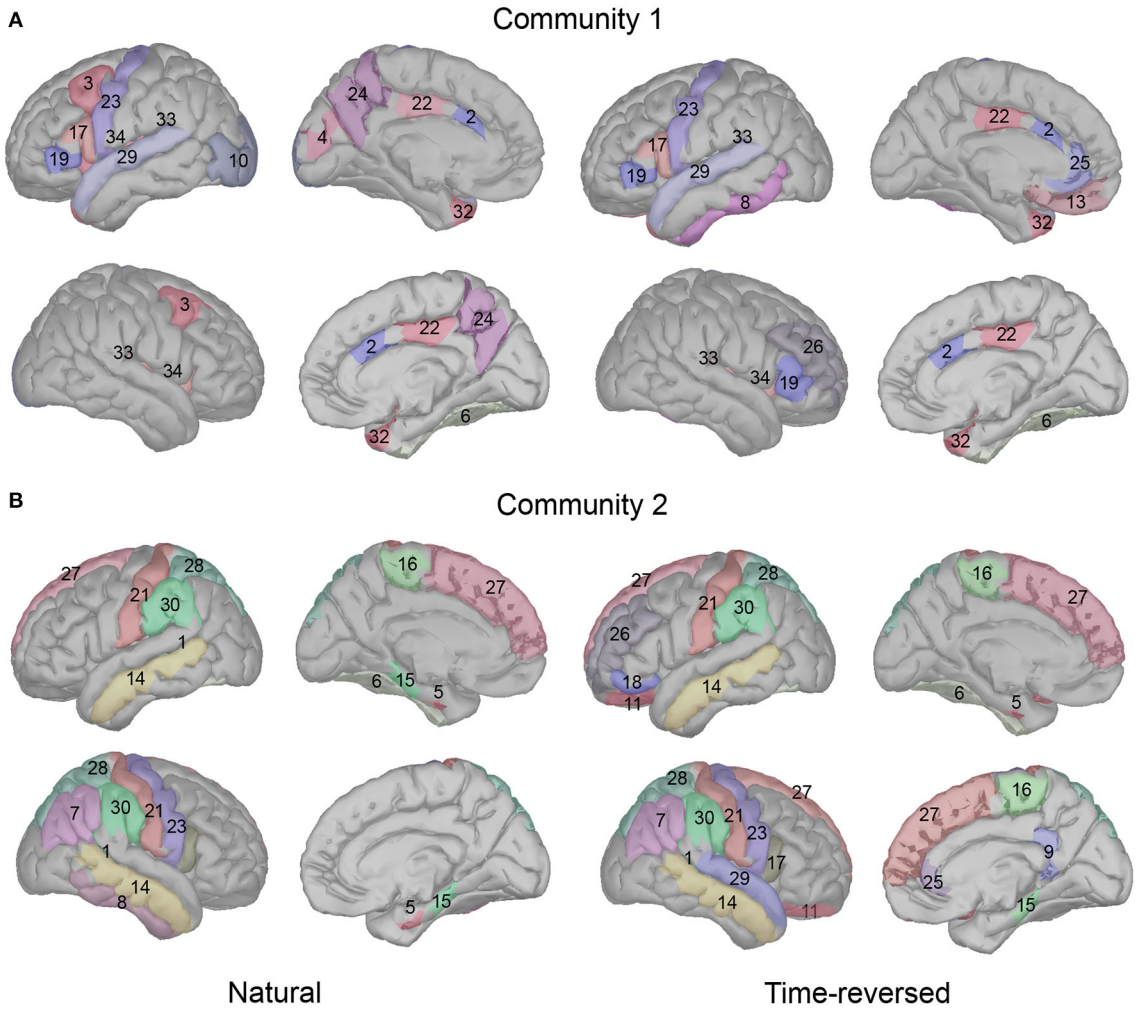

FIGURE 11 | Cortical regions in Desikan-Killiany cortical atlas for community 1 **(A)** and community 2 **(B)**.

areas such as Wernicke's and Broca's (Lerner et al., 2011; Simony et al., 2016).

On the basis of our results, we see a high consistency with previous speech processing research. We illustrate our community results for natural speech in **Figure 11**, which references the DKA surface. One can see that the brain areas that are activated in community 1 include the primary auditory cortex (transverse temporal), inferior frontal gyrus (IFG, which includes pars opercularis and pars triangularis), and insula cortex, which are related to phrase structure building and lexical selection, the middle frontal gyrus and precentral gyrus, which are related to phonological decoding, the temporal pole, which is related to semantic processing, the superior temporal gyrus, which is related to phonological encoding, word recognition, and syntactic processing, and the cingulate cortex, which is related to sentence comprehension (see meta-analysis in Walenski et al., 2019). In addition to these areas, the occipital cortex is also activated. Although the task in our experiment is speech perception, some studies have reported the importance of natural speech perception in the visual cortex (Micheli et al., 2020; Brandman et al., 2021).

In addition, long-time story comprehension also involves various cognitive processes such as memory, attention, and information integration. In our case, the precuneus, posterior-cingulate cortex (PCC), prefrontal cortex, and temporal pole are the main brain areas of the default mode network (DMN), which accumulates and integrates information over hundreds of seconds with our intrinsic information of memories during story perception (Simony et al., 2016; Yeshurun et al., 2021). And the activation in the fusiform gyrus, insula cortex, superior temporal gyrus, temporal pole, and lateral occipital cortex may be related to auditory attention process, because the similar regions are reported in a previous auditory attention task (Alho et al., 2015).

In community 2, the brain areas that were activated included the middle temporal gyrus and fusiform cortex, which are related to lexical-semantic processing, the superior frontal gyrus (supplementary motor area), which is related to effortful comprehension and phonological decoding, and

the supramarginal gyrus and parahippocampal gyrus, which are related to sentence-level processing. The parietal cortex is considered to be involved in both auditory and visual sentence comprehension. Additionally, the banks of the superior temporal sulcus and postcentral gyrus are important in complex sentence-level processing (Vigneau et al., 2006; Matchin and Hickok, 2016). Some research has reported that the paracentral lobule is activated more in 3-year-old than toddlers in the speech perception task, and it is pointed out that this area is important for language acquisition (Redcay et al., 2008). In addition, the entorhinal cortex is considered to be a high-level brain area for speech perception, and it has a direct connection with the auditory cortex; however, deafness may alter the brain's connectivity between the auditory cortex and entorhinal cortex (Kral et al., 2016).

We compared the difference in the community results between the natural and time-reversed speech. The activation for natural speech in community 1 was mainly the brain areas for the auditory attention (Salmi et al., 2009; Alho et al., 2015). However, in the time-reversed case, subjects hardly paid attention to the unintelligible speech the whole time without any positive top-down feedback. Therefore, the attention related regions were not activated in the time-reversed condition. Because of top-down processing, subjects were easily able to predict the following words or contexts for the natural speech, they used fewer brain resources for processing the upcoming speech stream, and only few auditory brain areas were activated for the natural speech condition in community 2. On the contrary, subjects used more brain resources for processing the time-reversed speech, especially in the frontal brain areas in the left hemisphere and even more brain areas in the right hemisphere.

6. CONCLUSION

In this paper, we proposed a functional hyper-alignment method with the additive average method to reduce the noises caused by individual physiological activities and/or electrode settings for subjects. Instead of using raw EEG signals, we reconstructed brain sources for estimating the temporal response functions in order to be able to study the brain networks underlying natural speech comprehension. The preliminary brain network was the pairwise connection of the brain areas, where links were defined by the correlation coefficient of the TRFs between paired areas. A multi-scale community detection was applied to the preliminary brain networks obtained from a natural speech comprehension task and time-reversed speech processing task to explore functional brain network communities. The results showed two clearly distinguishable functional network communities for semantically driven speech processing and for non-semantically driven (time-reversed speech) audio processing. The functional brain network can be explained on the basis of the achievements of past research. It was also verified that the multi-scale community detection method is suitable for source reconstruction-based brain network studies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Ethics Committee of Tianjin University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DZ performed the experiments, analyzed the data, and prepared the manuscript. GZ inspired the research idea and revised the manuscript. JD interpreted the results of the experiments and revised the manuscript. MU provided the code for extracting the speech feature based on cochlear filterbank and revised the manuscript. XL provided the idea and programmed the code for community detection. All authors approved the final version of the manuscript for submission.

FUNDING

This research was supported by the National Natural Science Foundation of China (No.61876126), a Grant-in-Aid for Scientific Fund for the Promotion of Joint International Research [Fostering Joint International Research (B); 20KK0233], and in part by JSPS KAKENHI Grant (20K11883).

REFERENCES

Alho, K., Salmi, J., Koistinen, S., Salonen, O., and Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. *Brain Res.* 1626, 136–145. doi: 10.1016/j.brainres.2014.12.050

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 493–498. doi: 10.1038/s41586-019-1119-1

Anwar, A. R., Hashmy, M. Y., Imran, B., Riaz, M. H., Mehdi, S. M. M., Muthalib, M., et al. (2016). "Complex network analysis of resting-state fMRI of the brain," in *The 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Orlando, FL: IEEE), 3598–3601. doi: 10.1109/EMBC.2016.7591506

Asadzadeh, S., Rezaii, T. Y., Beheshti, S., Delpak, A., and Meshgini, S. (2020). A systematic review of EEG source localization techniques and their applications on diagnosis of brain abnormalities. *J. Neurosci. Methods* 339:108740. doi: 10.1016/j.jneumeth.2020.108740

Beres, A. M. (2017). Time is of the essence: a review of electroencephalography (EEG) and event-related brain potentials (ERPS) in language research. *Appl. Psychophysiol. Biofeedb.* 42, 247–255. doi: 10.1007/s10484-017-9371-3

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2016). Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Tran. Neural Syst. Rehabil. Eng.* 25, 402–412. doi: 10.1109/TNSRE.2016.2571900

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Bolton, T. A. W., Jochaut, D., Giraud, A., and Ville, D. V. D. (2018). Brain dynamics in asd during movie-watching show idiosyncratic functional integration and segregation. *Hum. Brain Mapp*. 39, 2391–2404. doi: 10.1002/hbm.24009

Brandman, T., Malach, R., and Simony, E. (2021). The surprising role of the default mode network in naturalistic perception. *Commun. Biol*. 4, 1–9. doi: 10.1038/s42003-020-01602-z

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Lang. Linguist. Compass* 10, 299–313. doi: 10.1111/lnc3.12198

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018a). Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol*. 28, 3976–3983. doi: 10.1016/j.cub.2018.10.042

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018b). Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. *NeuroImage* 172, 162–174. doi: 10.1016/j.neuroimage.2018.01.042

Broderick, M. P., Anderson, A. J., Liberto, G. M. D., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol*. 28, 803–809. doi: 10.1016/j.cub.2018.01.080

Chow, W.-Y., and Phillips, C. (2013). No semantic illusions in the "semantic p600" phenomenon: ERP evidence from mandarin Chinese. *Brain Res*. 1506, 76–93. doi: 10.1016/j.brainres.2013.02.016

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. Cambridge, MA; London: MIT Press. doi: 10.7551/mitpress/9609.001.0001

Corey, D. M., Dunlap, W. P., and Burke, M. J. (1998). Averaging correlations: expected values and bias in combined Pearson RS and fisher's z transformations. *J. Gen. Psychol*. 125, 245–261. doi: 10.1080/00221309809595548

Crosse, M. J., Liberto, G. M. D., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a Matlab toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci*. 10:604. doi: 10.3389/fnhum.2016.00604

Das, N., Vanthornhout, J., Francart, T., and Bertrand, A. (2020). Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research. *Neuroimage* 204:116211. doi: 10.1016/j.neuroimage.2019.116211

de Cheveigné, A., Liberto, G. M. D., Arzounian, D., Wong, D. D. E., Hjortkjær, J., Fuglsang, S., et al. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage* 186, 728–740. doi: 10.1016/j.neuroimage.2018.11.026

Delorme, A., and Makeig, S. (2004). EEGlab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

der Maaten, L. V., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di Zhou, G. Z., Dang, J., Wu, S., and Zhang, Z. (2020). "Neural entrainment to natural speech envelope based on subject aligned EEG signals," in *INTERSPEECH 2020* (Shanghai). doi: 10.21437/Interspeech.2020-1558

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

Etard, O., and Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *J. Neurosci*. 39:5750. doi: 10.1523/JNEUROSCI.1828-18.2019

Forbes, C. E., Amey, R., Magerman, A. B., Duran, K., and Liu, M. (2018). Stereotype-based stressors facilitate emotional memory neural network connectivity and encoding of negative information to degrade math self-perceptions among women. *Soc. Cogn. Affect. Neurosci*. 13, 719–740. doi: 10.1093/scan/nsy043

Friston, K. J., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Hum. Brain Mapp*. 1, 153–171. doi: 10.1002/hbm.460010207

Gaskell, M. G., and Mirkovic, J. (2016). *Speech Perception and Spoken Word Recognition*. London; New York, NY: Routledge. doi: 10.4324/9781315772110

Gramfort, A., Papadopoulo, T., Olivi, E., and Clerc, M. (2010). OpenmEEG: opensource software for quasistatic bioelectromagnetics. *Biomed. Eng. Online* 9, 1–20. doi: 10.1186/1475-925X-9-45

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., et al. (2008). Review on solving the inverse problem in EEG source analysis. *J. Neuroeng. Rehabil*. 5, 1–33. doi: 10.1186/1743-0003-5-25

Hamilton, L. S., Edwards, E., and Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol*. 28, 1860–1871. doi: 10.1016/j.cub.2018.04.033

Handy, T. C. (2005). *Event-Related Potentials: A Methods Handbook*. Cambridge, MA; London: MIT Press.

He, Y., Lim, S., Fortunato, S., Sporns, O., Zhang, L., Qiu, J., et al. (2018). Reconfiguration of cortical networks in MDD uncovered by multiscale community detection with fMRI. *Cereb. Cortex* 28, 1383–1395. doi: 10.1093/cercor/bhx335

Howard, M. F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J. Neurophysiol*. 104, 2500–2511. doi: 10.1152/jn.00251.2010

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458. doi: 10.1038/nature17637

Janssen, N., Meij, M. v. d., López-Pérez, P. J., and Barber, H. A. (2020). Exploring the temporal dynamics of speech production with EEG and group ICA. *Sci. Rep*. 10, 1–14. doi: 10.1038/s41598-020-60301-1

Jin, D., Li, R., and Xu, J. (2019). Multiscale community detection in functional brain networks constructed using dynamic time warping. *IEEE Trans. Neural Syst. Rehabil. Eng*. 28, 52–61. doi: 10.1109/TNSRE.2019.2948055

Jrad, N., Kachenoura, A., Merlet, I., Bartolomei, F., Nica, A., Biraben, A., et al. (2016). Automatic detection and classification of high-frequency oscillations in depth-EEG signals. *IEEE Trans. Biomed. Eng*. 64, 2230–2240. doi: 10.1109/TBME.2016.2633391

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends Cogn. Sci*. 9, 512–518. doi: 10.1016/j.tics.2005.09.009

Khambhati, A. N., Medaglia, J. D., Karuza, E. A., Thompson-Schill, S. L., and Bassett, D. S. (2018). Subgraphs of functional brain networks identify dynamical constraints of cognitive control. *PLoS Comput. Biol*. 14:e1006234. doi: 10.1371/journal.pcbi.1006234

Klamer, S., Elshahabi, A., Lerche, H., Braun, C., Erb, M., Scheffler, K., and Focke, N. K. (2015). Differences between MEG and high-density EEG source localizations using a distributed source model in comparison to fMRI. *Brain Topogr*. 28, 87–94. doi: 10.1007/s10548-014-0405-3

Kral, A., Kronenberger, W. G., Pisoni, D. B., and O'Donoghue, G. M. (2016). Neurocognitive factors in sensory restoration of early deafness: a connectome model. *Lancet Neurol*. 15, 610–621. doi: 10.1016/S1474-4422(16)00034-X

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol*. 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci*. 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Liberto, G. M. D., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol*. 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Liu, K., Yu, Z. L., Wu, W., Gu, Z., Zhang, J., Cen, L., et al. (2019). Bayesian electromagnetic spatio-temporal imaging of extended sources based on matrix factorization. *IEEE Trans. Biomed. Eng*. 66, 2457–2469. doi: 10.1109/TBME.2018.2890291

Liu, M., Kuo, C.-C., and Chiu, A. W. L. (2011). "Statistical threshold for nonlinear granger causality in motor intention analysis," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Boston, MA), 5036–5039.

Mannepalli, T., and Routray, A. (2018). Certainty-based reduced sparse solution for dense array EEG source localization. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 172–178. doi: 10.1109/TNSRE.2018.2889719

Matchin, W., and Hickok, G. (2016). "Syntactic perturbation" during production activates the right IFG, but not Broca's area or the ATL. *Front. Psychol.* 7:241. doi: 10.3389/fpsyg.2016.00241

Micheli, C., Schepers, I. M., Ozker, M., Yoshor, D., Beauchamp, M. S., and Rieger, J. W. (2020). Electrocorticography reveals continuous auditory and visual speech tracking in temporal and occipital cortex. *Eur. J. Neurosci.* 51, 1364–1376. doi: 10.1111/ejn.13992

Millman, R. E., Johnson, S. R., and Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J. Cogn. Neurosci.* 27, 533–545. doi: 10.1162/jocn_a_00719

Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69:026113. doi: 10.1103/PhysRevE.69.026113

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Palmer, J. A., Kreutz-Delgado, K., and Makeig, S. (2012). *Amica: An Adaptive Mixture of Independent Component Analyzers With Shared Components*. Technical Report, Swartz Center for Computatonal Neursoscience; University of California San Diego.

Pan, X., Zou, J., Jin, P., and Ding, N. (2019). The neural encoding of continuous speech-recent advances in EEG and MEG studies. *Acta Physiol. Sin.* 71, 935–945. doi: 10.13294/j.aps.2019.0060

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr. Biol.* 25, 1649–1653. doi: 10.1016/j.cub.2015.04.049

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacol.* 24, 5–12. doi: 10.1002/med.10000

Peng, Z., Dang, J., Unoki, M., and Akagi, M. (2021). Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. *Neural Netw.* 140, 261–273. doi: 10.1016/j.neunet.2021.03.027

Peng, Z., Zhu, Z., Unoki, M., Dang, J., and Akagi, M. (2018). "Auditory-inspired end-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on spectral-temporal representation," in *2018 IEEE International Conference on Multimedia and Expo (ICME)* (San Diego, CA), 1–6. doi: 10.1109/ICME.2018.8486564

Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.* 72, 184–187. doi: 10.1016/0013-4694(89)90180-6

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Pirondini, E., Babadi, B., Obregon-Henao, G., Lamus, C., Malik, W. Q., Hämäläinen, M. S., et al. (2017). Computationally efficient algorithms for sparse, dynamic solutions to the EEG source localization problem. *IEEE Trans. Biomed. Eng.* 65, 1359–1372. doi: 10.1109/TBME.2017.2739824

Plechawska-Wojcik, M., Kaczorowska, M., and Zapala, D. (2018). "The artifact subspace reconstruction (ASR) for EEG signal correction. A comparative study," in *International Conference on Information Systems Architecture and Technology* (Cham: Springer), 125–135. doi: 10.1007/978-3-319-99996-8_12

Price, C. J. (2012). A review and synthesis of the first 20 years of pet and fMRI studies of heard speech, spoken language and reading. *NeuroImage* 62, 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Redcay, E., Haist, F., and Courchesne, E. (2008). Functional neuroimaging of speech perception during a pivotal period in language acquisition. *Dev. Sci.* 11, 237–252. doi: 10.1111/j.1467-7687.2008.00674.x

Salmi, J., Rinne, T., Koistinen, S., Salonen, O., and Alho, K. (2009). Brain networks of bottom-up triggered and top-down controlled shifting of auditory attention. *Brain Res.* 1286, 155–164. doi: 10.1016/j.brainres.2009.06.083

Schrader, S., Westhoff, A., Piastra, M. C., Miinalainen, T., Pursiainen, S., Vorwerk, J., et al. (2021). Duneuro-a software toolbox for forward modeling in bioelectromagnetism. *PLoS ONE* 16:e0252431. doi: 10.1371/journal.pone.0252431

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., et al. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* 7, 1–13. doi: 10.1038/ncomms12141

Smith, S. M. (2012). The future of fMRI connectivity. *Neuroimage* 62, 1257–1266. doi: 10.1016/j.neuroimage.2012.01.022

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., et al. (2013). Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* 17, 666–682. doi: 10.1016/j.tics.2013.09.016

Stropahl, M., Bauer, A.-K. R., Debener, S., and Bleichner, M. G. (2018). Source-modeling auditory processes of EEG data using EEGlab and brainstorm. *Front. Neurosci.* 12:309. doi: 10.3389/fnins.2018.00309

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., and Leahy, R. M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* 2011:879716. doi: 10.1155/2011/879716

Tanaka, H. (2020). Group task-related component analysis (GTRCA): a multivariate method for inter-trial reproducibility and inter-subject similarity maximization for EEG data analysis. *Sci. Rep.* 10, 1–17. doi: 10.1038/s41598-019-56962-2

Tang, C., Hamilton, L., and Chang, E. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science* 357, 797–801. doi: 10.1126/science.aam8577

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *J. Assoc. Res. Otolaryngol.* 19, 181–191. doi: 10.1007/s10162-018-0654-z

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Duffau, H., Crivello, F., Houde, O., et al. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage* 30, 1414–1432. doi: 10.1016/j.neuroimage.2005.11.002

Vorwerk, J., Engwer, C., Pursiainen, S., and Wolters, C. H. (2016). A mixed finite element method to solve the EEG forward problem. *IEEE Trans. Med. Imaging* 36, 930–941. doi: 10.1109/TMI.2016.2624634

Walenski, M., Europa, E., Caplan, D., and Thompson, C. K. (2019). Neural networks for sentence comprehension and production: an ALE-based meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* 40, 2275–2304. doi: 10.1002/hbm.24523

Weissbart, H., Kandylaki, K. D., and Reichenbach, T., (2020). Cortical tracking of surprisal during continuous speech comprehension. *J. Cogn. Neurosci.* 32, 155–166.

World Medical Association (2014). World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *J. Am. Coll. Dent.* 310, 2191–2194. doi: 10.1001/jama.2013.281053

Yeshurun, Y., Nguyen, M., and Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nat. Rev. Neurosci.* 22, 181–192. doi: 10.1038/s41583-020-00420-w

Yu, Q., Du, Y., Chen, J., Sui, J., Adalē, T., Pearlson, G. D., et al. (2018). Application of graph theory to assess static and dynamic brain connectivity: approaches for building brain graphs. *Proc. IEEE* 106, 886–906. doi: 10.1109/JPROC.2018.2825200

Zhang, G., and Liu, X. (2021). Investigation of functional brain network reconfiguration during exposure to naturalistic stimuli using graph-theoretical analysis. *J. Neural Eng.* 18:056027. doi: 10.1088/1741-2552/ac20e7

Zhang, G., Si, Y., and Dang, J. (2019). Revealing the dynamic brain connectivity from perception of speech sound to semantic processing by EEG. *Neuroscience* 415, 70–76. doi: 10.1016/j.neuroscience.2019.07.023

Zhang, Y., Ding, Y., Huang, J., Zhou, W., Ling, Z., Hong, B., et al. (2021). Hierarchical cortical networks of "voice patches" for processing voices in human brain. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2113887118. doi: 10.1073/pnas.2113887118

Zoefel, B., and VanRullen, R. (2016). EEG oscillations entrain their phase to high-level features of speech sound. *Neuroimage* 124, 16–23. doi: 10.1016/j.neuroimage.2015.08.054

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zhou, Zhang, Dang, Unoki and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.