NeuroImage 146 (2017) 333–340

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Time-domain analysis of neural tracking of hierarchical linguistic structures

Wen Zhang[a,1], Nai Ding[a,b,c,*,1]

[a] College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China
[b] Interdisciplinary Center for Social Sciences, Zhejiang University, Hangzhou, China
[c] Neuro and Behavior EconLab, Zhejiang University of Finance and Economics, Hangzhou, China

ARTICLE INFO

Keywords:
MEG
Canonical component analysis
Inter-subject correlation
Neural entrainment
Speech

ABSTRACT

When listening to continuous speech, cortical activity measured by MEG concurrently follows the rhythms of multiple linguistic structures, e.g., syllables, phrases, and sentences. This phenomenon was previously characterized in the frequency domain. Here, we investigate the waveform of neural activity tracking linguistic structures in the time domain and quantify the coherence of neural response phases over subjects listening to the same stimulus. These analyses are achieved by decomposing the multi-channel MEG recordings into components that maximize the correlation between neural response waveforms across listeners. Each MEG component can be viewed as the recording from a virtual sensor that is spatially tuned to a cortical network showing coherent neural activity over subjects. This analysis reveals information not available from previous frequency-domain analysis of MEG global field power: First, concurrent neural tracking of hierarchical linguistic structures emerges at the beginning of the stimulus, rather than slowly building up after repetitions of the same sentential structure. Second, neural tracking of the sentential structure is reflected by slow neural fluctuations, rather than, e.g., a series of short-lasting transient responses at sentential boundaries. Lastly and most importantly, it shows that the MEG responses tracking the syllabic rhythm are spatially separable from the MEG responses tracking the sentential and phrasal rhythms.

Introduction

In the human language, smaller linguistic units such as syllables and words can be recursively combined into larger linguistic structures such as phrases and sentences. How linguistic units of different sizes are represented in the brain is a fundamental question in cognitive neuroscience (Buiatti et al., 2009; Everaert et al., 2015; Garrett et al., 1966; Pallier et al., 2011; Peña and Melloni, 2012; Townsend and Bever, 2001). It is shown that when listening to continuous speech, cortical activity recorded by magnetoencephalography (MEG) and electroencephalography (EEG) follows the rhythms of acoustic features of speech that are related to syllabic and phonemic level processing (Di Liberto et al., 2015; Ding and Simon, 2012a, b; Gross et al., 2013; Kayser et al., 2015; Kerlin et al., 2010; Luo and Poeppel, 2007). Recently, it is further shown that cortical activity can concurrently follow higher level linguistic structures such as phrases and sentences using speech materials illustrated in Fig. 1a (Ding et al., 2016).

Neural tracking of phrasal and sentential structures, however, was only characterized in the frequency domain by analyzing the global field power, leaving several questions unanswered. First, when cortical activity become entrained/synchronized to the phrasal and sentential rhythms, how long does it take for entrained activity to reach a steady state? The buildup timecourse of entrained activity depends on at least 2 factors. First, it depends on the dynamic properties of the underlying neural sources. For example, the auditory Steady State Response (aSSR) to a sound repeating at 40 Hz builds up in ~300 ms, after ~12 cycles of the stimulus (Ross et al., 2002). Second, it depends on how much time the brain needs to parse the temporal structure of the input. For example, the aSSR may take up to 4 s to build up when the periodic stimulus is interfered by competing sounds (Elhilali et al., 2009). Here, we employ language materials that are easy to parse to focus on the buildup process constrained by the dynamical properties of the underlying neural network.

Second, frequency-domain analysis does not directly illustrate the response waveform. Therefore, it is not intuitive whether the sentential-rate response continuously changes over the timecourse of a sentence (Fig. 1b) or whether it only shows an abrupt change at sentential boundaries (Fig. 1c). If the neural response is a continuously changing slow oscillation, it could be interpreted as an integrator that accumulate information over the timecourse of a sentence (Pallier

[*] Corresponding author at: College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Zhejiang 310027, China.
E-mail address: ding_nai@zju.edu.cn (N. Ding).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.neuroimage.2016.11.016
Received 24 July 2016; Accepted 3 November 2016
Available online 15 November 2016

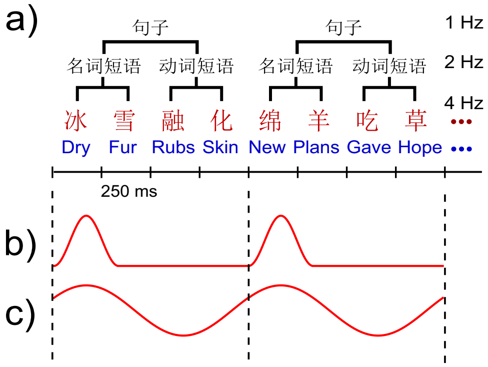
a)
句子
句子
1 Hz
名词短语
动词短语
名词短语
动词短语
2 Hz
4 Hz
冰 雪 融 化 绵 羊 吃 草 •••
Dry Fur Rubs Skin New Plans Gave Hope •••
250 ms
b)
c)

Fig. 1. Linguistic structure of the stimulus (a) and possible neural responses (b-c). (a) A sequence of Chinese syllables are presented isochronously at 4 Hz. Examples in English are also shown for illustrative purposes. All syllables are independently synthesized by a computer. Two syllables construct a phrase and two phrases construct a sentence. Therefore, the syllables, phrases, and sentences are presented at 4 Hz, 2 Hz, and 1 Hz respectively. This figure is adapted from Ding et al. (2016). (b-c) Two hypotheses about how cortical activity follows the sentential rhythm, whether it continuously changes over time (b) or occurs only briefly at sentential boundaries (c).

et al., 2011). In contrast, if the neural response is a short-lasting transient response at structural boundaries, it is more appropriately interpreted as a change/boundary detector.

Third, previous frequency-domain analysis was based on the global field power of MEG, leaving it unclear whether the neural tracking of different linguistic levels can be spatially separated using MEG? To answer the above mentioned 3 questions, we apply a time-domain analysis of MEG responses. Furthermore, although previous studies assume that the sentential- and phrasal-rate neural responses are entrained, i.e., phase locked, to the stimulus but the degree of phase locking is not quantified. Here, we quantify the consistency of the neural response phases over subjects by calculating the inter-subject phase coherence (Fisher, 1993).

A time-domain analysis of MEG responses, however, is difficult for several reasons. First, each neural generator, i.e. a current source, produces a source/sink pattern in the MEG signal. The MEG signals from the source and the sink have opposite polarities and the spatial locations of the source and sink patterns are not aligned across subjects due to the anatomical differences and the subjects' head position inside the MEG machine. Second, there are usually multiple neural generators contributing to the neural tracking of a continuous stimulus and these neural generators could have different response phases due to their positions in the neural processing hierarchy or their neurodynamical properties. As a result, if the macroscopic MEG responses are dominated by different neural generators in different subjects due to anatomical differences, they will show phase differences across subjects. Lastly, the MEG signal is a mixture of the responses from multiple neural generators and component analysis methods, such as the principal component analysis (PCA), independent component analysis (ICA), and denoising source separation (DSS), are often employed to separate different neural sources. The polarity of the response waveforms extracted by the component analysis, however, is usually arbitrary, which further increases the difficulty for grand averaging the response waveforms across subjects.

Although it is difficult to align the response phase across subjects in MEG, recent studies have shown that the neural response phase is relevant to perception (Henry and Obleser, 2012; Lakatos et al., 2008; Schroeder and Lakatos, 2009) and shows consistency across subjects during the processing of continuous natural stimuli (Dmochowski et al., 2014; Hasson et al., 2012; Hasson et al., 2004; Honey et al., 2012; Lankinen et al., 2014). For normal listeners, during speech comprehension, it is reasonable to assume and empirical studies support that common neurophysiological processes underlie the processing of the same unambiguous sentence.

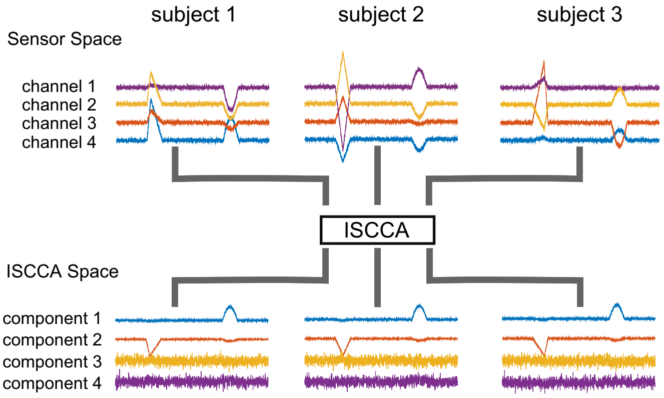

Fig. 2. Illustration of the basic function of the Inter-Subject Coherent Component Analysis (ISCCA). The ISCCA decomposes the multi-channel recordings from individual subjects into components, by maximizing the correlation between response waveforms across subjects. In this illustration, a 4-channel recording was simulated for 3 subjects. Each recording is a mixture of an early response, a late response, and white noise. The late response has the same waveform across subjects and is captured by the first ISCCA component. The waveform of the early response slightly varies across subjects and is captured by the second ISCCA component. The noise signal is captured by the 3rd and the 4th ISCCA components. The early response is simulated by a sawtooth signal. Its phase is identical within each subject across channels but varies across subjects. Both the early and the late responses have random polarity and amplitude in each channel. In this illustration, since the data have only 4 channels, the DSS dimension reduction step is omitted and the mCCA is applied to the 4-channel data directly.

To optimally extract neurophsyiological processes that are common across subjects, we propose an analysis method called the Inter-Subject Coherent Component Analysis (ISCCA). The ISCCA decomposes the multi-channel MEG recordings of each subject into components and maximizes the inter-subject correlation of each component. Each ISCCA component is extracted by a spatial filter and can be viewed as the recording from a virtual sensor spatially tuned to a cortical network that shows coherent neural activity over subjects. The ISCCA spatial filters are subject-specific and normalize individual differences in response topography. Since the ISCCA components are maximally correlated over subjects, they can be directly averaged for group level analysis. As an illustration, Fig. 2 shows that responses may have very different amplitude and polarity in different channels in the sensor space. However, when the responses are projected to the ISCCA space, responses that show coherence over subjects are attributed to the same ISCCA component with the same polarity, which facilitates group-level analysis of the response waveform.

In the following, we apply the ISCCA to extract MEG response components that have a coherent response waveform over subjects and analyze the time course of neural tracking of linguistic structures based on the grand averaged response waveform.

Materials and methods

Experimental procedures

Sixteen healthy young adults participated in the experiments and the data analyzed here were previously reported by Ding et al. (2016). In the experiment, the subjects listened to an isochronous sequence of syllables. These syllables were ordered so that neighboring 4 syllables constructed a sentence (Fig. 1a). Each sentence was composed of a noun phrase (2 syllables) followed by a verb phrase (2 syllables). The syllables were presented at a constant rate of 4 Hz and no pause was inserted between phrases or sentences. Therefore, the sentences were presented at 1 Hz and the phrases were presented at 2 Hz.

In each trial, 40 syllables were played and 28 trials were collected. To ensure attention, the subjects were instructed to detect semantically abnormal sentences such as “green frogs drove cars” by a button press at the end of the trial. Eight trials contained abnormal sentences and were removed from the analysis. Therefore, 20 trials were analyzed for each subject.

MEG recordings

The neuromagnetic signals were recorded using a 157-channel whole-head system (KIT). The signal was sampled at 1 kHz, with a 200-Hz low-pass filter and a 60-Hz notch filter applied online and a 0.5-Hz high-pass filter applied offline. Environmental magnetic fields were removed based on 3 reference sensors using the TSPCA method (de Cheveigné and Simon, 2007). The MEG signals were decimated to 200 Hz, with an FIR anti-aliasing filter (100 Hz cut-off frequency).

Spatial filtering

Spatial filters can be used to extract the neuromagnetic signals generated from specific neural sources (either point sources or networks) in the brain. Denote the MEG recording from a subject as $\boldsymbol{X}$ (channel×time) and a linear spatial filter as $\boldsymbol{a}$ (channel ×1). The output of the spatial filter is $\boldsymbol{y}=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}$, which is a weighted sum of the signals recorded from different channels. By adjusting the weight of each sensor, i.e., $\boldsymbol{a}$, a spatial filter can selectively enhance neural activity from some brain areas while suppressing neural activity from other areas. The output of a spatial filter can be viewed as the signal recorded by a virtual sensor that selectively records neural activity from a specific point source or neural network.

ISCCA

The ISCCA applies spatial filters to extract neural activity that is coherent over subjects. It designs spatial filters in two steps. The first step applies a dimension reduction matrix $\boldsymbol{D}$ to convert the multi-channel MEG recordings into a few components. This step is applied independently for different subjects, using the denoising source separation (DSS) (de Cheveigné and Parra, 2014; de Cheveigné and Simon, 2008). In the second step, an additional spatial filter $\boldsymbol{w}$ is applied to the DSS components. This spatial filter extracts the neural response component consistent over subjects by considering the data from all subjects simultaneously. The second step relies on the multi-set canonical component analysis (mCCA) (Kettenring, 1971). Using the two steps, the ISCCA factorizes a spatial filter $\boldsymbol{a}$ as $\boldsymbol{a}=\boldsymbol{D}\boldsymbol{w}$, where $\boldsymbol{D}$ is the DSS matrix (channel×component) and $\boldsymbol{w}$ is derived by mCCA (component ×1).

The DSS decomposes the MEG recordings to extract neural responses that are consistent over trials. It maximizes the ratio between the power of the averaged response and the power of single trial MEG recordings, and therefore it extracts neural responses that are consistent over trials. The DSS has been successfully applied to extract the neural activity entrained to speech (Ding et al., 2016; Ding and Simon, 2013). The DSS transforms the sensor-space MEG recordings into components and the transformation matrix $\boldsymbol{D}_0$ is derived as follows. If the covariance matrix for single trial MEG recordings is $\boldsymbol{C}_0$ (channel×channel) and the covariance matrix for the MEG response averaged over trials is $\boldsymbol{C}_1$ (channel×channel), the DSS spatial filters are the generalized eigenvectors of $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$. In other words, $\boldsymbol{C}_1\boldsymbol{D}_0=\Lambda\boldsymbol{C}_0\boldsymbol{D}_0$, where Λ is a diagonal matrix. The inverse of $\boldsymbol{D}$, denoted as $\boldsymbol{F}$, is the DSS mixing matrix (shown in Fig. 6c) and carries topographical information about each DSS component (de Cheveigné and Simon, 2008). If L components are kept for futher analysis, the dimension reduction matrix $\boldsymbol{D}$ is constructed by the first L columns of $\boldsymbol{D}_0$. The corresponding mixing matrix $\boldsymbol{F}_0$ is constructed by the first L rows of $\boldsymbol{F}$. In most of the analyses done in this article (except for Fig. 7), six DSS components were used for further analysis, in consistent with previous studies (Ding et al., 2016).

In the second step, another set of linear spatial filters is used to extract neural activity that shows strong coherence over subjects. Denote the DSS components from subject i as $\boldsymbol{X_i}$ (channel × time), where i = 1, 2,..., m. The response component extracted by a linear spatial filter $\boldsymbol{w_i}$ is then $\boldsymbol{y}_i=\boldsymbol{w}_i^T\boldsymbol{X}_i$. The correlation between the response components from two different subjects, i.e., $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, is $\boldsymbol{c_{ij}}=\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_j/\sqrt{\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j}$. The set of linear spatial filters that maximizes the total pairwise inter-subject correlation, i.e., $\sum_{i=1,j\neq i}^{m} c_{ij}$, can be obtained by solving the following optimization problem:

$$\begin{aligned} \max_{w_1,\dots w_N}\ \lambda = \sum_{i=1}^{m}\sum_{j\neq i}^{m}\boldsymbol{y}_i\boldsymbol{y}_j^T = \sum_{i=1}^{m}\sum_{j\neq i}^{m}\boldsymbol{w}_i^T\boldsymbol{X}_i\boldsymbol{X}_j^T\boldsymbol{w}_j \\ s.\,t. \quad \boldsymbol{y}_i\boldsymbol{y}_i^T = \boldsymbol{w}_i^T\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{w}_i = 1 \end{aligned} \qquad (1)$$

This optimization problem can only be solved numerically. However, if the constraint is relaxed to $\sum_{i=1}^{m}\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{w}_i=m$, the solution becomes the following generalized eigenvalue problem, which is known as the MAXVAR solution for the mCCA problem (Kettenring, 1971; Vía et al., 2007; Zhang et al., 2014).

$$(\boldsymbol{R}-\boldsymbol{S})\boldsymbol{w} = \lambda\boldsymbol{S}\boldsymbol{w} \qquad (2)$$

where

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{X}_1^{\mathrm{T}} & \cdots & \boldsymbol{X}_1\boldsymbol{X}_m^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_m\boldsymbol{X}_1^{\mathrm{T}} & \cdots & \boldsymbol{X}_m\boldsymbol{X}_m^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{X}_1^{\mathrm{T}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{X}_m\boldsymbol{X}_m^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} \boldsymbol{w}_{11} \\ \vdots \\ \boldsymbol{w}_m \end{bmatrix}$$

Multiple solutions exist for Eq. (2) and each solution $\mathbf{w}$ leads to an ISCCA spatial filter $\boldsymbol{a} = \boldsymbol{D}\boldsymbol{w}$. The ISCCA component for subject i is $\mathbf{z}_i=\boldsymbol{a}_i^T\boldsymbol{X}_i$. Since z_i is unitless, we normalize its mean to 0 and its variance to 1 for each subject. Essentially, the ISCCA extracts common response waveforms from multiple subjects and aligns the polarity of the waveform across subjects. The mCCA components are ranked based on the eigenvalue λ, which roughly corresponds to inter-subject correlation in response waveform. If there are only two subjects, i.e., m=2, the solution in (2) reduces to the CCA solution. The CCA has been widely applied to process MEG and EEG signals. A number of studies applied CCA to analyze the relationship between MEG responses and the sensory stimuli (Bin et al., 2009; Koskinen et al., 2012; Lin et al., 2006; Zhang et al., 2014) while this study and some other previous studies employ mCCA to analyze the relationship between the MEG responses of different subjects (Dmochowski et al., 2014; Lankinen et al., 2014).

If each solution to Eq. (2) is denoted as $\boldsymbol{w}^{(i)}$, then $\boldsymbol{W}$=[$\boldsymbol{w}^{(1)}$, $\boldsymbol{w}^{(2)}$, ..., $\boldsymbol{w}^{(K)}$] construct the demixing matrix of the mCCA, where K is the number of mCCA component being analyzed and in this case K equals the number of DSS component being analyzed. The ISCCA mixing matrix is calculated as $\boldsymbol{M}=\boldsymbol{W}^{-1}\boldsymbol{F}_0$. The ISCCA demixing matrix carries information about the topographic distribution of each ISCCA component (de Cheveigné and Simon, 2008) and is shown in Fig. 3c. The Matlab code is available at https://github.com/zjuzw/iscca for the mCCA and http://audition.ens.fr/adc/NoiseTools/ for the DSS.

Cross validation

To avoid artifacts due to overfitting, the ISCCA results were based on 10-fold cross-validation. The 20 trials were divided into 10 disjoint sets of equal size. In the cross validation procedure, one dataset was assigned as the test set while the other 9 datasets were assigned as the training set. This process was repeated 10 times with different dataset being the test set. For each training/test set, all trials were averaged.

The ISCCA spatial filter was derived based on the training set and applied to the test set. Such cross-validation procedure was repeated 10 times. When analyzing the response waveform, the results from the 10 cross-validation trials were averaged after aligning their polarity based on the following procedure: A PCA was used to extract the principal component across 10 folds as a benchmark. Each of the 10 cross-validation results were then correlated with the benchmark. If the correlation was negative, the polarity of the response waveform was

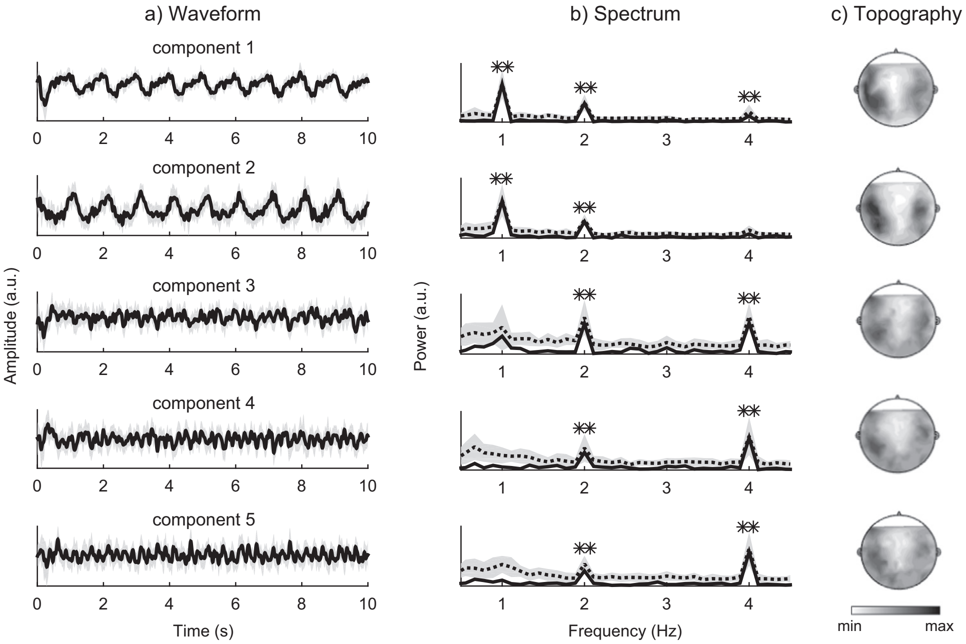

Fig. 3. The waveform and the spectrum of the first 5 ISCCA components. The component index is labeled on the right. (a) Response waveform. The first 2 ISCCA components show oscillations slower than the oscillations in the last 3 components. The grand average over subjects is shown by the fold black curve while the shaded gray area shows 1 SD over subjects on each side. The speech stimulus starts at time 0. See Fig. 5c for the response averaged over each 1-s duration sentence. (b) Response spectrum. The solid black curve shows the spectrum of the response waveform grand averaged over subjects, i.e., the black curve in panel (a). The shaded gray region is 1 SD over subjects on each side. The grand average of the response spectrum of individual subjects is shown by the dotted black curve. For the solid black curve, frequency bins showing significantly stronger power than neighboring bins are shown by stars ($P < 0.005$, F-test, $F(2,12) > 20$, FDR corrected). (c) Response topography averaged over subjects. Darker color indicates stronger power. The response generally shows a bilateral pattern but the detailed differences are difficult to quantify in the sensor space.

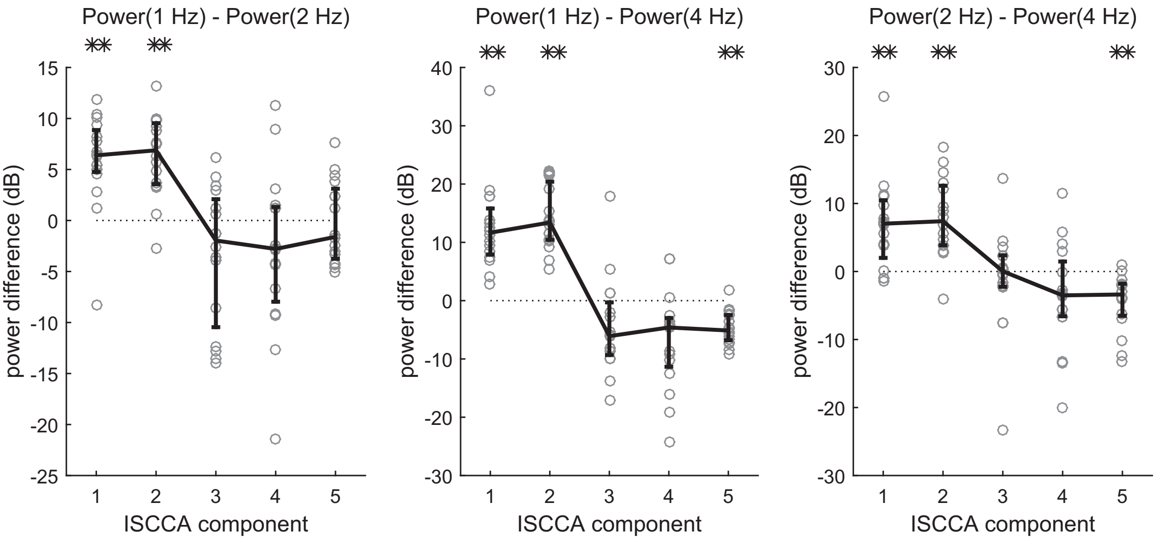

Fig. 4. Spatial separation of neural tracking of different linguistic structures. The power difference between frequencies significantly varies among different ISCCA components (1-way repeated measures ANOVA, $P < 0.005$). Data from individual subjects are shown by gray circles while the error bar is centered at the mean and covers the range between the 25th percentile and the 75th percentile. The power difference that is significantly larger or smaller than 0 is marked by stars (**$P < 0.005$, paired *t*-test, $t(15) > 4$, FDR corrected).

flipped by multiplying −1. The phase alignment was necessary since the cross-validation procedure was repeated 10 times independently and the 10 results could have arbitrary polarity.

Spectral analysis

In the spectral analysis, the response during the first sentence in each trial was removed, to avoid the transient responses related to the trial onset. The responses during the presentation of the remaining 9 sentences (9 s in duration) were transformed into the frequency domain using the Discrete Fourier Transform (DFT) without any window. The frequency resolution of the DFT was therefore 1/9 Hz.

When analyzing the phase coherence across subjects, the neural response phase was extracted based on the DFT coefficient at each frequency. The phase coherence across the 16 subjects was calculated as follows (Fisher, 1993),

$$R(f)^2 = \left(\frac{1}{16}\sum_{i=1}^{16}\cos(\theta_i(f))\right)^2 + \left(\frac{1}{16}\sum_{i=1}^{16}\sin(\theta_i(f))\right)^2 \quad (3)$$

where $\theta_i(f)$ denotes the response phase of subject i, at frequency f. $R(f)^2$ bounds between 0 and 1, with a larger value corresponding to higher phase coherence. During cross-validation, the inter-subject phase coherence was calculated for each test set and then averaged over 10

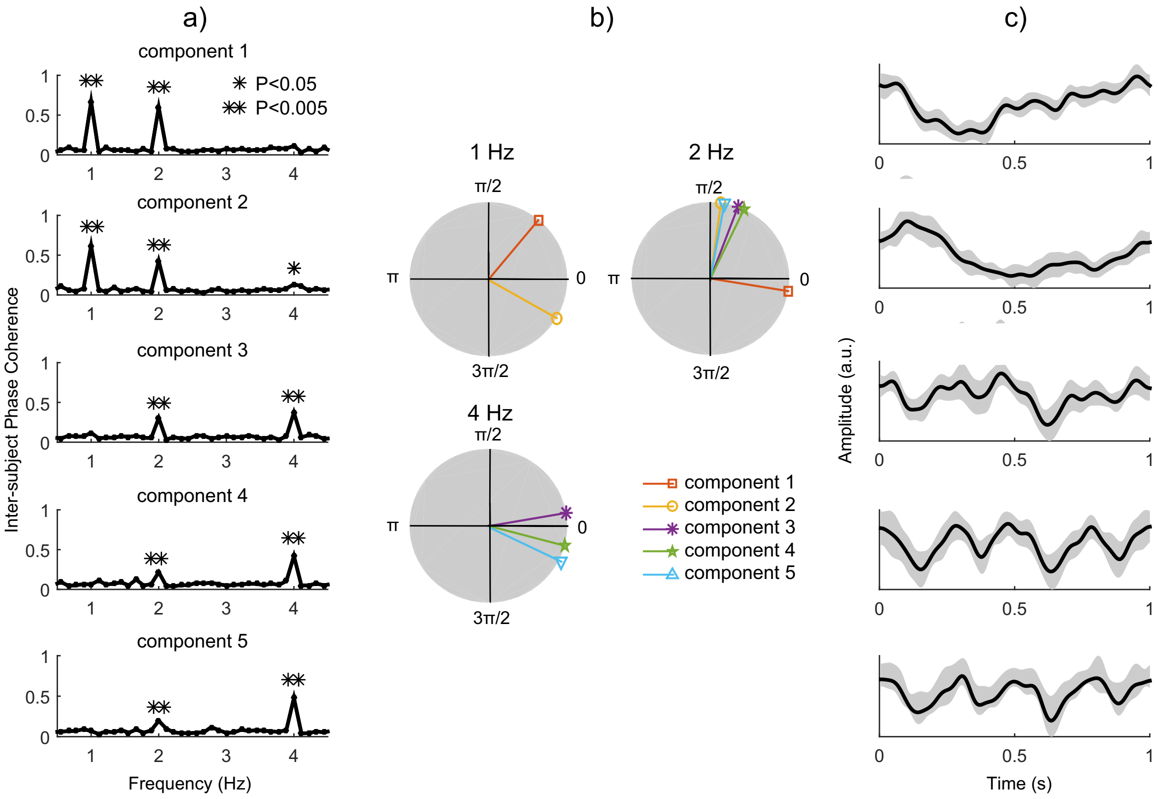

Fig. 5. Phase analysis for the first 5 ISCCA components. (a) Inter-subject phase coherence at different frequencies. The MEG response shows significant phase coherence over subjects at 1 Hz for the first two ISCCA components, at 2 Hz for all the components, and at 4 Hz for the last 3 components (statistical test described in Methods, FDR-corrected). (b) The response phases at 1, 2, and 4 Hz, shown on a unit circle. The response at each frequency shows consistent variance across ISCCA components, indicating multiple neural sources. Data are shown only when significant phase coherence (P < 0.005) is observed. (c) The response averaged over all sentences and subjects. The first sentence within each trial is not included in this analysis to avoid the transient response to sound onset. The first two components are both dominated by the 1 Hz response but have different response phases. Similarly, the last two components are dominated by the 4 Hz response but show distinct response phases.

test sets.

In this study, $R(f)^2$ calculated based on Eq. (3) is called the inter-subject phase coherence while the pairwise correlation between the response waveforms of different subjects is referred to as the inter-subject correlation.

Statistical analysis and significance tests

An F-test was used to test if the power at a target frequency was significantly higher than the power at neighboring frequencies. The power in neighboring frequencies was averaged over 6 bins (3 bins on each side). Neighboring frequency bins are separated by 1/9 Hz. Under the null hypothesis, i.e., the power at the target frequency being the same as the mean power at neighboring frequencies, the ratio between the power at the target frequency and the power averaged over neighboring frequency bins is subject to a F(2,12) distribution (John and Picton, 2000). All frequency bins ranged from 0.5 Hz to 4.5 Hz was submitted to the same test, followed by a FDR correction for multiple comparisons.

To assess the significance of inter-subject phase coherence, we estimated the false alarm rate numerically. The null distribution is simulated by 10,000 samples of phase angles uniformly distributed between $(-\pi, \pi)$ for each of the 16 subjects and each of the 10 test set for the 10-fold cross-validation procedure. The phase coherence across subjects was computed and averaged over 10 cross-validation test sets, following the same procedure of how the inter-subject phase coherence is calculated for real MEG data. The significance level of an inter-subject phase coherence value S is determined by the ratio of samples exceeding S in the 10,000 simulated samples. This significance test is a generalization of the Rayleigh Test for circular distributions (Fisher, 1993) and tests whether the response phase distribution deviates from a uniform distribution. The same test was applied for all frequency bins between 0.5 Hz to 4.5 Hz with a FDR correction.

Results

Response waveform

The data consists of 16 subjects listening to continuous speech stimuli with hierarchically embedded linguistic structures (Fig. 1). In the speech materials, the syllables, phrases, and sentences are presented at a constant rate of 4 Hz, 2 Hz, and 1 Hz, respectively. The multi-channel MEG responses are decomposed into components using the ISCCA based on a 10-fold cross-validation procedure. The waveforms of the first 5 ISCCA components are shown in Fig. 3a. The ISCCA components show clear periodicity and high consistency across subjects. Except for a transient response to the sound onset, the responses appear to be a steady state oscillation throughout the 10 s stimulus presentation.

To quantify if the neural response consistently builds up or adapts, we calculated the mean response power during the presence of each sentence (from the 2nd sentence to the 10th sentence, excluding the 1st sentence because of the transient response to sound onset). The response power does not significantly vary between the 2nd sentence and the 10th sentence ($P > 0.1$, 1-way ANOVA, no correction for multiple comparisons). In other words, the response to each sentence (excluding the first one) has roughly the same power.

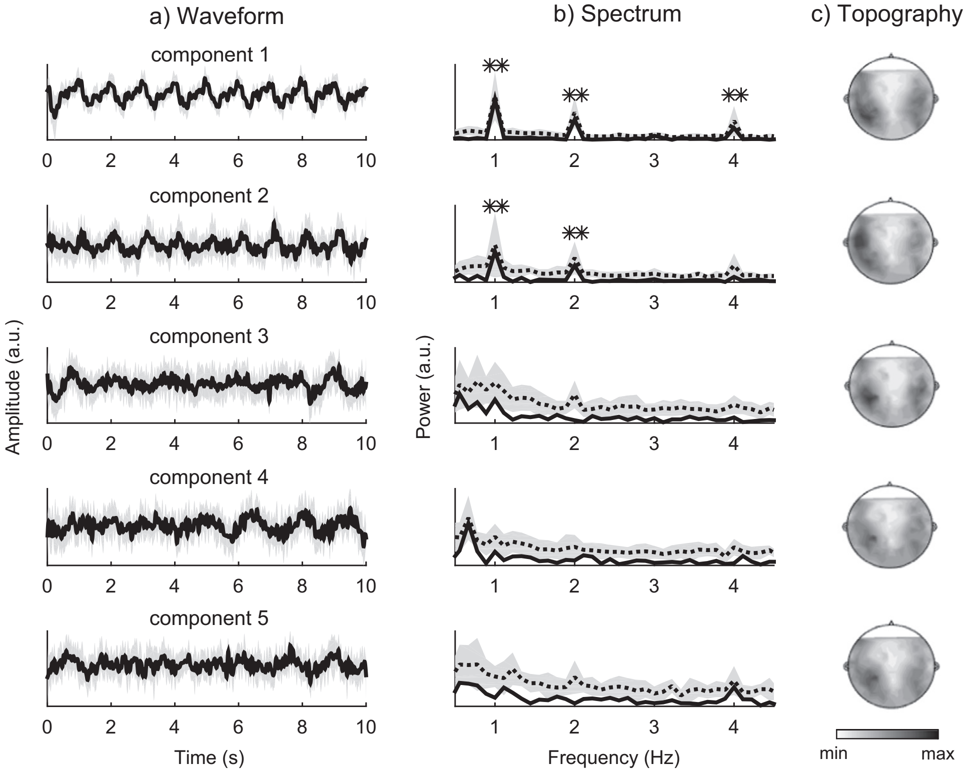

Fig. 6. The waveform and the spectrum of the first 5 DSS components. The response waveform and response spectrum are shown in panel (a) and (b) respectively. The black curves show the grand average while the gray region shows 1 SD over subjects on each side. More descriptions can be found in Fig. 3. Only the first two DSS components show significant responses at 1 and 2 Hz and the first component also shows a significant response at 4 Hz. For the solid black curve, frequency bins showing significantly stronger power than neighboring bins are shown by a star (P < 0.005, F-test, F(2,12) > 20, FDR corrected). (c) Response topography averaged over subjects. Darker color indicates stronger power. The response generally shows a bilateral pattern but the detailed differences are difficult to quantify in the sensor space.

Spatially separated neural tracking of different linguistic levels

As is evident from Fig. 3a, different ISCCA components fluctuate on different time scales. Frequency domain analysis further reveals that the first two ISCCA components are dominated by the 1-Hz and 2-Hz responses while the following three ISCCA components are dominated by 2-Hz and 4-Hz responses (Fig. 3b). Since each ISCCA component can be viewed as neural activity recorded from a specific neural network, these results demonstrate that some neural networks selectively follow the sentential and phrasal rhythms while other neural networks selectively follow the phrasal and syllabic rhythms. The topographical distribution of each ISCCA component is shown in Fig. 3c. Responses are observed in both hemispheres. Future studies are needed to quantify the spatial differences between components in the neural souce space, by integrating MEG with structual MRI.

The spatial separation of neural responses at different rates is further quantified in Fig. 4 for individual subjects. If two responses are spatially separable, we can selectively enhance one response over another by changing the spatial tuning of the spatial filter. In other words, for spatially separable neural responses, the power ratio between them could depend on the spatial filter. In contrast, if the responses are not spatially separable, a spatial filter can enhance or suppress both but will not affect their power ratio. Fig. 4 shows the power difference between the neural responses at two different frequencies. In this analysis, the response spectrum is normalized by the total power between 0.5 and 4.5 Hz for each subject. The power ratio significantly differs across components for each of the panels in Fig. 4 (P < 0.001, 1-way repeated measures ANOVA). Furthermore, the 1-Hz and 2-Hz responses are stronger than the 4-Hz response for the first 2 ISCCA component but weaker than the 4-Hz response for the 5th ISCCA component (P < 0.0005, paired *t*-test, t(15) > 4, FDR corrected), showing that spatial filters can selectively enhance the 1- and 2- Hz responses while suppressing the 4-Hz response.

Therefore, for individual subjects, the neural tracking of sentential, phrasal, and syllabic responses are spatially separable. Additionally, it is worth mentioning that the ISCCA is a method to optimize inter-subject correlation rather than separating neural tracking of different linguistic levels. By analyzing the response correlation across subjects, however, the neural responses to different linguistic levels are naturally separated into different spatial components.

Response phase

The response phase is further analyzed in Fig. 5. In the spectral domain, it is clear that high inter-subject phase coherence appears at 1, 2, or 4 Hz (Fig. 5a). At and only at these 3 frequencies, the response phase statistically significantly deviates from a uniform distribution. The response phase at each frequency varies across ISCCA components, as is clear from both the Fourier response phase (Fig. 5b) and the waveform averaged over sentences (Fig. 5c).

If different response components show different response phases at the same frequency for individual subjects, it indicates that the MEG responses at that frequency are generated from multiple neural sources that are spatially distinguishable by MEG (Simon and Wang, 2005). Therefore, that fact that different ISCCA components show different response phases at each of the frequency of interest, i.e., 1, 2, and 4 Hz, indicate multiple neural generators for the response tracking each linguistic level, consistent with previous ECoG results (Ding et al., 2016).

Validating the ISCCA procedure

In the ISCCA, the multi-channel MEG data from individual subjects are first processed by the DSS, for dimension reduction and denoising.

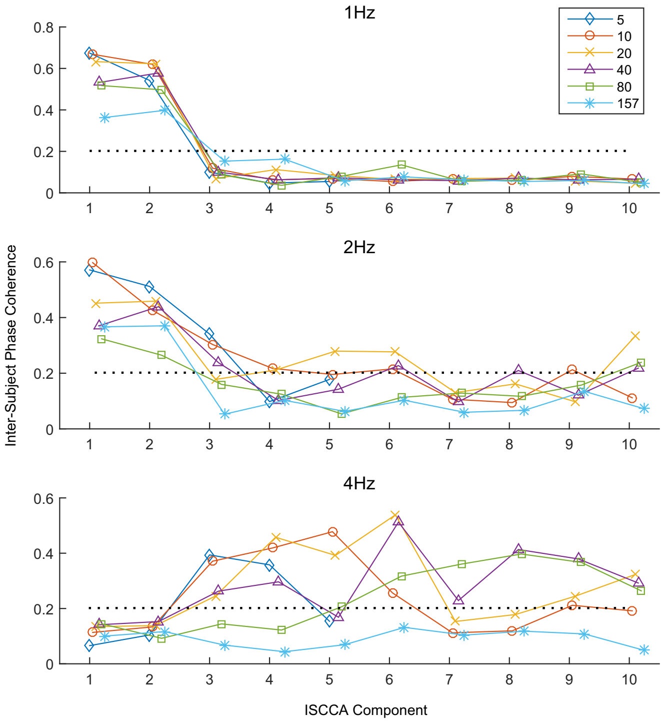

Fig. 7. The inter-subject phase coherence at 1, 2, and 4 Hz as a function of the number of DSS components kept for the mCCA analysis. The dotted line shows the significance level (P=0.05^{10} no correction for multiple comparision; P=0.05, Bonferonni correction for each DSS dimension). Since the neural responses are only significant at 1, 2, and 4 Hz, this figure summarizes all useful information of the inter-subject phase coherence spectrum for the first 10 ISCCA components. In general, the ISCCA based on 5 or 10 DSS components achieved the best performance. The number of DSS components kept for the mCCA does not alter the overall trend that the first two ISCCA components show a strong 1-Hz response while the following ISCCA components capture mainly the 4-Hz response.

Then, the data from all subjects are jointly optimized using the mCCA to extract components coherent across subjects. Here, we first investigate the necessity of mCCA and then investigate the effect of DSS dimension reduction.

Fig. 6 illustrates the waveform and spectrum of the first 5 DSS components. Since the waveform of each DSS component has arbitrary polarity, we align the response polarity over subjects by adjusting the polarity of each subject to maximize the mean pairwise inter-subject correlation. Even with polarity adjustment, as shown in Fig. 6, the response waveform still shows considerable variance across subjects, especially for the last 3 components analyzed here. More importantly, the DSS does not separate the neural tracking of different linguistic levels to different components. Therefore, the mCCA is necessary to reliably extract response waveforms and to separate neural tracking of different linguistic levels.

We then test if the DSS dimension reduction procedure is necessary. Fig. 7 shows how the inter-subject correlation at 1, 2, and 4 Hz for the first 10 ISCCA components and how the inter-subject correlation changes as a function of the number of DSS components retained for mCCA. It also shows the results when the sensor-space data is used for the mCCA directly. Clearly, the DSS dimension reduction increases the inter-subject correlation, and in general the ISCCA derived from the top 5 or 10 DSS components achieved the best performance. Therefore, both DSS and mCCA are necessary component for the ISCCA.

Discussion

This study demonstrates that neural tracking of different linguistic levels can be spatially dissociated using MEG and that the neural tracking of larger linguistic structures such as phrases and sentences emerge early, not requiring many repetitions of sentences. A new method, i.e. the ISCCA, is proposed to extract the timecourse of neural tracking of continuous stimulus and to decompose neural responses into components tracking different stimulus structures.

Spatially separable neural tracking of different linguistic levels

Each ISCCA component is obtained by spatially filtering the multichannel MEG recordings. Figs. 3b and 4 illustrate that, in an unsupervised manner, the ISCCA designs spatial filters can selectively enhance the neural tracking of one linguistic level (syllables, phrases, or sentences) while suppressing or not affecting the other responses. These results demonstrate that the neural networks tracking sentential and syllabic structures are spatially separable in the MEG sensor space. Furthermore, the fact that different ISCCA components show different response phases at the same frequency indicates that the neural response tracking each linguistic level is generated from multiple neural sources.

Consistent with these MEG results, previous ECoG results also show that neural tracking of different linguistic levels is generated from board but distinguishable neural networks (Ding et al., 2016). The ECoG results, however, show that the neural tracking of different linguistic levels is only separable on a fine spatial scale but in macroscopic cortical areas. Therefore, it is surprising that these responses are separable by MEG, in an unsupervised manner. On the other hand, separating the neural tracking of different linguistic levels in MEG does not necessarily require resolving the fine spatial separation of these networks. Instead it only requires distinguishing the “center of gravity” of these networks.

Waveform of low-frequency neural entrainment to speech

Neural activity can synchronize to the 1 Hz sentential rhythm in several ways (Zhou et al., 2016). One is that a transient, i.e., short-lasting, evoked response occurs at the boundaries between sentences (Fig. 1c). Another possibility is that the response continuously changes over the timecourse of a sentence (Fig. 1b). A third possibility is that the power of high-frequency neural activity, e.g., in the high-gamma or alpha bands, fluctuates at the sentential rate. Previous theoretical analysis suggests that a spectral peak at the sentential rate in the response spectrum indicates a roughly sinusoidal response at the sentential rate (Zhou et al., 2016). The current study, however, directly visualize the response waveform and shows that it is a slow wave fluctuating at the sentential rate. Furthermore, it is shown here that the response does not show significant build up or adaptation when the transient response to sound onset is removed. This means that, in contrast to the 40-Hz aSSR, which takes more than 10 cycles of the stimulus to stablize (Ross et al., 2002), neural entrainment to phrases/sentences stablizes from the 2nd cycle of the stimulus. Since the experiment being analyzed here employs only sentences with a simple and predictable syntactic structure, the current results only show that the neural generators of the phrasal/sentential-rate responses can quickly follow stimuli that are easy to parse. Future studies are needed to investigate whether neural parsing of more complex and less predictable syntactic structures is incremental or whether it takes time to build up (Townsend and Bever, 2001).

Inter-subject correlation of neural responses to continuous stimuli

When listening to the same speech recording or when watching the same movie, very slow fluctuations of cortical activity measured by fMRI show a high degree of correlation between subjects, in broad neural networks involved in speech/movie processing (Hasson et al., 2015; Hasson et al., 2004; Lerner et al., 2011). Recent MEG/EEG/ECoG studies have also revealed that electrophysiological activity below 10 Hz shows synchronization between subjects during speech listening and movie viewing (Chang et al., 2015; Dmochowski et al., 2014;

Honey et al., 2012; Lankinen et al., 2014). Inter-subject correlation can be observed in two types of experiments. One is that different subjects are recorded simultaneously, with possible interactions with each other (Dumas et al., 2010; Hari and Kujala, 2009). Another type of experiments, including the current study, is that different subjects are recorded in separate sessions although they are exposed to the same sensory stimulus. In these experiments, the synchronization between neural responses in different subjects arises from the neural synchronization to the common stimulus. Inter-subject correlation serves as a useful measure to detect stimulus-synchronous neural activity. Furthermore, since inter-subject correlation can be calculated based on single trials, it can be applied as long as the same stimulus is presented to each subject, not requiring repetitions of the same stimulus within a subject.

In summary, this study extracts the waveform of neural activity entrained to hierarchical linguistic structures by maximizing the inter-subject correlation. The ISCCA method separates neural tracking of different linguistic levels to different components, and provides a useful tool to extract stimulus-synchronous neural activity from single trial neural recordings.

Acknowledgments

We thank David Poeppel, Lucia Melloni, Lin Chen, Cheng Luo, Jianfeng Zhang and two anonymous reviewers for constructive comments. This work was supported by the National Natural Science Foundation of China 31500873 (ND), Fundamental Research Funds for the Central Universities (ND), Zhejiang Provincial Natural Science Foundation of China LR16C090002 (ND).

References

Bin, G., Gao, X., Yan, Z., Hong, B., Gao, S., 2009. An online multi-channel SSVEP-based brain–computer interface using a canonical correlation analysis method. J. Neural Eng. 6, 046002.

Buiatti, M., Peña, M., Dehaene-Lambertz, G., 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. NeuroImage 44, 509–551.

Chang, W.T., Jääskeläinen, I.P., Belliveau, J.W., Huang, S., Hung, A.Y., Rossi, S., Ahveninen, J., 2015. Combined MEG and EEG show reliable patterns of electromagnetic brain activity during natural viewing. NeuroImage 114, 49–56.

de Cheveigné, A., Parra, L.C., 2014. Joint decorrelation, a versatile tool for multichannel data analysis. NeuroImage 98, 487–505.

de Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift PCA. J. Neurosci. Methods 165, 297–305.

de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. J. Neurosci. Methods 171, 331–339.

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. 25, 2457–2465.

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164.

Ding, N., Simon, J.Z., 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA 109, 11854–11859.

Ding, N., Simon, J.Z., 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89.

Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33, 5728–5735.

Dmochowski, J.P., Bezdek, M.A., Abelson, B.P., Johnson, J.S., Schumacher, E.H., Parra, L.C., 2014. Audience preferences are predicted by temporal reliability of neural processing. Nature Communications.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., Garnero, L., 2010. Inter-brain synchronization during social interaction. PLoS One 5, e12166.

Elhilali, M., Xiang, J., Shamma, S.A., Simon, J.Z., 2009. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. PLoS Biol., 7.

Everaert, M.B., Huybregts, M.A., Chomsky, N., Berwick, R.C., Bolhuis, J.J., 2015. Structures, not strings: linguistics as part of the cognitive sciences. Trends Cogn. Sci. 19, 729–743.

Fisher, N.I., 1993. Statistical Analysis of Circular Data. Cambridge University Press, New York.

Garrett, M., Bever, T., Fodor, J., 1966. The active use of grammar in speech perception. Percept. Psychophys. 1, 30–32.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11, e1001752.

Hari, R., Kujala, M.V., 2009. Brain basis of human social interaction: from concepts to brain imaging. Physiol. Rev. 89, 453–479.

Hasson, U., Chen, J., Honey, C.J., 2015. Hierarchical process memory: memory as an integral component of information processing. Trends Cogn. Sci. 19, 304–313.

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 Cogn. Sci. 16, 114–121.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640.

Henry, M.J., Obleser, J., 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc. Natl. Acad. Sci. Vol. 109, pp. 20095–20100.

Honey, C.J., Thesen, T., Donner, T.H., Silbert, L.J., Carlson, C.E., Devinsky, O., Doyle, W.K., Rubin, N., Heeger, D.J., Hasson, U., 2012. Slow cortical dynamics and the accumulation of information over long timescales. Neuron 76, 423–434.

John, M.S., Picton, T.W., 2000. MASTER: a Windows program for recording multiple auditory steady-state responses. Comput. Methods Prog. Biomed. 61, 125–150.

Kayser, S.J., Ince, R.A., Gross, J., Kayser, C., 2015. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. J. Neurosci. 35, 14691–14701.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. 30, 620–628.

Kettenring, J.R., 1971. Canonical analysis of several sets of variables. Biometrika 58, 433–451.

Koskinen, M., Viinikanoja, J., Kurimo, M., Klami, A., Kaski, S., Hari, R., 2012. Identifying fragments of natural speech from the listener's MEG signals. Hum. Brain Mapp.. http://dx.doi.org/10.1002/hbm.22004.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113.

Lankinen, K., Saari, J., Hari, R., Koskinen, M., 2014. Intersubject consistency of cortical MEG signals during movie viewing. NeuroImage 92, 217–224.

Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915.

Lin, Z., Zhang, C., Wu, W., Gao, X., 2006. Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. IEEE Trans. Biomed. Eng. 53, 2610–2614.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.

Pallier, C., Devauchelle, A.-D., Dehaene, S., 2011. Cortical representation of the constituent structure of sentences. Proc. Natl. Acad. Sci. Vol. 108, pp. 2522–2527.

Peña, M., Melloni, L., 2012. Brain oscillations during spoken sentence processing. J. Cogn. Neurosci. 24, 1149–1164.

Ross, B., Picton, T.W., Pantev, C., 2002. Temporal integration in the human auditory cortex as represented by the development of the steady-state magnetic field. Hear. Res. 165, 68–84.

Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32, 9–18.

Simon, J.Z., Wang, Y., 2005. Fully complex magnetoencephalography. J. Neurosci. Methods 149, 64–73.

Townsend, D.J., Bever, T.G., 2001. Sentence Comprehension: The Integration of Habits and Rules. MIT Press, Cambridge, MA.

Vía, J., Santamaría, I., Pérez, J., 2007. A learning algorithm for adaptive canonical correlation analysis of several data sets. Neural Netw. 20, 139–152.

Zhang, Y., Zhou, G., Jin, J., Wang, X., Cichocki, A., 2014. Frequency recognition in SSVEP-based BCI using multiset canonical correlation analysis. Int. J. neural Syst. 24, 1450013.

Zhou, H., Melloni, L., Poeppel, D., Ding, N., 2016. Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. Front. Hum. Neurosci., 10.