Speech intelligibility changes the temporal evolution of neural speech tracking

Chen, Ya Ping; Schmidt, Fabian; Keitel, Anne; Rösch, Sebastian; Hauswald, Anne; Weisz, Nathan

Published in:
NeuroImage

DOI:
10.1016/j.neuroimage.2023.119894

Publication date:
2023

NeuroImage 268 (2023) 119894

Contents lists available at ScienceDirect

NeuroImage

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

Speech intelligibility changes the temporal evolution of neural speech tracking

Ya-Ping Chen[a,b,*], Fabian Schmidt[a,b], Anne Keitel[c], Sebastian Rösch[d], Anne Hauswald[a,b,1], Nathan Weisz[a,b,e,1]

[a] Centre for Cognitive Neuroscience, University of Salzburg, 5020 Salzburg, Austria
[b] Department of Psychology, University of Salzburg, 5020 Salzburg, Austria
[c] Psychology, School of Social Sciences, University of Dundee, DD1 4HN Dundee, UK
[d] Department of Otorhinolaryngology, Paracelsus Medical University, 5020 Salzburg, Austria
[e] Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, 5020 Salzburg, Austria

ARTICLE INFO

Keywords:
Vocoded speech
Temporal response function
Coherence
FOOOF
MEG

ABSTRACT

Listening to speech with poor signal quality is challenging. Neural speech tracking of degraded speech has been used to advance the understanding of how brain processes and speech intelligibility are interrelated. However, the temporal dynamics of neural speech tracking and their relation to speech intelligibility are not clear. In the present MEG study, we exploited temporal response functions (TRFs), which has been used to describe the time course of speech tracking on a gradient from intelligible to unintelligible degraded speech. In addition, we used inter-related facets of neural speech tracking (e.g., speech envelope reconstruction, speech-brain coherence, and components of broadband coherence spectra) to endorse our findings in TRFs. Our TRF analysis yielded marked temporally differential effects of vocoding: ~50–110 ms ($M50_{TRF}$), ~175–230 ms ($M200_{TRF}$), and ~315–380 ms ($M350_{TRF}$). Reduction of intelligibility went along with large increases of early peak responses $M50_{TRF}$, but strongly reduced responses in $M200_{TRF}$. In the late responses $M350_{TRF}$, the maximum response occurred for degraded speech that was still comprehensible then declined with reduced intelligibility. Furthermore, we related the TRF components to our other neural “tracking“ measures and found that $M50_{TRF}$ and $M200_{TRF}$ play a differential role in the shifting center frequency of the broadband coherence spectra. Overall, our study highlights the importance of time-resolved computation of neural speech tracking and decomposition of coherence spectra and provides a better understanding of degraded speech processing.

1. Introduction

Listeners process speech under a variety of adverse conditions in daily life (Mattys et al., 2012), which could be external (e.g., “cocktail party“) or internal (e.g., hearing damage). However, how the neural dynamics of speech processing in such challenging conditions evolve remains unclear. Several studies have shown that temporal and spectral modulations on speech signals are relevant for speech comprehension (Elliott and Theunissen, 2009; Obleser and Weisz, 2012; Santoro et al., 2014). Here, we focused on the effects of spectral modulation on speech comprehension. A popular approach to systematically manipulate speech intelligibility–and possibly mimicking the experience of cochlear implant (CI) users–in normal hearing individuals is by using noise-vocoded speech (Friesen et al., 2001; Rosen et al., 2013). This approach filters the speech signal into a given number of channels or bands (typically between 1 and 16 channels), resulting in reduced spectral information while temporal modulation of the speech is intact. The spectral information is reduced in a manner that the envelopes of the extracted frequency bands are convolved with noise filtered in the same passband as the source envelope. This manipulation allows to parametrically modulate speech intelligibility (Shannon et al., 1995) and to relate features of the signal to neural activity.

The high temporal resolution of electroencephalography (EEG) and magnetoencephalography (MEG) can be exploited to quantify how temporal fluctuations of speech and neural signals align together, a process often described as “neural speech tracking”. Most studies (e.g., Ding and Simon, 2012; Fiedler et al., 2019; Zion Golumbic et al., 2013), including those investigating the effects of vocoding, have focused on the speech envelope. Since the envelope of e.g. a target speaker can be easily obtained, neural tracking using M/EEG is an attractive option

* Corresponding author: Hellbrunnerstr. 34, A-5020 Salzburg, Austria.
E-mail address: ya-ping.chen@plus.ac.at (Y.-P. Chen).
[1] Anne Hauswald and Nathan Weisz are joint senior authors on this work.

https://doi.org/10.1016/j.neuroimage.2023.119894.
Received 7 July 2022; Received in revised form 13 December 2022; Accepted 20 January 2023
Available online 21 January 2023.

to study brain activity also during complex listening situations such as background noise or multi-speaker scenarios. Speech-brain coherence, a frequency-domain measure which reflects the degree to which the phase relationships of speech envelopes and brain signals are consistent across measurements, is likely the most established measure of neural tracking with low-frequency band (1–7 Hz) entailing crucial speech information as phrases, syllables and phonemes (Hauswald et al., 2020; Peelle et al., 2013; Schmidt et al., 2021). However, system identification approaches, such as temporal response functions (TRFs; Crosse et al., 2016; Ding et al., 2014; Kraus et al., 2021) and stimulus reconstruction (Nogueira et al., 2019; Verschueren et al., 2019), have been gaining a lot of popularity. TRFs especially yield interesting temporal information on the relationship between stimulus and neural activity, which goes beyond the "static" coherence measure. Interestingly, these measures are rarely reported together.

When relating neural activities to behavioral performance, the main focus usually is how speech intelligibility affects such neural speech tracking processes. As speech tracking can be assessed by different measures (see above: coherence, TRF, etc.), this is also the case for speech intelligibility: Some studies ask comprehension questions (Fuglsang et al., 2020), others for subjective difficulty reports (Obleser & Weisz, 2012), and again others for word or sentence recognition or reproduction (Hauswald et al., 2020; Verschueren et al., 2019). Speech intelligibility is commonly modulated not only via the study design, e.g. modulations of the stimulus characteristics [vocoding: Baltzell et al. (2017); Hauswald et al. (2020); speech in noise: Ding & Simon (2013)], multi-talker (Ding and Simon, 2012; Petersen et al., 2017; Tune et al., 2021) or a combination of those (Kraus et al., 2021; Rimmele et al., 2015), but can also be inferred via the study populations, e.g. normal hearing vs hearing-impaired (Decruy et al., 2020) and native speaker vs non-native speakers (Song and Iverson, 2018; Zou et al., 2019).

Regarding the influences of speech intelligibility on speech tracking, previous studies have shown mixed findings: (1) Some studies have shown positive correlation of speech tracking with speech comprehension: Investigating speech tracking in CI users, higher accuracy of speech envelope reconstruction was observed with better speech comprehension (Verschueren et al., 2019). Similarly, Peelle et al. (2013) and Schmidt et al. (2021) showed that the coherence decreased linearly as the speech intelligibility declined in normal hearing population. (2) Some have shown negative correlation of speech tracking with speech comprehension: When studying individuals with hearing loss, Decruy et al. (2020) observed higher envelope reconstruction accuracy in participants with mild to severe sensorineural hearing loss than age-matched normal hearing adults. Similarly, studying TRFs during vocoded speech has shown larger $M50_{TRF}$ amplitudes for less intelligible speech (Ding et al., 2014). (3) The others have shown non-linear relation between speech intelligibility or null results: Investigating coherence, Hauswald et al. (2020) showed an inverted-U-shaped pattern, with the highest coherence in the degraded but still comprehensible speech rather than the clear speech. Computing cross-correlation between speech envelopes and neural activities, Baltzell et al. (2017) found no effect of vocoding, whereas prior knowledge affected speech tracking. Overall, current findings using different "flavors" of neural tracking do not point to a consistent effect of vocoded speech.

Here, we utilized MEG and had our participants listen to speech at different vocoding levels. We applied the TRF method to explore how 6 levels of vocoded speech (Fig. 1) modulate the neural speech tracking with fine-grained temporal resolution. As expected, participants' behavioral performance declined as speech intelligibility deteriorated. Interestingly, our TRF results showed that degraded speech modulates the neural speech tracking at three intervals in differential ways: Responses at 50–110 ms ($M50_{TRF}$) and 175–230 ms ($M200_{TRF}$) generally captured whether speech was vocoded or not. Late responses around 315–380 ms ($M350_{TRF}$) were strongest at medium vocoding levels as compared to clear and less intelligible speech. To complement our findings in TRF, we also computed speech envelope reconstruction, speech-brain coherence, and periodic components (peak center frequency, peak bandwidth, peak height) and aperiodic components (exponent, offset) of broadband coherence spectra (Schmidt et al., 2021). Relating the TRF components to those neural measures, we found that $M50_{TRF}$ and $M200_{TRF}$ were correlated with the center frequency of the broadband coherence spectra in a different direction. These results suggest that speech intelligibility related fluctuation in TRF reflect shifts in which hierarchical speech features are tracked.

2. Materials and methods

All data used in this study were also reported as Study 2 in Hauswald et al. (2020). The duration of the MEG experiment was ~75 mins (preparation included) and the behavioral experiment was ~7 mins. The behavioral experiment was also conducted in the MEG chamber without recording of MEG signals.

2.1. Participants

Twenty-four individuals [11 females, mean age = 26.4 ± 5.7 (SD), age range = 18–45 years] were recruited in the MEG experiment. They were native German speakers, reported normal hearing and no history of psychological or neurological disease, and were eligible for MEG recordings (i.e., without ferromagnetic metals in or close to their bodies). Sixteen of these individuals [7 females, mean age = 27.3 ± 6.8 (SD), age range = 18–45 years] also participated in the behavioral experiment. All the participants provided informed consent, and were compensated monetarily or with course credit. The recruitment and experiment procedure was in accordance with the Declaration of Helsinki and approved by the Ethics Committees of the Department of Psychology, University of Salzburg.

2.2. Stimuli

In both the MEG and behavioral experiments, participants were instructed to listen to audio files from audiobooks and to maintain visual fixation on a cross centered on screen for the duration of each trial (Fig. 1A). We used a within-subjects design. There were six conditions in both experiments: original, 7-, 5-, 3-, 2-, or 1-channel (ch) noise-vocoded (Fig. 1B). The speech vocoding was done using the vocoder toolbox (Gaudrain, 2016) for MATLAB. For the 7-, 5-, 3-, 2-, 1-ch vocoding levels, the waveform of each audio stimulus was passed through two Butterworth filters with a range of 200–7000 Hz and then further bandpass-filtered into the corresponding frequency analysis bands. For each band, a noise carrier was generated, and the frequency of the noise was equal to the center frequency of the analysis filter. Amplitude envelope extraction was done with half-wave rectification and low-pass filtered at 250 Hz. The amplitude-modulated noise bands from all the channels were then combined to produce the vocoded speech. The root mean square of the resulting signal was adjusted to that of the original signal.

In the MEG experiment, 24 audio files were created from recordings of a female German native speaker reading Goethe's "Das Märchen" (von Goethe, 1795). Lengths of stimuli varied between 15 s and 180 s, with two stimuli of 15 s, 30 s, 60 s, 90 s, 120 s, 150 s, and 12 of 180 s. Each stimulus ended with a two-syllable noun within the last four words in the last sentence. The assignment of stimuli to conditions in the MEG experiment was controlled in order to obtain a similar overall length of stimulus presentation (~400 s) in each condition. The order of the stimuli did not follow the order of the original story. The order of conditions was pseudorandomized according to a Latin square design. Three stimuli were presented in one block after which a short break was offered, resulting in eight blocks. Each block was followed by a self-determined break. At the end of each auditory stimulus, participants were required to choose the noun in the last sentence from two-syllable nouns on the screen. Following the response, they could self-initiate the next trial via

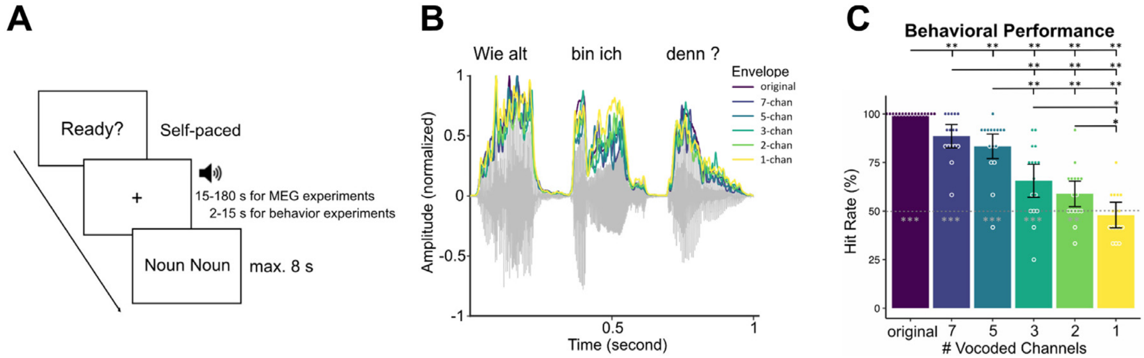

Fig. 1. (**A**) Example trial of the MEG and behavioral experiment. Participants started self-paced and listened to the audiobook with eye fixation on the cross. At the end of the trial, two nouns were presented on the screen. Participants had to choose which noun was the one they heard from the last sentence. (**B**) An exemplary original audio segment with the corresponding envelopes and with the envelopes from the other vocoded conditions. The envelope from the original audio is essentially identical to the envelopes from the vocoded audio on power level. (**C**) The hit rate declined as the speech intelligibility decreased. Bars represent 95% confidence intervals. $p_{fdr} < 0.05^{*}$, $p_{fdr} < 0.01^{**}$, $p_{fdr} < 0.001^{***}$.

a button press. The syllable rate of the 24 audio files varied between 3.93 and 4.55 Hz with a mean of 4.14 Hz, which was computed with a custom script from de Jong et al. (2021) using Praat (Boersma and Weenink, 2019).

In the behavioral experiment, 48 audio files were created from recordings of another female German native speaker reading a German version of Antoiné St. Exupery's "The little prince" (1943) written by Grete and Josef Leitgeb (1956). Each stimulus contained one sentence (length between 2 and 15 s) and ended with a two-syllable noun within the last four words. Similar to the MEG experiment, participants were asked to choose the last noun they heard between two nouns on the screen. Again, the order of conditions was pseudorandom.

Stimulus presentation was controlled using a MATLAB-based objective psychophysics toolbox (Hartmann and Weisz, 2020) built based on the Psychtoolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Auditory stimuli were presented binaurally using MEG-compatible pneumatic in-ear headphones (SOUNDPixx, VPixx technologies, Canada). The trigger-sound delay of 16 ms was measured via the Black Box Toolkit v2 and was corrected during preprocessing of MEG data. Participants' responses were acquired via a response pad (TOUCHPixx response box, VPixx technologies, Canada).

2.3. Data acquisition and analyses

2.3.1. Extraction of the acoustic speech envelope

For computation of the temporal response functions and stimulus reconstruction, we extracted the acoustic speech envelope from all auditory stimuli using the Chimera toolbox (Smith et al., 2002), with which five frequency bands in the range of 200 to 4000 Hz were constructed as equidistant on the cochlear map. Sound stimuli were band-pass filtered (forward and reverse) in 5 bands using a 4th-order Butterworth filter. For each band, envelopes were calculated as absolute values of the Hilbert transform and were averaged across bands to obtain the full-band envelope. Envelopes for all 6 conditions were processed with this procedure and used for the following TRF, stimulus reconstruction, and coherence analysis.

To determine the envelope modulation rate, custom Matlab scripts from Ding et al. (2017) were used. The audio files were chunked into 10-s duration segments, resulting in 306 segments per condition. After calculating a Fast Fourier Transformation (FFT), the global maximum value of each 306 power spectrum was taken and then averaged as the envelope modulation rate. The envelope modulation rate of the 24 audio files in the original, 7-ch, 5-ch, 3-ch, 2-ch, and 1-ch vocoded condition was 5.61 Hz, 5.53 Hz, 5.27 Hz, 5.23 Hz, 4.93 Hz, and 4.83 Hz, respectively. A main effect of vocoding was observed [$\chi^2(5) = 206.0$, $p = 1.83 \times 10^{-42}$, Kendall's $W = 0.134$, Friedman test], in line with Schmidt et al. (2021).

2.3.2. MEG acquisition and preprocessing

MEG signals were recorded at a sampling rate of 1000 Hz using a 306-channel Triux MEG system (Elekta-Neuromag Ltd., Helsinki, Finland) with 102 magnetometers and 204 planar gradiometers in a magnetically shielded room (AK3B, Vakuumschmelze, Hanau, Germany). The MEG signal was online high-pass and low-pass filtered at 0.1 Hz and 330 Hz respectively. Prior to the recording, individual head shapes were digitized for each participant including fiducials (nasion, bilateral preauricular points) and at least 300 points on the scalp using a Polhemus Fastrak system (Polhemus, Vermont, USA). A signal space separation algorithm implemented in the Maxfilter software (version 2.215) provided by the MEG manufacturer was used to remove external noise from the MEG signal (mainly 16.6 Hz from Austrian local train power and 50 Hz plus harmonics from power line) and realign data across different blocks to an individual common head position (based on the measured head position at the beginning of each block).

Data analysis was done using the Fieldtrip toolbox (Oostenveld et al., 2011) and in-house-built scripts. Firstly, a low-pass filter at 40 Hz using a finite impulse response (FIR) filter with Kaiser window was applied to continuous MEG data. Then, the data were resampled to 200 Hz to save computational power and were epoched into 2-second segments to increase the signal-to-noise ratio. Around 1% of those 2-second segments were excluded as the corresponding auditory stimuli contained silent periods of more than 1 second. With the Fieldtrip automatic artifact rejection algorithm, we rejected trials with z-value higher than 100 before conducting independent component analysis (ICA). Using ICA, we identified and removed components corresponding to blinks, eye movements, cardiac activities, and residual noise from local train power. On average 7.6 ± 2.7 (SD) components were removed. The automatic artifact rejection algorithm was applied again to reject trials with z-value higher than 50.

For the further TRF (Section 2.3.3) and envelope reconstruction analyses (Section 2.3.5), another band-pass filter between 0.5 and 8 Hz was applied to the data (FIR filter with Kaiser window). For the coherence analysis, we re-segmented the preprocessed data into 4-sec segments to increase frequency resolution.

2.3.3. Temporal response functions

The TRF analysis was done using the mTRF toolbox Version 2.0 (Crosse et al., 2016). TRFs were estimated by mapping speech features (e.g., envelopes, spectrograms) to neural responses of M/EEG. The mapping from stimulus to neural response is also known as “forward” modeling. The TRFs provide sensor-specific predictions and can indicate how a change of a stimulus feature (speech envelope) affects time-resolved neural activity at a certain location. In the context of a sensory system where the output is recorded by N recorded channels, we can assume that the instantaneous neural response $r(t,n)$, which samples at times $t = 1\ldots$ T and at channel n, can be modeled over a convolution of the stimulus property, $s(t)$, with a channel-specific TRF, $TRF(\tau,\ n)$. The TRF, $TRF(\tau,\ n)$, describes this transformation for a specified range of time lags, The response, $r(t,n)$, can be modeled as: $r(t,n) = \sum_{\tau} TRF(\tau,\ n)s(t-\tau) + \varepsilon(t,n)$. The $\varepsilon(t,n)$ is the residual response at each channel. Here, the TRF was estimated using ridge regression as follows, written in matrix format: $TRF = (S^T S + \lambda I)^{-1} S^T r$. Where S is the lagged time series of the stimulus, $\boldsymbol{I}$ is the identity matrix, and λ is the regularization parameter (Crosse et al., 2016). The TRF was estimated over time lags ranging from −150 to 450 ms. In a 5-fold cross-validation procedure with random selection of the trials, trained TRFs are used to predict the left-out MEG response based on the corresponding speech envelope. The predicted response is then compared with the recorded neural data to assess MEG prediction accuracy. To optimize the model for predicting speech envelope, the values of the regularization parameter (λ) between 1 and 10^6 was tuned using a leave-one-out cross-validation procedure. The λ value that produced the highest MEG prediction accuracy, averaged across trials and channels, was selected as the regularization parameter in each condition per participant. To increase signal to noise ratio, baseline normalization was applied using the time interval from −40 to 0 ms as baseline interval and computing the absolute change in TRF estimates with respect to the baseline interval. For further TRF sensor-level statistical analysis, we only used the gradiometers and calculated the combined planar gradient of the TRF estimates.

2.3.4. Source projection of temporal response functions

To transform the TRF sensor data into source space, we used a template structural magnetic resonance image (MRI) from Montreal Neurological Institute (MNI) and warped it to the individual head shape (Polhemus points) to match the individual fiducials and head shape landmarks. A 3-dimensional grid covering the entire brain volume of each participant with a resolution of 1 cm was created based on the standard MNI template MRI. We then used a mask to keep only the voxels corresponding to the gray matter (1457 voxels). The aligned brain volumes were further used to create realistic single-shell head models and lead field matrices (Nolte, 2003). By using the lead fields and a common covariance matrix (from all 2-second segments), common linearly constrained minimum variance (LCMV, Van Veen et al., 1997) beamformer spatial filter weights were computed based on an average covariance matrix estimated across all epochs. The regularization parameter was set to 20%. We then applied the spatial filter to the sensor-level TRF data.

2.3.5. Speech envelope reconstruction

The speech envelope reconstruction was also done by using the mTRF toolbox (Crosse et al., 2016). We used a backward decoding model which uses a linear filter or decoder that optimally combines the MEG signals of different sensors in order to reconstruct the speech envelope. The decoder, $g(\tau,\ n)$, represents the linear mapping from the neural response, $r(t,n)$, back to the stimulus, $s(t)$. Reconstructed stimulus (speech envelope), $\hat{s}(t)$ can be model as: Analogous to the TRF approach, the decoder is computed as follows: $g = (R^T R + \lambda I)^{-1} R^T s$, where R is the lagged time series of the response matrix, r.

The steps to obtain our measurement of neural speech tracking using this reconstruction method are the following. Firstly, the speech envelopes and MEG signals were down-sampled to 50 Hz to reduce the processing time. Secondly, we implemented a 5-fold cross-validation procedure and trained a linear decoder that combines the signals of all MEG channels and their time-shifted versions (integration window: −150–450 ms) on the MEG responses to the auditory stimuli. The decoder was then applied to the MEG responses to obtain reconstructed envelopes. The accuracy of reconstruction was measured by correlating the reconstructed envelope with the original envelope.

2.3.6. Speech-brain phase coherence

We directly projected preprocessed sensor space data to source space using LCMV beamformer filters to obtain time series data of each brain voxel (Van Veen et al., 1997). The procedure of source projection is comparable to those for the TRF analysis in Section 2.3.4. For calculating coherence between each brain voxel and the speech envelope, we then applied a frequency analysis to the 4-sec segments of all 6 conditions (original, 7-, 5-, 3-, 2-, 1-channel vocoded) using multi-taper frequency transformation (dpss taper: 1–25 Hz in 0.25 Hz steps, 4 Hz smoothing, no baseline correction).

In addition to canonical coherence analysis, we also looked into two sets of components of the coherence spectrum: periodic components (peak center frequency, peak bandwidth, peak height) and aperiodic components (aperiodic exponent, aperiodic offset) (Donoghue et al., 2020; Schmidt et al., 2021). These components were calculated using the FOOOF module with Python to compute aperiodic estimates and Gaussian model fits (Donoghue et al., 2020). The coherence data from 1 to 15 Hz was extracted from voxels in which the significant degradation effect was observed. The peak and aperiodic components of the coherence data were computed per voxel of each participant. The parameters used for modeling were from default setting (e.g., peak_width_limits: (0.5 12); max_n_peaks: inf; peak_threshold: 2.0). If R^2 (between the input spectrum and the full model fit) of the residual model or error of the full model fit differed from the rest by more than 3 standard deviations, the result from the voxel was dropped. Peak and aperiodic components were then averaged across the rest of the voxels for each participant.

2.3.7. Statistical analysis

For statistical comparisons, we first verified whether data distribution violated normality using the Shapiro-Wilk test (all $Ps > 0.5$) and whether the data contained outliers with a 1.5 interquartile range criterion. We used parametric tests for data that obey normal distribution and outliers free; otherwise, non-parametric tests were used. If not mentioned specifically, multiple comparisons were corrected by using the false discovery rate method (FDR, Benjamini and Hochberg, 1995).

For the behavioral hit rates, the Friedman test was used to test the speech degradation effect across conditions. Paired Wilcoxon signed-rank tests were used to compare the hit rates between conditions. One-sample Wilcoxon signed-rank tests were used to test the hit rates against chance level (50%).

For the TRF sensor-level data, non-parametric cluster-based permutation tests (Maris and Oostenveld, 2007) implemented in the FieldTrip toolbox (Oostenveld et al., 2011) were used. The dependent sample F-statistic (“depsamplesFunivariate”) was used for cluster formation when analyzing the main effect of speech degradation with 0.05 alpha level. The cluster-level statistic via the Monte Carlo approximation with 0.05 alpha level was calculated as the maximum of the cluster-level summed F-values of each cluster with a minimum of 3 neighboring channels. The Monte-Carlo estimation was based on 10,000 random partitions, and the time window of interest was defined from 0 to 400 ms.

For the TRF source-level data, we used time windows from the clusters showing significant effects with the cluster-based permutation test in sensor-level analysis and ran the F-statistic where we averaged over the time window for each condition and compared the average across conditions. For contrast between conditions, we extracted TRF estimates from voxels with significant effect and ran paired Wilcoxon signed-rank tests.

For the stimulus reconstruction data, the correlation coefficient which represents reconstruction accuracy was Fisher z transformed for further statistical testing. Repeated-measure one-way ANOVA was used to test the degradation effect. Paired t-tests were used to compare the hit rates between conditions.

For the coherence data, the dependent sample F-statistic was used for cluster formation when analyzing the main effect of speech degradation with a 0.05 alpha level. The cluster-level statistic via the Monte Carlo approximation with 0.05 alpha level was calculated as the maximum of the cluster-level summed F-values of each cluster with a minimum of 3 neighboring channels. The Monte-Carlo estimation was based on 10,000 random partitions and the frequency window of interest was defined from 2 to 7 Hz. Paired Wilcoxon signed-rank tests were used for comparison between conditions.

For the above statistical tests, the corresponding effect size was calculated. For repeated-measure one-way ANOVA, partial eta square (${\eta_p}^2$) was provided: $\eta_{p^2} = 0.01$ indicates a small effect; ${\eta_p}^2 = 0.06$ indicates a medium effect; $\eta_{p^2} = 0.14$ indicates a large effect. For paired t-tests, Cohen's d (d) was provided: $d = 0.2$ indicates a small effect; $d = 0.5$ indicates a medium effect; $d = 0.8$ indicates a large effect. For the Friedman test, the Kendall's W (W) was provided: $W = 0.1$ indicates a small effect; $W = 0.3$ indicates a medium effect; $W = 0.5$ indicates a large effect. For paired Wilcoxon signed-rank test and one-sample Wilcoxon signed-rank test, r was calculated as Z statistic divided by the square root of the sample size.: $r = 0.1$ indicates a small effect; $r = 0.3$ indicates a medium effect; $r = 0.5$ indicates a large effect.

To better understand functional characteristics of the TRF components, we calculated repeated measures correlations between TRF components and behavior hit rate, stimulus reconstruction accuracy as well as coherence results. Using repeated measures correlations allows us to analyze the intra-individual association for paired repeated measures (e.g. behavioral measures and speech-brain coherence here) assessed for each participant on two or more occasions (e.g. in six conditions here). Repeated measures correlations account for non-independence among within-individual measures using analysis of covariance (ANCOVA) to adjust for inter-individual variability and provides the best linear fit for each participant using regression lines with the same slope but varying intercepts. The *rmcorr* function in the *rmcorr* package (Bakdash and Marusich, 2017) was used to calculate the correlation coefficient, 1000-repetition bootstrapped 95% confidence interval, and p-value. Only the statistical tests using the Fieldtrip toolbox were conducted with MATLAB; the others were run with R software (version 3.6.2, R Development Core Team, 2019).

3. Results

3.1. Behavioral performance declines with decreased speech intelligibility

Sixteen of 24 healthy participants participated in both MEG and behavioral sessions, listening to audiobooks with 6-level of vocoded conditions (Fig. 1; original, 7-, 5-, 3-, 2-, and 1-channel vocoded). At the end of each audio presentation, participants were required to choose which noun is the last noun they heard from the two nouns on the screen. We pooled the behavioral responses for those participants who contributed data to both the MEG session and the behavioral session, which resulted in 12 behavioral responses in each condition for each participant (Fig. 1C). Participants' task performance decreased as the intelligibility of the audio stimuli dropped. The mean hit rate was 100.0% ± 0% (SD) for the original stimuli, 88.5% ± 11.3% for the 7-ch vocoded, 83.3% ± 14.9% for the 5-ch vocoded, 65.6% ± 19.0% for the 3-ch vocoded, 58.9% ± 14.1% for the 2-ch vocoded, and 47.9% ± 11.6% for the 1-ch vocoded. A speech degradation effect was found across the six conditions [$\chi^2(5) = 62.0$, $p = 4.69 \times 10^{-12}$, Friedman test, Kendall's $W = 0.775$]. Comparison within the six conditions showed that the hit rate for original stimuli was higher than all the other conditions (all $p_{fdr} < 0.01$, pairwise Wilcoxon signed-rank test, all effect size r > 0.80). The 7-ch vocoded condition had higher hit rates than 3-, 2-, and 1-ch conditions (all $p_{fdr} < 0.01$, all r > 0.25). The 5-ch condition had higher hit rates than 3-, 2-, and 1-ch conditions (all $p_{fdr} < 0.01$, all $r > 0.60$). The 3-ch condition had higher hit rates than the 1-ch condition ($p_{fdr} = 0.01$, $r = 0.67$). The 2-ch condition had higher hit rates than the 1-ch conditions ($p_{fdr} = 0.034$, $r = 0.53$). Except for the 1-ch vocoded condition ($p = 0.38$, $r = 0.23$, one-sample Wilcoxon signed-rank test), all the other conditions showed above-chance hit rates (all $p < 0.03$, all $r > 0.56$).

3.2. Temporal response functions show differential effects of vocoding on neural speech tracking

To investigate how a loss of spectral resolution influences speech tracking in a fine-grained temporal manner, we computed the temporal response functions (TRFs) between the speech envelope and the MEG signals in each condition from 24 participants.

The mean global field power in Fig. 2A characterizes temporal evolution of TRF estimates in each condition. To investigate the effects of degraded speech on TRF, we ran a cluster-based permutation test with a time window of interest from 0 to 400 ms and observed a main effect of speech degradation in six clusters. These six clusters revealed bilateral effects in three time windows respectively (Fig. 2A topographies): around 50–110 ms (termed as $M50_{TRF}$, $p_{cluster1} = 1.0 \times 10^{-4}$, $p_{cluster2} = 0.033$), around 175–230 ms (termed as $M200_{TRF}$, $p_{cluster3} = 0.003$, $p_{cluster4} = 0.021$), and around 315–380 ms (termed as $M350_{TRF}$, $p_{cluster5} = 0.001$, $p_{cluster6} = 0.035$).

We then projected the TRF sensor result to the source level (Fig. 2B) and extracted TRF estimates from voxels showing degradation effect (Fig. 2C). The statistical effect of $M50_{TRF}$ on the source level was mainly around bilateral temporal gyrus, and the amplitude of the original condition showed the smallest effect compared to 7-ch ($p_{fdr} = 2.21 \times 10^{-4}$, pairwise Wilcoxon signed-rank test, effect size $r = 0.77$), 5-ch ($p_{fdr} = 2.21 \times 10^{-4}$, $r = 0.79$), 3-ch ($p_{fdr} = 6.15 \times 10^{-4}$, $r = 0.71$), 2-ch ($p_{fdr} = 6.15 \times 10^{-4}$, $r = 0.71$), and 1-ch ($p_{fdr} = 2.21 \times 10^{-4}$, $r = 0.76$).

The statistical effect of $M200_{TRF}$ on the source level was around the bilateral parietal and temporal region. The source estimates overall attenuated as the intelligibility decreased. The original condition showed higher amplitude than 7-ch ($p_{fdr} = 0.003$, $r = 0.62$), 5-ch ($p_{fdr} = 0.002$, $r = 0.65$), 3-ch ($p_{fdr} = 1.78 \times 10^{-6}$, $r = 0.88$), 2-ch ($p_{fdr} = 4.47 \times 10^{-6}$, $r = 0.86$), and 1-ch ($p_{fdr} = 1.96 \times 10^{-5}$, $r = 0.82$). The 7-ch condition had higher amplitude than 3-ch ($p_{fdr} = 0.003$, $r = 0.62$), 2-ch ($p_{fdr} = 0.026$, $r = 0.47$) and 1-ch ($p_{fdr} = 0.006$, $r = 0.57$). The 5-ch condition had higher amplitude than 3-ch ($p_{fdr} = 8.34 \times 10^{-4}$, $r = 0.69$), 2-ch ($p_{fdr} = 0.011$, $r = 0.54$), and 1-ch ($p_{fdr} = 7.69 \times 10^{-4}$, $r = 0.71$).

The strongest responses of source of $M350_{TRF}$ were observed in the vocoded but still comprehensible conditions (i.e., 7-ch, 5-ch, and 3-ch conditions); the weakest response showed in the original condition. The 7-ch condition had a higher amplitude than 2-ch ($p_{fdr} = 0.009$, $r = 0.58$), 1-ch ($p_{fdr} = 0.002$, $r = 0.79$), and the original condition ($p_{fdr} = 3.70 \times 10^{-4}$, $r = 0.78$). The 5-ch condition had a higher amplitude than the 1-ch ($p_{fdr} = 0.009$, $r = 0.57$) and original condition ($p_{fdr} = 0.009$, $r = 0.60$). The 3-ch and 2-ch conditions had a higher amplitude than the original condition ($p_{fdr} = 0.009$ and 0.009, $r = 0.61$ and 0.57, respectively).

Overall, speech intelligibility did not modulate the temporal response function in the same manner over time. Modulation was observed at three intervals, and only the effect around the middle interval ($M200_{TRF}$) was related to speech intelligibility in a relatively straightforward manner, namely that only the amplitude of $M200_{TRF}$ decreased with the reduced intelligibility. These differential TRF patterns could be attributed to either general speech feature gain modulation or tracking in specific speech features e.g. amplitudes of speech envelope or syllables. In order to better understand these complex temporal patterns, we quantified other neural tracking measures and related them to the main TRF effects, which is described in the following sections.

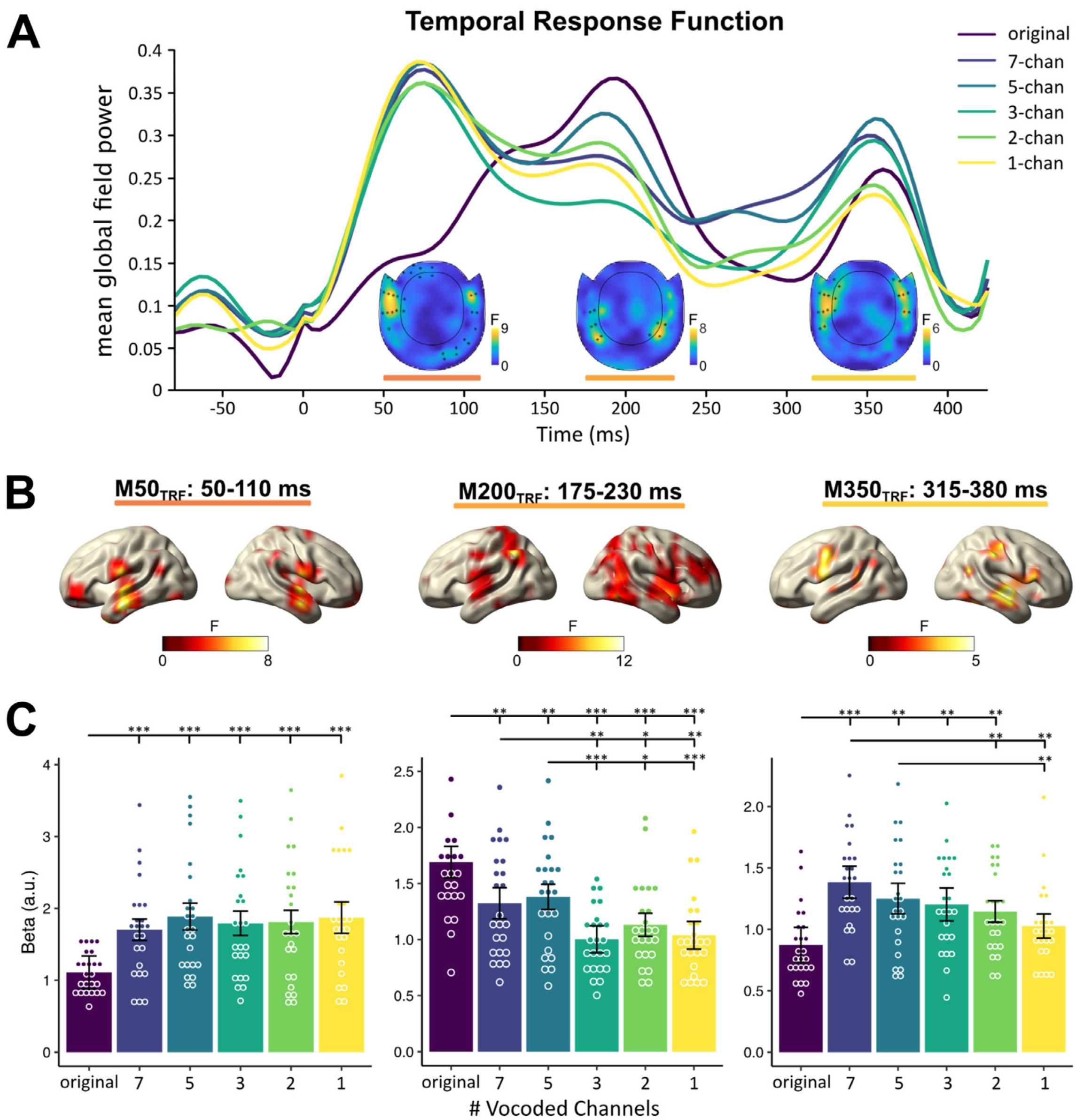

Fig. 2. Temporal response functions (TRFs) of six-level degraded speech. (**A**) Global mean field of the TRF on the sensor level of each condition. The orange-yellow horizontal lines denote the time windows which show significant speech degradation effects. The topographies show the region of the speech degradation effect at each latency respectively. Sensors showing significant effects are denoted with asterisks. (**B**) Source localizations of the degradation effects for each time window. (**C**) Individual TRF estimates of the six conditions extracted at voxels showing significant degradation effect. Bars represent 95% confidence intervals. $p_{fdr} < 0.05^{*}$, $p_{fdr} < 0.01^{**}$, $p_{fdr} < 0.001^{***}$.

3.3. Accuracy of the speech envelope reconstruction generally declines with decreased speech intelligibility

The envelope reconstruction accuracy was evaluated by correlation coefficient as reconstruction accuracy between the original envelope and the reconstructed envelope (Fig. 3A). A speech degradation effect on reconstruction accuracy was observed [repeated measure one-way ANOVA, F(5115) = 9.16, $p = 2.34 \times 10^{-7}$, $\eta_{p^2} = 0.29$]. The original condition had higher accuracy than the 2-ch [t(23) = 3.30, $p_{fdr} = 0.007$, Cohen's $d = 0.67$] and 1-ch conditions [t(23) = 4.87, $p_{fdr} = 4.88 \times 10^{-4}$, $d = 0.99$]. The 7-ch condition had higher accuracy than the 2-ch [t(23) = 3.48, $p_{fdr} = 0.005$, $d = 0.71$] and 1-ch conditions [t(23) = 3.82, $p_{fdr} = 0.003$, $d = 0.78$]. The 5-ch condition had higher accuracy than the 3-ch [t(23) = 3.58, $p_{fdr} = 0.005$, $d = 0.73$], 2-ch [t(23) = 4.47, $p_{fdr} = 8.65 \times 10^{-4}$, $d = 0.92$], and 1-ch conditions [t(23) = 5.39, $p_{fdr} = 2.68 \times 10^{-4}$, $d = 1.10$]. In sum, the reconstruction accuracy overall declined as the speech intelligibility decreased, while no differences were found among the original, 7-ch, and 5-ch conditions.

3.4. Speech-brain coherence declines with decreased speech intelligibility

Speech-brain coherence is likely the most common measure to quantify neural tracking. A degradation effect of speech was found around bilateral temporal, inferior parietal and inferior frontal regions from the broadband 2–7 Hz coherence analysis (Fig. 3B-2; cluster-corrected dependent-sample F-test; $p_{cluster1} = 9.99 \times 10^{-4}$, $p_{cluster2} = 0.018$). Within these areas, the strongest effect was found in the original condition, and no differences were found between 7-ch and 5-ch as well as between 3-ch and 2-ch condition (Fig. 3B-3). The original condition showed stronger coherence than the 7-ch ($p_{fdr} = 1.27 \times 10^{-4}$, pairwise Wilcoxon signed-rank test, effect size $r = 0.74$), 5-ch ($p_{fdr} = 0.005$, $r = 0.58$), 3-ch ($p_{fdr} = 2.23 \times 10^{-6}$, $r = 0.86$),

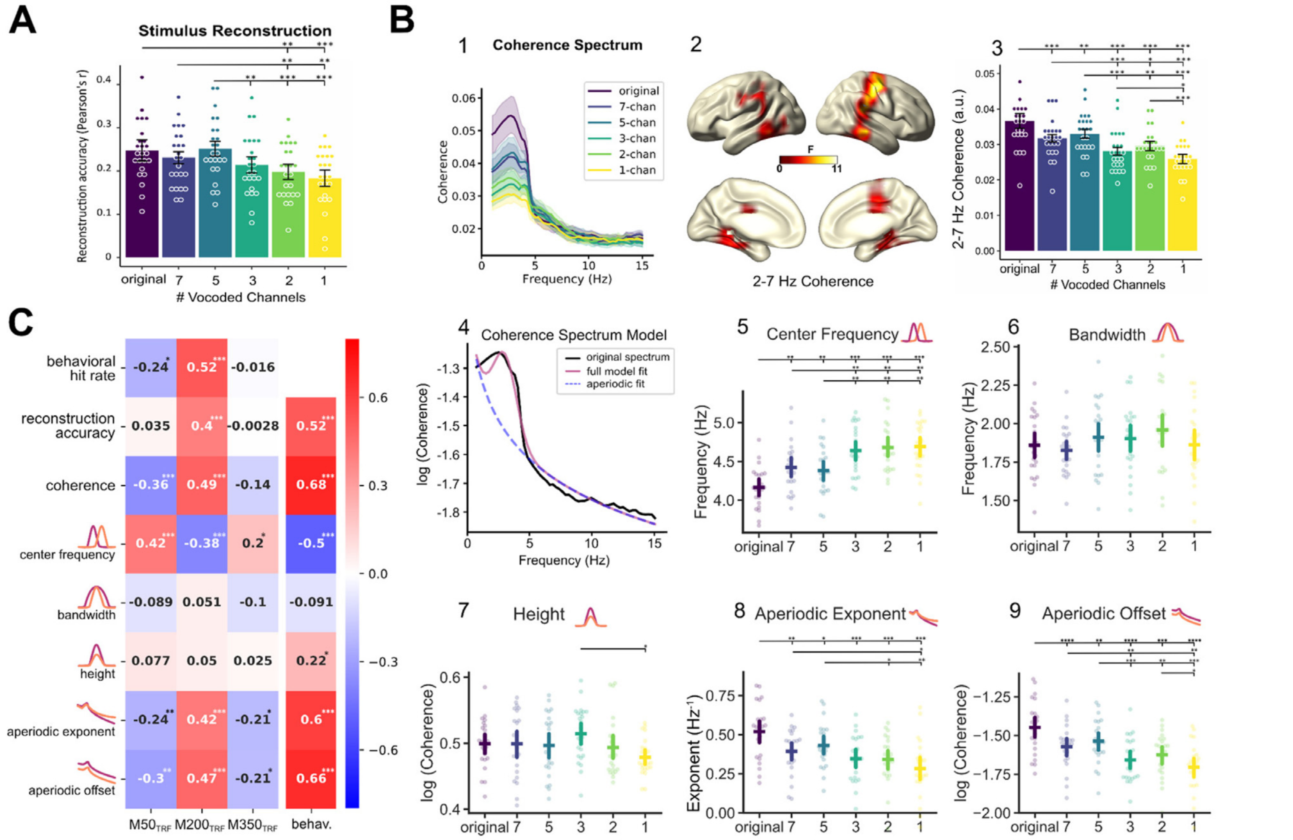

Fig. 3. Neural tracking of speech intelligibility measured via stimulus reconstruction and coherence. (**A**) Accuracy (Pearson's correlation coefficient) of stimulus reconstruction for each condition. (**B**) Coherence and components of the coherence spectrum in six conditions. **1**, Frequency spectrum of the coherence for the six conditions averaged across all voxels. **2**, Source localization of degradation effect on coherence across six conditions in bilateral temporal, inferior frontal, and parietal regions. **3**, Individual coherence values of the six conditions extracted from voxels showing significant effects. **4**, Example of model fitting for components of one coherence spectrum. **5–9**, Estimated peak center frequency, peak bandwidth, peak height, aperiodic exponent, and aperiodic offset of coherence spectrum in six conditions. Bars represent 95% confidence intervals. (**C**) Correlations between the TRF components and the behavioral hit rate, reconstruction accuracy, coherence as well as components of coherence spectrum show that $M200_{TRF}$ moderately correlated with hit rate, reconstruction accuracy, coherence, center frequency, exponent and offset which suggest that $M200_{TRF}$ can be another neural index of speech intelligibility. $p_{fdr} < 0.05^{*}$, $p_{fdr} < 0.01^{**}$, $p_{fdr} < 0.001^{***}$.

2-ch ($p_{fdr} = 4.08 \times 10^{-5}$, $r = 0.79$) and 1-ch condition ($p_{fdr} = 8.92 \times 10^{-7}$, $r = 0.88$). The 7-ch condition showed stronger coherence than the 3-ch ($p_{fdr} = 6.86 \times 10^{-5}$, $r = 0.76$), 2-ch ($p_{fdr} = 0.02$, $r = 0.48$) and 1-ch conditions ($p_{fdr} = 8.92 \times 10^{-7}$, $r = 0.88$). The 5-ch condition showed stronger coherence than the 3-ch ($p_{fdr} = 8.94 \times 10^{-6}$, $r = 0.83$), 2-ch ($p_{fdr} = 0.003$, $r = 0.61$), and 1-ch conditions ($p_{fdr} = 1.19 \times 10^{-6}$, $r = 0.87$). The 3-ch condition showed stronger coherence than the 1-ch ($p_{fdr} = 0.039$, $r = 0.43$). The 2-ch condition showed stronger coherence than the 1-ch ($p_{fdr} = 4.31 \times 10^{-5}$, $r = 0.78$). In sum, speech-brain coherence overall declined as the speech intelligibility decreased, while no differences were found between the 7-ch and 5-ch as well as between the 3-ch and 2-ch condition.

3.5. The peak center frequency, aperiodic exponent, and aperiodic offset are modulated by speech intelligibility

To complement our finding in the TRFs, we further extracted periodic components (peak center frequency, peak bandwidth, peak height) and aperiodic components (exponent, offset) from the broadband coherence spectrum.

We extracted the coherence data from 1 to 15 Hz from the voxels in which the significant degradation effect was observed from 2 to 7 Hz in the previous section. The peak and aperiodic components were computed per voxel of each participant (Fig. 3B-4). The peak and aperiodic components were then averaged across voxels which showed a significant degradation effect for each participant. Interestingly, we observed that the peak center frequency accelerated (mean center frequency from 4.17 Hz to 4.69 Hz) for less intelligible speech [$\chi^2(5) = 41.0$, $p = 9.49 \times 10^{-8}$, Friedman test, Kendall's $W = 0.78$] (Fig. 3B-5). This finding thereby fits with the idea proposed by Schmidt et al. (2021) that the speech tracking shifts from the more linguistic level (syllabic rate of the speech 4.3 Hz) to the more acoustic level (envelope modulation rate ∼5 Hz). The aperiodic components (exponent & offset) decreased as the speech intelligibility decreased [exponent: $\chi^2(5) = 35.2$, $p = 1.39 \times 10^{-6}$, $W = 0.29$; offset: $\chi^2(5) = 53.2$, $p = 3.07 \times 10^{-10}$, $W = 0.44$] (Fig. 3B-8, 3B-9). A borderline degradation effect was observed in the peak height [$\chi^2(5) = 11.2$, $p = 0.048$, $W = 0.09$] (Fig. 3B-7), and no degradation effect was found in the peak bandwidth [$\chi^2(5) = 6.43$, $p = 0.26$, $W = 0.05$] (Fig. 3B-6). In sum, our results show that with stronger vocoding, the peak center frequency shifted toward the acoustic level and the aperiodic components decreased.

3.6. The TRF finding highly correlates with behavioral and other neural measures of neural speech tracking

To relate our findings in TRF components with the other measures, we computed the repeated measures correlation coefficients between TRF components, behavior measurement, and the other neural indices (stimulus reconstruction accuracy, coherence, and components of coherence spectrum). The repeated measures correlation coefficients between the TRF components and the other measures are shown in Fig. 3C.

While the higher $M50_{TRF}$ was mildly correlated with the lower behavioral hit rates [$r_{rm}(79) = -0.24$ (−0.41 to −0.05), $p = 0.033$], the enhancement in $M50_{TRF}$ was correlated with the center frequency shifting to the acoustic level [$r_{rm}(119) = 0.42$ (0.25 to 0.56), $p = 1.85 \times 10^{-6}$]. Besides, the higher $M50_{TRF}$ was also mildly correlated with the lower coherence [$r_{rm}(119) = -0.36$ (−0.52 to −0.19), $p = 6.34 \times 10^{-5}$], and the lower aperiodic components [exponent: $r_{rm}(119) = -0.24$ (−0.37 to −0.07), $p = 0.008$; offset: $r_{rm}(119) = -0.30$ (−0.44 to −0.12), $p = 0.001$].

Of the three TRF components, $M200_{TRF}$ can best predict behavioral performance [$r_{rm}(79) = 0.52$ (95% CI: 0.40 to 0.65), $p = 5.25 \times 10^{-7}$]. Furthermore, the attenuation in $M200_{TRF}$ was correlated with the center frequency shifting from the linguistic to the acoustic level [$r_{rm}(119) = -0.38$ (−0.50 to −0.24), $p = 1.74 \times 10^{-5}$], which was opposite to $M50_{TRF}$ (Figure S1). In addition, $M200_{TRF}$ was also positively correlated with the reconstruction accuracy [$r_{rm}(119) = 0.40$ (0.29 to 0.52), $p = 4.42 \times 10^{-6}$], the coherence [$r_{rm}(119) = 0.49$ (0.35 to 0.61), $p = 1.21 \times 10^{-8}$], and the aperiodic components [exponent: $r_{rm}(119) = 0.42$ (0.26 to 0.54, $p = 2.09 \times 10^{-6}$; offset: $r_{rm}(119) = 0.47$ (0.33 to 0.59), $p = 4.88 \times 10^{-8}$].

Compared to $M50_{TRF}$ and $M200_{TRF}$, the nonlinear pattern of $M350_{TRF}$ showed mild correlation to only three neural measures, which was correlated positively with the center frequency [$r_{rm}(119) = 0.20$ (0.03 to 0.36), $p = 0.030$] and negatively with the aperiodic components [exponent: $r_{rm}(119) = -0.21$ (−0.34 to −0.07), $p = 0.019$; offset: $r_{rm}(119) = -0.21$ (−0.35 to −0.06), $p = 0.022$].

We also related the behavior results to the other neural measures: The better behavior hit rates were highly correlated with the higher envelope reconstruction accuracy [$r_{rm}(79) = 0.52$, 95% CI: 0.35 to 0.67, $p = 8.70 \times 10^{-7}$], the higher coherence [$r_{rm}(79) = 0.68$ (0.54 to 0.76), $p = 4.52 \times 10^{-12}$], the higher aperiodic components [exponent: $r_{rm}(79) = 0.60$ (0.46 to 0.71), $p = 2.99 \times 10^{-9}$; offset: $r_{rm}(79) = 0.66$ (0.56 to 0.77), $p = 1.64 \times 10^{-11}$] but with the lower center frequency [$r_{rm}(79) = -0.50$ (−0.65 to −0.34), $p = 2.16 \times 10^{-6}$]. The correlation with the peak height was weak but statistically significant [$r_{rm}(79) = 0.22$ (0.03 to 0.42), $p = 0.046$].

Together, among the three TRF components, only $M200_{TRF}$ was highly correlated with the behavioral results as well as the reconstruction accuracy, speech-brain coherence, peak center frequency, and aperiodic components. Interestingly, both $M50_{TRF}$ and $M200_{TRF}$ were correlated with the shifting center frequency but in the opposite direction. These findings depicted how vocoded speech affects neural speech tracking over time.

4. Discussion

Neural speech tracking is modulated by the spectral details of speech. However, the reported pattern of the intelligibility modulation has been mostly computed in a "static" manner [e.g. stimulus reconstruction (Verschueren et al., 2019) and speech-brain coherence (Hauswald et al., 2020)]. Here, we aimed to explore how the human brain represents different levels of degraded speech in a time-resolved manner. For this reason, we used original (i.e. clear and unaltered) speech and 5 levels of vocoded speech as stimuli and computed time-resolved temporal response functions (TRFs). Secondly, we related TRF findings with behavioral and other neural measures (stimulus reconstruction, speech-brain coherence, and components of broadband coherence spectra) of neural speech tracking. Overall, the behavioral performance declined with decreased speech intelligibility as expected. Our results showed that various levels of degraded speech differentially modulate the TRF around three time-windows: 50–110 ms ($M50_{TRF}$), 175–230 ms ($M200_{TRF}$), and 315–380 ms ($M350_{TRF}$). Interestingly, both $M50_{TRF}$ and $M200_{TRF}$ are correlated with the shifting center frequency of the coherence spectrum, however, in the opposite direction. These findings suggest that the distinctive effects on TRF are linked to altered tracking of speech features.

4.1. Neural tracking and speech intelligibility

In our neural measures, the modulation of the $M200_{TRF}$, the speech brain coherence and the aperiodic components of the broadband coherence spectra support the notion that the neural speech tracking declines as the speech intelligibility decreases.

The $M200_{TRF}$ is the TRF component that reflects the behavioral response best, i.e. captures the speech intelligibility (for the discussion of $M50_{TRF}$ and $M350_{TRF}$, see Section 4.2). Its amplitude decreases as the speech intelligibility declines (Fig. 2C-central). This relationship is also shown by the similarity of the correlational patterns of the $M200_{TRF}$ and the behavioral performance with the other measures (see Fig. 3C middle and right column). This finding of pointing to the relevance of speech intelligibility (measured behaviorally) of the $M200_{TRF}$ is consistent with other studies, e.g. in an EEG study with multi-talker and vocoded speech design, greater TRF deflection around 200 ms was also observed in an attended-nonvocoded condition compared to an attended-vocoded condition (Kraus et al., 2021). Using conventional ERP analysis, greater P200 amplitude has been also observed in clear speech compared to degraded speech (Strauß et al., 2013) and in words pronounced with native speech than words pronounced with foreign-accented speech (Romero-Rivas et al., 2015). Besides, the P200 amplitude has been positively correlated with the successful extraction of phonetic/phonological information from vowels or words (De Diego Balaguer et al., 2007; Reinke et al., 2003). All the above results indicate that the higher amplitude of this component is, the higher speech intelligibility is. Furthermore, the $M200_{TRF}$ effect is located mainly around inferior parietal cortex, superior temporal gyrus, and inferior frontal cortex, consistent with prior fMRI work on the effect of speech intelligibility (Obleser and Kotz, 2010).

The association between the $M200_{TRF}$ and the peak center frequency of the coherence spectrum also supports the notion that $M200_{TRF}$ can be an index of speech intelligibility (Schmidt et al., 2021). The shifting center frequency reflects the neural speech tracking from the (linguistic-level) syllabic rate of the speech (4.3 Hz) to the (acoustic-level) envelope modulation rate (~5 Hz). This is not merely due to the physical characteristics of the acoustic stimulus as the reduction of spectral details (from clear to 1-ch vocoded) in the audio stimuli results in the decrease of peak frequency in the amplitude modulation of the stimuli (from 5.61 to 4.83 Hz). However, the center frequency of the speech brain coherence, on the other hand, increased (from 4.17 to 4.69 Hz). This phenomenon also highlights that the brain is not exclusively driven by the rhythm of external stimuli.

The speech-brain coherence was also highly correlated with speech comprehension, in line with Schmidt et al. (2021). However, the coherence result was different from Hauswald et al. (2020), from which we re-analyzed the data. In Hauswald et al. (2020), the coherence result showed an inverted U-shaped pattern: The highest coherence showed in the marginally intelligible (5-ch vocoded) condition. After re-analyzing the data, we found that the difference could result from distinct high-pass filter settings [0.1 Hz here and in Schmidt et al. (2021) instead of 1 Hz in Hauswald et al. (2020)]. The cutoff of the 1-Hz high-pass filter in Hauswald et al. (2020) was too close to the target frequency (2 Hz) which induced phase shift in MEG signals around 2-Hz results (Yael et al., 2018). Separating periodic/aperiodic contribution in the coherence is not common (Schmidt et al., 2021). Works in the related field should develop an increased awareness that aperiodic part can affect the narrow-band coherence (also where the coherence peaks are) and to refrain from prematurely interpreting results from an "oscillations" perspective (e.g. enhanced/decreased rhythmic entrainment).

A speech degradation effect was found in the aperiodic components (exponent & offset), consistent with Schmidt et al. (2021). Here, we demonstrate that the aperiodic components were highly correlated with the behavioral performance, which suggests that the aperiodic components can be reliable biomarkers for speech processing. Overall, we again highlight the importance of both the periodic and aperiodic components of the coherence spectra as they can provide more detailed information of speech processing than the coherence alone. As the periodic and aperiodic components of the frequency power spectrum have been shown to be biological markers in aging (Dave et al., 2018; Voytek et al., 2015) and mental disorders (Molina et al., 2020; Ostlund et al., 2021; Robertson et al., 2019), it is possible that this can also be the case for components of the coherence spectrum. Future studies could test whether those components could offer more insight into more clinically-relevant participants, e.g. in individuals with hearing impairment.

An important idea in this study is that we manipulated speech intelligibility by varying the spectral information only. The assertion that more spectral information resulted in better speech intelligibility has been supported by our behavioral results and previous studies (Peelle et al., 2013; Schmidt et al., 2021). It is worth to note that speech features show dependencies along both spectral and temporal modulation dimensions (Santoro et al., 2014, 2017; Zulfiqar et al., 2020). Further study is needed to look into how less spectral information sabotages speech intelligibility.

4.2. Neural speech tracking could be affected by other potential factors

The modulation of spectral details in speech signals not only affects speech intelligibility but also influences e.g. attention load, memory load, and perceptual learning (Mattys et al., 2012). Our findings in $M50_{TRF}$ and $M350_{TRF}$ also argue that speech intelligibility can be modulated by other factors.

$M50_{TRF}$, the earliest effect occurred around 50 ms and clearly differentiated all vocoded conditions from the non-vocoded one (Fig. 2C-left). The finding of this early component is in line with previous studies showing that the TRF as well as ERP amplitudes of this early peak were greater in vocoded conditions compared to intact conditions (Ding et al., 2014; Kraus et al., 2021; Strauß et al., 2013). Besides, a comparable effect was also found for louder conditions when compared to soft-spoken conditions (Verschueren et al., 2021). While this $M50_{TRF}$ peak thus seems to follow physical manipulations of the speech signal, Ding & Simon (2012) could not find attentional modulations earlier than ~100 ms. Taken together, we propose that this early effect might be modulated by sensory gain on acoustic properties of the auditory stimuli rather than task difficulty as there is marked distinction between clear (normal) and spectral-degraded (abnormal) speech. In addition, our source result showing the prominent effect around bilateral primary auditory cortices further support the notion that the speech processing is rather early at this stage (de Heer et al., 2017; Khalighinejad et al., 2021). It is worth noting that a similar effect on this early component was also observed for age both in pure tones (Herrmann et al., 2022) and in speech (Brodbeck et al., 2018), with the older participants showing an increased amplitude compared to the younger ones. The potential neural mechanism in these cases can be the altered balance between inhibitory and excitatory neural mechanisms in the cortex during aging. Furthermore, we also found a positive correlation between the $M50_{TRF}$ amplitude and center frequency of coherence, which means that higher $M50_{TRF}$ amplitudes go along with higher center frequencies in speech-brain coherence, and both of these are related to reduced speech intelligibility. This finding highlights that $M50_{TRF}$, which reflects the distinctions of acoustic information between original and vocoded speech, can also play a role in shifting center frequency for degraded speech.

The effect of speech degradation on TRF around 350 ms ($M350_{TRF}$) shows an inverted U-shaped pattern across six conditions: the maximal amplitude found in 7-ch condition, followed by 5-ch, 3-ch, 2-ch, 1-ch, and original (Fig. 2C-right). The $M350_{TRF}$ effect was source-localized mainly around the left inferior frontal cortex. This $M350_{TRF}$ shares temporal and source characteristics with the N400 which is seen as a potential marker of semantic integration (for review, see Kutas and Federmeier, 2011; Lau et al., 2008). It is possible that the strongest effect observed in the 7-ch vocoded condition resulted from that the speech was still understandable and predictable while the acoustic input was violated. The effect reduced in the clear speech as the acoustic input was as predicted. Furthermore, the effect declined in the more degraded speech as the speech became less comprehensible and less predictable. Accordingly, a similar nonlinear N400 effect was also found in a previous study manipulating 6 levels of signal to noise ratio (Jamison et al., 2016). Furthermore, lexical-semantic processing is modulated when the spectral properties of the speech are disrupted as shown by Obleser & Kotz (2011) and Strauß et al. (2013). Taken together, previous research and our results seem to indicate that the $M350_{TRF}$ is an index of late-stage speech processing (e.g. semantic integration), whereas more research directly addressing this question is needed.

Based on our time-resolved findings in TRF components, we propose that envelope reconstruction is not a robust measure for the modulation of speech intelligibility on neural speech tracking as the result of envelope reconstruction is rather static compared to those of temporal response functions, especially in which we have shown three components (M50TRF, M200TRF, and M350TRF) do not correlate uniformly with intelligibility. This argument is supported by results from ours and previous studies (Decruy, Lesenfants, et al., 2020; Presacco et al., 2019). In our results, no differences were observed among the conditions that can be understood very easily to mildly challenging (original, 7-ch, and 5-ch conditions). In Decruy et al. (2020) and Presacco et al. (2019), higher reconstruction accuracy did not fully reflect better speech understanding especially when comparing hearing impaired populations to healthy control groups.

4.3. Conclusions

In this study, we demonstrate that at least three neural processing stages ($M50_{TRF}$, $M200_{TRF}$, and $M350_{TRF}$) are affected when processing continuous degraded speech. Only $M200_{TRF}$ highly correlates with behavior and other neural measures of speech processing (i.e., speech brain coherence). We also indicate that the effect of $M50_{TRF}$ and $M200_{TRF}$ can reflect shifts in the neural speech tracking from more acoustic level to more linguistic level. This argument is also supported by the effect of center frequency of coherence as it decreased when speech intelligibility was enhanced. In general, using the temporal response function combined with parametrization of coherence spectra, we demonstrate the temporal dynamics of neural speech tracking and provide potential explanations for the inconsistent literature.

Data/Code availability

The data and code necessary for generating the figures and computing statistics will be shared in the corresponding author's gitlab repository: https://gitlab.com/yapingchen. Access to raw data will be made available upon reasonable request to the corresponding author.

Declaration of Competing Interests

The authors declare no competing financial interests.

Credit authorship contribution statement

Ya-Ping Chen: Conceptualization, Data curation, Formal analysis, Writing – original draft, Visualization. **Fabian Schmidt:** Conceptualization, Data curation, Formal analysis, Writing – review & editing. **Anne Keitel:** Funding acquisition, Conceptualization, Supervision, Writing – review & editing. **Sebastian Rösch:** Funding acquisition, Writing – review & editing. **Anne Hauswald:** Funding acquisition, Conceptualization, Data curation, Formal analysis, Supervision, Writing – review & editing. **Nathan Weisz:** Funding acquisition, Conceptualization, Supervision, Writing – review & editing.

Data Availability

Data will be made available on request.

Acknowledgements

This research was supported by the Austrian Science Fund (FWF, P31230). We thank the participants in this study as well as Manfred Seifter and Bernadette Eckart for their support during data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.119894.

References

Bakdash, J.Z., Marusich, L.R., 2017. Repeated measures correlation. Front. Psychol. 8. doi:10.3389/fpsyg.2017.00456.

Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. J. Neurophysiol. 118 (6), 3144–3151. doi:10.1152/jn.00023.2017.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57 (1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer.* (6.1.04). http://www.praat.org/

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10 (4), 433–436. doi:10.1163/156856897x00357.

Brodbeck, C., Presacco, A., Simon, J.Z., 2018. 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.

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

Dave, S., Brothers, T.A., Swaab, T.Y., 2018. 1/f neural noise and electrophysiological indices of contextual prediction in aging. Brain Res. 1691, 34–43. doi:10.1016/j.brainres.2018.04.007.

Decruy, L., Lesenfants, D., Vanthornhout, J., Francart, T., 2020a. Top-down modulation of neural envelope tracking: the interplay with behavioral, self-report and neural measures of listening effort. Eur. J. Neurosci. 52 (5), 3375–3393. doi:10.1111/ejn.14753.

Decruy, L., Vanthornhout, J., Francart, T., 2020b. Hearing impairment is associated with enhanced neural tracking of the speech envelope. Hear. Res. 393, 107961. doi:10.1016/j.heares.2020.107961.

De Diego Balaguer, R., Toro, J.M., Rodriguez-Fornells, A., Bachoud-Lévi, A.-C., 2007. Different neurophysiological mechanisms underlying word and rule extraction from speech. PLoS ONE 2 (11), e1175. doi:10.1371/journal.pone.0001175.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E., 2017. The hierarchical cortical organization of human speech processing. J. Neurosci. 37 (27), 6539–6557. doi:10.1523/JNEUROSCI.3267-16.2017.

de Jong, N.H., Pacilly, J., Heeren, W., 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. Assess. Educ. 28 (4), 456–476. doi:10.1080/0969594X.2021.1951162.

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

Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C., Poeppel, D., 2017. Temporal modulations in speech and music. Neurosci. Biobehav. Rev. 81, 181–187. doi:10.1016/j.neubiorev.2017.02.011.

Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. 109 (29), 11854–11859. doi:10.1073/pnas.1205381109.

Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33 (13), 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013.

Donoghue, T., Haller, M., Peterson, E.J., Varma, P., Sebastian, P., Gao, R., Noto, T., Lara, A.H., Wallis, J.D., Knight, R.T., Shestyuk, A., Voytek, B., 2020. Parameterizing neural power spectra into periodic and aperiodic components. Nat. Neurosci. 23 (12), 1655–1665. doi:10.1038/s41593-020-00744-x.

Elliott, T.M., Theunissen, F.E., 2009. The modulation transfer function for speech intelligibility. PLoS Comput. Biol. 5 (3), e1000302. doi:10.1371/journal.pcbi.1000302.

Fiedler, L., Wöstmann, M., Herbst, S.K., Obleser, J., 2019. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. Neuroimage 186, 33–42. doi:10.1016/j.neuroimage.2018.10.057.

Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110 (2), 1150–1163. doi:10.1121/1.1381538.

Fuglsang, S.A., Märcher-Rørsted, J., Dau, T., Hjortkjær, J., 2020. Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. J. Neurosci. 40 (12), 2562–2572. doi:10.1523/JNEUROSCI.1936-19.2020.

Gaudrain, E. (2016). *Vocoder, v1.0.* Zenodo. doi:10.5281/zenodo.48120.

Hartmann, T., Weisz, N., 2020. An introduction to the objective psychophysics toolbox. Front. Psychol. 11, 2789. doi:10.3389/fpsyg.2020.585437.

Hauswald, A., Keitel, A., Chen, Y.-P., Rösch, S., Weisz, N., 2020. Degradation levels of continuous speech affect neural speech tracking and alpha power differently. Eur. J. Neurosci. 00, 1–15. doi:10.1111/ejn.14912.

Herrmann, B., Maess, B., Johnsrude, I.S., 2022. A neural signature of regularity in sound is reduced in older adults. Neurobiol. Aging 109, 1–10. doi:10.1016/j.neurobiolaging.2021.09.011.

Jamison, C., Aiken, S.J., Kiefte, M., Newman, A.J., Bance, M., Sculthorpe, -P.L., 2016. Preliminary investigation of the passively evoked N400 as a tool for estimating speech-in-noise thresholds. Am. J. Audiol. 25 (4), 344–358. doi:10.1044/2016_AJA-15-0080.

Khalighinejad, B., Patel, P., Herrero, J.L., Bickel, S., Mehta, A.D., Mesgarani, N., 2021. Functional characterization of human Heschl's gyrus in response to natural speech. Neuroimage 235, 118003. doi:10.1016/j.neuroimage.2021.118003.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in psychtoolbox-3. Perception 36 (14), 1–16.

Kraus, F., Tune, S., Ruhe, A., Obleser, J., Wöstmann, M., 2021. Unilateral acoustic degradation delays attentional separation of competing speech. Trends Hear. 25, 1–13. doi:10.1177/23312165211013242.

Kutas, M., 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.

Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (De)constructing the N400. Nat. Rev. Neurosci. 9 (12), 920–933. doi:10.1038/nrn2532.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190. doi:10.1016/j.jneumeth.2007.03.024.

Mattys, S.L., Davis, M.H., Bradlow, A.R., Scott, S.K., 2012. Speech recognition in adverse conditions: a review. Lang. Cogn. Process. 27 (7–8), 953–978. doi:10.1080/01690965.2012.705006.

Molina, J.L., Voytek, B., Thomas, M.L., Joshi, Y.B., Bhakta, S.G., Talledo, J.A., Swerdlow, N.R., Light, G.A., 2020. Memantine effects on electroencephalographic measures of putative excitatory/inhibitory balance in Schizophrenia. Biol. Psychiatry 5 (6), 562–568. doi:10.1016/j.bpsc.2020.02.004.

Nogueira, W., Dolhopiatenko, H., Schierholz, I., Büchner, A., Mirkovic, B., Bleichner, M.G., Debener, S., 2019. Decoding selective attention in normal hearing listeners and bilateral cochlear implant users with concealed ear EEG. Front. Neurosci. 13. doi:10.3389/fnins.2019.00720.

Nolte, G., 2003. The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys. Med. Biol. 48 (22), 3637–3652. doi:10.1088/0031-9155/48/22/002.

Obleser, J., Kotz, S. A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb. Cortex 20 (3), 633–640. doi:10.1093/cercor/bhp128.

Obleser, J., Kotz, S.A., 2011. Multiple brain signatures of integration in the comprehension of degraded speech. Neuroimage 55 (2), 713–723. doi:10.1016/j.neuroimage.2010.12.020.

Obleser, J., Weisz, N., 2012. Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. Cereb. Cortex 22 (11), 2466–2477. doi:10.1093/cercor/bhr325.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci., 156869 doi:10.1155/2011/156869, *2011*.

Ostlund, B.D., Alperin, B.R., Drew, T., Karalunas, S.L., 2021. Behavioral and cognitive correlates of the aperiodic (1/f-like) exponent of the EEG power spectrum in adolescents with and without ADHD. Dev. Cogn. Neurosci. 48, 100931. doi:10.1016/j.dcn.2021.100931.

Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23 (6), 1378–1387. doi:10.1093/cercor/bhs118.

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10 (4), 437–442.

Petersen, E.B., Wöstmann, M., Obleser, J., Lunner, T., 2017. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J. Neurophysiol. 117 (1), 18–27. doi:10.1152/jn.00527.2016.

Presacco, A., Simon, J.Z., Anderson, S., 2019. Speech-in-noise representation in the aging midbrain and cortex: effects of hearing loss. PLoS ONE 14 (3). doi:10.1371/journal.pone.0213899.

R Development Core Team, 2019. R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing.

Reinke, K.S., He, Y., Wang, C., Alain, C., 2003. Perceptual learning modulates sensory evoked response during vowel segregation. Cogn. Brain Res. 17 (3), 781–791. doi:10.1016/S0926-6410(03)00202-7.

Rimmele, J.M., Zion Golumbic, E., Schröger, E., Poeppel, D., 2015. The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. Cortex 68, 144–154. doi:10.1016/j.cortex.2014.12.014.

Robertson, M.M., Furlong, S., Voytek, B., Donoghue, T., Boettiger, C.A., Sheridan, M.A., 2019. EEG power spectral slope differs by ADHD status and stimulant medication exposure in early childhood. J. Neurophysiol. 122 (6), 2427–2437. doi:10.1152/jn.00388.2019.

Romero-Rivas, C., Martin, C.D., Costa, A., 2015. Processing changes when listening to foreign-accented speech. Front. Hum. Neurosci. 9, 167. doi:10.3389/fnhum.2015.00167.

Rosen, S., Souza, P., Ekelund, C., Majeed, A.A., 2013. Listening to speech in a background of other talkers: effects of talker number and noise vocoding. J. Acoust. Soc. Am. 133 (4), 2431–2443. doi:10.1121/1.4794379.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. 10 (1), e1003412. doi:10.1371/journal.pcbi.1003412.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc. Natl. Acad. Sci. 114 (18), 4799–4804. doi:10.1073/pnas.1617622114.

Schmidt, F., Chen, Y.-.P., Keitel, A., Rösch, S., Hannemann, R., Serman, M., Hauswald, A., Weisz, N., 2021. Neural speech tracking shifts from the syllabic to the modulation rate of speech as intelligibility decreases. BioRxiv doi:10.1101/2021.03.25.437033, 2021.03.25.437033.

Shannon, R.V., Zeng, F.-.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304. doi:10.1126/science.270.5234.303.

Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90. doi:10.1038/416087a.

Song, J., Iverson, P., 2018. Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. Cognition 179, 163–170. doi:10.1016/j.cognition.2018.06.001.

Strauß, A., Kotz, S.A., Obleser, J., 2013. Narrowed expectancies under degraded speech: revisiting the N400. J. Cogn. Neurosci. 25 (8), 1383–1395. doi:10.1162/jocn_a_00389.

Tune, S., Alavash, M., Fiedler, L., Obleser, J., 2021. Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. Nat. Commun. 12 (1), 4533. doi:10.1038/s41467-021-24771-9.

Van Veen, B.D., van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44 (9), 867–880. doi:10.1109/10.623056.

Verschueren, E., Somers, B., Francart, T., 2019. Neural envelope tracking as a measure of speech understanding in cochlear implant users. Hear. Res. 373, 23–31. doi:10.1016/j.heares.2018.12.004.

Verschueren, E., Vanthornhout, J., Francart, T., 2021. The effect of stimulus intensity on neural envelope tracking. Hear. Res. 403, 108175. doi:10.1016/j.heares.2021.108175.

von Goethe, J.W. 1795. Das Märchen.

Voytek, B., Kramer, M.A., Case, J., Lepage, K.Q., Tempesta, Z.R., Knight, R.T., Gazzaley, A., 2015. Age-Related changes in 1/f neural electrophysiological noise. J. Neurosci. 35 (38), 13257–13265. doi:10.1523/JNEUROSCI.2332-14.2015.

Yael, D., Vecht, J.J., Bar-Gad, I., 2018. Filter-based phase shifts distort neuronal timing information. eNeuro 5 (2). doi:10.1523/ENEURO.0261-17.2018, ENEURO.0261-17.2018.

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

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., Zheng, J., Ding, N., 2019. Auditory and language contributions to neural encoding of speech features in noisy environments. Neuroimage 192, 66–75. doi:10.1016/j.neuroimage.2019.02.047.

Zulfiqar, I., Moerel, M., Formisano, E., 2020. Spectro-temporal processing in a two-stream computational model of auditory cortex. Front. Comput. Neurosci. 13. doi:10.3389/fncom.2019.00095.