Behavioral/Cognitive

Attentional Modulation of the Cortical Contribution to the Frequency-Following Response Evoked by Continuous Speech

Alina Schüller,[1] **Achim Schilling,**[2] **Patrick Krauss,**[2,3] **Stefan Rampp,**[4,5,6] **and Tobias Reichenbach**[1]

[1]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen-Nürnberg, 91054 Erlangen, Germany, [2]Neuroscience Laboratory, University Hospital Erlangen, 91058 Erlangen, Germany, [3]Pattern Recognition Lab, Department Computer Science, Friedrich-Alexander-University Erlangen-Nürnberg, 91054 Erlangen, Germany, [4]Department of Neurosurgery, University Hospital Erlangen, 91058 Erlangen, Germany, [5]Department of Neurosurgery, University Hospital Halle (Saale), 06120 Halle (Saale), Germany, and [6]Department of Neuroradiology, University Hospital Erlangen, 91058 Erlangen, Germany

Selective attention to one of several competing speakers is required for comprehending a target speaker among other voices and for successful communication with them. It moreover has been found to involve the neural tracking of low-frequency speech rhythms in the auditory cortex. Effects of selective attention have also been found in subcortical neural activities, in particular regarding the frequency-following response related to the fundamental frequency of speech (speech-FFR). Recent investigations have, however, shown that the speech-FFR contains cortical contributions as well. It remains unclear whether these are also modulated by selective attention. Here we used magnetoencephalography to assess the attentional modulation of the cortical contributions to the speech-FFR. We presented both male and female participants with two competing speech signals and analyzed the cortical responses during attentional switching between the two speakers. Our findings revealed robust attentional modulation of the cortical contribution to the speech-FFR: the neural responses were higher when the speaker was attended than when they were ignored. We also found that, regardless of attention, a voice with a lower fundamental frequency elicited a larger cortical contribution to the speech-FFR than a voice with a higher fundamental frequency. Our results show that the attentional modulation of the speech-FFR does not only occur subcortically but extends to the auditory cortex as well.

Key words: **cortical response; selective attention; speech processing; speech-FFR**

Significance Statement

Understanding speech in noise requires attention to a target speaker. One of the speech features that a listener can use to identify a target voice among others and attend it is the fundamental frequency, together with its higher harmonics. The fundamental frequency arises from the opening and closing of the vocal folds and is tracked by high-frequency neural activity in the auditory brainstem and in the cortex. Previous investigations showed that the subcortical neural tracking is modulated by selective attention. Here we show that attention affects the cortical tracking of the fundamental frequency as well: it is stronger when a particular voice is attended than when it is ignored.

Introduction

Selective attention is a fundamental cognitive process that allows us to focus on relevant information while filtering out distracting signals. Referred to as the "cocktail-party effect," in complex acoustic settings such as a busy pub or restaurant, selective attention enables us to focus on a particular speaker among other competing voices to selectively process that speech signal and extract linguistic information and meaning (Cherry, 1953; McDermott, 2009).

Recent research has used continuous natural speech to explore which speech features are involved in selective attention to speech. These investigations have mostly focused on low-

Received July 6, 2023; revised Sep. 7, 2023; accepted Sep. 21, 2023.

Author contributions: A. Schüller and T.R. designed research; A. Schüller performed research; A. Schilling, P.K., and S.R. contributed unpublished reagents/analytic tools; A. Schüller and T.R. analyzed data; A. Schüller, A. Schilling, P.K., S.R., and T.R. wrote the paper.

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation), as follows: Grants KR 5148/2-1 (Project 436456810), KR 5148/3-1 (Project 510395418), and GRK 2839 (Project 468527017) to P.K.; and Grant SCHI 1482/3-1 (Project 451810794) to A. Schüller; and by the Emerging Talents Initiative of the University Erlangen-Nuremberg (Grant 2019/2-Phil-01 to P.K.).

The authors declare no competing financial interests.

Correspondence should be addressed to Tobias Reichenbach at tobias.j.reichenbach@fau.de.

https://doi.org/10.1523/JNEUROSCI.1247-23.2023

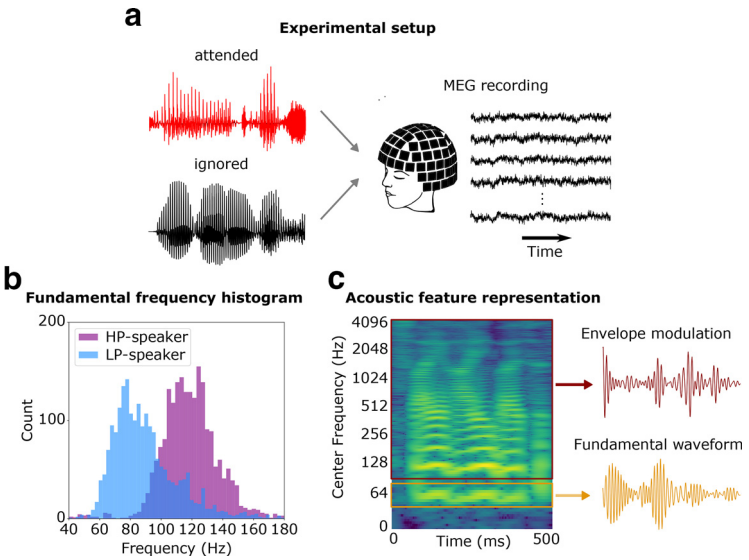

Figure 1. Experimental setup and acoustic stimuli. ***a***, Two audiobooks (one attended and one ignored) were presented simultaneously while MEG was recorded. ***b***, One of the two male speakers had a lower fundamental frequency and hence pitch (LP), and the other speaker a higher one (HP). ***c***, We quantified the speech-FFR through two acoustic features: the fundamental waveform that reflected the portion of the speech spectrogram around the fundamental frequency, and the envelope modulation of the higher harmonics.

frequency responses in the auditory cortex. In particular, they found that the neural tracking of rhythms in speech set by the rate of syllables and words, in the delta (1–4 Hz) and theta (4–8 Hz) frequency bands, is modulated by selective attention to one of two competing speakers (Kerlin et al., 2010; Ding and Simon, 2012; Power et al., 2012; Horton et al., 2013; Ding and Simon, 2014). In addition, the power of high-frequency responses in the gamma band, between 70 and 150 Hz, tracks the low-frequency speech rhythms as well (Mesgarani and Chang, 2012; Synigal et al., 2020). These findings were obtained using different measurement techniques, in particular an invasive one, electrocorticography (Mesgarani and Chang, 2012), as well as noninvasive ones such as magnetoencephalography (MEG; Ding and Simon, 2012) and electroencephalography (EEG; Kerlin et al., 2010; Horton et al., 2013; Etard and Reichenbach, 2019).

As evidence of the robustness of the attentional effect on the neural tracking, attention to a specific voice could be accurately decoded from single trials with short speech stimuli that lasted ~1 min using MEG (Ding and Simon, 2012) and EEG (Mirkovic et al., 2015; O'Sullivan et al., 2015; Fiedler et al., 2017). The decoding accuracy further improved with optimization of statistical modeling, allowing for accurate decoding from recordings <30 s (Biesmans et al., 2017; Van Eyndhoven et al., 2017). Additionally, the ability to detect changes in attentional focus within tens of seconds was demonstrated using EEG data, and even faster when combined with sparse statistical modeling techniques in MEG data, as observed in the study by Miran et al. (2018).

In addition to the cortical tracking of the low-frequency speech rhythms, a high-frequency neural response to a high-frequency speech features has been investigated as well (Chandrasekaran and Kraus, 2010; Kraus et al., 2017). It emerges in response to the fundamental frequency and its higher harmonics of the voiced parts of speech. The frequency range of the response is that of the fundamental frequency of speech, typically between 100 Hz and 300 Hz. Because of its similarities to the frequency-following response (FFR) to a pure tone, we refer to it as the "speech-FFR" in the following.

Using EEG, we have recently shown that the speech-FFR is modulated by selective attention to one of two competing speakers (Forte et al., 2017; Etard et al., 2019; Saiz-Alía et al., 2019). In particular, we measured a larger speech-FFR to a particular speaker when that speaker was attended compared with when they were ignored. The latency of the speech-FFR that we analyzed was ~10 ms. Together with the topographic map that showed large contributions from the mastoid channels and the channels at the vertex, this demonstrated a subcortical origin of the response. The attention decoding accuracy was significant even for short segments of EEG data, down to a few seconds in duration (Etard et al., 2019).

However, studies using MEG recently showed the presence of cortical contributions to the speech-FFR (Coffey et al., 2016, 2017a,b; Hartmann and Weisz, 2019; Kulasingham et al., 2020; Gorina-Careta et al., 2021). These contributions have been verified through EEG measurements as well (Bidelman, 2015). In response to continuous speech, the cortical portion of the speech-FFR has been found to occur at latencies of ~30–40 ms (Kulasingham et al., 2020; Schüller et al., 2023). EEG and MEG have thereby been found to play partly complimentary roles when assessing the speech-FFR: EEG measures mostly the subcortical contributions, while MEG records predominantly the cortical portions.

Whether the cortical contribution to the speech-FFR is modulated by selective attention has, however, not yet been investigated. Here we set out to close that gap in knowledge. In particular, by using the sensitivity of MEG to cortical sources as well as an auditory stimulus that contained two competing continuous speech signals, we aimed to investigate how attention modulates cortical responses during continuous speech perception.

Materials and Methods

Experimental design and data analysis. We used MEG recordings of neural responses to two competing talkers (Fig. 1*a*). The speech signals consisted of two audiobooks, and participants regularly switched their attention between the two speakers. The MEG data were analyzed through first performing source reconstruction and then relating the source-reconstructed neural activity to two high-frequency speech features using linear regression. We thus obtained temporal response

functions (TRFs) that described the speech-FFR. We then compared the TRFs between the condition in which the corresponding speaker was attended to the condition where they were ignored.

Participants. We recruited 22 healthy, right-handed, native German speakers (10 females, 12 males; age range, 19–29 years) with no history of neurologic disease or hearing impairment. The study was granted ethical permission by the ethics board of the University Hospital Erlangen (Registration 22–361-S).

Speech stimuli. The participants listened to ~40 min of acoustic stimuli consisting of two German audiobooks (referred to as "story-audiobooks") narrated by two competing male speakers. Two additional German audiobooks (referred to as "noise-audiobooks") of the same narrators were used to serve as background noise. The first story-audiobook was "*Frau Ella*" by Florian Beckerhoff, and the second story-audiobook was "*Den Hund überleben*" by Stefan Hornbach. As the first noise-audiobook we used "*Darum*" by Daniel Glattauer, and as the second noise-audiobook, "*Looking for hope*" by Colleen Hoover (translated to German by Katarina Ganslandt). The first story-audiobook and the first noise-audiobook were narrated by Peter Jordan, and the second story-audiobook and the second noise-audiobook were narrated by Pascal Houdus. All audiobooks were published by Hörbuch Hamburg and are available in stores.

Peter Jordan's voice had, on average, a lower pitch (LP) than the voice of Pascal Houdus (Fig. 1*b*). In particular, the fundamental frequency of Peter Jordan's voice varied between 70 and 120 Hz, whereas that for Pascal Houdus occurred between 100 and 150 Hz. We therefore refer to Peter Jordan as the "LP speaker" in the following, and to Pascal Houdus as the "higher-pitch (HP) speaker."

The stimuli containing the two competing speakers were presented diotically, at approximately the same sound-pressure level of 67 dB(A) (A-weighted decibels) throughout the experiment. To facilitate the listening process for the participants, we kept the original chapters of the story-audiobooks and divided them into segments of lengths according to the original chapter lengths. The resulting chapter lengths were between 3 and 5 min long. The background noise was generated by randomly picking audio segments from the noise-audiobooks that were of the same length as the chapters from the story-audiobook. In total, we used ~37 min of acoustic stimuli.

MEG data acquisition. Two speech stimuli were presented simultaneously, with the first chapter of the first story-audiobook played alongside an unrelated story narrated by the speaker of the second noise-audiobook in the background. Subsequently, the first chapter of the second story-audiobook was played alongside an unrelated story narrated by the speaker of the first noise-audiobook in the background. This pattern continued so that both story-audiobooks were told in a subsequent but alternating manner.

To assess the impact of selective attention on the cortical response, the participants were instructed to selectively attend only one of the two acoustic stimuli, the story-audiobook (Fig. 1*a*). In detail, the participants switched their attention between the two speakers with every chapter, starting by attending the LP speaker for the first chapter of the first audiobook, then attending the HP speaker in the following, for the first chapter of the second audiobook, then again attending the LP speaker for the second chapter of the first audiobook, and so on. To instruct the participants about which speaker they should attend, the story-audiobook always started alone and ~5 s later the noise-audiobook started. After each chapter, before attention switched, participants were visually presented with three single-choice questions, each of which had four response options, to assess whether they correctly attended the intended speaker. The total stimulation protocol lasted ~50–55 min.

MEG data were recorded using a 248 MEG system (4D-Neuroimaging) with a sampling frequency of 1017.25 Hz. The participants were in a supine position with their eyes open during the recordings. An analog bandpass filter (1.0–200 Hz) was applied online to eliminate unwanted frequency components. To correct for environmental noise, a calibrated linear weighting of 23 reference sensors (manufacturing algorithm, 4D-Neuroimaging) was applied, and five landmark positions were recorded using an integrated digitizer (Polhemus). Additionally, before each measurement, head shape digitization was performed. For further analysis, a 50 Hz notch filter (firwin; transition bandwidth, 0.5 Hz) was applied offline using MNE-Python (Gramfort et al., 2014) to remove power line interference, and the data were downsampled to a sampling frequency of 1000 Hz to facilitate subsequent processing and underwent offline digital bandpass filtering in the approximate range of the fundamental frequencies of the two speakers (between 70 and 120 Hz for the LP voice, between 100 and 150 Hz for the HP voice; linear digital Butterworth filter, second-order, critical frequencies obtained by dividing the lower and upper cutoff frequency by the Nyquist frequency, applied forward and backward).

The speech signal was presented during the MEG recordings using a custom-designed setup, as described in detail in our previous study on linguistic responses (Schilling et al., 2021) and in our previous study on the speech-FFR (Schüller et al., 2023). The setup involved a stimulation computer connected to an external USB sound device with five analog outputs. Two of these outputs were connected to an audio amplifier. The first output was linked in parallel to an analog input channel of the MEG data logger, recording the mixed audio stimulus as presented to the subject. To achieve precise alignment between the speech stimulus and the MEG recording, we used cross-correlation of the speech stimulus with the audio reference recording obtained from the analog input channel of the MEG data logger, yielding an alignment accuracy of 1 ms.

Acoustic stimulus representations. We used the following two acoustic speech features to capture the high-frequency neural response at the fundamental frequency of the speech signals: the fundamental waveform $f_0(t)$ and the higher-mode envelope modulation $e(t)$ (Fig. 1*c*). The fundamental waveform was extracted using the probabilistic YIN (pYIN) algorithm (Mauch and Dixon, 2014). As an adaptation of the YIN algorithm (De Cheveigné and Kawahara, 2002), pYIN is a method used for the estimation of the fundamental frequency (f_0). By integrating the YIN algorithm and Viterbi decoding (Forney, 1973), pYIN enhances the accuracy and reliability of the f_0 estimation. The use of YIN initially generates a set of potential f_0 candidates, whereas the subsequent application of Viterbi decoding refines these candidates, producing a more precise estimation of the f_0 contour. This combined approach provides an effective methodology for extracting fundamental frequency information from audio signals.

We extracted the range of the f_0 for each chapter of the LP audiobook separately, resulting in slightly different f_0 ranges between the chapters. For each chapter, we accordingly used slightly different corner frequencies for the f_0 filtering, corresponding to the range of the fundamental frequency in the particular chapter. The lower corner frequencies were between 65 and 70 Hz, and the upper corner frequencies were between 115 and 125 Hz. The same procedure was followed for the HP audiobook, yielding lower corner frequencies between 95 and 100 Hz and upper corner frequencies between 140 and 155 Hz. These estimated frequency bands matched well with the fundamental frequency histograms obtained for both speakers (Fig. 1*b*).

As already demonstrated in previous studies (Kulasingham et al., 2020; Kegler et al., 2022; Schüller et al., 2023), the envelope modulation of the higher harmonics in an acoustic signal contributes to the neural response at the fundamental frequency even more than the fundamental frequency itself. To describe these envelope modulations, we used a computational model of the auditory periphery that draws inspiration from both psychoacoustics and neurophysiology. It aims to replicate and simulate the fundamental processes and mechanisms underlying auditory perception by taking into account the tonotopic organization of the cochlea, the frequency-tuning properties of auditory-nerve fibers, and the neural responses observed in various auditory nuclei along the ascending pathway. The model was originally implemented within the NSL auditory-cortical toolbox in MATLAB (Chi et al., 2005). For this study, we used the parts of this toolbox that describe the early auditory processing. The MATLAB code was translated to Python within our group.

In detail, the model first used a bank of constant-Q bandpass filters; that is, specialized filters that have a varying frequency resolution across the auditory spectrum. These filters were followed by nonlinear compression and derivative mechanisms implemented across different scales, yielding a sharpening of frequency resolution. We then considered the 86 frequency bands for each speaker that corresponded to the higher

harmonics >300 Hz. For each of these frequency bands, we computed the amplitude envelope and then bandpass filtered the envelopes in the range of the fundamental frequency, between 70 and 120 Hz for the LP speaker, and between 100 and 150 Hz for HP speaker. The bandpass filtered envelopes were then averaged across the 86 frequency bands, resulting in a single signal, the high-frequency envelope modulation feature.

Neural source estimation. The source reconstruction process was implemented using the MNE-Python software package (Gramfort et al., 2014). Given the absence of subject-specific MR scans, we used the FreeSurfer template MRI called "fsaverage" as a substitute (Fischl, 2012). Previous studies have shown that using an average brain template can yield results comparable to individual MR scans in source localization analyses (Holliday et al., 2003; Douw et al., 2018). Its validity has not only been established in our recent study on early subcortical MEG responses to continuous speech (Schüller et al., 2023), but also in previous studies that investigated high-frequency neural mechanisms of speech processing (Kulasingham et al., 2020).

To account for individual differences, we collected information about the head position of each subject with respect to the MEG scanner at the beginning and end of each measurement, using five marker coils. Additionally, we digitized the shape of each subject's head using a Polhemus system. The so-obtained subject-specific information was then used to align the fsaverage brain template with each individual's head by applying rotation, translation, and uniform scaling.

To create a volumetric source space for the average brain, we used a regular grid with neighboring grid points spaced at intervals of 5 mm. This volume source space was subjected to the application of the FreeSurfer 'aparc+aseg' parcellation, to subsequently define a specific region of interest (ROI) for source estimation.

We established a cortical ROI, representing the auditory cortex in both the right and left hemispheres. Specifically, the "aparc" labels "transversetemporal," "middletemporal," "superiortemporal," "supramarginal," "insula," and "bankssts" were used to identify the cortical ROI. This cortical subdivision resulted in a total of 525 source locations with arbitrary orientations.

To generate a volume conductor model that accurately represents the shape of the head for source reconstruction, even in the absence of subject-specific MR scans, we used the boundary element model provided by FreeSurfer for the fsaverage brain template. Using the volume source space and the forward solution, we computed a linearly constrained minimum variance (LCMV) beamformer (Bourgeois and Minker, 2009). The LCMV beamformer is a spatial filter that applies a set of weights to scan through each source location in the predefined source space grid. It independently estimated the MEG activity at each source point.

To perform the source reconstruction, we used a data covariance matrix estimated from a 1 min segment of MEG data acquired during audio stimulation, as well as a noise covariance matrix derived from 3 min prestimulus empty room recordings. The beamformer filter was then applied to the preprocessed MEG data of each subject. This process resulted in the estimation of a three-dimensional current dipole vector, with magnitude and direction, at each source location. We thus source reconstructed the preprocessed MEG data of every subject.

Temporal response functions. To investigate the latency and origin of the measured neural signal, we used a linear forward model that aimed to predict the neural activity $y_t^{(v)}$ of each source point (voxel) v at time t from a linear combination of the acoustic stimuli f_t and e_t. The acoustic stimuli were shifted by different time delays τ. The resulting weights $\alpha_\tau^{(v)}$ and $\beta_\tau^{(v)}$ of this linear equation are the TRFs, one for each source point. A TRF provides a quantitative representation of how the output system changes over time because of changes in the input. Thus, the TRF for each voxel describes the corresponding neural response to each acoustic feature across a range of time delays, from $\tau_{\min}$ to $\tau_{\max}$ as follows:

$$y_t^{(v)} = \sum_{\tau=\tau_{\min}}^{\tau_{\max}} (\alpha_\tau^{(v)} f_{t-\tau} + \beta_\tau^{(v)} e_{t-\tau}).$$

We used regularized ridge regression to estimate the TRFs. In this approach, the regularization parameter λ can be defined as $\lambda = \lambda_n \cdot e_m$, where λ_n represents the normalized regularization parameter, and e_m corresponds to the mean eigenvalue of the covariance matrix. After applying fivefold cross-validation to estimate the appropriate regularization parameter for each subject, we found that the same regularization parameter $\lambda = 1$ could be used for all subjects, since the highest model accuracy appeared close to this value for all subjects. The forward model and the TRF estimation used in this study were implemented in Python based on the algorithms developed previously in our group by Etard et al. (2019) and Kegler et al. (2022).

We considered a range of time delays of $\tau_{\min} = -20$ ms to $\tau_{\max} = 120$ ms with an increment of 1 ms since the sampling rate was 1000 Hz. This resulted in 141 time lags in total. We calculated both of the voxel-wise TRFs on the subject level to capture subject-specific diversity, as well as TRFs on the population average to provide representative results as well as statistical inference on population level.

Attentional modulation of the cortical response. To investigate the impact of attention on the cortical contribution to the speech-FFR, we computed two pairs of TRFs for each acoustic feature and each subject.

The first pair of TRFs represented the neural response to the LP speaker. One TRF was estimated when the LP voice was attended by the subject (referred to as the LP-A condition), while the other TRF was constructed when the LP was ignored (referred to as the LP-I condition).

Similarly, the second pair of TRFs was designed to represent the neural response to the HP speaker. The first TRF in this pair was computed when the subject directed their attention to the HP voice, referred to as the HP-A condition. The second TRF in the pair was created when the subject ignored the HP speaker, referred to as the HP-I condition.

Previous studies found that the cortical portion of the speech-FFR, as assessed through TRF amplitude, occurs predominantly at delays between 30 and 40 ms (Coffey et al., 2016; Kulasingham et al., 2020; Schüller et al., 2023). Since maxima and minima in the TRFs are both mapped to maxima when computing the absolute amplitude, the maxima in the TRF amplitudes occur at a rate of twice the fundamental frequency. Several maxima and minima therefore emerge in the TRF amplitude between 30 and 40 ms. Since the latencies between subjects may slightly vary in a range of a few milliseconds, it is likely that at a time lag of 34 ms, for instance, one subject might display a maximum in the TRF amplitude and another subject a minimum. A direct comparison between the amplitudes of the TRFs in the attended and ignored conditions at one specific time lag may hence lead to inaccurate results. To avoid this problem, we computed the envelopes of the TRF amplitudes. The envelope represents the magnitude of the signal over time, capturing the overall modulation pattern without relying on precise temporal alignment. Using envelope comparisons allowed us to capture the essential characteristics of the cortical response to different speakers while minimizing the impact of small temporal variations. This approach provided a more stable and interpretable basis for evaluating the attentional effects on the cortical response.

Significance of the cortical response. To assess the statistical significance of the neural responses at the population level, we conducted statistical tests by comparing the calculated TRFs to noise models. The noise models were generated for each subject by reversing the acoustic feature in time. Because of the disparity between the reversed acoustic feature and the MEG signal, the noise models were not able to identify a significant brain response at any time lag.

To evaluate the statistical significance of the TRFs, we used a bootstrapping approach with the single-subject noise models for each audio feature. This process involved resampling the noise models across time lags and subjects, averaging them across subjects and vertices, and computing magnitudes over time lags in a manner consistent with the actual TRFs. By repeating this procedure 10,000 times, we generated a distribution of noise model magnitudes across time lags. We then determined the proportion of values from this noise distribution that exceeded the magnitude of the actual TRF for each model. This allowed us to estimate empirical p-values for each time lag. To account for multiple comparisons, the estimated p-values were corrected using the Bonferroni method.

Lateralization of the cortical responses. To investigate the potential lateralization of cortical activity at time delays where significant

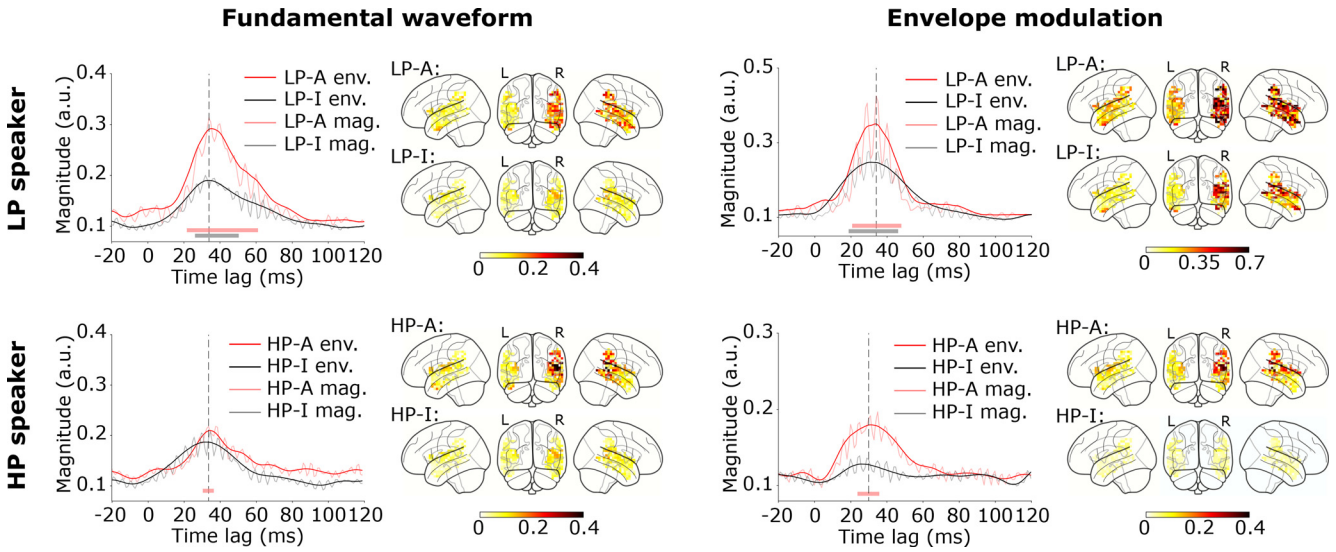

Figure 2. Cortical responses to both acoustic features of the LP and the HP voice. The voxel-averaged and subject-averaged envelopes of the TRF amplitudes are significant at time lags of ~34 ms (dashed gray lines) for both features and both in the LP-A condition (LP-A env., red) and in the LP-I condition (LP-I env., black). The TRF magnitudes display oscillation at $2f_0$ and are displayed as well (LP-A mag., pink; LP-I mag., gray). Significant time delays are indicated through the colored bars at the bottom of the plots. The corresponding brain plots show right-lateralized activity at 34 ms (attended, top row; ignored, bottom row). The same applies for the response to the fundamental waveform in the HP-A and HP-I conditions. The envelope modulation for the HP voice shows significant activity at time lags of ~30 ms (dashed gray line) when the speaker was attended, but not when he was ignored.

responses emerged, we conducted a two-tailed Wilcoxon signed-rank test. This test focused on assessing the differences in magnitudes of the population average TRFs between the right and left cortical ROIs at the time delays of significant responses.

Significance of the attentional modulation. To assess potential significant differences in the envelopes of the TRF magnitudes between the attended and the ignored conditions for individual subjects, we conducted a two-tailed Mann–Whitney rank test at the subject level. We therefore split the source-reconstructed MEG data as well as the acoustic stimuli into 10 segments of equal length (these segments were thus not the same as the audiobook chapters) and calculated for each of them the TRF amplitude and the corresponding envelope. We then extracted the envelope value of each split-TRF at a certain latency time, which was chosen based on the peak of the population-average TRF magnitude. We thus generated a distribution of magnitude values for each subject and each acoustic feature, which was then used for the statistical testing.

We additionally used a two-tailed Wilcoxon signed-rank test to investigate whether the population-average envelope of the TRF magnitude of each feature deviated significantly between the attended and the ignored condition.

Finally, we used a two-tailed Wilcoxon signed-rank test to examine whether the population-average TRFs obtained for the LP speaker and the HP speaker differed significantly.

Results

Evaluation of comprehension questions

To test whether the participants attended the target speaker, we presented three single-choice questions (in total, 15 questions per audiobook narrator) at the end of each chapter. Each question had four response options, resulting in a 25% chance level. Regarding the LP speaker, the questions were answered with an accuracy of 81±3% (mean and SEM). The response accuracy for questions related to the HP speaker was comparable at 80 ± 2%.

Cortical responses to the LP and the HP speaker

We measured neural responses to two competing male speakers, with distinct, but partly overlapping, fundamental frequencies (Fig. 1*b*). The participants were asked to switch the attention from the LP to the HP speaker and back after each chapter of the corresponding audiobook. The recorded MEG data were then analyzed separately, first by source reconstructing the preprocessed MEG signals and subsequently calculating source-level TRFs in a cortical ROI.

We then computed four linear forward models, each of which contained two acoustic features, the fundamental waveform and the envelope modulation. The model for the first condition, LP-A, captured the neural responses when the lower-pitch speaker was attended, while the model for the second condition, LP-I, represented the response to the ignored lower-pitch speaker. The models for the third and fourth condition, HP-A and HP-I, were calculated analogously to describe the neural responses when the HP speaker was attended or ignored, respectively.

We first verified that, for both acoustic features, the population-average TRFs showed significant neural activity in the auditory cortex. For the fundamental waveform, the TRF showed significant activity between 21 and 61 ms, peaking at 35 ms ($p < 0.0001$), when the LP speaker was attended and between 26 and 50 ms, peaking at 33 ms ($p < 0.0001$), when the LP voice was ignored (Fig. 2, top left). The source activation at the peak time lags in both the LP-A condition and the LP-I condition showed a right lateralization (LP-A, $p = 5.5 \cdot 10^{-12}$; LP-I, $p = 1.2 \cdot 10^{-7}$).

For the envelope modulation, the population-average TRFs for the LP-A condition yielded significant cortical activity at time lags between 21 and 47 ms, peaking at 34 ms ($p < 0.0001$). Regarding the LP-I condition, the TRFs showed significant activity at delays ranging from 19 to 46 ms, peaking at 31 ms ($p < 0.0001$; Fig. 2, top right). As for the fundamental waveform, the source activation also revealed a right-lateralized dominance at the peak time lags in both the LP-A and the LP-I condition (LP-A, $p = 1.5 \cdot 10^{-8}$; LP-I, $p = 2.3 \cdot 10^{-13}$).

For the HP voice, the envelopes of the TRF amplitudes showed a similar behavior. When attended, the TRF for the fundamental waveform yielded a significant response between 32 and 40 ms, peaking at 35 ms (Fig. 2, bottom left; $p < 0.0001$). The TRF in the HP-I condition showed no significant time lags. The source activation for the HP-A condition emerged similarly to the LP voice and was right lateralized at the peak time lags (HP-A, $p = 5.2 \cdot 10^{-8}$).

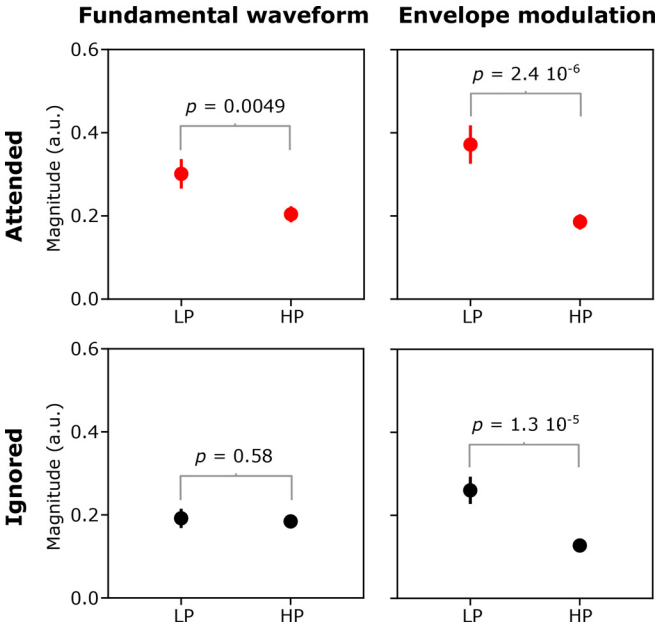

Figure 3. Comparison of the responses to the LP voice to the responses to the HP speaker. We show the peak of the envelope of the TRF magnitudes. The responses to the fundamental waveform (top left) and to the envelope modulation (top right) were significantly higher for the LP voice than for the HP voice when these voices were attended (two-tailed Wilcoxon signed-rank test). When the voices were ignored, the difference between the responses to the LP voice and the HP voice were significant for the envelope modulation (bottom left), but not for the fundamental waveform (bottom right, two-tailed Wilcoxon signed-rank test). The error bars indicate the SEM when averaging across subjects.

For the envelope modulation of the HP speaker, a significant response only emerged when the HP voice was attended, in a range from 24 to 36 ms, peaking at 35 ms ($p < 0.0001$; Fig. 2, bottom right). The TRF in the HP-I condition did not show any significant time lags. The source activation in the HP-A condition at the peak time lags exhibited a right lateralization (HP-A, $p = 1.9 \cdot 10^{-6}$).

Neural responses for individual subjects in each of the four conditions LP-A, LP-I, HP-A, and HP-I will be presented below.

Differences between responses to the LP voice and the HP voice

We wondered whether the neural responses to the two speakers differed, independent of any putative attentional modulation. We hence investigated whether the neural response to the LP speaker was systematically higher or lower than that to the HP speaker.

To assess those differences in the responses to the LP compared with the HP voice for each acoustic feature, we extracted from each subject the envelope of the TRF magnitude at the latency of interest (Fig. 2, dashed gray lines). We then applied a two-tailed Wilcoxon signed-rank test to assess whether there was a significant difference in the TRF amplitudes between the LP and the HP speakers. We did this separately for the attended and the ignored conditions.

Regarding the response to the fundamental waveform, we found a significantly higher envelope of the TRF amplitude for the LP speaker than for the HP speaker when the respective speaker was attended ($p = 0.0049$; Fig. 3, top left), but not when it was ignored ($p = 0.58$; Fig. 3, bottom left).

For the response to the envelope modulation, the comparison yielded a significant difference when both speakers were attended ($p = 2.4 \cdot 10^{-6}$; Fig. 3, top right), as well as when both speakers were ignored ($p = 1.3 \cdot 10^{-5}$; Fig. 3, bottom right). In each case, the neural response to the LP voice was higher than that to the HP voice.

Attentional modulation of the cortical contribution to the speech-FFR: response to the LP speaker

To investigate how attention modulates the cortical response, we used a classic paradigm of auditory selective attention paradigm, where participants attended one of two competing speech signals while ignoring the other one. We then assessed the neural responses to both speakers, and compared the responses to the same speaker when they were attended to the condition in which they were ignored.

To quantify the neural responses, we extracted the peak value of each of the four population-averaged envelopes of the TRF magnitudes (Fig. 2, dashed gray lines). For the statistical analysis, we moreover calculated 10 split-TRFs for 10 shorter segments for each subject and each condition (LP-A, LP-I, HP-A, and HP-I), extracted the value of the corresponding envelope of the TRF magnitudes at the prior extracted latency time of interest, and applied a two-tailed Mann–Whitney rank test on the difference between the values obtained for the attended condition and the ignored condition. This was done for the responses to the LP voice and the HP voice separately. It is important to note here that Figures 4*a*, 5*a*, 6*a*, and 7*a* all show the values of the attended and ignored TRF envelopes that resulted from averaging across the 10 split-TRFs for each subject. In contrast, Figures 4*b*, 5*b*, 6*b*, and 7*b* provide the TRFs and corresponding envelopes for assorted subjects, which were calculated by taking all data of one subject and not the average of the split-TRFs.

Regarding the responses to the fundamental waveform of the LP voice, we found that for most of the subjects (17 of 22) the envelope of the TRF magnitude at the latency time of interest, 34 ms, when the speaker was attended, significantly exceeded the envelope when the speaker was ignored (Fig. 4*a*, two-tailed Mann–Whitney rank test). On the population level, the envelope of the TRF magnitudes in the LP-A condition was also significantly higher than in the LP-I condition ($p < 0.001$, two-tailed Wilcoxon signed-rank test; Fig. 2, top left). In particular, the neural response in the attended condition was 26±3% higher than in the ignored condition (mean ± SEM).

To investigate the time course of the envelope of the TRF magnitudes for individual subjects, as well as the location of the neural sources in the brain, we display this information for two exemplary subjects (Fig. 4*b*). Subject 9 exhibits significant attentional modulation at the time lag of 34 ms as well as at earlier and later time lags. The neural activity appears predominantly in the primary auditory cortex and is much smaller when the speaker is ignored. Subject 3 shows a much smaller difference in the neural responses between the attended and the ignored condition.

The subject-specific neural responses to the envelope modulation, at a delay of 34 ms, revealed a similar behavior, with 14 of 22 subjects displaying a larger neural response in the LP-A condition than in the LP-I condition. The population-average neural response exhibited this behavior as well (Fig. 5*a*, $p < 0.001$, two-tailed Wilcoxon signed-rank test). It was 20±2% higher in the attended than in the ignored condition (mean ± SEM).

Figure 5*b* presents further data on the neural response for the same two exemplary subjects as in Figure 4*b*. Subject 9 again showed a significant difference between the LP-A and the LP-I conditions, whereas subject 3 did not. Compared with the neural

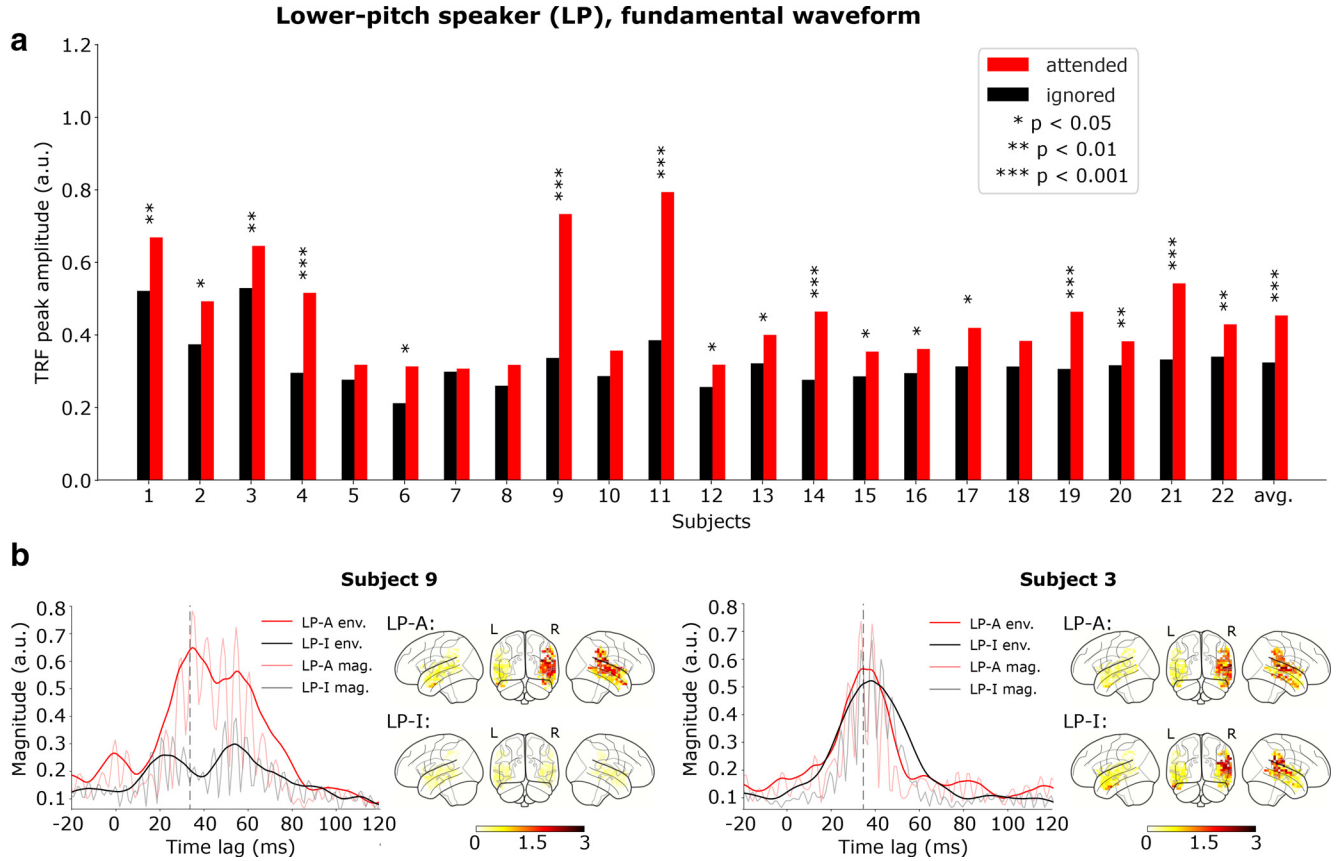

Figure 4. Responses to the fundamental waveform of the LP speaker. ***a***, Attentional modulation of the cortical contribution to the speech-FFR. For 17 of 22 subjects, the peak in the envelope of the TRF magnitudes at a delay of 34 ms showed a significant difference between the attended (red) and the ignored (black) condition (*, $0.01 \leq p < 0.05$; **, $0.001 \leq p < 0.01$; ***, $p < 0.001$). The same behavior emerged regarding the population-average response (avg.). ***b***, Cortical TRFs and corresponding voxel magnitudes for the LP-A condition (LP-A env., red; LP-A mag., pink; top brainplots) and the LP-I condition (LP-I env., black; LP-I mag., gray; bottom brainplots) speaker for the exemplary subject 9 (left) and subject 3 (right). There is a large effect of attention for subject 9 and a much smaller one for subject 3.

response to the fundamental waveform (Fig. 4*b*), subject 9 exhibited a less noisy signal, especially in the LP-I condition.

Regarding the population-average neural responses to the LP speaker (Fig. 2, top right), in both the LP-A condition and in the LP-I condition the response to the envelope modulation showed less noise and a higher magnitude and voxel activation than the response to the fundamental waveform (Fig. 2, top left).

Attentional modulation of the cortical contribution to the speech-FFR: response to the HP speaker

We applied the same analysis to investigate the attentional modulation of the neural response to the HP voice. We found that, for the response to the fundamental waveform, only a minority of subjects displayed significant attentional modulation (7 of 22; Fig. 6*a*). For six subjects, the response in the HP-A condition significantly exceeded the one in the HP-I condition. For subject 22, however, the response in the HP-I condition was significantly larger than that in the HP-A condition (Fig. 6*b*), although both TRFs were noisy. Exemplary subject 7 yielded noisy TRFs both in the HP-A and in the HP-I conditions with no significant attentional modulation (Fig. 6*b*).

The population average (Figs. 2, 6*a*, avg.) displayed a significant cortical response at a delay of 34 ms as well as a significant effect of attention ($p < 0.01$, two-tailed Wilcoxon signed-rank test). The response in the attended condition exceeded the one in the ignored condition by 10±3% (mean ± SEM).

The neural responses to the envelope modulation, at a peak delay of 30 ms, revealed a similar pattern. Among the 22 subjects, 6 exhibited a significant difference when comparing cortical responses in the HP-A condition to those in the HP-I condition.

The population average of the cortical response displayed a clear peak centered at 30 ms in the HP-A condition, but no clear peak in the HP-I condition (Fig. 2, bottom right). It was significantly larger in the HP-A condition than in the HP-I (Fig. 7*a*, $p < 0.01$, two-tailed Wilcoxon signed-rank test). The difference in the neural response between the two conditions was 8±4% (mean ± SEM).

Regarding the two exemplary subjects 7 and 22, subject 22 showed a significantly smaller cortical response when the HP speaker was attended compared with when he was ignored (Fig. 7*b*). Subject 7, in contrast, did not display a significant effect of attention.

Correlation between attention and participant behavior

The attentional modulation of the cortical contribution to the speech-FFR that we observed was, on a qualitative level, relatively consistent across the individual subjects. However, the amount of the attentional modulation differed considerably between the participants. We wondered whether this variability could be explained partly by the performance of the subjects in the comprehension questions, and thus be linked to behavior.

We therefore investigated whether there was a correlation between the participants' task performance, specifically their

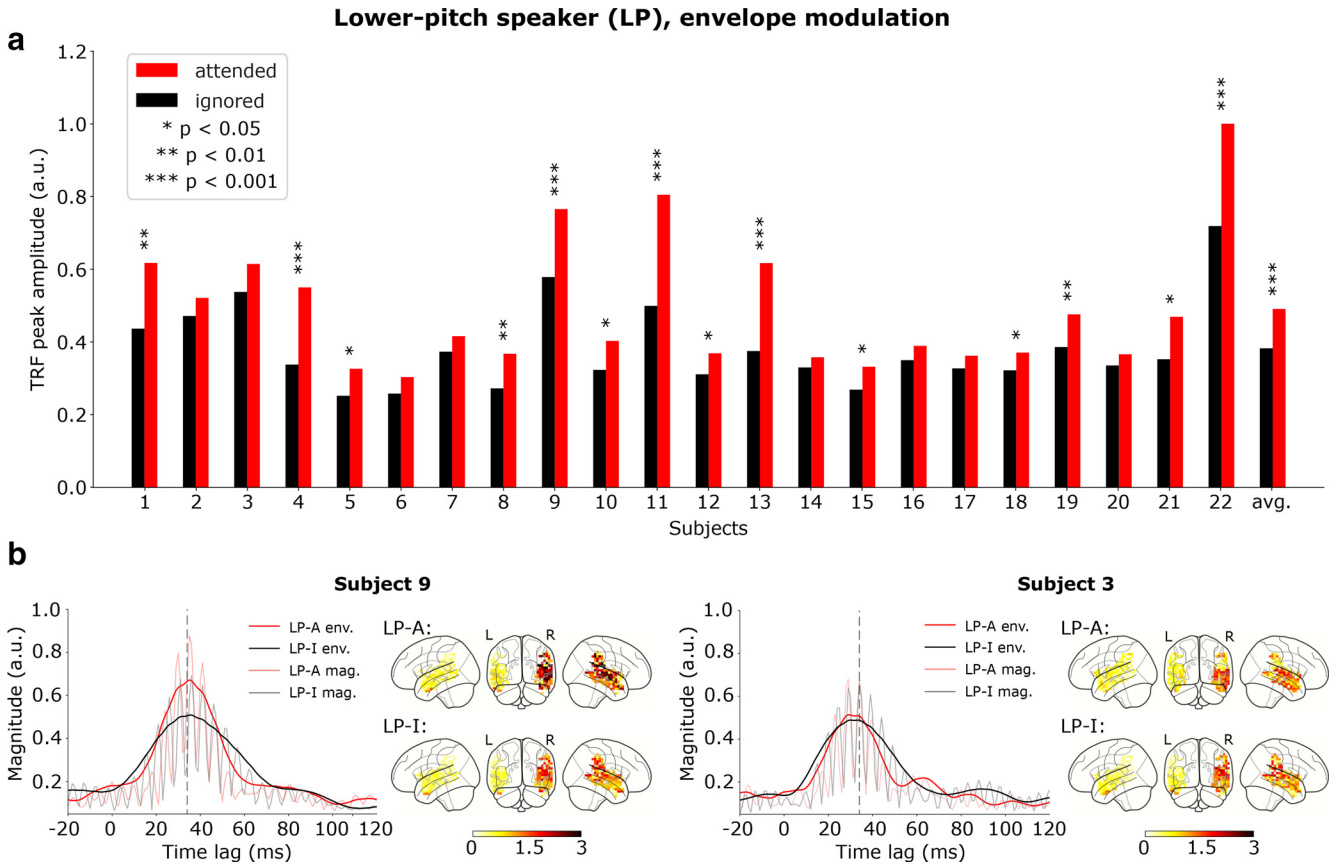

Figure 5. Responses to the envelope modulation for the LP speaker. ***a***, Attentional modulation of the cortical contribution to the speech-FFR. For 14 of 22 subjects, the envelope of the TRF magnitude at a delay of 34 ms (peak of the envelope) showed a significant difference between the attended condition (red) and the ignored condition (black; *, $0.01 \leq p<0.05$; **, $0.001 \leq p<0.01$; ***, $p<0.001$). The population-average TRF (avg.) shows the same attentional modulation. ***b***, The cortical TRFs and the corresponding voxel magnitudes for the LP-A condition (LP-A env., red; LP-A mag., pink; top brainplots) and the LP-I condition (LP-I env., black; LP-I mag., gray; bottom brainplots) for exemplary subject 9 (left) and subject 3 (right). The channel-averaged TRFs for subject 9 show a strong attentional modulation, while the one for subject 3 is insignificant.

percentage of correct answers, to the amount of attention modulation observed in their speech-FFR. We quantified the attentional modulation through the attentional modulation index I that we defined as follows:

$$I = \frac{A_{att} - A_{ign}}{A_{att} + A_{ign}}, \quad (1)$$

where A_{att} refers hereby to the neural response in the attended condition, and A_{ign} to that in the ignored condition.

Each participant listened to five chapters of each audiobook, after each of which three single-choice questions with four response options were presented. The total number of responses per audiobook and participant was hence 15, and the chance level was 25%. We therefore calculated a correct answer score for each subject, which reflected the percentage of correctly answered questions per audiobook. We then correlated the correct answer score with the attentional modulation index I of the fundamental waveform responses (Fig. 8, left) and of the envelope modulation responses (Fig. 8, right), respectively, for both audiobook speakers (Fig. 8: LP, blue dots; HP, brown squares), using Pearson's correlation coefficient.

We found that all participants answered the questions with a score far exceeding the chance level of 25%. However, there were no significant correlations between the attentional modulation index I and participant behavior, neither for the fundamental waveform of either speaker (LP: $r = 0.04$, $p = 0.87$; HP: $r = -0.16$, $p = 0.49$) nor for the envelope modulation (LP: $r = 0.2$, $p = 0.37$; HP: $r = -0.09$, $p = 0.69$).

Discussion

Our study investigated the attentional modulation of the cortical contributions to the speech-FFR, using MEG recordings. We therefore used a well established paradigm in which participants focused selectively on one of two competing voices. We then examined the cortical contribution to the speech-FFR through two acoustic features, the fundamental waveform and the envelope modulation. The neural responses were computed through source estimation followed by ridge regression that related the source-reconstructed neural activities to the two speech features. We then assessed the effect of selective attention on these two neural responses.

Our results first verified that both the fundamental waveform and the envelope modulation elicited significant cortical responses. We identified high neural activation in the auditory cortex region for both acoustic features, at latencies ranging from 30 to 35 ms. These neural responses were observed despite the presence of a competing speaker and were presented both for the attended and the ignored voice. These findings are in line with previous research using MEG, where cortical contributions to the speech-FFR in response to short speech tokens were observed (Coffey et al., 2016). In two recent studies focusing on cortical and subcortical contributions to the speech-FFR elicited by continuous speech, cortical responses within similar

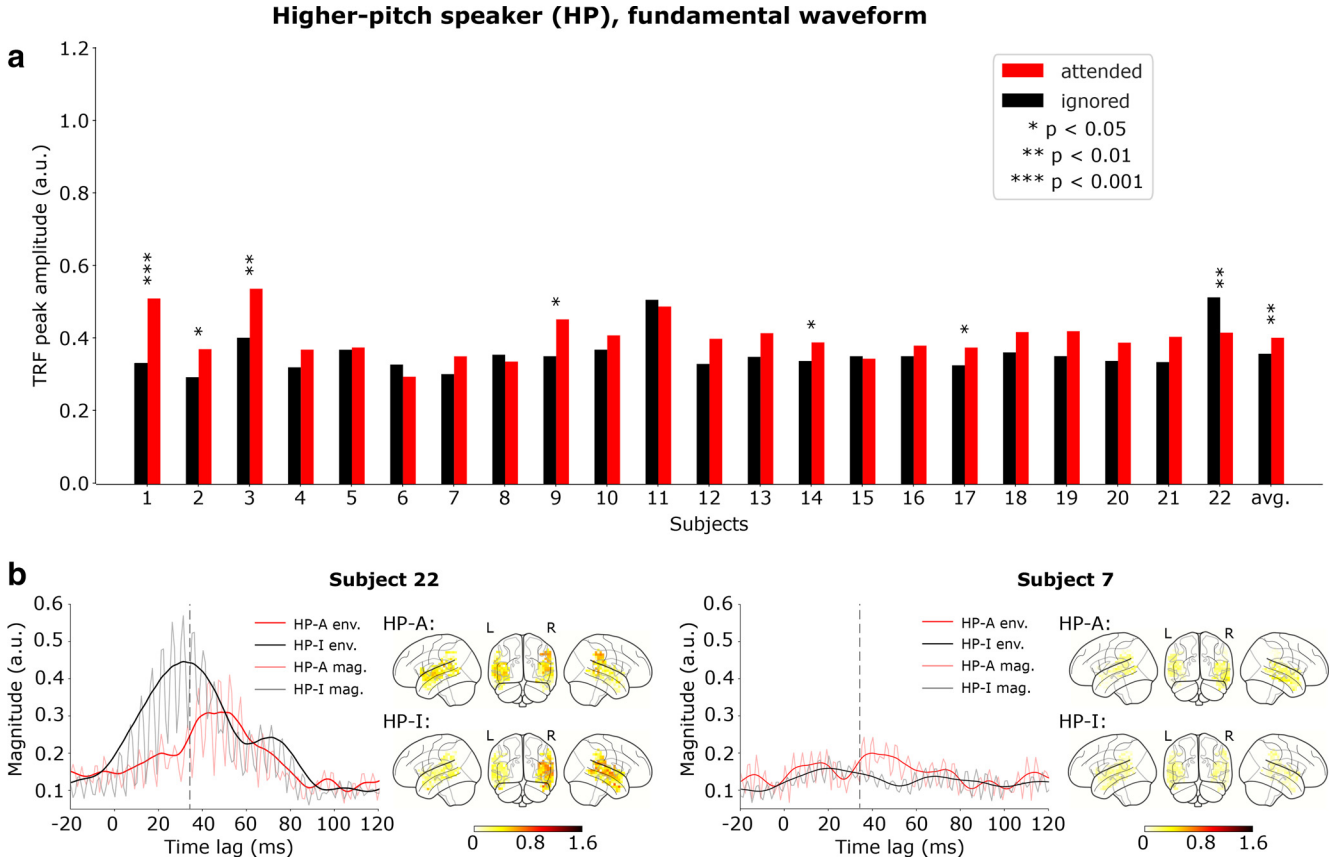

Figure 6. Cortical responses to the fundamental waveform of the HP speaker. ***a***, Attentional modulation of the cortical response. For 7 of 22 subjects, the peak envelope of the TRF magnitude, at a delay of 34 ms, showed a significant difference between the attended (red) and the ignored (gray) conditions (∗, $0.01 \leq p < 0.05$; ∗∗, $0.001 \leq p < 0.01$; ∗∗∗, $p < 0.001$). The population-average TRF (avg.) displayed the attentional effect as well. ***b***, The time course of the TRF magnitudes and the corresponding envelopes as well as the corresponding voxel magnitudes for the HP-A condition (HP-A env., red; HP-A mag., pink; top brainplot) and the HP-I condition (HP-I env., black; HP-I mag., gray; bottom brainplots) for exemplary subject 22 (left) and subject 7 (right). The cortical response of subject 22 showed a significant effect of selective attention, but not the response of subject 7.

latency ranges (30–40 ms) were reported (Kulasingham et al., 2020; Schüller et al., 2023). The subcortical contribution to the speech-FFR, in contrast, has been found to occur at time lags of up to 12 ms and is hence temporally well separated from the later cortical activity, as revealed both by EEG and MEG measurements (Forte et al., 2017; Etard et al., 2019; Schüller et al., 2023).

In contrast to many previous studies on selective attention to competing speakers, we did not use a male and a female voice, but two male voices. We could thereby demonstrate that selective attention can work also when both competing speakers have the same gender. To differentiate the two voices, we used their pitch, which was higher for one speaker (HP, ∼120 Hz) and lower for the other (LP, ∼80 Hz). We found that the responses to the LP speaker were larger than those to the HP speaker, indicating that the cortical responses to the speech-FFR decline with increasing fundamental frequency. This finding parallels previous EEG as well as modeling results on the subcortical contribution to the speech-FFR that also showed larger subcortical responses for lower-pitch voices (Saiz-Alía et al., 2019; Saiz-Alía and Reichenbach, 2020; Van Canneyt et al., 2021). It presumably reflects the decline of phase locking in the auditory pathway with increasing frequency. We note that the drop in the response between the LP and the HP speaker also matches the pattern of fine structure found in EEG-recorded FFRs (Tichko and Skoe, 2017). The observed fine structure was explained from the interference of different subcortical sources at latencies between 0 and 25 ms, which should therefore not have interfered with the later cortical response found here.

We further found, regarding the response to the LP speaker, that the envelope modulation elicited a larger response than the fundamental waveform. This result emerged both for the attended and the ignored condition, and aligns with a previous finding that the cortical contribution to the speech-FFR is dominated by the response to the envelope modulation (Kulasingham et al., 2020). Moreover, a recent EEG study found a similar behavior regarding the subcortical contribution to the speech-FFR (Kegler et al., 2022). Regarding the HP speaker, however, our results showed either a similar response to the envelope modulation and to the fundamental waveform (attended condition) or a higher response to the latter speech feature (ignored condition). This finding could reflect the weaker and thus noisier response to the HP speaker compared with the LP speaker.

The cortical response that we measured was right lateralized, which aligns with prior findings regarding MEG-measured cortical responses to short speech tokens (Coffey et al., 2016) and to continuous speech (Kulasingham et al., 2020; Brodbeck and Simon, 2022; Schüller et al., 2023). This right-lateralized pattern may reflect the important role of the right hemisphere in voice perception, as demonstrated for instance by functional magnetic resonance imaging (fMRI; Lattner et al., 2005). This right-lateralized pattern also aligns with an fMRI study involving sung stimuli by Albouy et al. (2020), where right-lateralized responses in the brain were indicated for processing speech and melody.

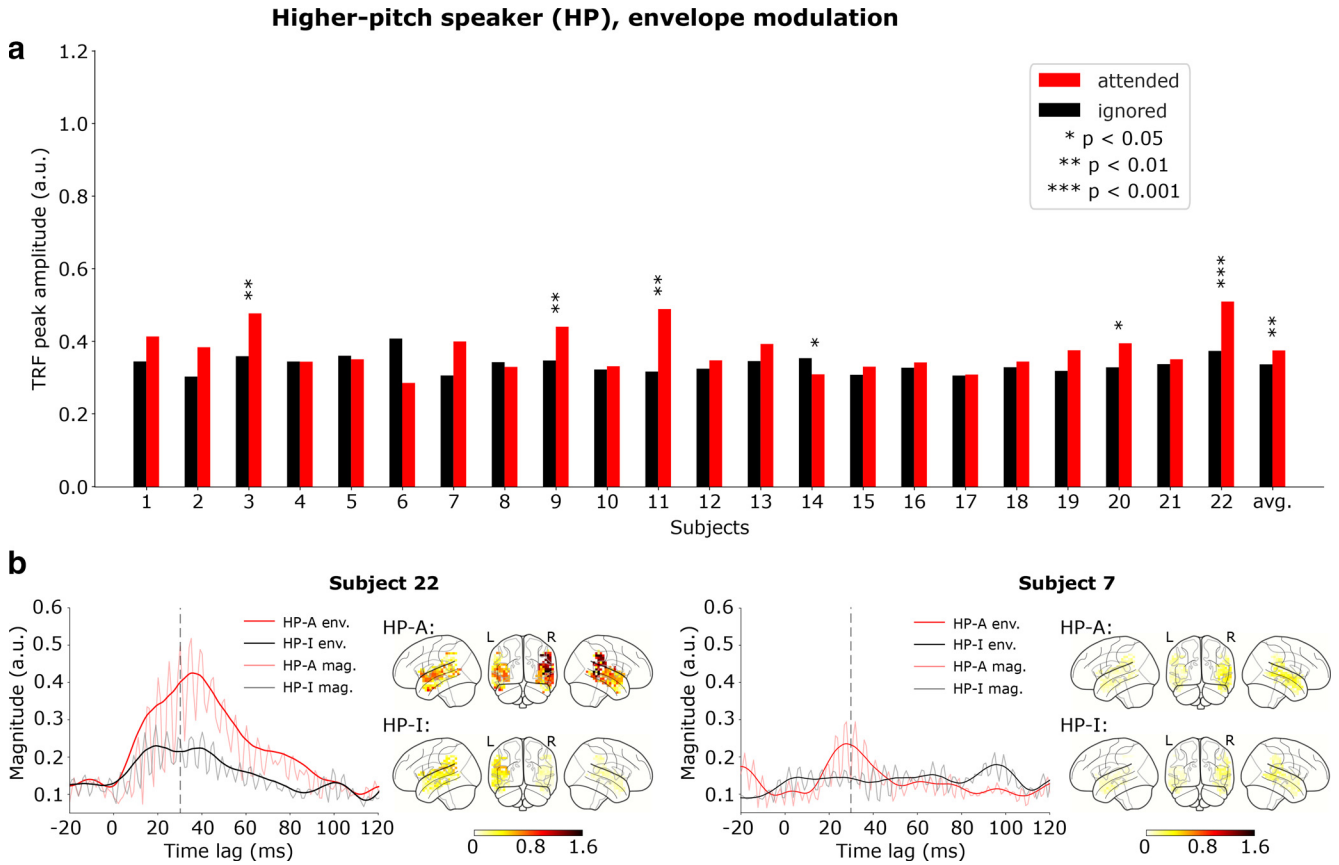

Figure 7. Cortical responses to the envelope modulation of the HP speaker. ***a***, Attentional modulation of the cortical contribution to the speech-FFR. For 6 of 22 subjects, the peak envelope of the TRF magnitude, at a delay of 30 ms, differed significantly between the attended (red) condition and the ignored (black) condition (*, $0.01 \leq p < 0.05$; **, $0.001 \leq p < 0.01$; ***, $p < 0.001$). The population average (avg.) showed the attentional modulation as well. ***b***, Envelopes of the TRF magnitudes as well as the TRF magnitudes themselves and the corresponding voxel magnitudes for the HP-A condition (HP-A env., red; HP-A mag., pink; top brainplot) and the HP-I condition (HP-I env., black; HP-I mag., gray; bottom brainplot) for exemplary subjects 22 (left) and subject 7 (right). The cortical response of subject 22 showed a significant attentional effect, whereas that of subject 7 did not.

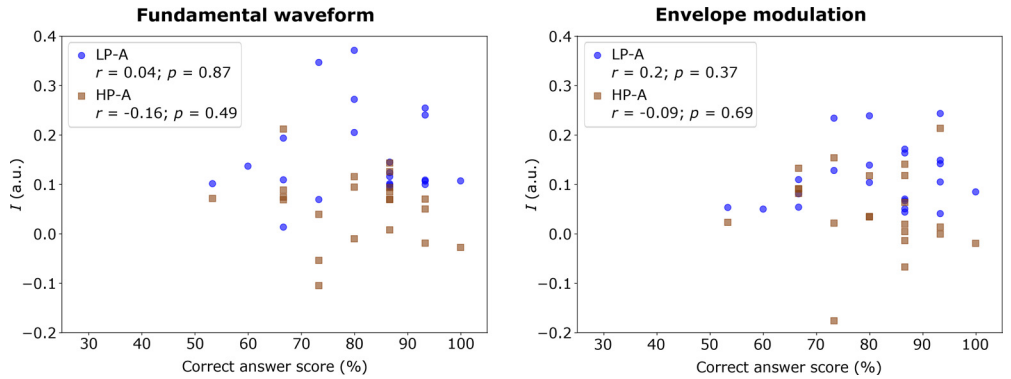

Figure 8. Correlation analysis of attention and participant behavior. No significant correlation emerged between the correct answer score of the different participants and their attentional modulation index *I*, either for the fundamental waveform feature (left) or for the envelope modulation feature (right). Significant correlations were absent both for the LP speaker (blue dots) as well as for the HP speaker (brown squares). The correlation was investigated by calculating Pearson's correlation coefficient *r* as well as the according *p*-value.

In another recent investigation on MEG responses to a single voice without background noise, we found that the cortical contribution to the speech-FFR was strongest in the area of the primary auditory cortex (Schüller et al., 2023). Here we observed localized responses near the primary auditory cortex for the HP voice, in the attended condition, while the responses to the HP voice in the ignored condition, as well as those to the LP voice, were relatively uniformly distributed across the cortical ROI. We assume that the lower signal-to-noise ratio of the response in the ignored condition as well as the background noise because of the two competing speakers obstructed a more precise source localization.

Importantly, our results showed a systematic modulation of the cortical contribution to the speech-FFR by selective attention. The clearest results thereby emerged for the responses to the LP speaker, where we observed consistent attentional effects at the peak latency of 34 ms, both for the fundamental waveform and the envelope modulation and on the level of individual subjects

as well as on the population level. The LP-A condition led to larger cortical responses than the LP-I condition in all subjects and for both acoustic features, and significant differences between the LP-A and the LP-I condition were observed in more than half of the subjects (Figs. 4*a*, 5*a*). Additionally, the neural responses to the fundamental waveform and to the envelope modulation peaked at the same latency, a behavior that was also found on the level of individual subjects (Figs. 4*b*, 5*b*). The population average analysis confirmed attentional modulation on a group level for both acoustic features, with the LP-A condition leading to significantly larger responses than the LP-I condition.

Attentional effects on the neural responses to the HP voice were less prominent than for the LP speaker. While we found significant differences between the HP-A condition and the HP-I condition, these effects were not entirely consistent across subjects. For both acoustic features, less than half of the subjects showed a significant difference between the HP-A condition and the HP-I condition (Figs. 6*a*, 7*a*). The subject-level TRFs were noisier than those for the LP speaker. In four subjects, we found larger cortical responses in the HP-I condition than in the HP-A condition. Importantly, however, the population averages nonetheless yielded a significantly larger response in the attended condition compared with the ignored condition.

The consistent attentional modulation that we found on the population level, as well as, in many instances, on the level of individual subjects, aligns with a recent study that, independent of ours, obtained similar results (Vrishab et al., 2023). The attentional modulation further aligns with our previous findings regarding attentional modulation of the subcortical contribution to the speech-FFR, as well as the possible involvement of the active process in the inner ear (Forte et al., 2017; Etard et al., 2019; Saiz-Alía et al., 2019, 2021). Moreover, the differences between the neural responses in the attended and the ignored conditions were comparable to those observed on the subcortical level. It hence remains an open question whether the attentional modulation of the cortical contribution to the speech-FFR is entirely driven by the subcortical contribution, or whether further attentional processes act on the cortical response. Recording of the speech-FFR through combined EEG and MEG may in the future be able to clarify this issue through simultaneous high-fidelity measurements of the subcortical and cortical portions of the speech-FFR.

The investigation of top-down effects is also important regarding the possible influencing of the speech-FFR by higher cortical processing. Effects of selective attention indeed become more pronounced along the auditory pathway: while acoustic features such as amplitude modulations of an ignored speech stream are still encoded in early neural processing, lexical information such as that related to phonemes and words is only extracted for an attended speech signal (Brodbeck et al., 2018; Teoh et al., 2022). Moreover, invasive recordings have shown that attentional modulation increases from primary to secondary auditory cortex, and that higher auditory areas can selectively represent an attended speaker (Mesgarani and Chang, 2012; Golumbic et al., 2013; O'Sullivan et al., 2019). How these higher-level processes feed back to the lower-level cortical areas and how the early attentional modulation reported here emerges there are major open research questions.

Subjects differ in their ability to understand speech in noise. In a previous study, we found that the attentional modulation of subcortical speech-FFRs measured with EEG was correlated to speech-in-noise comprehension (Saiz-Alía et al., 2019). Subjects that had worse abilities to understand speech in background noise had higher attentional modulation of the neural response, perhaps because they needed to rely more on this neural mechanism. Here we observed significant variability in the attentional modulation of the cortical contribution to the speech-FFR as well. However, we did not find a significant correlation between this neural measure and subjects' ability to answer the comprehension questions (Fig. 8). This might be because of the comparatively small sample size of 22 subjects, as well as the fact that all were young, healthy adults and displayed accordingly only moderate variability in their speech-in-noise comprehension. It will, in the future, be instructive to include a larger number of participants of different ages and speech-in-noise abilities to see whether the attentional modulation of the cortical contribution to the speech-FFR can explain some of the speech-in-noise variability, which might, for instance, also partly result from cochlear synaptopathy (Kujawa and Liberman, 2009; Liberman et al., 2016; Prendergast et al., 2017).

We would like to highlight that the experimental conditions in our study were highly controlled, with participants switching their attention between two specific speakers, in a supervised environment. However, in real-life situations, attention is a highly dynamic and multifaceted process influenced by various factors such as environmental distractions, cognitive load, and individual differences. In addition, attentional states are often influenced by subjective experiences, cognitive processes, and behavioral cues that may not be directly observable from neural activity. Integrating multimodal approaches such as recently achieved will in the future provide a more comprehensive understanding of attentional processes (Xie et al., 2023). It will thereby be of particular interest to investigate whether the early cortical response described here will be affected by such additional factors, or whether the latter will rather affect cortical processing at a later stage.

References

Albouy P, Benjamin L, Morillon B, Zatorre RJ (2020) Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. Science 367:1043–1047.

Bidelman GM (2015) Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient abr. Hear Res 323:68–80.

Biesmans W, Das N, Francart T, Bertrand A (2017) Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Trans Neural Syst Rehabil Eng 25:402–412.

Bourgeois J, Minker W (2009) Linearly constrained minimum variance beamforming. In: Time-domain beamforming and blind source separation: speech input in the car environment, pp 27–38. New York: Springer.

Brodbeck C, Simon JZ (2022) Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. Front Neurosci 16:828546.

Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.e5.

Chandrasekaran B, Kraus N (2010) The scalp-recorded brainstem response to speech: neural origins and plasticity. Psychophysiology 47:236–246.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975–979.

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906.

Coffey EB, Herholz SC, Chepesiuk AM, Baillet S, Zatorre RJ (2016) Cortical contributions to the auditory frequency-following response revealed by MEG. Nat Commun 7:11070.

Coffey EB, Musacchia G, Zatorre RJ (2017a) Cortical correlates of the auditory frequency-following and onset responses: EEG and fMRI evidence. J Neurosci 37:830–838.
Coffey EB, Chepesiuk AM, Herholz SC, Baillet S, Zatorre RJ (2017b) Neural correlates of early sound encoding and their relationship to speech-in-noise perception. Front Neurosci 11:479.
De Cheveigné A, Kawahara H (2002) YIN, a fundamental frequency estimator for speech and music. J Acoust Soc Am 111:1917–1930.
Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859.
Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:311.
Douw L, Nieboer D, Stam CJ, Tewarie P, Hillebrand A (2018) Consistency of magnetoencephalographic functional connectivity and network reconstruction using a template versus native MRI for co-registration. Hum Brain Mapp 39:104–119.
Etard O, Reichenbach T (2019) Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. J Neurosci 39:5750–5759.
Etard O, Kegler M, Braiman C, Forte AE, Reichenbach T (2019) Decoding of selective attention to continuous speech from the human auditory brainstem response. Neuroimage 200:1–11.
Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J (2017) Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng 14:036020.
Fischl B (2012) Freesurfer. Neuroimage 62:774–781.
Forney GD (1973) The viterbi algorithm. Proc IEEE 61:268–278.
Forte AE, Etard O, Reichenbach T (2017) The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. Elife 6:e27203.
Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980–991.
Gorina-Careta N, Kurkela JL, Hämäläinen J, Astikainen P, Escera C (2021) Neural generators of the frequency-following response elicited to stimuli of low and high frequency: a magnetoencephalographic (MEG) study. Neuroimage 231:117866.
Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS (2014) MNE software for processing MEG and EEG data. Neuroimage 86:446–460.
Hartmann T, Weisz N (2019) Auditory cortical generators of the frequency following response are modulated by intermodal attention. Neuroimage 203:116185.
Holliday IE, Barnes GR, Hillebrand A, Singh KD (2003) Accuracy and applications of group meg studies using cortical source locations estimated from participants' scalp surfaces. Hum Brain Mapp 20:142–147.
Horton C, D'Zmura M, Srinivasan R (2013) Suppression of competing speech through entrainment of cortical oscillations. J Neurophysiol 109:3082–3093.
Kegler M, Weissbart H, Reichenbach T (2022) The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information. Front Neurosci 16:915744.
Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". J Neurosci 30:620–628.
Kraus N, Anderson S, White-Schwoch T (2017) The frequency-following response: a window into human communication In: The frequency-following response (Kraus N, Anderson S, White-Schwoch T, Fay RR, Popper AN, eds), pp 1–15. Cham, Switzerland: Springer International.
Kujawa SG, Liberman MC (2009) Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss. J Neurosci 29:14077–14085.
Kulasingham JP, Brodbeck C, Presacco A, Kuchinsky SE, Anderson S, Simon JZ (2020) High gamma cortical processing of continuous speech in younger and older listeners. Neuroimage 222:117291.
Lattner S, Meyer ME, Friederici AD (2005) Voice perception: sex, pitch, and the right hemisphere. Hum Brain Mapp 24:11–20.
Liberman MC, Epstein MJ, Cleveland SS, Wang H, Maison SF (2016) Toward a differential diagnosis of hidden hearing loss in humans. PLoS One 11:e0162726.
Mauch M, Dixon S (2014) pYIN: a fundamental frequency estimator using probabilistic threshold distributions. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 659–663. Piscataway, NJ: IEEE.
McDermott JH (2009) The cocktail party problem. Curr Biol 19:R1024–R1027.
Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236.
Miran S, Akram S, Sheikhattar A, Simon JZ, Zhang T, Babadi B (2018) Real-time tracking of selective auditory attention from M/EEG: a Bayesian filtering approach. Front Neurosci 12:262.
Mirkovic B, Debener S, Jaeger M, De Vos M (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng 12:046007.
O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706.
O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM, Sheth SA, Mehta AD, Mesgarani N (2019) Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron 104:1195–1209.e3.
Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? a late locus of selective attention to natural speech. Eur J Neurosci 35:1497–1503.
Prendergast G, Guest H, Munro KJ, Kluk K, Léger A, Hall DA, Heinz MG, Plack CJ (2017) Effects of noise exposure on young adults with normal audiograms. I: electrophysiology. Hear Res 344:68–81.
Saiz-Alía M, Reichenbach T (2020) Computational modeling of the auditory brainstem response to continuous speech. J Neural Eng 17:036035.
Saiz-Alía M, Forte AE, Reichenbach T (2019) Individual differences in the attentional modulation of the human auditory brainstem response to speech inform on speech-in-noise deficits. Sci Rep 9:14131.
Saiz-Alía M, Miller P, Reichenbach T (2021) Otoacoustic emissions evoked by the time-varying harmonic structure of speech. eNeuro 8:
Schilling A, Tomasello R, Henningsen-Schomers MR, Zankl A, Surendra K, Haller M, Karl V, Uhrig P, Maier A, Krauss P (2021) Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. Lang Cogn Neurosci 36:167–186.
Schüller A, Schilling A, Krauss P, Reichenbach T (2023) Early subcortical response at the fundamental frequency of continuous speech measured with MEG. bioRxiv 546296. https://doi.org/10.1101/2023.06.23.546296.
Synigal SR, Teoh ES, Lalor EC (2020) Including measures of high gamma power can improve the decoding of natural speech from EEG. Front Hum Neurosci 14:130.
Teoh ES, Ahmed F, Lalor EC (2022) Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment. J Neurosci 42:682–691.
Tichko P, Skoe E (2017) Frequency-dependent fine structure in the frequency-following response: the byproduct of multiple generators. Hear Res 348:1–15.
Van Canneyt J, Wouters J, Francart T (2021) Neural tracking of the fundamental frequency of the voice: the effect of voice characteristics. Eur J Neurosci 53:3640–3653.
Van Eyndhoven S, Francart T, Bertrand A (2017) EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. IEEE Trans Biomed Eng 64:1045–1056.
Vrishab C, Kulasingham JP, Simon JZ (2023) Cortical responses time-locked to continuous speech in the high-gamma band depend on selective attention. bioRxiv 549567. https://doi.org/10.1101/2023.07.20.549567.
Xie Z, Brodbeck C, Chandrasekaran B (2023) Cortical tracking of continuous speech under bimodal divided attention. Neurobiol Lang (Camb) 4:318–343.