communications biology

Article

https://doi.org/10.1038/s42003-024-06913-z

# FMRI speech tracking in primary and non-primary auditory cortex while listening to noisy scenes

Check for updates

Lars Hausfeld [1,2] ✉, Iris M. H. Hamers [1,4] & Elia Formisano [1,2,3]

Invasive and non-invasive electrophysiological measurements during "cocktail-party"-like listening indicate that neural activity in the human auditory cortex (AC) "tracks" the envelope of relevant speech. However, due to limited coverage and/or spatial resolution, the distinct contribution of primary and non-primary areas remains unclear. Here, using 7-Tesla fMRI, we measured brain responses of participants attending to one speaker, in the presence and absence of another speaker. Through voxel-wise modeling, we observed envelope tracking in bilateral Heschl's gyrus (HG), right middle superior temporal sulcus (mSTS) and left temporo-parietal junction (TPJ), despite the signal's sluggish nature and slow temporal sampling. Neurovascular activity correlated *positively* (HG) or *negatively* (mSTS, TPJ) with the envelope. Further analyses comparing the similarity between spatial response patterns in the *single speaker* and *concurrent speakers* conditions and envelope decoding indicated that tracking in HG reflected both relevant and (to a lesser extent) non-relevant speech, while mSTS represented the relevant speech signal. Additionally, in mSTS, the similarity strength correlated with the comprehension of relevant speech. These results indicate that the fMRI signal tracks cortical responses and attention effects related to continuous speech and support the notion that primary and non-primary AC process ongoing speech in a push-pull of acoustic and linguistic information.

Speech and other continuous sound streams are increasingly used to examine human auditory processing under naturalistic listening conditions. Using "cocktail-party-like" scenes as stimuli, recent investigations have linked temporally-resolved neural signals, as measured with electrocorticography (ECoG), magnetoencephalography (MEG) and electroencephalography (EEG), to continuously changing features of the input[1–4]. A robust finding is that the envelope of incoming speech is "tracked" by these signals and that, in case of multiple concurrent sounds, the envelope of the relevant (i.e., attended) speech is tracked more reliably compared to the non-relevant one[5–9]. These effects have been localized to primary and secondary auditory cortical regions in Heschl's gyrus and sulcus (HG, HS), superior temporal gyrus (STG), and planum temporale (PT)[10–12] and, more recently, to subcortical areas[13,14]. However, as these techniques offer limited coverage (ECoG) and/or spatial resolution (EEG/MEG), it has been problematic to distinguish the specific contribution of different auditory brain regions to the neural tracking. The contribution of areas beyond auditory cortex requires further study as well.

To address these issues, here we present ongoing speech stimuli of two speakers (*v1* and *v2*) while measuring brain activity with high-field functional magnetic resonance imaging (fMRI), at high spatial resolution and with whole-cortex coverage. MRI poses challenges for performing auditory studies that increase with field strength mostly due to its noisy and magnetic environment, in particular when presenting long, continuous sound stimuli. Nevertheless, behavioral and EEG speech tracking results from simultaneous MRI and EEG measurements[15] suggested that participants were able to listen selectively to one speaker. Moreover, the EEG-based tracking of the speech envelope inside the MRI scanner was found to be correlated with tracking outside the scanner across participants. It remains unclear, however, whether the hemodynamic signal, an indirect and sluggish measure of neural activity, follows the speech envelope similar to electromagnetic neural signals.

In this study, we investigated fMRI neural tracking of the speech envelope by presenting unique 5 min blocks of task-relevant speech (Fig. 1A) both with and without concurrent speech (referred to as *single*

[1]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands. [2]Maastricht Brain Imaging Centre, 6200 MD Maastricht, The Netherlands. [3]Maastricht Centre for Systems Biology, Faculty of Science and Engineering, 6200 MD Maastricht, The Netherlands. [4]Present address: Department of Biomedical Sciences of Cells & Systems, Section Cognitive Neurosciences, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. ✉e-mail: lars.hausfeld@maastrichtuniversity.nl

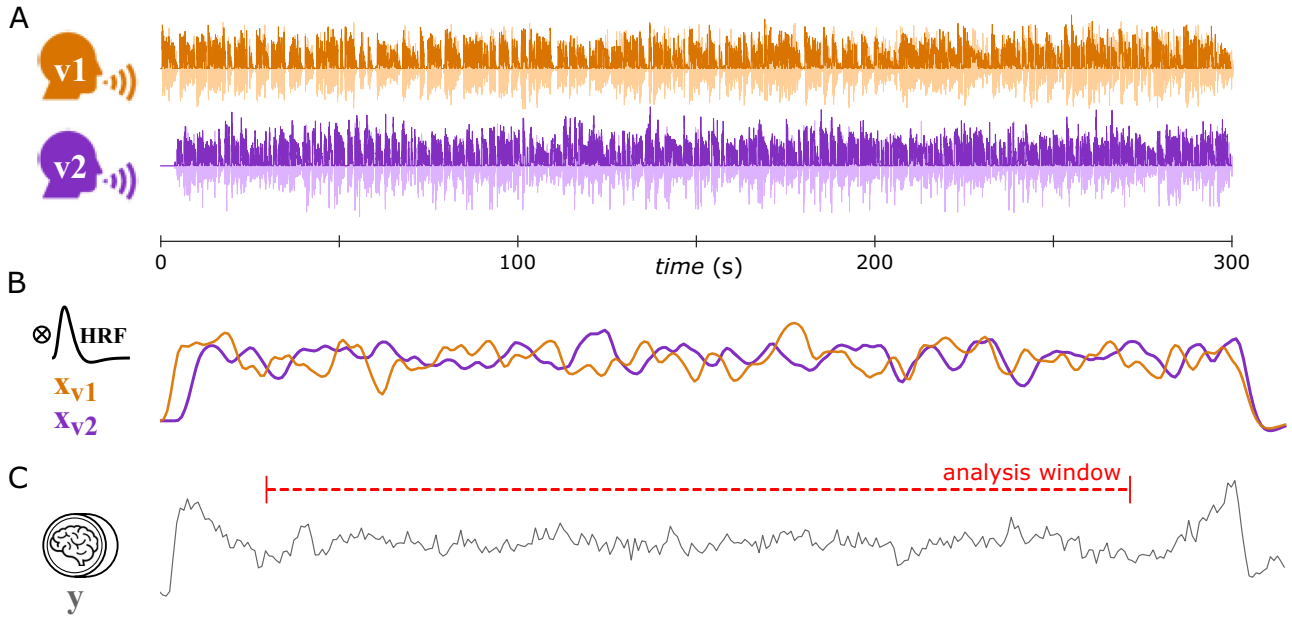


**Fig. 1 | Experimental stimuli and fMRI data. A** Speech waveforms during one trial (300 s) for the auditory scene condition containing speech of speakers v1 and v2. Light shading shows the speech waveform; the saturated lines denote the envelope of these waveforms. Note the initial silent period for speaker v2 to cue and support the tracking of the task-relevant speaker (here: v1). **B** To analyze whether the BOLD signal tracked the speech envelope, the envelopes of the relevant and non-relevant speech (see **A**) were convolved with a hemodynamic response function (HRF, indicated in black). The resulting HRF-convolved envelopes used as predictors in the analysis are depicted as orange and purple lines. **C** BOLD signal for one example region located on HG. Due to the onset and offset effects at the beginning and end of sounds, the speech tracking analysis was limited to 240 s excluding the initial and final 30 s of each block (red dashed line).

*speaker* and *auditory scene* condition, respectively; *Methods: Sound Stimuli, Experimental Design*). We measured the participants' ($n = 15$) brain activity with 7T-fMRI (Supplementary Fig. 1; *Methods: Functional MRI*). To avoid potential top-down effects due to repeated sound presentations, each relevant and non-relevant speech segment was presented only once to participants throughout the experiment. In addition, to capture the influence of the envelope dynamics, we limited the maximum length of silent periods in the stimuli to 300 ms, which mitigates the strong effect of comparing blood-oxygen level-dependent (BOLD) responses to sound vs. no-sound periods.

Previous tracking studies employed high-temporal precision measurements (EEG, MEG, sampling rate > 100 Hz) that allowed analyzing the time series with complex models (fitting > 100 parameters). Similar analyses in fMRI are difficult, as the sampling rate is two orders of magnitude lower (TR = 1 s = 1 Hz in this study) and the time-series comparably smaller. To enable speech tracking with fMRI, we acquired long time-courses by presenting listening blocks of 5 min and derived spatial maps of speech tracking within a voxel-by-voxel General Linear Model (GLM)[16] framework using envelope time courses convolved with the hemodynamic response (HRF). This analysis was applied to single speaker and auditory scene conditions to extract a coefficient of envelope tracking for each voxel. Assuming that selective attention reflects the attended speech in a scene similarly to the presentation of an isolated speech, we follow a template approach based on multi-voxel patterns[17,18] of these tracking coefficients. First, regions-of-interest (ROIs) are defined for clusters with significant tracking of the envelope in single speaker conditions that also define the template. These templates are then compared to the tracking patterns of relevant and non-relevant speech during auditory scenes in the same ROIs.

## Results

### Participants follow audiobooks during MRI acquisition

We asked participants to selectively listen to the (relevant) speaker. Participants were asked to answer questions about the audiobook's content and provide a subjective rating on their selective listening performance after each 5 min segment. The accuracy of responses to content questions indicated that participants were able to listen selectively to the single speaker and auditory scene stimuli (single speaker: 0.760 ± 0.071 [mean ± s.d.], $t(14) = 27.736$, $p < 0.001$ [vs. theoretical chance level of 0.25], Cohen's $d = 7.16$; auditory scene: 0.607 ± 0.175, $t(14) = 7.877$, $p < 0.001$, $d = 2.03$). This was confirmed by the participants' subjective ratings (single speaker: 7.72 ± 0.82 [mean ± s.d.]; auditory scene: 5.46 ± 1.67; ratings between 1 and 9: 1 = "could not follow the relevant speaker at all", 9 = "could follow as well as if presented without noise"). As expected, presenting a second speaker rendered the listening task more difficult (single speaker vs. auditory scene: $t(14) = 3.672$, $p = 0.0003$, $d_{av} = 1.15$), with participants rating their selective listening performance higher during the single speaker vs. auditory scene condition ($t(14) = 9.208$, $p < 0.001$, $d_{av} = 1.65$). In addition, accuracy and rating scores correlated across participants for the single speaker (Spearman's rank correlation, one-tailed; $\rho = 0.520$, $p = 0.024$) and auditory scene condition ($\rho = 0.445$, $p = 0.048$) suggesting that answers on content questions reflected perceived listening performance.

### Listening to audiobooks activates the speech comprehension network

The initial investigation of the fMRI data time courses during listening blocks (see Fig. 1C) revealed early (expected; sound onset) and late (unexpected; preceding sound offset) BOLD signal increases. To remove these tracking-unrelated effects at on- and offset when analyzing the tracking of speech, we restricted our tracking analysis to the central 4 min period of listening blocks by cutting the first and final 30 s (Fig. 1C, *Methods: fMRI Data Analysis—Tracking*).

Furthermore, in order to relate tracking results during these central 4 min sections to activation levels relative to baseline (i.e. no sound presentation), we also performed an activation-based GLM analysis (using HRF-convolved boxcar predictors for sound onset, offset and the central sections; *Methods: fMRI Data Analysis—Activation*). The results showed a sustained BOLD response to listening blocks compared to baseline, in regions typically involved in speech processing, for both single speaker and auditory scene conditions (Supplementary Fig. 2). Significant activation was observed in auditory cortical regions (HG, STG), superior temporal sulcus (STS) and inferior frontal gyrus (IFG); significant deactivation was found in

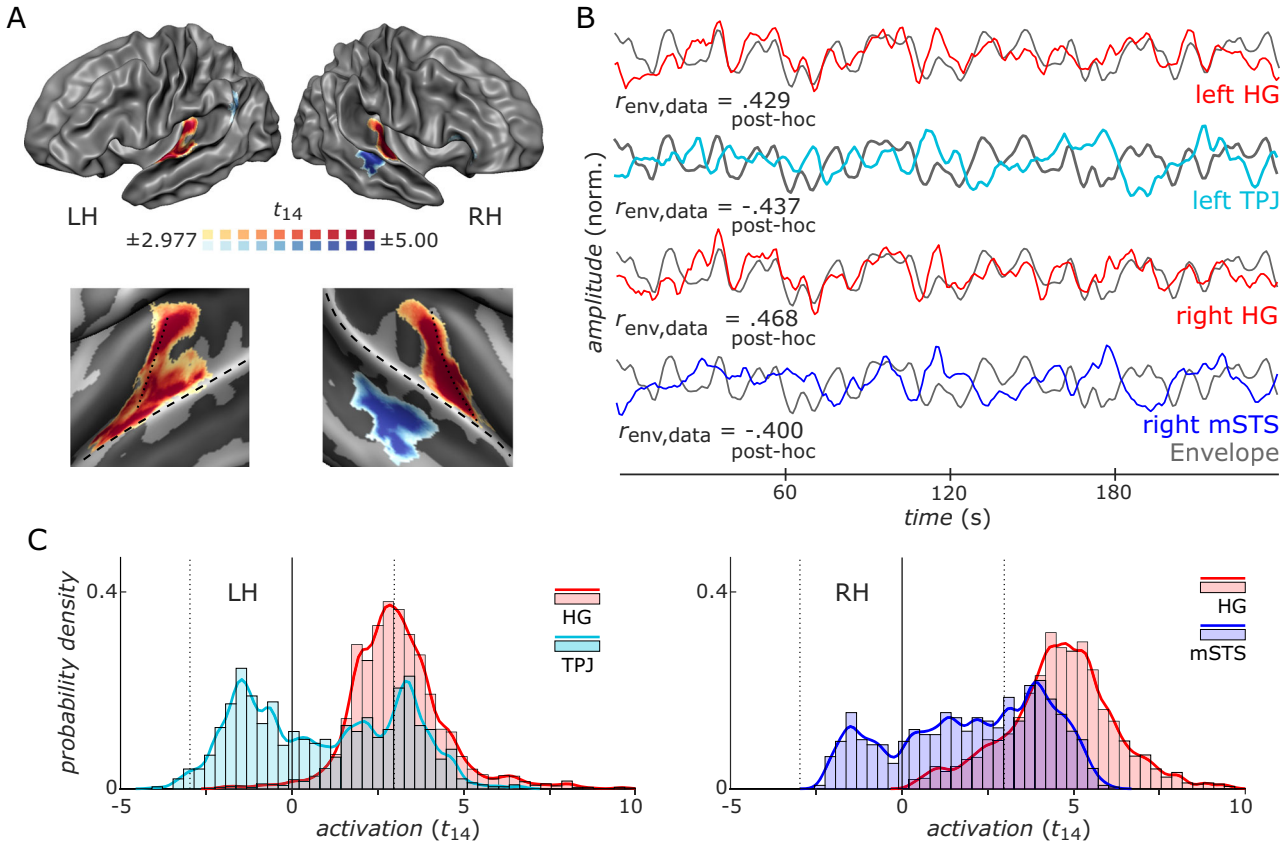


**Fig. 2 | Overview speech tracking single speaker condition. A** Color-coding depicts regions showing significant *positive* (warm colors) and *negative* (cold colors) speech tracking in the single speaker condition across participants ($n = 15$) in the left (LH) and right hemisphere (RH). Upper panels show lateral views of reconstructed average gray-white matter boundary after cortex-based alignment. Lower panels show enlarged views of temporal cortex on the inflated boundary. Dotted and dashed black lines indicate HG and STG, respectively. **B** Example BOLD signal time courses (240-s duration, *z*-normalized) averaged across participants during the same block in left and right HG (red lines) and right TPJ and middle STS (blue lines) showing significant speech tracking. *Positive tracking* in HG is indicated by the positive temporal correlation between MRI signal and speech envelope (gray lines) whereas *negative tracking* in left TPJ and right mSTS is indicated by their anti-correlation. Significant regions of the group analysis were back-projected to single-participant volume space where the most significant 20% of voxels (non-directional) were selected to create individual time courses. r-values in the lower left of each panel indicate the temporal correlation of MRI data and envelope time courses in this example. Note that these correlations are expected given the informed voxel selection and presented here to provide an intuitive interpretation of positive and negative tracking. **C** Distribution of statistical values for the "traditional" activation-based analysis of sustained activity in the single-speaker condition as compared to pre-stimulus baseline for the HG and mSTS regions that tracked the speech envelope (see panel A). Significant activation in tracking regions was found in bilateral HG and right mSTS. The distribution of *t*-values is indicated by solid lines and light-colored bars that show the estimate of the probability density function and the normalized histogram, respectively. Vertical dotted lines in histograms denote the significance threshold ($p < .01$, two-tailed). Statistical maps in (**A**) are thresholded at $p < 0.01$ (two-tailed) and corrected for multiple comparisons by cluster size ($p < 0.05$). LH left hemisphere, RH right hemisphere.

the temporoparietal junction (TPJ), Insula, middle frontal gyrus (MFG) and inferior central sulcus.

### BOLD responses show positive and negative tracking of speech

To analyze envelope tracking, we generated predictors for the GLM by convolving the extracted envelopes of the relevant speech and, for auditory scenes, non-relevant speech with a canonical HRF (Fig. 1B; *Methods: fMRI Data Analysis—Tracking*). Applying this GLM tracking analysis to the single speaker condition, we found significant fitting and positive parameter estimates (β-values) in bilateral contiguous regions along the HG/HS and on STG both anterior and posterior to HS (Fig. 2A). We also found regions with significant fitting and negative β-values in (left) temporo-parietal junction (TPJ) and (right) middle superior temporal sulcus (mSTS). These results reveal that low temporal resolution BOLD-fMRI responses track speech amplitude envelopes, extending previous results obtained with high temporal resolution neural measures (EEG, ECoG, MEG).

The BOLD signal time courses of these regions can be interpreted as showing a positive and negative temporal correlation with the envelope of the speech sound (upper and lower panel of Fig. 2B, respectively) which we refer to as *positive tracking* and *negative tracking*, respectively. Further analyses showed that areas in bilateral HG displaying positive tracking also displayed significant sustained positive activity (activation-based analysis, see Supplementary Fig. 2) with regard to pre-stimulus baseline (i.e., no sound presenation) (Fig. 2C, red color). Interestingly, we also found significant positive sustained activity for areas within left TPJ and right mSTS that displayed negative tracking (Fig. 2C, blue color).

This indicates that a positive BOLD response to sound with respect to pre-stimulus baseline can show both positive and negative tracking of the speech envelope. To verify that the negative correlations were not due to the specific choice of hemodynamic response model, we varied the HRF model for a wide range of values of the time-to-peak parameter (3.5–7 s) (Supplementary Fig. 3A). Results showed that HRF models with shorter time-to-peak fitted better in medial HG, whereas HRF models with longer time-to-peak fitted better in TPJ and mSTS. Importantly, the tracking of the envelope in TPJ and mSTS remained negative for the entire range of HRF models considered (Supplementary Fig. 3B, C). Here, we used a time-to-peak parameter of 4.5 s, providing a compromise between fitting BOLD responses in both positively and negatively correlated regions.

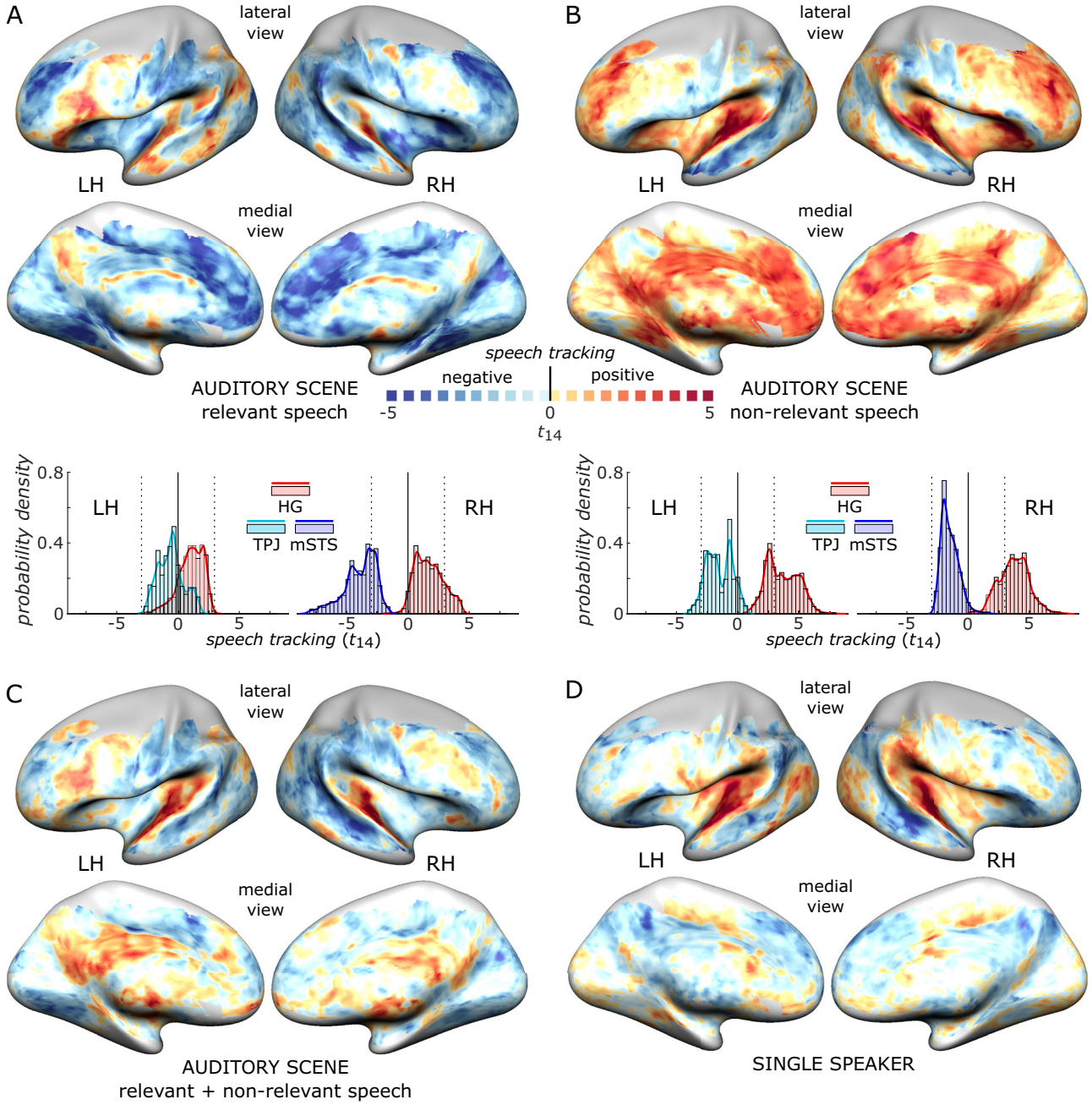


**Fig. 3 | Overview tracking in auditory scenes for relevant and non-relevant speech.** **A** Coeffcients for the tracking of the speech envelope of relevant speech are presented on inflated hemispheres (unthresholded) across participants ($n = 15$). Histograms below hemispheres indicate the distribution of coefficients within the regions-of-interest determined by the single speaker condition (Fig. 2A). **B** same as (**A**) but for non-relevant speech. **C** same as in (**A**) but showing coefficients for the contrast relevant + non-relevant speech. **D** For comparison, the unthresholded tracking results for the single speaker condition are shown (see Fig. 2A). Positive and negative tracking on inflated hemispheres is indicated by red and blue colors, respectively; maps are not thresholded. Vertical dotted lines in histograms denote the significance threshold ($p < 0.01$, two-tailed). LH: left hemisphere, RH: right hemisphere.

### FMRI tracking of relevant and non-relevant speech in auditory scenes

Similar to the single speaker condition, we obtained tracking maps for the auditory scene condition by including amplitude envelope predictors of the non-relevant speech in addition to the relevant speech (*Methods: Spatial Pattern Analysis of Tracking Maps*). The tracking maps for relevant and non-relevant speech for the auditory scene condition are presented in Fig. 3. For the relevant speech (Fig. 3A), we find positive tracking notably in bilateral HG and left IFG while negative tracking can be seen in a network including bilateral superior and middle frontal gyrus (SFG and MFG), anterior cingulate cortex, STG/S, anterior Insula and right TPJ. For the non-relevant speech (Fig. 3B), we find strong positive tracking on the bilateral temporal plane, including HG, as well as SFG, MFG and cingulate cortex. The tracking in ROIs defined by the single speaker condition, show the same directionality with positive tracking in HG, negative tracking in left TPJ and right mSTS for both relevant and non-relevant speech (histograms in Fig. 3A and B).

### Tracking patterns reveal (non-) relevant speech processing in HG and mSTS regions

Previous research highlighted that spatial activation patterns across the (auditory) cortical surface represent auditory objects including speech streams and that these spatiotemporal representations are significantly

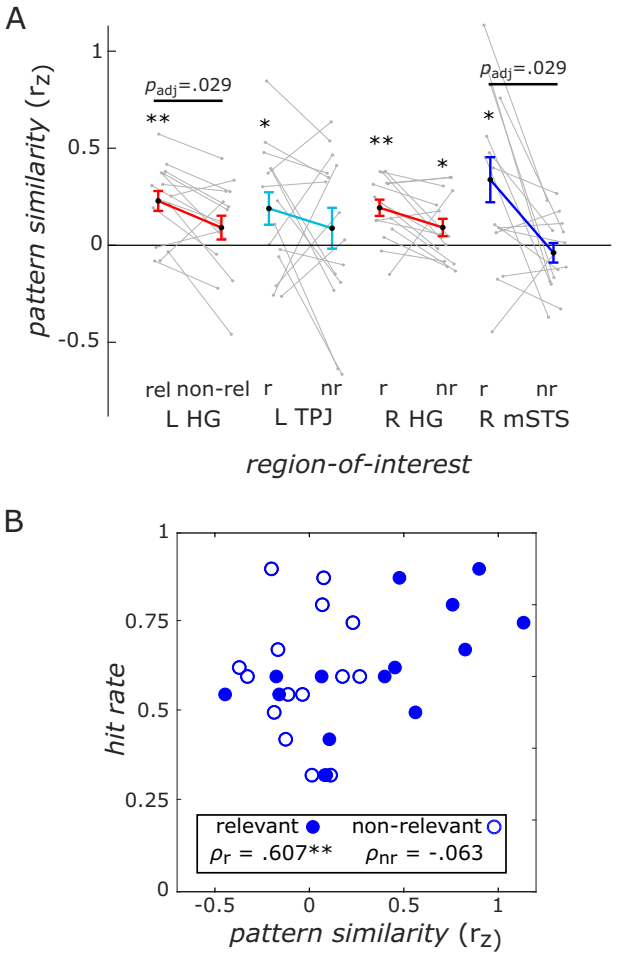


**Fig. 4 | Analysis of pattern similarity between tracking maps of single speaker and auditory scene conditions in regions-of-interest. A** Similarity of spatial patterns of speech tracking between the single speaker and auditory scene condition for the HG (red) and mSTS (blue) regions determined by significant tracking for the single speaker conditions (see Fig. 1A) across participants ($n = 15$). Gray lines indicate single participants and thick colored lines their average (error bar ± s.e.m.). Asterisks denote significant pattern similarity between single speaker and scene conditions (**$p_{adj} < 0.01$, *$p_{adj} < 0.05$, two-tailed; FDR-adjusted $p$-values; MCC across 8 tests) and straight lines show differences in pattern similarity (MCC across 4 tests) between the relevant speech (r/rel) and the non-relevant speech (nr/non-rel). **B** Scatter plots showing pattern similarity in the right mSTS region and behavioral performance. Circles represent data points of individual participants for relevant (filled) and non-relevant speech (open). A non-parametric correlation analysis across participants (**$p_{adj} < 0.05$, two-tailed; MCC across 4 tests) supported a positive relationship between pattern similarity for tracking patterns for relevant speech and participants' responses. Analyses for non-relevant tracking patterns and the other ROIs were non-significant.

affected by selective attention in multi-talker scenes[6,11,18–21]. We thus performed a spatial pattern similarity analysis to investigate the effects of selective attention in the regions-of-interest identified during the single speaker condition (i.e., bilateral HG and right mSTS regions). Specifically, we compared spatial maps for tracking (i.e., voxel-wise parameter estimates for speech envelope predictors) obtained from the single-speaker condition (Fig. 2A) with those from the auditory scene condition (Fig. 3A and B).

Overall, we find that the pattern similarity between tracking maps for single speaker and auditory scene conditions was significant in the left and right HG regions for relevant speech (Fig. 4A, red; $p_{adj} < 0.002$; $t(14) > 4.460$, Cohen's $d > 1.15$; multiple comparison corrected [MCC] across 8 tests via False-Discovery-Rate [FDR][22]) and for non-relevant speech in right but not left HG (right: $p_{adj} = 0.047$; $t(14) = 2.054$, $d = 0.53$; left: $t(14) = 1.480$, $p_{unc} = 0.210$). In left TPJ, the similarity of tracking patterns for relevant ($p_{adj} = 0.040$; $t(14) = 2.264$, $d = 0.58$) but not non-relevant speech was significant ($p_{unc} = 0.210$; $t(14) = 0.832$, $d = 0.21$). In the right mSTS region, we found significant similarity of tracking patterns for relevant speech (Fig. 4A, blue; $t(14) = 2.910$, $p_{adj} = 0.015$, $d = 0.75$) but not for non-relevant speech ($t(14) = -0.771$, $p_{unc} = 0.773$, $d = 0.20$). This indicates similar BOLD speech tracking

maps when listening to a single speaker or an auditory scene of two concurrent speakers for relevant speech in the HG, left TPJ, and right mSTS regions and for non-relevant speech in right HG.

### Tracking patterns reveal dominant processing of relevant speech

Across regions, we found that the similarity of tracking maps was modulated by speech relevance ($F(1,14) = 20.18$, $p < 0.001$, $\eta_p^2 = 0.59$; repeated-measures ANOVA) but not region ($F(2.23,28.00) = 0.028$, $p = 0.981$, $\eta_p^2 < 0.01$); these factors did not interact significantly ($F(2.00,28.00) = 2.02$, $p = 0.151$, $\eta_p^2 = 0.13$). Post-hoc tests showed that, in the left hemisphere, the similarity of tracking patterns was modulated by relevance in the HG region ($\Delta r = 0.137$; $t(14) = 2.784$, $p_{adj} = 0.029$, $d_{av} = 0.63$; MCC across 4 tests) but not TPJ ($\Delta r = 0.102$; $t(14) = 0.893$, $p_{adj} = 0.387$, $d_{av} = 0.28$). In the right hemisphere, we found that speech relevance did not affect tracking map similarity in right HG ($\Delta r = 0.101$; $p_{adj} = 0.088$, $t(14) = 1.997$, $d_{av} = 0.61$). In contrast, the similarity of tracking patterns between single speaker and auditory scene conditions was higher for relevant than non-relevant speech in right mSTS ($\Delta r = 0.377$, $p_{adj} = 0.029$; $t(14) = 3.132$, $d_{av} = 1.17$).

### Tracking patterns link to relevant speech comprehension in mSTS region

To examine whether the similarities of spatial tracking patterns were linked to behavior, we performed a non-parametric correlation analysis between fMRI (i.e., tracking patterns for relevant and non-relevant speech) and behavioral data (accuracy of answers to content questions). Across participants, both left and right HG regions did not show significant correlation (left HG: $\rho = -0.118$, $p_{unc} = 0.675$; right HG: $\rho = 0.102$, $p_{unc} = 0.717$; Spearman's rank correlation $\rho$, two-tailed); similarly, left TPJ tracking accuracy was not correlated to behavioral performance ($\rho = 0.272$, $p_{unc} = 0.326$). In contrast, the negative tracking region in right mSTS showed a significant brain-behavior relationship (Fig. 4B). More specifically, the tracking pattern similarity for single speech and relevant speech was positively correlated with the response accuracy to content questions ($\rho = 0.607$, $p_{unc} = 0.016$; $p_{adj} = 0.066$, MCC across 4 tests) while this did not hold for the tracking pattern similarity for non-relevant speech ($\rho = -.063$, $p_{unc} = 0.824$). A post-hoc test showed that the brain-behavior correlation in right mSTS was significantly stronger for pattern similarities for relevant vs. non-relevant speech ($p = 0.037$, permutation test, one-tailed, $n_{perm} = 10^5$).

Overall, these results show that speech relevance modulated the similarity of tracking maps in HG and mSTS regions from the auditory scene to the single speaker condition with higher similarity for the relevant vs. non-relevant speech. Furthermore, our observations indicate that right mSTS reflects relevant but not non-relevant speech and that this similarity of tracking maps for relevant speech was linked to its comprehension.

### Tracking patterns decode relevant speech envelopes

To compare fMRI-based speech tracking to the speech tracking with neuroelectric measurements[5–7], we trained ridge-regression models on data from the single speaker condition for each ROI to decode the speech envelope for data of the auditory scene condition (Fig. 5). In line with the pattern-based analysis (Fig. 4) the results indicate successful attention decoding, i.e. the predicted envelope was more similar to the relevant vs. non-relevant speech in bilateral HG and right mSTS. In addition, the analysis showed increasing decoding performance reaching its maximum (rate > 0.70) in left HG and right mSTS at the full trial windows, which is longer than for EEG, for which similar or higher decoding is observed for shorter windows[7,23,24]. Corroborating the pattern-based analysis (Fig. 4), the correlations of predicted and presented envelopes (Fig. 5C, D) revealed that

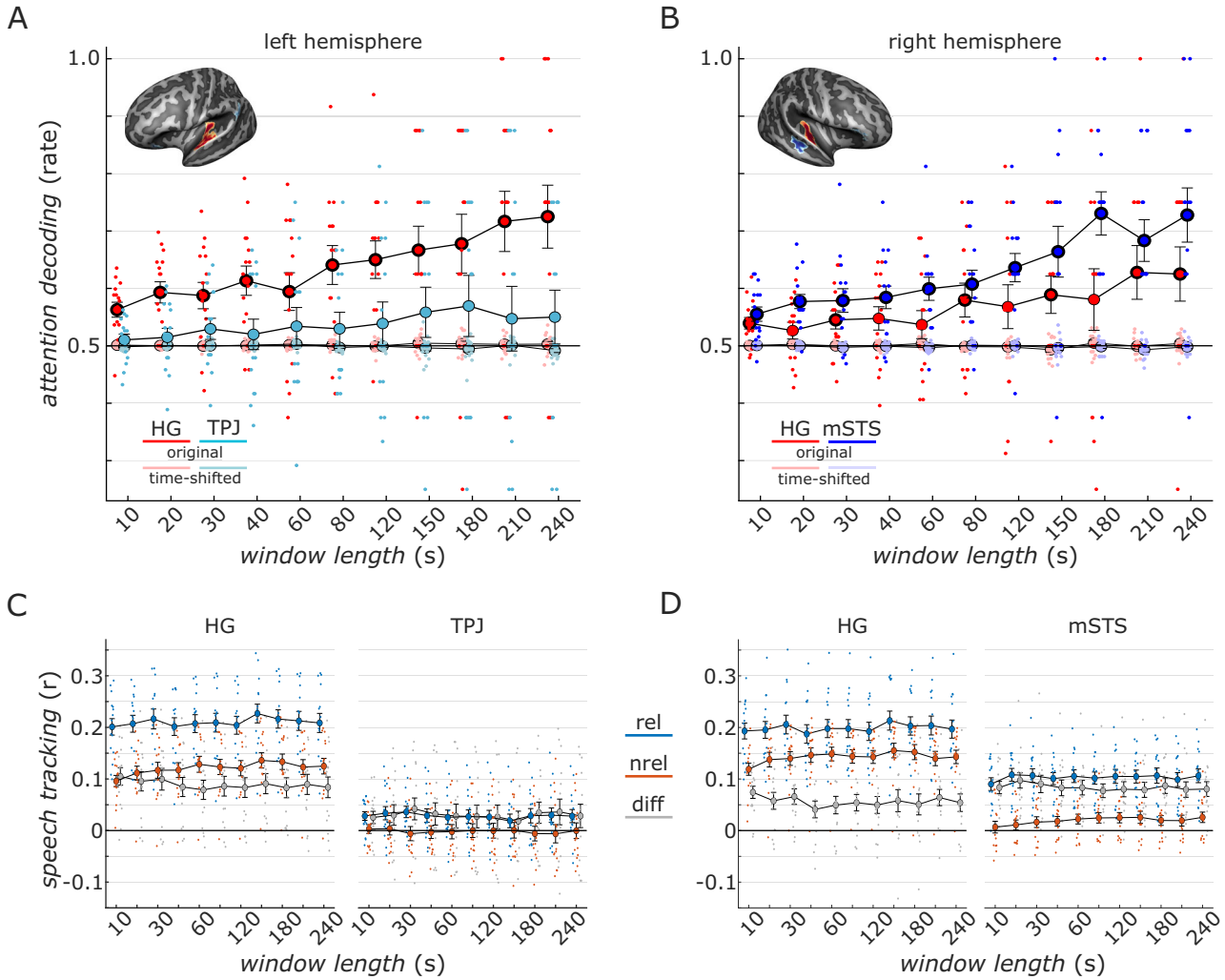


**Fig. 5 | Overview auditory attention decoding. A**, **B** show the accuracy of auditory attention decoding during auditory scenes based on regions and models determined with data from the single speaker condition across participants ($n = 15$). Red lines indicate decoding as a function of window length for the positive tracking in left and right HG while turquoise/blue lines denote decoding for the negative tracking in left TPJ/right mSTS. The lighter colors denote correlations with the time-shifted envelope. Thick black rims of circles: attention decoding significantly different from decoding of time-shifted envelopes (pale red and blue lines; $p < 0.05$, one-tailed, FDR-corrected across 44 tests [11 window lengths x 4 regions); error bars: mean ± s.e.m.; theoretical chance level: 0.5; insets: regions projected on inflated group-aligned surfaces; dots show individual data points. Results are shown separately for right and left hemisphere regions in A) and B), respectively. Note that decoding was done on non-repetitive sections of the full 240 s segment and assessed for different lengths. **C**, **D**) present the correlation of relevant (blue) and non-relevant speech (orange) with predicted envelope time courses. Note that across windows, correlations and their spread are stable, but that they are more variable for individual instances leading to lower decoding performance for shorter windows in (**A**, **B**). Gray lines: average correlation difference between relevant and non-relevant speech tracking; differences are significant for all windows and regions except left TPJ ($p < 0.05$, FDR-corrected across 44 tests); error bars: mean ± s.e.m.; dots show individual data points. Results are shown separately for right and left hemisphere regions in (**A**, **B**), respectively. Note that models are trained to lead to positive correlations with envelopes, independent of positive or negative tracking regions.

non-relevant speech was represented in bilateral HG ($r_{nrel} > 0.10$), albeit to a lesser degree as relevant speech ($r_{rel} > 0.20$); this was not observed in right mSTS where non-relevant speech showed minimal correlation with decoded envelopes ($r_{nrel} < 0.03$).

### Positive tracking of continuous and negative tracking of binarized envelopes

While we found that the BOLD signal followed the HRF-convolved amplitude envelopes, it remains speculative whether acoustic or other linguistic and cognitive processing fluctuations co-occurring with these envelopes (characterized by low-pass filter characterics; Supplementary Fig. 5) are tracked by the fMRI signal. In a first step to investigate the specificity of the tracking with regard to slower and faster amplitude modulations, we created binary sound envelopes (i.e. '0' for silent periods, '1' for periods with sounds) and performed a whole-brain speech tracking analysis of the single speaker condition (see Fig. 2A) with the (HRF-convolved) continuous and binary envelope predictors (Supplementary Fig. 4A). The two predictors have a medium correlation ($r_{cont,bin} = 0.68 \pm 0.24$; median ± inter-quartile range across trials) showing that these are similar but might capture different variance components of the BOLD time-courses. We found that the continuous envelope predictor explained better the positive tracking in bilateral HG (Supplementary Fig. 4B) while the binary predictor explained in particular the negative tracking in right mSTS as well as bilateral STG (Supplementary Fig. 4C). This finding provides evidence of differential speech tracking of naturalistic speech in these regions possibly reflecting the processing of more abstract linguistic or cognitive aspects in mSTS vs. HG.

## Discussion

In this study, we measured cortical responses to continuous speech stimuli using high-field fMRI. In a first step, we showed that the hemodynamic

response follows or "tracks" the speech envelope amplitude. More specifically, we found that the BOLD signal tracks the ongoing speech envelope of a single speaker in bilateral HG, STG and STS. These findings resemble the speech tracking observed by direct and temporally resolved neural measures (ECoG, MEG and EEG), which showed robust tracking of the speech envelope amplitude.

Interestingly, our results showed positive tracking of the convolved speech envelope by the left and right HG regions and negative tracking by the left TPJ and right mSTS regions. Both positive and negative tracking was found in areas that showed increased sustained activity in response to speech sounds (i.e., the speech envelope modulated the signal "on the plateau" of the positive BOLD activation). We interpret the positive tracking in the HG region to reflect the ongoing envelope amplitude of the speech stream. This is in line with previous ECoG studies showing that activity in HG and middle STG is correlated with responses to speech[6,11,25]. The negative tracking observed in the right mSTS region might reflect cortico-cortical top-down signals that aid in following relevant speech in particular during periods of low speech audibility, i.e., when the task is difficult (e.g., due to the fMRI noise and lower intensity of relevant speech).

Layer-resolved high-field MRI acquisitions might help to better define the role of this region in terms of top-down and bottom-up input when listening to a single speaker or an auditory scene[26]. Whether these regions of BOLD speech tracking coincide with the sources of the speech tracking observed with neuro-electromagnetic signals remains an open question. The relationship between neural activity, electric and hemodynamic signals is complex[27–29]. Concurrent measures of speech tracking by EEG and fMRI[15] would allow linking, within participants, the observed results by hemodynamic and neuroelectric measures more directly and shed light on the underlying neural processes.

After having established that hemodynamic signals follow continuous speech signals, we performed an analysis of multi-voxel patterns that revealed high similarity between spatial tracking patterns of relevant speech in an auditory scene and speech presented without concurrent distractor in HG and the STS regions. In right mSTS, this similarity of the tracking patterns was (positively) correlated with the assessment of speech comprehension across participants.

Examining the spatial patterns of speech tracking maps for the HG region, we found a high similarity between the single speaker and auditory scene conditions for both the relevant and—to a lesser extent—non-relevant speech. These results indicate that the overall incoming speech signal, containing the relevant and non-relevant speech, is reflected in the HG region including medial and lateral HG/HS and adjacent STG. In these regions, the pattern similarity was higher for relevant speech in comparison to non-relevant speech (Fig. 4A). This is in line with previous observations showing attentional modulation of envelope representations with high selectivity for attended vs. unattended speech using ECoG[6,11] and MEG[5,12]. Although being significantly lower compared to relevant speech, the pattern similarity of tracking maps for non-relevant speech suggests residual information about non-relevant speech in the HG region.

In contrast, for the right mSTS region, we found significantly higher pattern similarity of tracking maps of relevant speech vs. non-relevant speech tracking and no significant similarity for tracking maps of non-relevant speech. In addition, we observed that participants with a higher pattern similarity between single speaker and relevant speech of the auditory scene performed better in a comprehension task about the audiobook's content. This result indicate that right mSTS processes exclusively information reflecting the relevant speech, implying that the cocktail party is resolved at this stage. This fits with previous results in which pattern similarity in mSTS/STG only represents relevant speech but not non-relevant speech or music; the effect in this region was category specific such that relevant speech but not music showed significant pattern similarity[18]. Another possible explanation for these results is that activation in this area might reflect increased top-down control of selective listening at a temporal scale of linguistic units or cognitive adaptations[30–32] with decreasing amplitude of the relevant speaker and thus increasing energetic masking by scanner noise and, in auditory scenes, additional energetic and informational masking by the non-relevant speaker. These explanations are not mutually exclusive such that both bottom-up and top-down contributions important for selective listening are represented in this region's signals. This region partially overlaps with electrophysiological recording sites in STG, which suggested responses to sustained features of the speech signal (i.e., speech envelope[25]) in line with the current findings. Results of a recent fMRI study using continuous speech stimuli suggested that the HG region mostly represented spectral information (related to the envelope amplitude), while the mSTS region was mostly correlated with semantic features[33]. While this is in agreement with our findings in the HG region, the significant tracking of the envelope in mSTS presumably following phonological or semantic features might be explained by correlations between these features and the amplitude envelope. Additional differences, for example in analyses or data acquisition (3 T vs. 7 T MRI, 1 Hz vs. 0.5 Hz sampling) could explain these observations. However, overall, the current and previous results support that the mSTS region links bottom-up acoustic and top-down linguistic processing of relevant speech during auditory scenes, possibly similar to the posterior middle temporal gyrus MTG suggested to lie between auditory-phonological and semantic processing regions[34] and in line with longer temporal integration for processing information at that time scale[35,36].

The correlation between tracking map similiarites in mSTS and participants' performance on answering content questions suggests that activity in this region reflects the behavioral outcome. It might be linked to previous observations associating MEG and EEG-based speech tracking with speech intelligibility and comprehension[7,9,37–39] although, if analyzed, this brain-behavior association is not always found[40]. The STS has been linked to intermediate linguistic representations[31,34] and EEG-based speech tracking exploiting linguistic features was correlated with the performance in speech comprehension tasks across participants[41]. Our findings corroborate these observations by suggesting a neural source of this link between behavior and linguistically informed speech tracking.

Our analyses showed similar tracking patterns for non-relevant speech in the HG regions. Previous studies using EEG, MEG and ECoG have indicated that background sounds including speech are represented in the auditory system in particular at early latencies of processing[42–45], which is in line with the current finding suggesting information of non-relevant speech being represented in earlier areas in auditory cortex and being represented less in higher areas in the auditory processing hierarchy like STG and STS.

Methodologically more similar to the M/EEG-based speech reconstruction approach[5,7,23], we predicted speech envelopes during auditory scenes based on linear combinations of voxel time courses and the results match the spatial pattern analysis. Specifically, high reconstruction performance was observed in left and right HG for both relevant and, to a lower extent, non-relevant speech. Although showing lower reconstruction, the performance in right mSTS was high for relevant speech and at chance for non-relevant speech resulting in similar classification performance for HG and right mSTS for attended speech.

Exactly what information is represented (i.e., what is being "tracked") by the BOLD time course remains an open question. While we have shown that the BOLD signal follows the amplitude envelope after applying the HRF, it is unlikely that fast acoustic or linguistic features (e.g., phonemes, syllables, or words represented in the theta band[46]) are *directly* represented in the signal given the sluggish hemodynamic response. However, there is support that information of dynamic, transient signals in the low delta frequency band (<1 Hz) might be represented, in particular, by signal changes on the plateau of sustained responses[47]. The observed fMRI tracking is limited to properties of the BOLD response capturing slow fluctuations that might not only reflect acoustic signals but also higher-level linguistic as well as attention and other cognitive effects required to listen selectively (as indicated by language and domain general networks in unthresholded tracking maps for auditory scenes; Fig. 3A, B). In a first attempt to better understand the observed tracking, we applied a simple threshold to create a binary representation of the speech envelope and found that the continuous envelope best explains signals in HG while the binarized predictor explains

data in STG and STS (Supplementary Fig. 4). This is in line with an interpretation that stresses more acoustic representation in early auditory cortex and a region representing acoustic-linguistic representations in mSTS[31,34]. To summarize, further modeling of data collected during naturalistic listening of continuous speech is required for establishing how acoustic, linguistic, or cognitive processing is related to the ongoing BOLD signal[33,48–50]; comparing results from fMRI-based and neuro-electric speech tracking need to take these consideration into account.

To conclude, our results showed that speech tracking, a robust phenomenon observed with high temporal resolution and neuro-electric signals, can be observed with low temporal sampling and high-field fMRI BOLD responses. Furthermore, we found opposite tracking of speech in HG and mSTS regions and tracking of non-relevant speech in HG but not mSTS. Speech tracking in the mSTS was linked to speech comprehension. These results indicate neural processes potentially related to stronger feedback and linguistic integration processing in mSTS compared to HG aiding successful listening in noise. In addition, these results provide support for neural signals in mSTS that reflect a processing stage at which the cocktail-party is resolved and relevant speech is analyzed. Utilizing indirect measurements of neural activity, fMRI speech tracking is likely sensitive not only to speech features but also to time-varying (non-acoustic) changes in attention, cognitive demand and working memory that may co-occur at these slower time scales.

## Methods

### Participants

Fifteen students (native German speakers) of Maastricht University (13 female, 2 male, mean age ± [s.d.]: 24.1 ± [3.8] years, age range: [19–33] years), after signing the written informed consent, took part in the experiment and received course credit or gift vouchers for their participation. The local ethics committee of the Faculty of Psychology and Neuroscience (*Ethics Review Committee Psychology and Neuroscience*) at Maastricht University approved the experimental procedures of the study (#167_09_05_2016).

### Sound stimuli

We presented participants with speech (audiobook excerpts[43]) of one female (*v1*; $f_0 = 159 \pm 8.3$ Hz, mean ± s.d.) and one male speaker (*v2*; $f_0 = 107 \pm 7.3$ Hz). The fundamental frequency $f_0$ for each excerpt was determined by averaging $f_0$ contours obtained with the YIN algorithm[51]. Sounds were played on top of MRI scanner noise and delivered via an MR-compatible sound system (Sensimetrics S14, Sensimetrics Corporation, Malden, MA) diotically by in-ear earphones. Sound stimuli were presented at a high but comfortable level that was individually adjusted at the beginning of the experiment. Sound intensity of the two audiobooks was equalized based on root-mean-square (RMS), i.e. v1-speech was presented at a signal-to-noise ratio (SNR) of 0 $dB_{RMS}$ with regard to v2-speech. To avoid clicks, the onset and offset of each speech signal were ramped (linear ramps of 0.1 s). Auditory stimuli were digitized using a sampling rate of 44.1 kHz and 16 bits. For all sound stimuli, silent periods (e.g., during words or sentences) were adjusted to a duration of at most 300 ms using Praat[52]. Audiobook excerpts did not repeat during this experiment, i.e., each of the excerpts was only presented once throughout the experiment either as relevant or non-relevant speech. Each of the excerpts had a duration of 5 min matching the length of a trial.

### Experimental design

The design included two conditions, 1) the *single speaker* condition, i.e. the presentation of speech of one audiobook, and 2), and the *auditory scene* condition, i.e. the concurrent presentation of speech of two audiobooks. To obtain sufficient samples for each presentation, each trial lasted 5 min. During the auditory scene condition, the relevant speech started 4.5 s before the non-relevant speech to provide listeners with an auditory cue indicating the to-be-attended speech (see Fig. 1). In total, we presented 12.5 min trials across 6 functional runs (see "Functional MRI") of which 4 trials (20 min) in the single speaker condition and 8 trials (40 min) in the auditory scene condition; none of the audiobook excerpts were repeated. Stimuli were presented with Presentation (v20, Neurobehavioral Systems, Berkely, CA)

### Functional MRI

Brain imaging was performed with a 7-Tesla Siemens Siemens Magnetom scanner with a whole brain coil at the Maastricht Brain Imaging Center (Maastricht, The Netherlands). Anatomical scans were acquired during each session with an MP2RAGE sequence[53] (voxel size: 0.65 mm isotropic; 240 slices; FoV: 208 mm; TR: 5000 ms; TE: 2.51 ms; GRAPPA 2) and masked with the second inversion contrast. For each participant, 6 functional runs of 722 ± 10 volumes (mean ± s.d.; range [698 756]) with large cortical coverage were collected using an echo-planar imaging (EPI) sequence with multiband 3 acceleration (57 slices; voxel size: 1.5 mm isotropic; FoV = 192 × 192 mm; TR = 1000 ms; TE = 19 ms; GRAPPA 2). For correcting EPI distortions two sets of five images were acquired in opposite phase encoding directions (i.e., anterior-posterior and posterior-anterior) between the third and fourth functional runs.

The two conditions (i.e., single speaker and auditory scene) were presented in different runs (single speaker in runs 1 and 4, auditory scene in runs 2, 3, 5 and 6) each containing one block of the v1-task and v2-task with alternating first condition counter-balanced across participants (Supplementary Fig. 1). Participants were asked to selectively listen to speech of v1 (*v1-task*) or v2 (*v2-task*). Presentations for the two conditions included a 15 s rest period followed by the 5 min presentation of the sound stimulus and was followed by another rest period of 10 s (a fixation cross was presented throughout in the center of the visual display through a mirror at the back of the scanner). Subsequently, participants indicated their subjective task performance ("How well did you follow the relevant voice?", range: 1 [could not follow the relevant speaker at all]—9 [could follow as well as if presented without noise]) and responded to five questions on the content of the (relevant) audiobook (4-alternative-forced-choice task; answer alternatives indicated by A, B, C, D) by button press[43].

### Behavioral data analysis

Ratings of self-assessed selective listening performance and responses to content questions were extraced as rating (1–9) and hit rates (#correct responses among all content questions). Hit rates were compared to theoretical chance level (0.25) with one-tailed one-sample *t*-tests. Comparisons between behavioral outcomes for single speaker and auditory scene conditions were performed via two-tailed paired *t*-tests. Effect sizes were estimated using Cohen's *d*. For one-tailed *t*-tests, Cohen's *d* was estimated as $d = \frac{M-.25}{SD} = \frac{t}{\sqrt{N}}$; for paired *t*-tests, we computed Cohen's d using the averaged standard deviation in the denominator[54] $d_{av} = \frac{M_1-M_2}{s_{av}}$, $s_{av} = \sqrt{0.5 \cdot \left(SD_1^2 + SD_2^2\right)}$.

### fMRI data preprocessing

Preprocessing of both functional and anatomical data was performed with BrainVoyager (v21.4, Brain Innovation, Maastricht, The Netherlands). FMRI data preprocessing consisted of slice-scan-time correction, motion correction, EPI distortion correction, and temporal high-pass filtering (0.015 Hz ≈ 11 cycles per run). EPI distortions were corrected using BrainVoyager's COPE plugin[55] (v1.1). Functional runs were individually aligned to anatomical scans and transformed to Talairach space[56]. The functional data was spatially smoothed (4 mm FWHM) and individual maps were projected onto the group-aligned surface provided by cortex-based alignment[57] to create group maps. Other data processing and analyses were performed in Matlab (version R2022b; The MathWorks Inc, Natick, MS, US). Functional maps were restricted to regions that were included in all functional runs of all participants and not affected by regions outside functional coverage during spatial smoothing.

### fMRI data analysis—activation

To detect cortical regions responding to the presentation of and selectively listening to audiobooks in comparison to pre-stimulus baseline, the

functional data of the single speaker and auditory scene conditions was analyzed voxel-wise using a general linear model[16] (GLM). Because we observed strong onset and offset effects and were interested in the sustained activation for the tracking analysis in the central 4 min (see fMRI Tracking Analysis), the listening blocks were modeled by three boxcar predictors reflecting the onset (0–30 s), sustained (31–270 s) and offset responses (271–300 s) to the stimuli. In addition, confound predictors reflecting participant responses, and constant predictors for functional runs were included. The resulting maps were multiple-comparison (MC) corrected by cluster size (initial threshold $p < 0.01$, cluster size threshold $p < 0.05$) using Monte-Carlo simulations[39].

### fMRI data analysis—tracking

To analyze how well the BOLD signal could be estimated from the speech envelope, we performed a voxel-wise GLM analysis using the middle portions of trials and the speech envelopes.

The speech envelopes were estimated for each audiobook excerpt using the Hilbert-transform and subsequent low-pass filtering (FIR filter, 8 Hz cut-off). This envelope was then convolved with the canonical HRF and, in a final step, downsampled to match the sampling rate of the BOLD signal (1 Hz). For an additional analysis (Supplementary Fig. 4), we created binary envelope predictors by thresholding the speech envelopes before HRF convolution and downsampling by setting the predictor to 1 for non-zero values; these binary envelopes were added to continuous envelope predictors thus entering the same GLM. Before the tracking analysis, the functional data was cut to 4 min per block by removing the initial and final 30 s to avoid confounds from onset or offset effects observed during data exploration (Fig. 1C). To avoid concatenating trials, we performed the tracking analyses for the central portion of single trials (4 min duration). Subsequently, the functional data was analyzed—voxel-by-voxel—for the tracking of speech envelopes by making use of the GLM framework[16]. More specifically, we modeled BOLD voxel time courses by $y = X\,\beta_{track} + \boldsymbol{\varepsilon}$ where y ($n$ x 1) denotes the voxel time course of $n$ TR (or samples), X ($n$ x $p$) a design matrix of model time courses for $p$ predictors, $\beta_{track}$ ($p$ x 1) model coefficients and ε ($n$ x 1) the error term). The design matrix included the main predictors for the speech envelope time courses and confound predictors reflecting participant's motion and an offset (constant). For the single speaker condition, one predictor reflecting the presented speech was included. For the auditory scene condition, two predictors were included, one for relevant speech and one for non-relevant speech. These speech predictors were included for each voice separately, i.e. we computed a fixed effects GLM with two predictors for v1 and v2 across the four trials (single speaker condition) or four predictors for relevant and non-relevant v1 and v2 across the eight trials (auditory scene condition). The coefficients $\beta_{track}$ were averaged for further analysis across v1 and v2. The significance of the speech tracking (i.e., model coefficients for envelope predictors) at the group-level was MC-corrected by cluster size (initial threshold $p < 0.01$, cluster size threshold $p < 0.05$) using Monte-Carlo simulations[39].

### Spatial pattern analysis of tracking maps

To investigate BOLD activity during listening to auditory scenes, we analyzed the spatial patterns of envelope coefficients (see fMRI tracking Analysis) using a template approach[17,18]. For this pattern similarity analysis, the tracking maps (i.e., coefficients $\beta_{track}$ for the relevant and non-relevant HRF-convolved speech envelope predictors) in the left and right HG, left TPJ, and right mSTS regions from the analysis of single speaker tracking were set as templates. The pattern similarity was computed between the templates and the tracking maps obtained for the auditory scene presentations for relevant and non-relevant speech using Pearson's correlation and Fisher's transformation $r_z = 0.5 \cdot \ln\left(\frac{1-r}{1+r}\right)$. Circularity ("double-dipping"[58]) cannot explain the observed effects of the spatial pattern analysis because of the univariate-based region-of-interest definition, the independent runs used for single speaker and auditory scene conditions, and the orthogonal comparison between relevant and non-relevant speech. Statistical testing of these scores was done via a repeated-measure ANOVA (Greenhouse-Geisser corrected) with factors relevance (relevant and non-relevant speech) and region-of-interest (left HG, left TPJ, right HG, and right mSTS), two-tailed paired sample $t$-tests (to compare results between relevant and non-relevant speech) and one-tailed one-sample $t$-tests (to compare results to theoretical chance level $r_z = 0$). $T$-test results were corrected for multiple comparisons by FDR[14]. For the ANOVA, effect sizes were estimated using partial η squared, $\eta_p^2$. For one-sample and paired $t$-tests, we computed Cohen's $d$ and $d_{av}$ (see Behavioral Data Analysis)[54].

Brain-Behavior correlations across participants were computed using the non-parametric Spearman's rank correlation coefficient $\rho$ (two-tailed). We correlated individual listening performance (hit rates of content questions) with the pattern similarity of tracking maps between speech envelopes for the single speaker and auditory scene conditions. To test whether the brain-behavior correlation was higher for pattern similarity of relevant vs. non-relevant speech, a permutation test was performed (one-tailed, $n_{perm} = 10^5$) by reshuffling the behavioral scores across participants.

### Speech envelope decoding

To better compare the speech tracking of fMRI data to EEG-based speech tracking, we used ridge regression to predict the speech envelope using the data time courses of auditory scenes for each of the ROIs derived from the single speaker trials (Fig. 2). The regression coefficents $\beta_{ridge}$ were determined using data from the single speaker condition by $\beta_{ridge} = (X_{single}{}^T X_{single} + \lambda I)^{-1} \cdot X_{single}{}^T y$ where $X_{single}$ ($v$ x $p$) denotes the matrix of $v$ voxel time courses of $n$ TR (or samples) for the single speaker condition, y ($n$ x 1) represents the HRF-convolved envelope, and I ($v$ x $v$) the identity matrix. The regressions models were optimized for the voxel set size (based on absolute tracking coefficients in single speaker condition; $n_{vox\%}$ = {5,10,15,20,30,40,50,75,100}) and the regularization parameter λ was optimized by selecting the parameter, $\lambda = \{10^{-5}, 10^{-4}, \ldots, 10^{0}, \ldots, 10^{5}\}$, that maximized the prediction of the envelopes in the training set. The speech envelope for the auditory scene is estimated by $\hat{y} = X_{scene}\,\beta_{ridge}$ where $X_{scene}$ indicates voxel time courses acquired during the auditory scene condition of the ROI's voxels. Subsequently, the estimated envelope time courses were compared to the relevant and non-relevant (HRF-convolved) envelopes using Pearson's correlation. Decoding performance was computed as the amount of trials for which $r_{rel} > r_{nrel}$ across all trials. To investigate duration needed for successful decoding, 11 windows of {10, 20, 30, 40, 60, 80, 120, 150, 180, 210 and 240} seconds were created. For windows up to 120 s, the 240-s time-courses were divided into non-overlapping windows (e.g., 24 windows for 10 s, 6 windows for 40 s) and the decoding performances were averaged across windows; windows >120 s included one window starting from the trial's first TR/sample. Across participants, the decoding performances were compared to the average decoding performance estimated for each participant by time-shifting the envelopes (Matlab's circshift.m; $n_{shift} = 60$ with $step_{shift} = 4$ TR) via paired $t$-tests. These results were MC-corrected by FDR[22].

### Statistics and reproducibility

Statistics include mass-univariate tests for tracking coefficients of individual voxels for speech in the single-speaker condition and relevant and non-relevant speech in the auditoy scene condition. For group statistics (random effects across $n = 15$ participants), the individual coefficient maps are projected on the aligned surfaces and subject to two-tailed $t$-tests. To test for significance, after initial thresholding, a monte-carlo based cluster-based multiple comparison correction was applied. Multivariate tests were performed by comparing maps of tracking coefficients between the single-speaker and auditory scene condition. The obtained Fisher-transformed correlation values were analyzed by a repeated-measure ANOVA (factors relevance: relevant and non-relevant speech) and region-of-interest), two-tailed paired sample $t$-tests (relevant vs. non-relevant speech) and one-tailed one-sample $t$-tests (vs. theoretical chance level $r_z = 0$); corrected for multiple comparisons by false-discovery rate. For speech envelope decoding, we computed models based on the songle-speaker condition and applied these to data of the auditory scene condition. Attention decoding performance (i.e., relevant > non-relevant) was tested across participants via one-tailed

paired *t*-tests vs. decoding on surrogate data using time-shifted envelopes, corrected for multiple comparisons across regions and interval length by false discovery rate. To ensure reproducibility, we provide the analyzed processed and resulting source data in the repository and as supplemental data. Data and are available online and upon request.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The processed and analyzed data, stimulus envelopes as well as source data are available in the zenodo repository (https://doi.org/10.5281/zenodo.13359542)[59]; source data are also available as supplemental data.

## Code availability

Matlab analysis code is available in the zenodo repository (https://doi.org/10.5281/zenodo.13359542). Relevant software packages for this project are BrainVoyager (v21.4; data preprocessing and MRI visualization), BrainVoyager's COPE plugin[55] (v1.1; EPI distortion correction), Matlab (version R2022b; data processing, analysis, and figures), and neuroelf (v1.1, https://neuroelf.net; MRI data import and analysis).

Received: 24 May 2023; Accepted: 17 September 2024;
Published online: 30 September 2024

## References

1. Crosse, M. J., Di Liberto, G. M., Bednar, A. & Lalor, E. C. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **10**, 604 (2016).
2. Obleser, J. & Kayser, C. Neural entrainment and attentional selection in the listening brain. *Trends Cogn. Sci.* **23**, 913–926 (2019).
3. Brodbeck, C. & Simon, J. Z. Continuous speech processing. *Curr. Opin. Physiol.* **18**, 25–31 (2020).
4. Wöstmann, M., Fiedler, L. & Obleser, J. Tracking the signal, cracking the code: speech and speech comprehension in non-invasive human electrophysiology. *Lang. Cogn. Neurosci.* **32**, 855–869 (2017).
5. Ding, N. & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl Acad. Sci. USA* **109**, 11854–11859 (2012).
6. Zion Golumbic, E. M. et al. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77**, 980–991 (2013).
7. O'Sullivan, J. A. et al. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **25**, 1697–1706 (2015).
8. Rimmele, J. M., Zion Golumbic, E., Schröger, E. & Poeppel, D. The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex* **68**, 144–154 (2015).
9. Petersen, E. B., Wöstmann, M., Obleser, J. & Lunner, T. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *J. Neurophysiol.* **117**, 18–27 (2017).
10. Ding, N. & Simon, J. Z. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* **107**, 78–89 (2012).
11. O'Sullivan, J. et al. Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron* **104**, 1195–1209.e3 (2019).
12. Brodbeck, C., Hong, L. E. & Simon, J. Z. Rapid transformation from auditory to linguistic representations of continuous speech. *Curr. Biol.* **28**, 3976–3983.e5 (2018).
13. Forte, A. E., Etard, O. & Reichenbach, T. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *Elife* **6**, e27203 (2017).
14. Maddox, R. K. & Lee, A. K. C. Auditory brainstem responses to continuous natural speech in human listeners. *eNeuro* **5**, ENEURO.0441–17.2018 (2018).
15. Puschmann, S. et al. The right temporoparietal junction supports speech tracking during selective listening: evidence from concurrent EEG-fMRI. *J. Neurosci.* **37**, 11505–11516 (2017).
16. Friston, K. J. et al. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).
17. Peelen, M. V., Fei-Fei, L. & Kastner, S. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature* **460**, 94–97 (2009).
18. Hausfeld, L., Riecke, L. & Formisano, E. Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. *Neuroimage* **173**, 472–483 (2018).
19. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. 'Who' is saying 'what'? brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
20. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).
21. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
22. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
23. Mirkovic, B., Debener, S., Jaeger, M. & De Vos, M. Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *J. Neural Eng.* **12**, 046007 (2015).
24. Das, N., Bertrand, A. & Francart, T. EEG-based auditory attention detection: boundary conditions for background noise and speaker positions. *J. Neural Eng.* **15**, 066017 (2018).
25. Hamilton, L. S., Edwards, E. & Chang, E. F. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol.* **28**, 1860–1871.e4 (2018).
26. De Martino, F. et al. The impact of ultra-high field MRI on cognitive and computational neuroimaging. *Neuroimage* **168**, 366–382 (2018).
27. Haufe, S. et al. Elucidating relations between fMRI, ECoG, and EEG through a common natural stimulus. *Neuroimage* **179**, 79–91 (2018).
28. Kayser, C. A comparison of hemodynamic and neural responses in cat visual cortex using complex stimuli. *Cereb. Cortex* **14**, 881–891 (2004).
29. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–157 (2001).
30. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
31. Wilson, S. M., Bautista, A. & McCarron, A. Convergence of spoken and written language processing in the superior temporal sulcus. *Neuroimage* **171**, 62–74 (2018).
32. Quillen, I. A., Yen, M. & Wilson, S. M. Distinct neural correlates of linguistic and non-linguistic demand. *Neurobiol. Lang.* **2**, 202–225 (2021).
33. de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L. & Theunissen, F. E. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
34. Matchin, W. & Hickok, G. The cortical organization of syntax. *Cereb. Cortex* **30**, 1481–1498 (2020).
35. Luo, H. & Poeppel, D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54**, 1001–1010 (2007).
36. Poeppel, D. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time. *Speech Commun.* **41**, 245–255 (2003).
37. Ding, N. & Simon, J. Z. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* **33**, 5728–5735 (2013).

38. Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L. & Francart, T. Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *Hear. Res.* **380**, 1–9 (2019).
39. Etard, O. & Reichenbach, T. 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 (2019).
40. Fiedler, L., Wöstmann, M., Herbst, S. K. & Obleser, J. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *Neuroimage* **186**, 33–42 (2019).
41. Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J. & Lalor, E. C. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* **28**, 803–809.e3 (2018).
42. Brodbeck, C., Jiao, A., Hong, L. E. & Simon, J. Z. Neural speech restoration at the cocktail party: auditory cortex recovers masked speech of both attended and ignored speakers. *PLoS Biol.* **18**, 1–22 (2020).
43. Hausfeld, L., Riecke, L., Valente, G. & Formisano, E. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *Neuroimage* **181**, 617–626 (2018).
44. Khalighinejad, B., Herrero, J. L., Mehta, A. D. & Mesgarani, N. Adaptation of the human auditory cortex to changing background noise. *Nat. Commun.* **10**, 2509 (2019).
45. Puvvada, K. C. & Simon, J. Z. Cortical representations of speech in a multitalker auditory scene. *J. Neurosci.* **37**, 9189–9196 (2017).
46. Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J. & Lorenzi, C. A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* **142**, 1976–1989 (2017).
47. Lewis, L. D., Setsompop, K., Rosen, B. R. & Polimeni, J. R. Fast fMRI can detect oscillatory neural activity in humans. *Proc. Natl Acad. Sci.* **113**, E6679–E6685 (2016).
48. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* **25**, 2457–2465 (2015).
49. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
50. Shain, C., Blank, I. A., Fedorenko, E., Gibson, E. & Schuler, W. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *J. Neurosci.* **42**, 7412–7430 (2022).
51. de Cheveigné, A. & Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **111**, 1917–1930 (2002).
52. Boersma, P. & Weenink, D. Praat: doing phonetics by computer. *Ear. Hear.* **32**, 266 (2019).
53. Marques, J. P. et al. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage* **49**, 1271–1281 (2010).
54. Cumming, G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis* 1st edn, Vol. 536 (Routledge, 2012).
55. Jezzard, P. & Balaban, R. S. Correction for geometric distortion in echo planar images from B0 field variations. *Magn. Reson. Med.* **34**, 65–73 (1995).
56. Talairach, J. & Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System—an Approach to Cerebral Imaging* 1st edn, Vol. 132 (Thieme Medical Publishers, 1988).
57. Goebel, R., Esposito, F. & Formisano, E. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum. Brain Mapp.* **27**, 392–401 (2006).
58. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
59. Hausfeld, L., Hamers, I. M. H. & Formisano, E. Data from: FMRI speech tracking in primary and non-primary auditory cortex while listening to noisy scenes [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.13359542 (2024).


## Acknowledgements

We would like to thank Federico De Martino for help with data acquisition and comments on the manuscript, Anne Bach for help with participant recruitment, and our participants for their collaboration and time. This work was supported by Maastricht University, the Dutch Province of Limburg (E.F.), and the Netherlands Organization for Scientific Research (NWO; VENI grant 451-17-033 to L.H., Open Competition grant 406.20.GO.030 to E.F).

## Author contributions

L.H.: Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing—original draft, Visualization, Funding acquisition, Supervision; I.M.H.H.: Data curation, Investigation, Visualization; E.F.: Conceptualization, Writing—review and editing, Supervision.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-06913-z.

**Correspondence** and requests for materials should be addressed to Lars Hausfeld.

**Peer review information** *Communications Biology* thanks Johanna Rimmele, Jonathan H. Venezia, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Sahba Besharati, George Inglis and Tobias Goris. A peer review file is available.