Listening to speech in noisy scenes:
Antithetical contribution of primary and non-primary auditory cortex

Lars Hausfeld [a,b,*], Elia Formisano [a,b,c]

Author affiliations:

[a] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[b] Maastricht Brain Imaging Centre, 6200 MD Maastricht, The Netherlands

[c] Maastricht Centre for Systems Biology, 6200 MD, Maastricht, The Netherlands

[*] Corresponding author: Lars Hausfeld (lars.hausfeld@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Maastricht University, P.O. Box 616, 6200 MD Maastricht

Abstract

Invasive and non-invasive electrophysiological measurements during “cocktail-party”-like listening indicate that neural activity in human auditory cortex (AC) “tracks” the envelope of relevant speech. Due to the measurements’ limited coverage and/or spatial resolution, however, the distinct contribution of primary and non-primary auditory areas remains unclear. Using 7-Tesla fMRI, here we measured brain responses of participants attending to one speaker, without and with another concurrent speaker. Using voxel-wise modeling, we observed significant speech envelope tracking in bilateral Heschl’s gyrus (HG) and middle superior temporal sulcus (mSTS), despite the sluggish fMRI responses and slow temporal sampling. Neural activity was either positively (HG) or negatively (mSTS) correlated to the speech envelope. Spatial pattern analyses indicated that whereas tracking in HG reflected both relevant and (to a lesser extent) non-relevant speech, right mSTS selectively represented the relevant speech signal. These results indicate that primary and non-primary AC antithetically process ongoing speech suggesting a push-pull of acoustic and linguistic information.

Introduction

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 ECoG, MEG or EEG, to continuously changing features of the input (Crosse et al., 2015). 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 (Ding & Simon, 2012; O'Sullivan et al., 2015; Zion Golumbic et al., 2013). These effects have been localized to primary and secondary auditory cortical regions in Heschl's gyrus (HG), superior temporal gyrus (STG), and planum temporale (PT) (Ding & Simon, 2012; O'Sullivan et al., 2019) and, more recently, to subcortical areas (Forte et al., 2017; Maddox & Lee, 2018). 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. However, behavioral and EEG speech tracking results from simultaneous MRI and EEG measurements (Puschmann et al., 2017) 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 electro-magnetic neural signals.

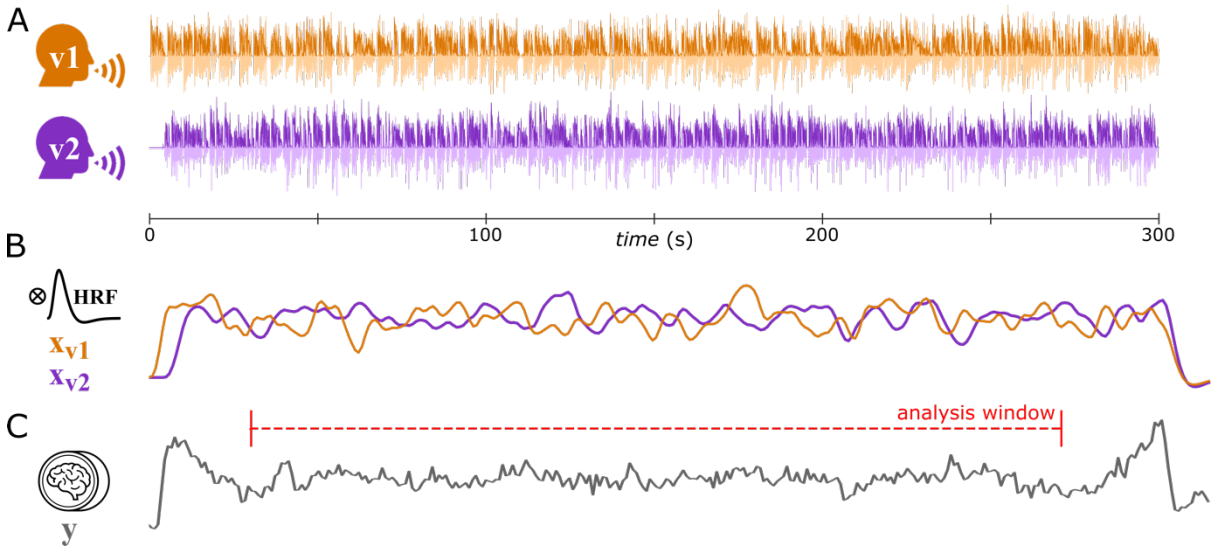

Figure 1. Example Stimuli and Processing. A) Speech waveforms during one trial (300s) for the auditory scene condition containing speech of speakers v1 and v2. Light shading shows the speech waveform; the saturated lines denote the envelope of these waveforms. Note the initial silent period for speaker v2 introduced 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 above) were convolved with a hemodynamic response function (HRF, indicated in black). 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.

In this study, we investigated fMRI neural tracking of the speech envelope by presenting 5-min blocks of task-relevant speech (Figure 1A) both with and without concurrent speech (referred to as *single speaker* and *auditory scene* condition, respectively) and measuring the participants' (N = 10) brain activity with 7T-fMRI. To avoid potential top-down effects due to repeated sound presentations, each relevant and non-relevant speech segment was presented only once to participants. In addition, to capture the influence of the envelope dynamics, we limited the 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.

While previous tracking studies employed high-temporal precision measurements (>100 Hz) that allowed analyses with complex models (fitting >100 parameters), fMRI's sampling rate (TR = 1 s = 1 Hz in this study) is two magnitudes lower, resulting in comparably shorter data time-courses and thus makes it difficult to apply similar analyses. To approach this, we acquired long data time-courses by presenting listening blocks of 5min (vs. ≤ 60 s in EEG studies). We quantified the tracking of speech voxel-by-voxel using the General Linear Model (GLM) framework that presumes a canonical

hemodynamic response (e.g., Boynton et al., 1996). A follow-up multi-voxel pattern analysis (MVPA) of these spatial tracking patterns characterized the representation of relevant and non-relevant speech.

Results

Participants follow audiobooks during MRI acquisition

We asked participants to selectively listen to the (relevant) speaker. To probe participants, questions about the audiobook's content and a subjective rating on their selective listening performance were presented after each 5-min segment. The accuracy of responses to content questions indicated that participants were able to listen selectively to single speaker and auditory scene stimuli (single speaker: 0.765 ± .071 [mean ± s.d.], $t(9) = 22.968$, $p < .001$ [vs. theoretical chance level of 0.25], Cohen's $d = 7.26$; auditory scene: 0.650 ± 0.154, $t(9) = 8.232$, $p < .001$, $d = 2.60$). This was confirmed by the participants' subjective ratings (single speaker: 8.80 ± 0.87 [mean ± s.d.]; auditory scene: 5.73 ± 1.83; 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: $t(9) = 2.365$, $p = .042$, $d_{av} = 0.96$), with participants rating their selective listening performance higher during the single speaker vs. auditory scene condition ($t(9) = 6.442$, $p < .001$, $d_{av} = 2.15$).

Listening to Audiobooks Activates the Speech Comprehension Network

A first analysis showed sustained BOLD signal (de)activation in response to listening blocks compared to baseline in regions typically involved in speech processing, for both single speaker and auditory scene conditions (Figure S1). Significant activation was observed in auditory cortical regions (HG, STG and STS) and frontal cortex (IFG); significant deactivation was found in the temporo-parietal junction (TPJ), Insula, frontal cortex (MFG) and inferior central sulcus.

The initial investigation of the fMRI data time courses during listening blocks (cf. Figure 1C) revealed early (expected; sound onset) and late (unexpected; preceding sound offset) BOLD signal increases. To mitigate these tracking-unrelated effects at on- and offset when analyzing the tracking of

speech, we restricted our analysis to the central 4-min period of listening blocks by cutting the first and final 30 s (Figure 1C).

BOLD responses Reveal 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 (Figure 1B). 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 posterior to HS (Figure 2A). We also found regions with significant fitting and negative β-values in middle superior temporal sulcus (mSTS). These results reveal that low temporal resolution BOLD-fMRI responses track speech amplitude envelopes, extending previous results obtained with direct and high temporal resolution neural measures (EEG, ECoG, MEG).

To relate the results of this analysis to those of previous speech tracking studies, 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 Figure 2B, respectively) to which we refer to as *positive* and *negative tracking*. Interestingly, further analyses showed that areas in bilateral HG displaying positive tracking also displayed significant sustained positive activity with regard to pre-stimulus baseline (i.e., no sound)(Figure 2C, red color). For areas in mSTS displaying negative tracking, we found positive sustained activity for most vertices of the right but not the left hemisphere, where most vertices instead presented weak overall response (Figure 2C, blue color). This indicates 1) that significant activation in response to speech vs. pre-stimulus baseline is not a prerequisite for the significant tracking of the speech envelope and 2) that a positive BOLD response to sound with respect to pre-stimulus baseline can show negative tracking. Varying the HRF model indicated that our observed results were robust to a wide range of values of the time-to-peak parameter (3.5-7 s) and showed a tendency for a shorter time-to-peak for medial HG when compared to STS for efficient speech tracking (Figure S2). The results presented here were obtained with a time-to-peak parameter of 4.5 s, providing a compromise between fitting BOLD responses in both the HG and mSTS regions.

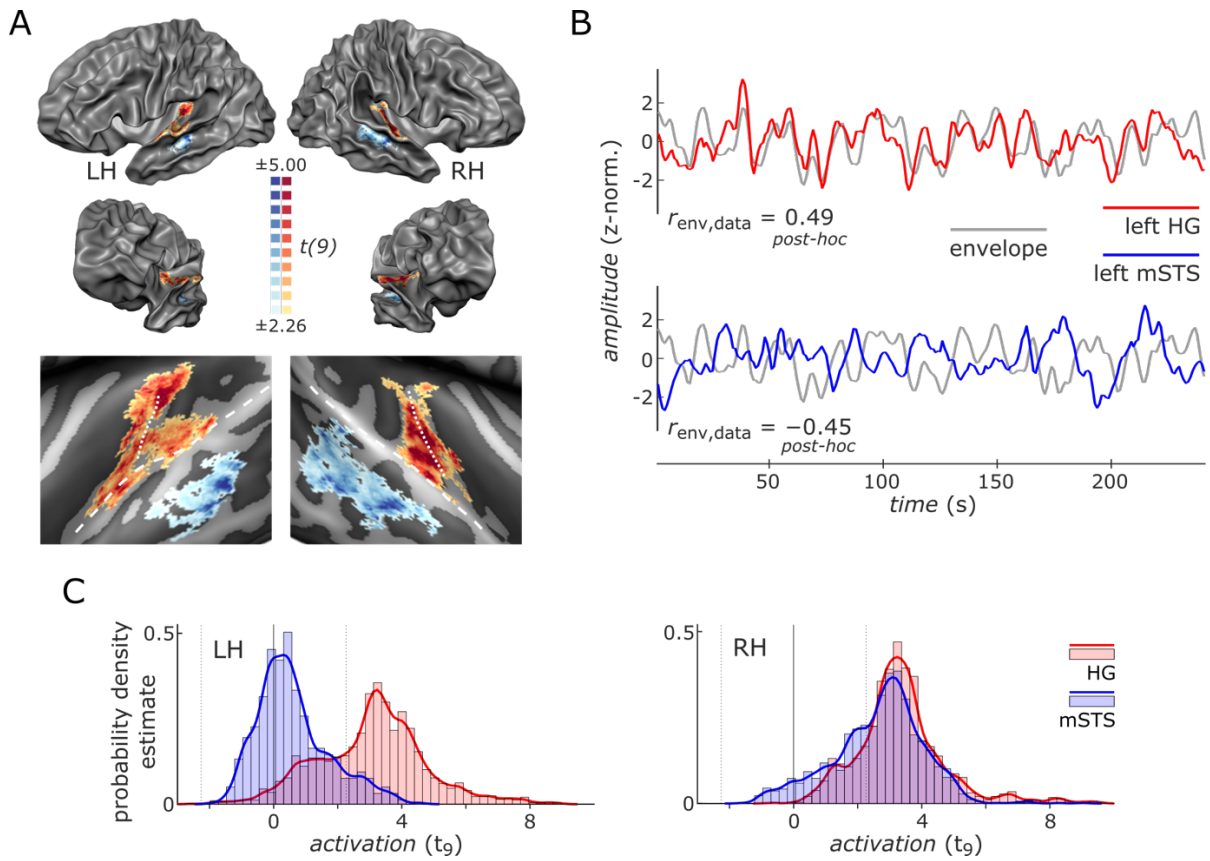

Figure 2. Overview Speech Tracking Single Speaker Condition. A) Color-coding depicts regions showing significant *positive* (warm colors) and *negative* (cold colors) speech tracking in the single speaker condition across participants in the left (LH) and right hemisphere (RH). Upper and middle panels show lateral and frontal views of reconstructed average grey-white matter boundary after cortex-based alignment. Lower panels show enlarged views of temporal cortex on the inflated boundary. Dotted and dashed white 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 HG (red line) and left middle STS (blue line) showing significant speech tracking. *Positive tracking* in HG is indicated by the positive temporal correlation between MRI data and speech envelope (grey line) whereas *negative tracking* in middle STS 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 lower left of each panel indicate the temporal correlation of MRI data and envelope time courses in this example. For visualization only, data time courses were smoothed using a moving average including the prior and subsequent sample. Note that this (anti-)correlation is 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 regions that tracked the speech envelope (see panel A). Significant activation in tracking regions was found in bilateral HG and right mSTS but only in few locations in left mSTS. The distribution of *t*-values is indicated by thick curved lines and light-colored bars that show the estimate of the probability density function and the normalized histogram, respectively. Statistical maps are thresholded at $p < .05$ (two-tailed) and corrected for multiple comparisons by cluster-size ($p < .05$).

Tracking Patterns Reveal (Non-)Relevant Speech Processing in HG and mSTS Regions

Previous research highlighted that activation patterns across the (auditory) cortical surface represent auditory objects including speech streams and that these spatio-temporal representations are

significantly affected by selective attention in multi-talker scenes (Formisano et al., 2008; Hausfeld, Riecke, & Formisano, 2018; Mesgarani & Chang, 2012; O'Sullivan et al., 2019; Pasley et al., 2012; Zion Golumbic et al., 2013). 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., left and right HG and 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 (see Figure 2A) with those from the auditory scene condition, obtained including amplitude envelope predictors of the non-relevant speech in addition to the relevant speech (see Methods).

Overall, our multivariate results show that the pattern similarity of tracking maps across tasks was significant in the left and right HG for both relevant and irrelevant speech (Figure 3A; $q < .05$; $t(9) > 3.908$, $p < .002$, Cohen's $d > 1.24$) as well as in the right and left mSTS regions for relevant speech (Figure 3B; $q < .05$; $t(9) > 2.802$, $p < .010$, $d > 0.88$) but not for non-relevant speech ($t(9) < 0.963$, $p < .002$, $d < 0.31$; MCC across 8 tests). This indicated 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 and mSTS and non-relevant speech in the HG regions.

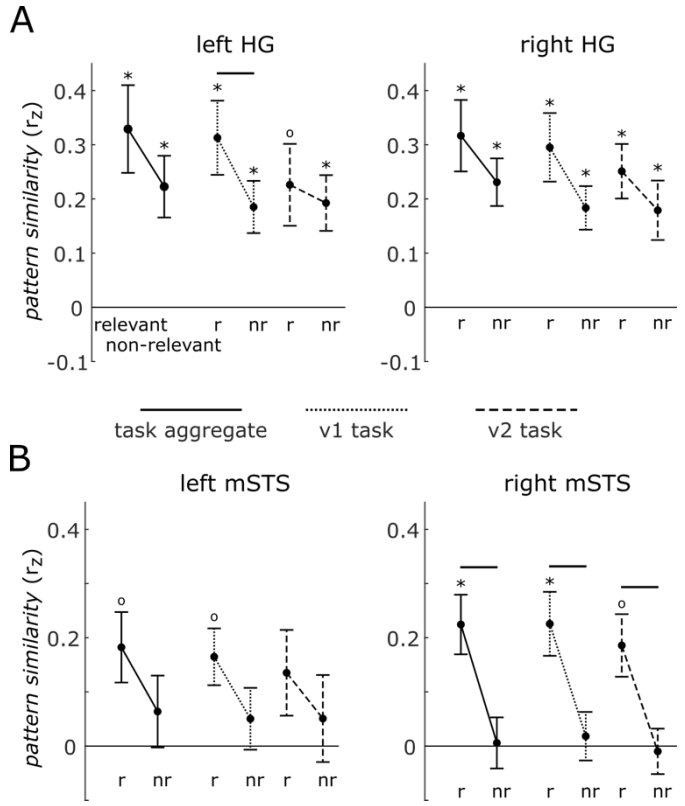

Figure 3. Analysis of Pattern Similarity between Tracking Maps of Single Speaker and Auditory Scene Conditions in Regions-of-Interest. A) Similarity of spatial patterns of speech tracking between the single speaker and auditory scene condition for the HG region in the left and right hemisphere. Regions of interest were identified as the HG regions found for the single speaker conditions (see Fig. 1A). B) Same as in A but for the mSTS regions in the left and right hemisphere. Symbols denote significant pattern similarity (*$p < .01$, °$p < .05$, two-tailed) and straight lines show differences in pattern similarity (solid lines: $p < .01$, dashed lines: $p < .05$, two-tailed) between the relevant speech (r) and the non-relevant speech (nr).

Tracking Patterns Reveal Dominant Processing of Relevant Speech in Right mSTS Region

Moreover, when analyzing the tracking map similarity in the left hemisphere, we found that it was marginally affected by both speech relevance (relevant and non-relevant; $F(1,9) = 4.11$, $p = .073$, $\eta_p^2 = 0.31$) and region (HG and mSTS; $F(1,9) = 4.43$, $p = .065$, $\eta_p^2 = 0.33$) but not by a relevance-by-region interaction ($F(1,9) = 0.01$, $p = .933$, $\eta_p^2 < 0.01$; repeated-measures ANOVA). For the right hemisphere, we found similarly that the tracking map similarity was affected by both relevance ($F(1,9) = 15.11$, $p = .004$, $\eta_p^2 = 0.63$) and region ($F(1,9) = 34.77$, $p < .001$, $\eta_p^2 = 0.79$). In contrast to the left hemisphere, these factors interacted significantly ($F(1,9) = 9.33$, $p = .014$, $\eta_p^2 = 0.51$; repeated-measures ANOVA). Post-hoc analyses showed that the pattern similarity between tracking maps in the right mSTS region of the single speaker condition were more similar to tracking maps for relevant speech as compared to non-relevant speech ($q < .05$; $t(9) = 4.491$, $p = .002$, $d_{av} = 1.35$). Note that right HG showed a tendency for the modulation of the similarity of tracking maps with relevance ($t(9) = 4.491$, $p = .063$, $d_{av} = 0.48$); this modulation, however, was less strong as compared to right mSTS (HG: $\Delta r = 0.086$; mSTS: $\Delta r = 0.219$; $t(9) = -3.056$, $p = .014$, $d_{av} = 0.94$).

Overall, these results suggested that speech relevance modulated the similarity of tracking maps to the single speaker condition with higher similarity for the relevant vs. non-relevant speech and that the pattern similarity of tracking maps was higher in the HG vs. the mSTS region. In addition, our observations indicate that neural responses in right mSTS reflect a processing stage at which the cocktail-party problem is resolved such that relevant speech is predominantly represented (Hausfeld et al., 2018).

Processing in Lateral Clusters of HG Region is Modulated by Speech Relevance

To further characterize the pattern similarity and effects of speech relevance within the HG and mSTS regions, we performed the same pattern analysis for sub-regions, as defined by data-driven clustering of their fMRI time courses (Figure 4). For the HG region, these results indicated that, for the two hemispheres, in particular the more lateral portions including lateral and posterior HG and adjacent posterior STG (red cluster) showed both a high pattern similarity and their modulation by speech

relevance. While we still observed significant similarity of tracking maps, more medial clusters of the HG region (yellow, orange) did not show a modulation by speech relevance. For the mSTS region, the analyses of clusters robustly indicated two cluster in right mSTS, both of which had significant pattern similarity for the relevant speech only and a significant effect of speech relevance. Conversely, the left mSTS region was not divided into different clusters.

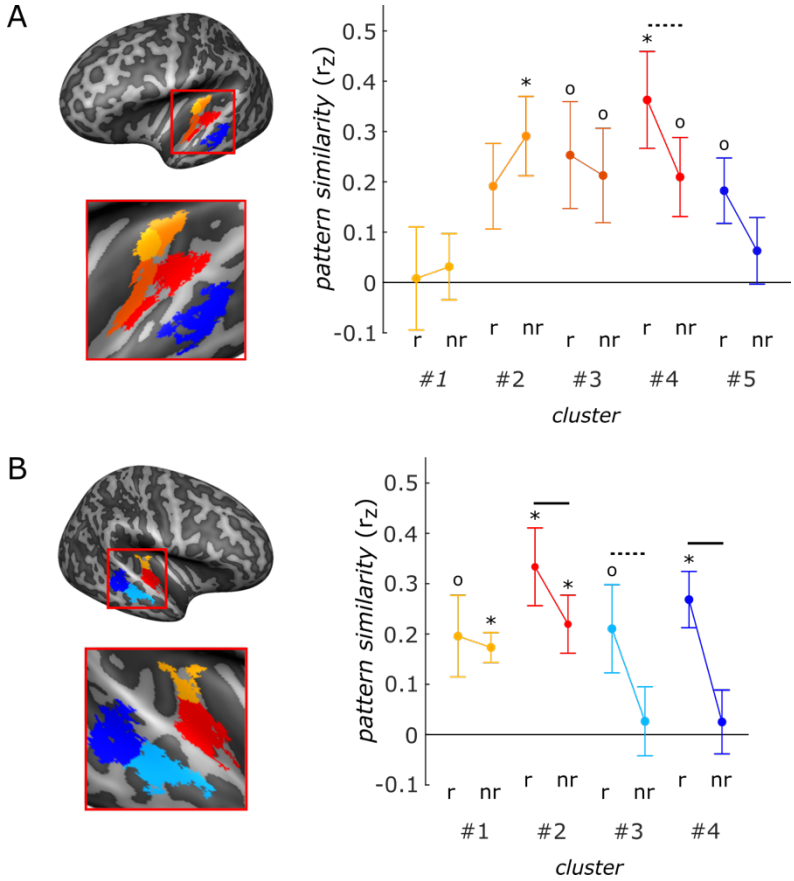

Figure 4. Analysis of Pattern Similarity between Tracking Maps of Single Speaker and Auditory Scene Conditions in data-driven Clusters. A) Panels show the five clusters (limited to auditory scene map "across tasks") in the left hemisphere and the spatial similarity for tracking maps of relevant and non-relevant speech (r and nr, respectively). Colored areas on inflated brain representations denote the loci of the five clusters denoted as clusters #1 - #5 right graphs. The red square indicates the area of the cut-out below. B) Same as A but for the right hemisphere separated into four clusters. Clusters were determined by *k*-means clustering in the left (A) and right hemisphere (B); *k* was determined by leave-participant-out bootstrapping and the silhouette index for internal cluster validation. For indicators of statistical significance, see Figure 2.

Discussion

In this study, we measured neural response to continuous speech stimuli using high field fMRI. 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.

175 Interestingly, our results showed positive and negative tracking of the speech envelope by the HG
176 and mSTS regions, respectively. Both positive and negative tracking was found in areas that showed,
177 apart from left mSTS, increased sustained activity in response to speech sounds (i.e., the speech
178 envelope modulated the signal "on the plateau" of the positive BOLD activation). We interpret the
179 positive tracking in the HG region to reflect the ongoing envelope amplitude of the speech stream. This
180 is in line with previous ECoG studies showing that activity in HG and middle STG is correlated with
181 responses to speech (Hamilton et al., 2018; O'Sullivan et al., 2019). We interpret the negative tracking
182 observed in the mSTS region to reflect cortico-cortical top-down signals that aid in following relevant
183 speech in particular during periods of low speech audibility, i.e., when the task is more difficult (e.g.
184 due to the fMRI noise and lower intensity of relevant speech). Layer-resolved high-field MRI
185 acquisitions might help better defining the role of this region in terms of top-down and bottom-up input
186 when listening to a single speaker or an auditory scene (e.g., De Martino et al., 2018). Whether these
187 regions of BOLD speech tracking coincide with the sources of the speech tracking observed with neuro-
188 electromagnetic signals remains an open question. The relation between neural activity, electric and
189 hemodynamic signals is complex (Haufe et al., 2018; Kayser, 2004; Logothetis et al., 2001). Concurrent
190 measures of speech tracking by EEG and fMRI would allow linking, within participants, the observed
191 results by hemodynamic and neuroelectric measures more directly and shed light on the underlying
192 neural processes (Puschmann et al., 2017).
193 Examining the spatial patterns of speech tracking maps for the HG region, we found a high
194 similarity between the single speaker and auditory scene conditions for both the relevant and non-
195 relevant speech. These results indicate that the overall incoming speech signal, containing the relevant
196 and non-relevant speech, is reflected in the HG region including medial and lateral HG/HS and adjacent
197 STG. We observed a tendency that the pattern similarity was higher for relevant speech in comparison
198 to non-relevant speech (Figure 3A) in particular in a lateral cluster of the HG region of both hemispheres
199 (Figure 2), which contained lateral HG and adjacent left posterior STG. This is in line with previous
200 observations showing faithful envelope representations as well as attentional modulation and high
201 selectivity for attended speech in this region using ECoG (O'Sullivan 2019), MEG (Ding & Simon,
202 2012) and fMRI (de Heer et al., 2017). Note that the pattern similarity of tracking maps for non-relevant

speech, although being significantly lower, was still significant suggesting residual information about non-relevant speech in these clusters. 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. This result might indicate that right mSTS processes exclusively information reflecting the relevant speech implying that the cocktail party is resolved at this stage. This fits 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 speech specific such that relevant music did not show significant pattern similarity (Hausfeld, Riecke, & Formisano, 2018). Another possible explanation for these results is that activation in this area reflects increased top-down control of selective listening at a temporal scale of envelope changes (i.e., phonemes, syllables and words) with decreasing amplitude of the relevant speaker and thus increasing energetic masking. 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)(Hamilton et al., 2018) 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 (de Heer et al., 2017). While this is in agreement with our findings in the HG region, the significant tracking of the envelope in mSTS presumably tracking semantic features might be explained by correlations between semantic features and amplitude envelope. Additional differences, for example in analyses or data acquisition (3T vs. 7T MRI, 1Hz vs 0.5Hz 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.

Our analyses showed tracking of 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 (Brodbeck et al., 2020; Hausfeld, Riecke, Valente, et al., 2018; Khalighinejad et al., 2019; Puvvada & Simon, 2017), 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.

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. These results suggest neural processes potentially related to stronger feedback and linguistic integration processing in mSTS as compared to HG. In addition, these results provide support for neural signals in mSTS that reflect a processing stage at which the cocktail-party is resolved.

Materials and Methods

Participants

Ten students (native German speakers) of Maastricht University (9 female, mean age [range]: 23 [21 27] 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)(Hausfeld, Riecke, Valente, et al., 2018) of one female (*v1*; f_0 = 159 ± 8.3 Hz, mean ± s.d.) and one male speaker (*v2*; f_0 = 107 ± 7.3 Hz). The fundamental frequency f_0 for each excerpt was determined by averaging f_0 contours obtained with the YIN algorithm (de Cheveigné & Kawahara, 2002). 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 (Boersma & Weenink, 2019).

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 block lasted 5 min. During the auditory scene condition, the target speech started 4.5 s before the distractor speech to provide listeners with an auditory cue indicating the target speech (see Figure 1).

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 (Marques et al., 2010; voxel size: 0.65mm 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 whole cortex coverage were collected using an echo-planar imaging (EPI) sequence with multiband 3 acceleration (57 slices; voxel size: 1.5 mm isotropic; FoV = 192 x 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 run.

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 (Figure S3). 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 10s (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 (Hausfeld, Riecke, Valente, et al., 2018).

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, temporal high-pass filtering (11 cycles per run ≈ 0.015 Hz). EPI distortions were corrected using BrainVoyager's COPE plugin (v1.1; Jezzard and

Balaban, 1995). Functional runs were individually aligned to anatomical scans and transformed to Talairach space (Talairach and Tournoux, 1988). The functional data was spatially smoothed (4 mm FWHM) and individual maps were projected onto the group aligned surface via cortex-based alignment (Goebel et al., 2006) to create group maps.

fMRI Activation Analysis

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 using a general linear model (GLM). Because we observed strong onset and offset effects and were interested in the sustained activation for the tracking analysis (see fMRI Tracking Analysis), listening blocks were modelled by three predictors reflecting the onset, sustained and offset responses.

fMRI Tracking Analysis

Before analysis, the functional data was cut to 4min per block by removing the initial and final 30s in order to avoid confounds from onset or offset effects observed during data exploration (Figure 1C). Subsequently, the functional data was analyzed - voxel-by-voxel - for the tracking of speech envelopes by making use of the GLM framework. More specifically, we modelled fMRI voxel time courses by $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ (where y denotes a voxel time course, X a design matrix, β coefficients and ε the error term). The design matrix included the main predictors for the speech envelope time courses and confound predictors reflecting participant's motion, level difference between first and second block within each run 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.

Spatial Pattern Analysis of Tracking Maps

To investigate BOLD activity during listening to auditory scenes, we analyzed the spatial patterns. For this pattern similarity analysis, the tracking maps in the HG and mSTS regions from the analysis of single speaker tracking were used as templates. The pattern similarity was computed between the

tracking maps in the same regions obtained from BOLD activity during auditory scene presentations for each relevant voice using Pearson's correlation and Fisher's transformation $r_z = 0.5 \cdot ln\left(\frac{1-r}{1+r}\right)$. Statistical testing of these scores within and across regions was done via repeated-measure ANOVAs as well as paired *t*-tests. For ANOVAs and paired *t*-tests, effect sizes were estimated using partial η square (η_p^2) and Cohen's d using the averaged variance in the denominator $d_{av} = \frac{\mu_1 - \mu_2}{\sqrt{0.5 \cdot (\sigma_1^2 + \sigma_2^2)}}$, respectively.

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.).

Author Contributions

L.H.: Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Visualization, Funding acquisition, E.F.: Conceptualization, Writing – review & editing, Supervision

Declaration of Interests

The authors declare no competing interests.

References

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by Computer* (6.0.55, retrieved 13 June 2019). http://www.praat.org

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1. *The Journal of Neuroscience*, *16*(13), 4207–4221. https://doi.org/10.1523/JNEUROSCI.16-13-04207.1996

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party : Auditory cortex recovers masked speech of both attended and ignored speakers. *BioRxiv*. https://doi.org/10.1101/866749

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930. https://doi.org/10.1121/1.1458024

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

De Martino, F., Yacoub, E., Kemper, V., Moerel, M., Uludağ, K., De Weerd, P., Ugurbil, K., Goebel, R., & Formisano, E. (2018). The impact of ultra-high field MRI on cognitive and computational neuroimaging. *NeuroImage*, *168*(March 2017), 366–382. https://doi.org/10.1016/j.neuroimage.2017.03.060

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). “Who” Is Saying “What”? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970–973.

https://doi.org/10.1126/science.1164318

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *ELife*, *6*, e27203. https://doi.org/10.7554/eLife.27203

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, *28*(12), 1860-1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Haufe, S., DeGuzman, P., Henin, S., Arcaro, M., Honey, C. J., Hasson, U., & Parra, L. C. (2018). Elucidating relations between fMRI, ECoG, and EEG through a common natural stimulus. *NeuroImage*, *179*(May), 79–91. https://doi.org/10.1016/j.neuroimage.2018.06.016

Hausfeld, L., Riecke, L., & Formisano, E. (2018). Acoustic and higher-level representations of naturalistic auditory scenes in human auditory and frontal cortex. *NeuroImage*, *173*, 472–483. https://doi.org/10.1016/j.neuroimage.2018.02.065

Hausfeld, L., Riecke, L., Valente, G., & Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, *181*, 617–626. https://doi.org/10.1016/j.neuroimage.2018.07.052

Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine*, *34*(1), 65–73. https://doi.org/10.1002/mrm.1910340111

Kayser, C. (2004). A Comparison of Hemodynamic and Neural Responses in Cat Visual Cortex Using Complex Stimuli. *Cerebral Cortex*, *14*(8), 881–891. https://doi.org/10.1093/cercor/bhh047

Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nature Communications*, *10*(1), 2509. https://doi.org/10.1038/s41467-019-10611-4

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. https://doi.org/10.1038/35084005

Maddox, R. K., & Lee, A. K. C. (2018). Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners. *Eneuro*, *5*(1), ENEURO.0441-17.2018. https://doi.org/10.1523/ENEURO.0441-17.2018

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*. https://doi.org/10.1016/j.neuron.2019.09.007

Pasley, B. N., David, S. V, Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1), e1001251. https://doi.org/10.1371/journal.pbio.1001251

Puschmann, S., Steinkamp, S., Gillich, I., Mirkovic, B., Debener, S., & Thiel, C. M. (2017). The Right Temporoparietal Junction Supports Speech Tracking During Selective Listening: Evidence from Concurrent EEG-fMRI. *Journal of Neuroscience*, *37*(47), 11505–11516. https://doi.org/10.1523/JNEUROSCI.1007-17.2017

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *Journal of Neuroscience*, *37*(38), 9189–9196. https://doi.org/10.1523/JNEUROSCI.0938-17.2017

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Supplementary Figures

Figure S1

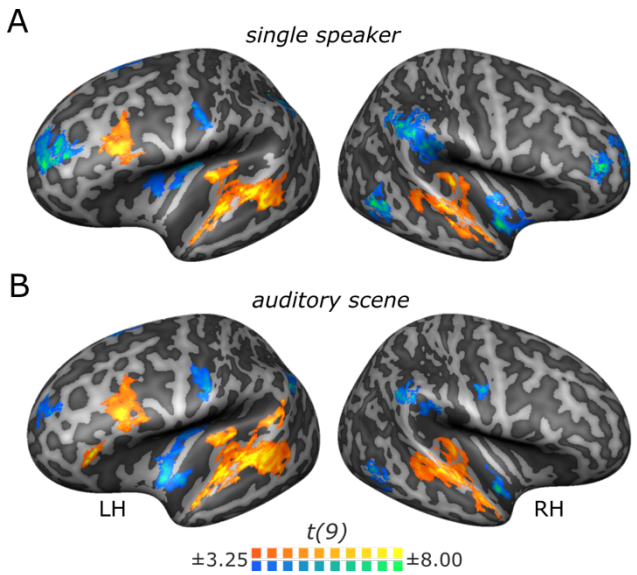

Figure S1. Sustained Activation and Deactivation during Listening Blocks. A) Cortical Maps show regions with significant sustained responses during listening blocks as compared to baseline periods (i.e., no sound) during the single speaker condition for the right and left hemisphere. B) Same as A) but for the auditory scene condition. Statistical maps are thresholded at $p < .01$ (two-tailed) and corrected for multiple comparisons by cluster-size. GLMs included, in addition, predictors coding for onset and offset responses (maps not shown).

422 ***Figure S2***

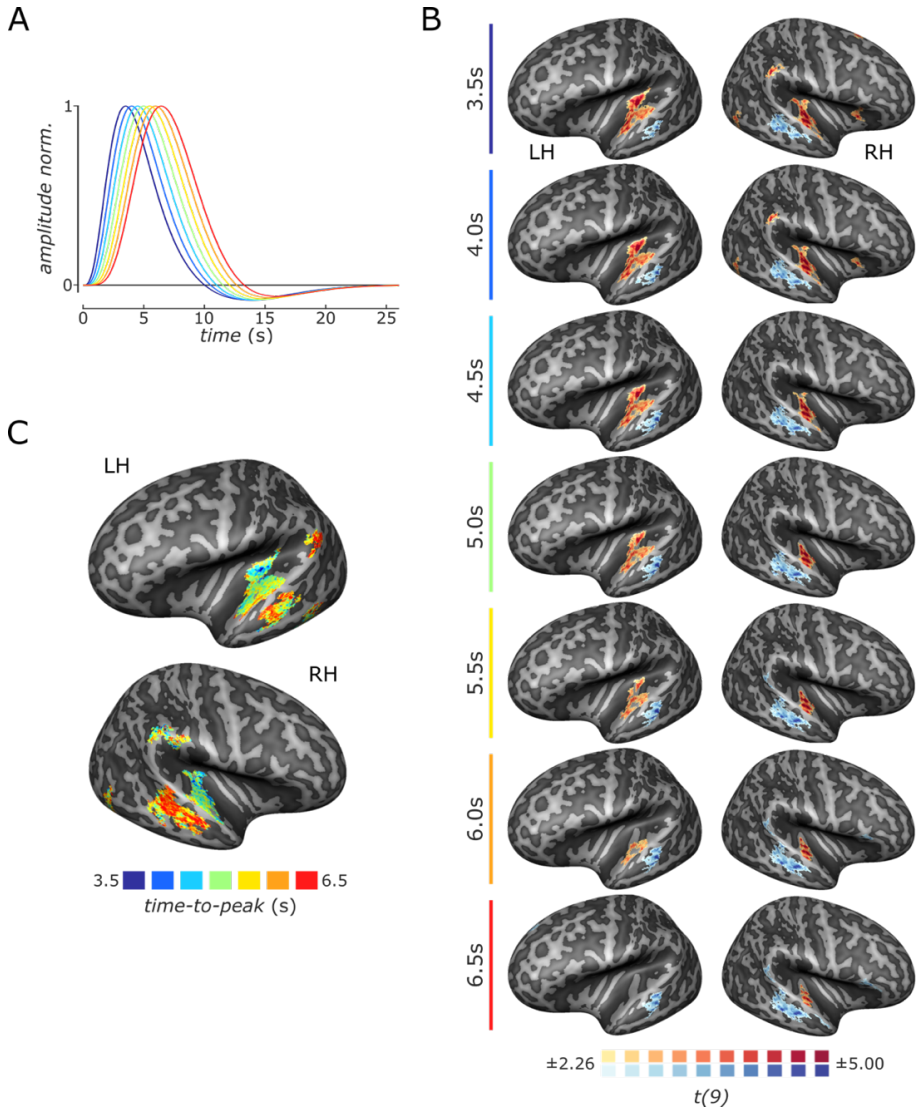

Figure S2. Time-to-Peak Dependency of Speech Tracking. A) Curves present 7 hemodynamic response functions (HRFs) created by varying the time-to-peak latency, a crucial parameter when specifying the HRF for analysis via GLM. Colors indicate the HRFs with different time-to-peak parameters (see panel C) for color coding). B) Results of speech tracking for the single speaker condition for analyses with the different HRFs. Color-coded regions show significant speech tracking across participants ($p < .05$, multiple comparison corrected by cluster size). Maps are presented on inflated reconstructions of the average grey-white matter boundary after cortex-based alignment. C) Overview of the time-to-peak parameter resulting in best speech tracking performance. These maps (upper panel: left hemisphere; lower panel: right hemisphere) are restricted to loci showing significant speech tracking for at least one model.

423

Figure S3

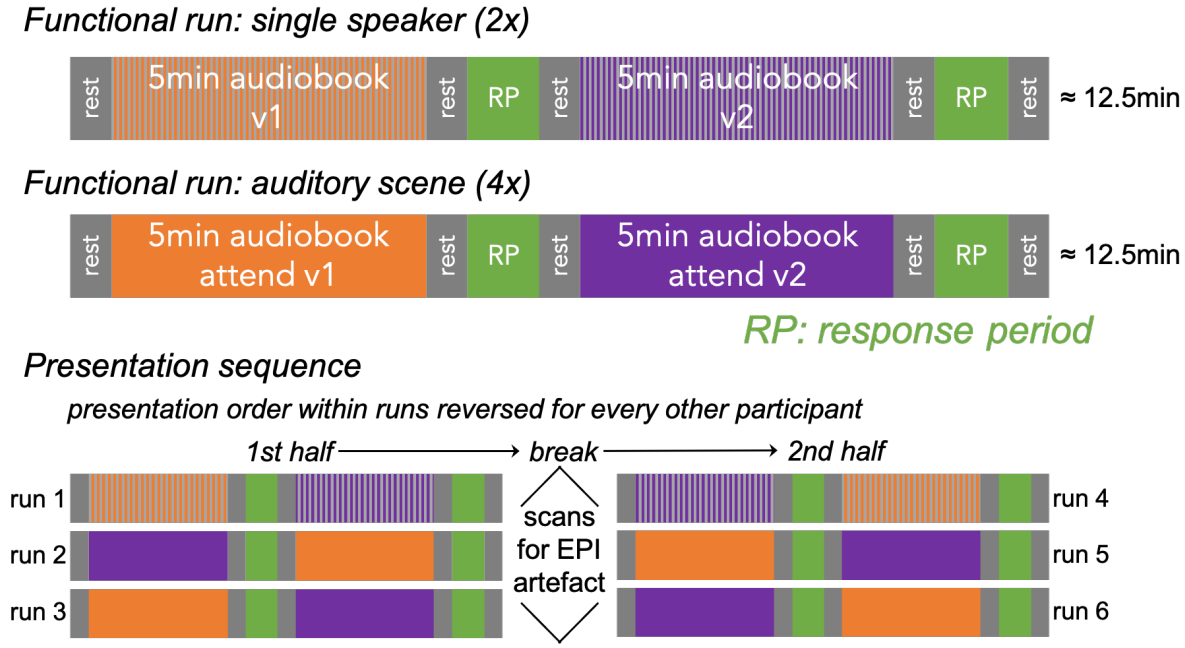

Figure S3. Overview of MRI-Design. The upper plot shows examples of functional runs for the single speaker (repeated twice during each scanning sessions) and auditory scene conditions (repeated four times). After presentation of the single speaker or auditory scene stimuli (5 min) and a rest period (15 s), participants were asked to answer 5 questions on the presented content and to rate their listening performance. After another rest period, another single speaker or auditory scene conditions was presented followed by the response period. The lower plot shows the overview one scanning session. These were split in two halves by scans used for EPI artefact correction after 3 functional runs.