Behavioral/Cognitive

The Right Temporoparietal Junction Supports Speech Tracking During Selective Listening: Evidence from Concurrent EEG-fMRI

Sebastian Puschmann,[1,2,3] Simon Steinkamp,[2,4] Imke Gillich,[2] Bojana Mirkovic,[3,5] Stefan Debener,[3,5] and Christiane M. Thiel[2,3]

[1]Montreal Neurological Institute, Cognitive Neuroscience Unit, McGill University, Montreal, Quebec H3A 2B4, Canada, [2]Biological Psychology Laboratory, Department of Psychology, European Medical School, Carl von Ossietzky Universität, 26111 Oldenburg, Germany, [3]Cluster of Excellence Hearing4All, 26111 Oldenburg, Germany, [4]Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Jülich Research Center, 52425 Jülich, Germany, and [5]Neuropsychology Lab, Department of Psychology, European Medical School, Carl von Ossietzky Universität, 26111 Oldenburg, Germany

Listening selectively to one out of several competing speakers in a "cocktail party" situation is a highly demanding task. It relies on a widespread cortical network, including auditory sensory, but also frontal and parietal brain regions involved in controlling auditory attention. Previous work has shown that, during selective listening, ongoing neural activity in auditory sensory areas is dominated by the attended speech stream, whereas competing input is suppressed. The relationship between these attentional modulations in the sensory tracking of the attended speech stream and frontoparietal activity during selective listening is, however, not understood. We studied this question in young, healthy human participants (both sexes) using concurrent EEG-fMRI and a sustained selective listening task, in which one out of two competing speech streams had to be attended selectively. An EEG-based speech envelope reconstruction method was applied to assess the strength of the cortical tracking of the to-be-attended and the to-be-ignored stream during selective listening. Our results show that individual speech envelope reconstruction accuracies obtained for the to-be-attended speech stream were positively correlated with the amplitude of sustained BOLD responses in the right temporoparietal junction, a core region of the ventral attention network. This brain region further showed task-related functional connectivity to secondary auditory cortex and regions of the frontoparietal attention network, including the intraparietal sulcus and the inferior frontal gyrus. This suggests that the right temporoparietal junction is involved in controlling attention during selective listening, allowing for a better cortical tracking of the attended speech stream.

Key words: attention; auditory perception; cognition; electroencephalography; functional magnetic resonance imaging; multimodal imaging

Significance Statement

Listening selectively to one out of several simultaneously talking speakers in a "cocktail party" situation is a highly demanding task. It activates a widespread network of auditory sensory and hierarchically higher frontoparietal brain regions. However, how these different processing levels interact during selective listening is not understood. Here, we investigated this question using fMRI and concurrently acquired scalp EEG. We found that activation levels in the right temporoparietal junction correlate with the sensory representation of a selectively attended speech stream. In addition, this region showed significant functional connectivity to both auditory sensory and other frontoparietal brain areas during selective listening. This suggests that the right temporoparietal junction contributes to controlling selective auditory attention in "cocktail party" situations.

Introduction

Listening selectively to one speaker in the presence of background noise or competing speech streams poses an ultimate hearing challenge. Since the early work by Cherry (1953), the sensory, cognitive, and neural processes contributing to solving this "cocktail party problem" have been investigated (McDermott, 2009; Anderson and Kraus, 2010; Rönnberg et al., 2013; Bidel-

Received April 13, 2017; revised Aug. 28, 2017; accepted Sept. 5, 2017.

Author contributions: S.P., S.S., I.G., B.M., S.D., and C.M.T. designed research; S.P., S.S., and I.G. performed research; S.P., S.S., and B.M. analyzed data; S.P. and C.M.T. wrote the paper.

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft Cluster of Excellence "Hearing4all" Grant EXC 1077/1 and Research Grant PU590/1). We thank Markus Garrett for help with EEG data acquisition.

The authors declare no competing financial interests.

Correspondence should be addressed to Sebastian Puschmann, Montreal Neurological Institute, Cognitive Neuroscience Unit, 3801 Rue University, Montreal, QC H3A 2B4, Canada. E-mail: sebastian.puschmann@mail.mcgill.ca.

DOI:10.1523/JNEUROSCI.1007-17.2017

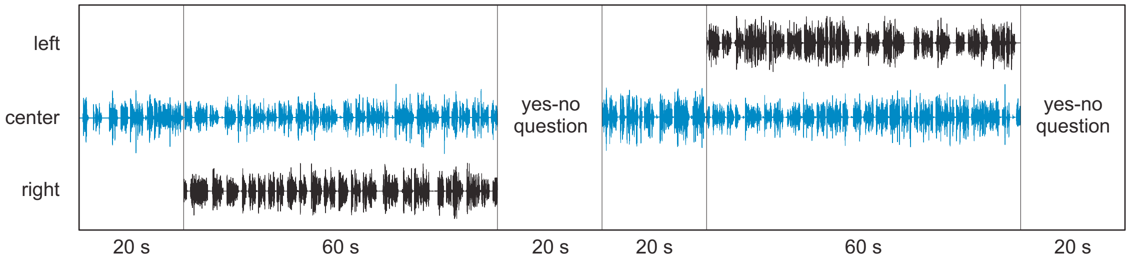

Figure 1. Selective listening task. Participants were instructed to listen attentively to a continuous speech stream. The stream was presented diotically over headphones, thus appearing to originate form a central sound source (blue). A second, competing speech stream originating from the left or right side had to be ignored (black). Selective listening blocks of 60 s duration were interspersed with 20 s single speaker blocks, in which only the to-be-attended stream were present. Yes/no questions on the content of the to-be-attended stream served as control for a correct allocation of selective attention.

man and Dexter, 2015; Bronkhorst, 2015; Coffey et al., 2017). On the neural level, it has been shown that competing auditory streams are encoded individually at the level of the auditory cortex (Simon, 2015). Focusing attention selectively to one out of several competing sound streams results in an increased representation of the attended stream in the auditory cortex while simultaneously decreasing responses to the irrelevant input (Teder et al., 1993; Bidet-Caulet et al., 2007; Ross et al., 2010; Xiang et al., 2010; Power et al., 2011; Ding and Simon, 2012a; Mesgarani and Chang, 2012). Therefore, the attended stream dominates the ongoing auditory cortex activity. Recent studies successfully demonstrated that the relative cortical representation of competing speech streams in multispeaker situations can be assessed reliably from electrophysiological responses using stimulus reconstruction approaches (Ding and Simon, 2012b; Zion Golumbic et al., 2013; Crosse et al., 2016b). Stimulus reconstruction accuracies are typically higher for selectively attended than for ignored speech streams, allowing for a robust classification of the listener's focus of attention (Mirkovic et al., 2015). Further, individual differences in stimulus reconstruction accuracies have been linked to selective listening performance, suggesting that this measure can be used to quantify the strength of attention directed to the to-be-attended speech stream (O'Sullivan et al., 2015).

Two frontoparietal brain networks have been related to controlling attention (Corbetta et al., 2008; Vossel et al., 2014). Although both the dorsal and the ventral attention networks, which are involved in the voluntary allocation of attention and stimulus-driven reorienting, respectively, have been primarily identified in the visual domain, there is increasing evidence that similar brain regions also control auditory attention (Green et al., 2011; Kim, 2014; Kong et al., 2014; Lee et al., 2014; Michalka et al., 2015). Consistent with this, core regions of the dorsal, but also of the ventral, attention network have been found to be activated during selective listening (Hill and Miller, 2010; Seydell-Greenwald et al., 2014). Several studies suggest a role of the intraparietal sulcus and superior parietal lobe for controlling spatial auditory attention (Kerlin et al., 2010; Ahveninen et al., 2013; Huang et al., 2014). However, the frontal eye fields and the right temporoparietal junction have also been reported to be activated in this context (Lee et al., 2012; Larson and Lee, 2014).

Although prior research demonstrated both attentional modulations in the auditory sensory tracking of speech streams and frontoparietal brain activity during selective listening, the relationship between both observations is not well understood. We therefore combined an EEG-based stimulus reconstruction approach with concurrently acquired fMRI data to study the relationship between sustained blood oxygenation level-dependent (BOLD) responses in frontoparietal brain areas and attentional modulations in the cortical tracking of to-be-attended and to-be-ignored speech streams derived from EEG. We applied a natural selective listening setting, in which participants listened selectively to a continuous centrally presented speech stream while ignoring a competing stream presented to their left or right. Immediate questions and a delayed recall of the story content served as control for a correct allocation of attention.

We hypothesized that in brain regions involved in controlling auditory selective attention in multispeaker situations, sustained BOLD activation levels should covary with the degree of attention directed to the to-be-attended speech stream, and thus with the strength of its cortical tracking. We further expected that regions showing such a relationship should also show increased functional connectivity to auditory sensory areas during selective listening.

Materials and Methods

Subjects

Twenty-three healthy, young volunteers (15 female; mean age: 24 ± 2 years; age range: 19–29 years) were recruited at the University of Oldenburg for participation in the experiment. All subjects were right-handed and had normal hearing (hearing thresholds <20 dB HL; frequency range: 125–8 kHz) and no history of neurological or psychiatric disease. Written informed consent was obtained from all participants. All experimental procedures were performed in agreement with the Declaration of Helsinki and were approved by the ethics committee of the University of Oldenburg.

Procedure

The experiment consisted of two sessions taking place on separate days. During the first session, participants performed a complete run of the selective listening task in a laboratory environment while scalp EEG was recorded from 96 channels. This served both to familiarize participants with the task and to record clean EEG data from each subject for a comparison of speech envelope reconstruction quality with the data obtained during the fMRI measurements. Subsequently, an additional behavioral dichotic deviance detection task was applied to gather information about individual spatial selective listening abilities. During the second session, EEG and fMRI data were acquired concurrently while subjects performed the selective listening task inside the MRI scanner. Subsequently, resting-state fMRI time courses were obtained with eyes open. After finishing the fMRI recordings, participants were asked to recall freely the content of the to-be-attended story, to rate their subjective selective listening performance, and to indicate their distraction by the to-be-ignored speech stream and the scanner background noise during the task.

Selective listening task

The selective listening task is depicted in Figure 1. During two-speaker blocks, participants had to listen attentively to a continuous speech stream (indicated in blue) presented diotically over headphones while ignoring a second competing speech stream (indicated in black) presented to the left or the right of the to-be-attended stream. The spatial separation between competing streams generated a naturalistic listening situation in which the central sound source is attended and eased stream segregation in the adverse listening environment of the MRI scanner. The duration of the two-speaker blocks was set to 60 s to ensure a robust reconstruction of the speech streams within the MRI environment. This block duration corresponds to previous experiments using similar EEG-based stimulus reconstruction approaches (Mirkovic et al., 2015; O'Sullivan et al., 2015). Yes/no questions on the content of the to-be-attended story served as control that subjects followed the to-be-attended stream. Questions were displayed on the screen after preselected two-speaker blocks and were related to the contents of the last 20 s of that block only to minimize memory effects. Questions had to be answered within 20 s. After having entered a button response, the question disappeared from the screen and the fixation dot was presented for the remaining time. No feedback on the correctness of the given answer was provided. Each question block was followed by a single-speaker block of 20 s duration, in which only the to-be-attended stream was present. These blocks allowed participants to refocus attention on the to-be-attended stream. In both single- and two-speaker blocks, a fixation dot was presented centrally on the screen to stabilize eye gaze.

In total, the experiment consisted of 24 two-speaker blocks, 18 single-speaker blocks, and 18 question blocks. In half of the two-speaker blocks, the to-be-ignored stream was presented to the left and, in the other half, to the right of the attended speech stream. A pseudorandomized block order was generated and kept identical for all subjects to allow for comparability of performance and EEG measures. The total task duration was 36 min. After finishing the task, participants were asked to recall freely the content of the to-be-attended story. Performance was scored based on the number of correctly recalled main events of the storyline.

Excerpts taken from audio books in German language served as stimulus material for the selective listening task. In the EEG-fMRI session, "The Cold Heart" and "The Story of Little Muck," both fairy tales written by Wilhelm Hauff, served as to-be-attended and to-be-ignored speech streams. For the laboratory EEG session, "The Five Orange Pips" by Sir Arthur Conon Doyle and Mark Twain's "The Million Pound Note" were used as the to-be-attended and to-be-ignored streams. Each pair of stories was narrated by two nonidentical male speakers. The to-be-attended stories were kept constant for all participants to ensure a good comparability of individual results.

To reduce the possibility that the listener's attention was drawn to the to-be-ignored stream during prolonged silent gaps of the to-be-attended story, silent periods exceeding 400 ms in duration were trimmed to 400 ms in all streams. Sounds were cut into consecutive pieces of 20 or 60 s duration with 5 ms $\sin^2$ onset and offset ramps. The to-be-attended stream was presented diotically over headphones, thus appearing to originate from a central sound source. The to-be-ignored stream was shifted in azimuth relatively of the to-be-attended stream by introducing interaural time differences of -1 or $+1$ ms between both headphone channels. This resulted in the to-be-ignored stream being perceived to originate from the left or right of the listener.

The mean presentation level of the competing stories was equalized. In the laboratory session, the combined sound level in the two-speaker phases was set to 75 dB(A). In the EEG-fMRI session, subjects could choose their own sound comfort level to compensate for individual differences in scanner background noise attenuation by ear protection within the MRI environment.

Dichotic deviance detection task

Complementing the EEG-based measures of speech envelope tracking during selective listening, individual spatial selective listening performance was assessed in an additional behavioral task. Here, a dichotic deviance detection task was applied, in which participants were asked to attend selectively to a stream of pure tones presented to the left or right ear and to respond to rare deviant stimuli via a button press. A second to-be-ignored stream of pure tones was presented to the other ear. Deviants occurred in both streams with equal probability. Task blocks containing no contralateral auditory input served as baseline to control for general deviant detection performance. In total, the task consisted of four two-stream blocks and four baseline blocks; the to-be-attended ear was balanced across conditions. The block order was kept constant across participants.

Pure tones with a duration of 50 ms, including 5 ms $\sin^2$ onset and offset ramps and a frequency of 1000 or 1100 Hz, served as standard and deviant tones, respectively, for both streams. Each stream consisted of 275 standard and 30 deviant tones per task block. Stimulus order was randomized, but there had to be at least three standard tones between succeeding deviants. In both the baseline and two-stream blocks, the interstimulus interval within a stream varied randomly between 100 and 300 ms with the restriction that the stimulus presentation rate was kept at $<$5 Hz. In the two-stream blocks, stimuli were presented alternately in the left and right ear streams. This resulted in an effective interstimulus interval of 25–125 ms across streams.

Similar dichotic deviance detection designs have been used previously to study the neural and behavioral differences in auditory selective attention (Gomes et al., 2012). Based on previous work, we expected that individual differences in task performance are primarily related to the auditory system's ability to focus attention in space and to suppress responses to deviants occurring in the contralateral ear (Sussman et al., 2003; Ahveninen et al., 2011). If individual differences in speech envelope tracking are directly related to the degree of attentional modulations of auditory sensory processing, then speech tracking accuracies should correlate with deviance detection performance.

Data acquisition

During the laboratory session, EEG data were acquired from 96 electrodes using a BrainAmp amplifier system (Brainproducts) and a customized, infracerebral electrode cap with an equidistant electrode layout (Easycap). Data were recorded with a sampling rate of 500 Hz using the nose tip as reference and analog filtered between 0.016 and 250 Hz. Electrode impedances were maintained $<$20 kΩ before data acquisition. Behavioral responses were recorded using a custom-made response pad. Sounds were presented to the participants via insert earphones (E-A-RTONE 3A; E-A-R Auditory Systems).

fMRI data were acquired on a 3 T Siemens MAGNETOM Verio MRI scanner with a 12-channel head array. Key presses were recorded using an MR-compatible response keypad (Nata Technologies). Acoustic stimuli were delivered via MR-compatible insert earphones (MR confon HP AT01; MR Confon). Participants were equipped with over-ear hearing protectors during the experiment to attenuate scanner background noise. During the selective listening task, 1100 T2*-weighted gradient echo planar imaging volumes with BOLD contrast were obtained (time of repetition [TR] = 2000 ms, time of echo [TE] = 30 ms, flip angle α = 80°, field of view [FoV] = 200 × 200 mm^2, voxel-size = 3.1 × 3.1 × 3.0 mm^3). Volumes consisted of 31 transverse slices with a gap of 0.9 mm in between and were recorded in an ascending order. The same imaging parameters were used to obtain a resting-state time series consisting of 261 T2* volumes. A high-resolution structural volume was acquired for each participant using a T1-weighted magnetization-prepared rapid acquisition gradient echo sequence (TR = 1900 ms, TE = 2.52 ms, α = 9°, FoV = 256 × 256 mm^2, voxel-size = 1 × 1 × 1 mm^3).

During fMRI data acquisition, EEG was recorded from 64 electrodes using an MRI-compatible amplifier system (BrainAmp MR Plus; Brainproducts). The scalp electrodes were placed in a customized, equidistant layout with electrodes AFz serving as ground and Cz as online recording reference. Eye movements were monitored by an electrode placed below the left eye and the electrocardiogram was recorded by an electrode placed on the left lower back. The hardware clock of the EEG system and the MRI scanner's master clock were synchronized. The data were recorded with a sampling rate of 5000 Hz and analog filtered between 0.016 and 250 Hz. Electrode impedances were maintained at $<$20 kΩ before data acquisition.

Speech envelope reconstruction

EEGLAB (Delorme and Makeig, 2004; RRID: SCR_007292) was used for the preprocessing of the EEG data. MRI artifact correction was performed using the FMRIB plug-in for EEGLAB (for a description of the methods, see Niazy et al., 2005). Scanner gradient artifacts were removed

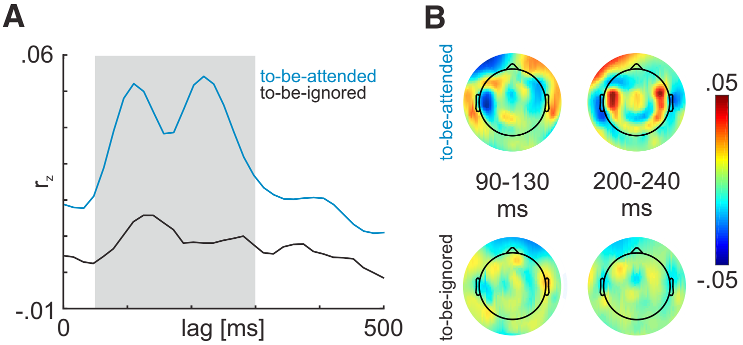

Figure 2. Speech envelope reconstruction. ***A***, Systematic investigation of the speech envelope reconstruction accuracy r_z for individual time lags ranging from 0–500 ms after sound presentation. Shown are the mean group accuracies for the to-be-attended and the to-be-ignored speech streams. The final data analysis was based on the time window from 50–300 ms, which contained the highest reconstruction accuracies for the to-be-attended stream (gray). ***B***, Topographic maps of the mean regression weights for each EEG sensor within the two time intervals showing the highest reconstruction accuracies for the to-be-attended stream.

using a slice template obtained from averaging over 30 consecutive fMRI volumes (*fmrib_fastr* function). Cardio-ballistic artifacts were detected and cleaned from the EEG using the functions *fmrib_qrsdetect* and *fmrib_pas*. After applying these correction steps, EEG data were downsampled to 500 Hz for further processing.

The further preprocessing and the analysis of the EEG datasets obtained in the laboratory and the MRI environment were kept identical to ensure a good comparability of results. Datasets were re-referenced to a common average reference. Artifacts related to eye blinks and lateral eye movements were pruned from the datasets using independent components analysis. For this procedure, a copy of the rereferenced EEG data was offline filtered from 1–40 Hz and epoched into continuous 2 s intervals. Principal component analysis was conducted to reduce the dimensionality of the EEG data to 45 dimensions and 45 independent components (ICs) were computed using the extended infomax algorithm implemented in EEGLAB. The demixing matrix obtained from this procedure was applied to the original unfiltered EEG dataset and ICs reflecting eye blinks, lateral eye movements, and cardiovascular activity were removed. No further data cleaning was performed.

For the speech envelope reconstruction, the EEG data were offline filtered from 1–8 Hz, epoched from 0–60 s relative to the onset of each two-speaker block, and downsampled to 64 Hz to reduce computational demands. EEG time courses were transferred into standardized z-scores to equalize means and SDs across channels and trials. EOG and nose electrode channels were removed from the dataset.

Amplitude envelopes of the speech signals were obtained using a Hilbert transform, followed by low-pass filtering with an 8 Hz cutoff frequency (O'Sullivan et al., 2015). Filtering was performed using a third-order Butterworth filter and the *filtfilt* function in MATLAB for a zero-phase digital filtering of the data. Audio streams were subsequently cut into the corresponding 60 s intervals, downsampled to 64 Hz, and z-normalized.

Decoders for reconstructing the speech envelope of the to-be-attended and to-be-ignored speech streams from the concurrently measured EEG data were implemented using in-house MATLAB routines applied previously by Mirkovic and colleagues (2015, 2016). The underlying methods have been described in detail by O'Sullivan et al. (2015) and Crosse et al. (2016b). In short, a ridge regression was used for a backward mapping from the recorded EEG data to the sound envelope fluctuation, allowing a reconstruction of the speech envelope time course from the concurrently measured cortical EEG responses. The ridge regression was based on all scalp electrodes and multiple consecutive time lags, thus taking into account the delayed cortical response to the audio input. Decoders were computed for each trial and for both speech streams, resulting in 24 sets of regression weights for each participant and each stream. The speech envelope reconstruction was performed using a leave-one-out cross-validation procedure on the subject-level (Mirkovic et al., 2015). This means that the reconstruction of the speech envelope of the to-be-attended or to-be-ignored stream in a selected trial was based on the mean regression weights obtained for this stream in all other but this trial. The leave-one-out cross-validation procedure ensured that the reconstruction does not depend on trial-specific properties of the recorded EEG data, but rather was related to a general and trial-independent mapping between sound envelope and EEG response. To quantify the reconstruction accuracy, Pearson's correlation was calculated between the reconstructed and the original speech envelope in each trial. For further analysis, the obtained correlation coefficients were converted to normally distributed r_z values using Fisher's z transformation.

The time lags entering the ridge regression were restricted to the interval showing the most robust reconstruction of the to-be-attended stream. Reconstruction accuracies r_z were calculated as stated above for each individual time lag from 0–500 ms. A grid search and a leave-one-out cross-validation were used to identify the best tuning parameter λ for each lag and each subject with the goal of obtaining the highest reconstruction accuracy r_z for the to-be-attended stream (grid values: $\lambda = 10^{-3}, 10^{-2}, \ldots 10^{5}$). The most robust relationship between EEG time courses and the speech envelope of the to-be-attended stream emerged within the time window from 50–300 ms (Fig. 2*A*), so this interval was selected to estimate the final decoder. The tuning parameter $\lambda = 100$ was found to result in the best overall reconstruction accuracy within the selected time window for both the to-be-attended and the to-be-ignored stream. Figure 2*B* depicts the resulting set of regression weights for the to-be-attended and the to-be-ignored stream averaged across the peak time windows from 90–130 ms and 200–240 ms.

To control for the generalizability of the obtained regression weights across subjects in the EEG-fMRI session, r_z values were recomputed using the same settings, but with a leave-one-out cross-validation procedure across subjects. The reconstruction for a given subject was performed using the mean reconstruction weights across all trials from all but this subject.

fMRI data analysis

MRI data were processed and analyzed using SPM12 (FIL, Welcome Trust Centre for Neuroimaging, UCL, London; RRID: SCR_007037). The first 10 volumes of the functional time series were discarded to account for T1 equilibration effects. To correct for head motion, the functional time series were realigned spatially to the first image of the time series. The structural T1-weighted volume was registered to a mean functional image and segmented to obtain spatial normalization parameters. Using these parameters, functional and structural images were normalized to the Montreal Neurological Institute (MNI) template brain. Finally, normalized functional volumes were smoothed with a 3D Gaussian kernel of 8 mm full-width-half-maximum.

A general linear model and a random-effects analysis were used for the statistical analysis of the fMRI data. The single-subject level contained five task regressors, accounting for onset responses and sustained activity during single- and two-speaker blocks, respectively, as well as question blocks. Onset responses were modeled as stick functions and sustained activity was modeled as blocks of 20 or 60 s duration. Parametric regressors modeling the spatial position of the to-be-ignored speaker were added to the two-speaker onset and block regressors to explain BOLD signal variance related to the changing spatial position of the distracting speech input. Signal changes related to head movement were accounted for by including the six movement parameters computed in the SPM12 realign procedure. The time series in each voxel were high-pass filtered to 1/256 Hz and modeled for temporal autocorrelation across scans with an AR(1) process. The high-pass filter cutoff value was chosen with respect to the long block durations and the rather long intervals between subsequent blocks of the same trial type [single-speaker: 126 ± 38 s (M ± SD); two-speaker: 89 ± 10 s; question: 120 ± 37 s].

Given that the envelope reconstruction accuracy for the to-be-attended stream showed considerable fluctuations over the time course of the experiment, we investigated whether this effect is coupled to trialwise fluctuations in BOLD signal strength. For this, an additional single-subject model was computed in which the individual single-trial reconstruction accuracies r_z for the to-be-attended stream were added as an additional parametric regressor to model systematic BOLD amplitude changes during two-speaker blocks. The spatial location of the to-be-ignored speaker was not accounted for in this model. However, it was confirmed that adding additional modulators for the spatial location did not alter the relationship between r_z and single-trial BOLD amplitudes qualitatively.

Functional connectivity analysis

Because the fMRI analysis indicated a significant relationship between the mean envelope reconstruction accuracy r_z for the to-be-attended stream and the BOLD amplitude in the right temporal parietal junction during selective listening, the task-related functional connectivity pattern of this brain region was investigated. To isolate real task-related effects from permanent intrinsic connections of this region, the individual resting-state connectivity pattern of the right temporal parietal junction was subtracted from the pattern computed using the task data.

Data preprocessing for the task and resting-state time series were kept similar to the main data analysis, but involved an additional slice-timing correction to the middle slice of the volume. The CONN toolbox for SPM (Whitfield-Gabrieli and Nieto-Castanon, 2012; RRID: SCR_009550) was used to compute task and resting-state functional connectivity between a right temporal parietal junction seed (i.e., an 8 mm sphere around the group activation peak voxel) and whole-brain voxel time courses for each subject. The resting-state data were pruned from physiological and motion artifacts by regressing out signal time courses obtained from white matter and CSF as well as the six realignment parameters accounting for head motion. For the task dataset, BOLD responses related to the experimental stimulation were regressed out using the five task regressors of the main fMRI analysis. This aimed to reduce correlations of voxel-time courses caused by the external sensory stimulation. A linear detrending of the data was performed and time series were band-pass filtered between 0.008 and 0.09 Hz. A hemodynamic response function weighting was used for computing task-related functional connectivity within selective listening blocks.

The seed-to-voxel functional connectivity analysis resulted in single-subject connectivity maps containing information about the temporal correlation between fMRI time courses in the seed region and each voxel in the brain during resting state and during the two-speaker phases of the selective listening task. For further analysis, the resting-state connectivity was subtracted from the task-based functional connectivity map.

Structural MRI analysis

It has been speculated that individual differences in the overall stimulus reconstruction accuracy may be related to anatomical differences between subjects (O'Sullivan et al., 2015). Here, the relationship between mean r_z values and brain structure was controlled for on a global level in terms of the total intracranial volume and on a local level using a voxel-based morphometry analysis (Mechelli et al., 2005). Gray matter, white matter, and CSF tissue maps were computed from the individual anatomical T1 scan using the segmentation tools provided in SPM12. The total intracranial volume was obtained via a summation of all voxels belonging to one of these tissue types. The voxel-based morphometry was conducted using the DARTEL tools included in SPM12 to improve the intersubject alignment of the gray matter tissue maps (Ashburner, 2007). The gray matter images were normalized to the MNI space and smoothed with a 3D Gaussian kernel of 8 mm full-width-half-maximum. The normalized gray matter intensity was scaled by the determinant of the Jacobian transformation in each voxel, so that voxel intensities provide a quantitative measure local gray matter volume (Peelle et al., 2012).

Experimental design and statistical analysis

Four of the 23 participants showed severe head movements during the fMRI data acquisition (criteria: a total displacement >3 mm and/or multiple scan-to-scan movements >1 mm) and/or an inferior speech envelope reconstruction quality (criterium: the reconstructed sound envelope of the to-be-attended stream correlates with the original sound envelope in $<2/3$ of all trials) and were therefore removed from the dataset. One of the remaining subjects did not finish the behavioral dichotic deviance detection task, which served as a measure of individual spatial selective listening abilities. Two subjects showed very low accuracies in this task (accuracy in the two-stream condition was >3 SDs below the mean of all other subjects) and were not considered for the analysis. Therefore, the statistical analysis of all EEG, fMRI, and behavioral measures but the deviance task was based on 19 datasets (13 female, 24 ± 2 years). Analyses involving the latter task relied on 16 datasets only (10 female, 24 ± 3 years).

Speech envelope reconstruction. Our study is the first to apply a speech envelope reconstruction approach on EEG data acquired within an MRI scanner. Therefore, as an initial step, the overall quality of the speech envelope reconstruction accuracy within such an adverse environment was determined. Mean reconstruction accuracies r_z of the EEG-fMRI session for the to-be-attended and the to-be-ignored stream were compared with values obtained in the laboratory session of the experiment using paired *t* tests and Pearson's correlation. To further test for the robustness and generalizability of the regression weights computed in the fMRI dataset, results of the within-subject decoding procedure were compared with the results obtained using the across-subject cross-validation using paired *t* tests and Pearson's correlation.

Although speech envelope reconstruction accuracies are generally considered to be fairly stable across a wide range of signal-to-noise ratios (Ding and Simon, 2013), there exist also experimental data showing decreased reconstruction accuracies with increasing levels of background noise (Presacco et al., 2016). In addition, hearing loss has been reported to affect speech envelope tracking (Petersen et al., 2017). To control for such effects in our data, Pearson's correlation between mean r_z(att) values and the subjects' mean hearing loss (i.e., hearing threshold averaged over both ears and all tested frequencies from 125–8000 Hz) and the individual sound presentation level in the EEG-fMRI session were computed. Pearson's correlation was performed between r_z(att) and the intracranial volume of each participant to test for a global anatomical relationship between envelope tracking accuracy and brain size. A group-level regression analysis in SPM12 then aimed to assess potential positive or negative correlations between mean individual r_z(att) values and the local gray matter volume. The total intracranial volume served as covariate of no interest for this analysis to control for the effect of brain size on local gray matter volume.

fMRI analysis. The statistical analysis of the task fMRI data concentrated on the selective listening phase only. The contrasts two-speaker onset > baseline and two-speaker block > baseline were studied on the group level using simple *t* test models to reveal brain activations related to block onsets and sustained activity during selective listening. Differences in BOLD response amplitudes related to the spatial location (left vs right) of the to-be-ignored stream during onset and sustained selective listening were further investigated using simple *t* tests on the parametric regressors added to the two-speaker onset and two-speaker block regressors.

To test for a linear relationship between BOLD activation levels and the individual mean envelope reconstruction performance at the onset of selective listening blocks and during sustained selective listening, the contrasts two-speaker onset > baseline and two-speaker block > baseline were entered into group-level regression models with the mean r_z(att) values serving as a covariate. To control for possible relationships between mean r_z(att) values and individual differences in the degree of spatial modulations in BOLD response amplitudes, similar analyses were conducted for the parametric regressors modeling the spatial location of the to-be-ignored stream during the onset of the two-speaker blocks as well as during sustained selective listening.

Complementing the analyses on the relationship between mean BOLD response amplitudes and mean r_z(att) values, the relationship between trial-to-trial fluctuations in r_z(att) and single-trial BOLD response amplitudes during sustained selective listening blocks was analyzed. For this, the parametric regressor modeling this relationship was tested on the group level with a simple *t* test model.

The general pattern of task-related functional connectivity of the right temporoparietal junction during selective listening (i.e., task − resting-state connectivity) was studied using a simple *t* test model in SPM. To control for a linear relationship between functional connectivity and the average envelope reconstruction accuracy for the to-be-attended stream, a group-level regression analysis was performed with the mean r_z(att) value serving as a covariate.

Behavioral data analysis. For the selective listening task performed during EEG-fMRI recordings, the percentage of correctly answered questions on the contents of the to-be-attended story as well as the number of correctly recalled events in the delayed recall task were calculated. Pearson's correlation between both scores was computed to test whether both measures are related. Individual ratings of distraction by the to-be-ignored speech stream and scanner background noise were compared using the Wilcoxon signed-rank test.

For the deviance detection task, the number of correct and erroneous button presses was computed for each participant and both the two-stream and the baseline condition. Button presses were considered correct when occurring within 150–1000 ms after a to-be-attended deviant. Individual deviant detection accuracy was quantified for both task conditions using the F_1 score, which is the harmonic mean of sensitivity and precision (Van Rijsbergen, 1979).

The accuracy in both conditions was compared using a paired *t* test. To determine whether the accuracy in the deviance detection task was related to performance during selective listening to speech in the EEG-fMRI experiment, Pearson's correlation was computed between F1 scores obtained in the two-stream condition and the number of correctly answered questions on the story content during the main task.

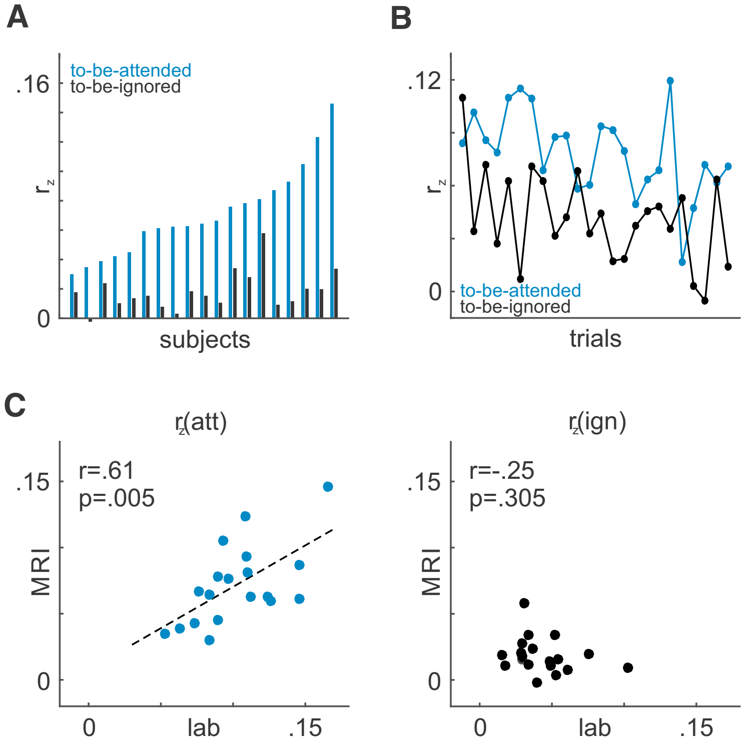

Figure 3. Speech envelope reconstruction accuracies obtained during the simultaneous EEG-fMRI measurements: The graphs on the top depict the mean speech envelope reconstruction accuracies r_z of each participant for the to-be-attended (blue) and the to-be-ignored (black) speech stream (***A***), as well an example of single-trial reconstruction accuracies in an individual subject across the experiment (***B***). The lower figures shows mean r_z values obtained in the EEG-MRI experiment compared with EEG laboratory data obtained in the same subjects (***C***). Reconstruction accuracies were generally reduced during EEG-fMRI measurements. For the to-be-attended stream, reconstruction accuracies r_z(att) were positively correlated between both environments. No such relationship was found for the to-be-ignored stream.

To assess relationships between behavioral selective listening performance and individual envelope reconstruction accuracies, Pearson's correlation between mean r_z(att) values and both the percentage of correctly answered yes-no-questions during the selective listening task and F_1 scores obtained in the behavioral deviance detection task was performed. Consistent with previous results obtained by O'Sullivan et al. (2015), a positive relationship between selective listening performance and speech envelope reconstruction accuracy (i.e., better tracking of the to-be-attended speech stream is related to superior behavioral performance) was expected, so only positive correlations were tested for.

For the fMRI and voxel-based morphometry analyses, a statistical threshold of $p < 0.05$, corrected for familywise errors (FWEs) on the cluster level, was applied. The single-voxel threshold for cluster identification was set to $p < 0.001$ (uncorrected).

Planned statistical comparisons of the behavioral data and speech envelope reconstruction accuracies were reported as statistically significant when passing a threshold of $p < 0.05$, corrected for multiple comparisons using an FDR correction (Benjamini–Hochberg method, 6 tests in total). Reported *p*-values were adjusted accordingly. Tests controlling for the technical quality of the speech envelope reconstruction or the influence of external factors on r_z scores were reported on an uncorrected level.

Results

Speech envelope reconstruction

Applying a stimulus reconstruction approach on the EEG data obtained during the fMRI experiment, we obtained a mean speech envelope reconstruction accuracy of $r_z(\text{att}) = 0.07 \pm 0.03$ (M ± SD) for the to-be-attended speech stream and $r_z(\text{ign}) = 0.02 \pm 0.01$ for the to-be-ignored stream. As expected based on other selective listening studies, the mean reconstruction accuracy for the to-be-attended stream was significantly higher than for the to-be-ignored stream (paired *t* test: $t_{(18)} = 8.3$; $p < 0.001$). Figure 3, *A* and *B*, shows individual differences in mean speech envelope reconstruction accuracies across subjects as well as trial-to-trial fluctuations in r_z during the fMRI experiment for one representative listener.

Because this experiment is the first to apply a stimulus reconstruction method on EEG data acquired in the adverse environment of an MRI scanner, we compared these results with envelope reconstruction accuracies obtained in a controlled laboratory setting to control for the general quality of the reconstruction. For both the to-be-attended and the to-be-ignored stream, mean reconstruction accuracies r_z(att) were significantly increased in the laboratory environment ($r_z(\text{att}) = 0.10 \pm 0.03$; $r_z(\text{ign}) = 0.04 \pm 0.02$) as indicated by paired *t* tests between both settings (att: $t_{(18)} = 5.2, p < 0.001$; ign: $t_{(18)}=3.9; p < 0.001$). This suggests a negative impact of the MRI environment on the speech envelope reconstruction. Importantly however, the mean reconstruction accuracy for the to-be-attended stream obtained for each subject was significantly correlated with the mean envelope reconstruction accuracy of the to-be-attended speech stream in the lab (Pearson's $r = 0.61, p = 0.005$; Fig. 3*C*, left column). No such relationship was found for the to-be-ignored stream (Pearson's $r = -0.25$; $p = 0.305$; Fig. 3*C*, right column).

To gain the best individual speech envelope reconstruction accuracies, single-subject results were obtained using a leave-one-out cross-validation procedure on the single-subject level. We tested for the generalizability of the obtained ridge regression weights across subjects by comparing the r_z values with reconstruction accuracies resulting from a cross-validation across subjects. In the latter case, the speech envelope reconstruction was performed using average regression weights from other subjects only. Reconstruction accuracies for both speech streams were generally lower when performing the cross-validation over subjects (att: $r_{z,\text{group}} = 0.04 \pm 0.02$, $t_{(18)} = 6.14$, $p < 0.001$; ign: $r_{z,\text{group}} = 0.01 \pm 0.01$, $t_{(18)} = 3.22$, $p = 0.005$). For the to-be-attended stream, r_z values were positively correlated between both cross-validation approaches (Pearson's $r = 0.75$; $p < 0.001$), whereas no such relationship was found for the to-be-ignored stream (Pearson's $r = 0.36$; $p = 0.127$).

Although individual results obtained for the to-be-attended stream seemed to be robust across sessions and could be replicated using the cross-validation procedure across subjects, no such relationships were found for the to-be-ignored stream. This suggests that the EEG trace of the to-be-ignored stream could not be picked up reliably in the MRI environment, possibly due to increased noise levels in the EEG recordings. We therefore based all further data analyses on r_z values obtained for the to-be-attended stream only.

To control for potential effects of hearing loss or individual stimulus presentation level on the obtained speech envelope reconstruction accuracies in the EEG-fMRI session, we computed Pearson's correlation between individual r_z(att) scores and both the participants average hearing thresholds (Pearson's $r = 0.09$, $p = 0.705$) and the individual stimulus presentation level (Pearson's $r = -0.24$, $p = 0.324$). None of these tests revealed a significant relationship, suggesting that individual differences in r_z(att) are not linked to these factors. In addition, it was speculated previously that differences in speech envelope reconstruction accuracy may be related to individual anatomical differences. Here, we investigated possible relationships on a global and local anatomical level. Individual mean r_z(att) values were not related to the total intracranial volume (i.e., the brain size; Pearson's $r = -0.17$, $p = 0.494$). Similarly, a voxel-based morphometry analysis, which tested for a systematic relationship between r_z(att) and the local gray matter volume, revealed no significant positive or negative effects (at $p < 0.05$, FWE corrected on the cluster level).

BOLD responses related to selective listening

We analyzed BOLD responses related to the onset of two-speaker blocks and sustained activity during selective listening phases. Brain areas activated during block onset phases encompassed the primary and secondary auditory cortices and adjacent parts of the frontoparietal operculum, the right middle frontal gyrus, the right temporoparietal junction and adjacent parts of the inferior parietal lobe, the precuneus, the mid- and posterior cingulate cortex, and mainly right but also left posterior parietal cortex. The spatial location of to-be-ignored stream modulated onset-related BOLD amplitudes in auditory sensory areas. BOLD responses were higher in bitaleral Heschl's gyrus and in the right anterior planum temporale when the to-be-ignored stream was presented to the left. The inverse contrast yielded no significant results.

During selective listening, sustained BOLD responses were found in the primary and secondary auditory cortices and adjacent parts of the frontoparietal operculum, the left and right precentral and postcentral gyri, the medial portion of the supplementary motor area, parts of visual cortex, as well as in the lobules V and VI of the cerebellum (all activations at $p < 0.05$, FWE corrected for multiple comparisons on the cluster level). Several brain regions showed modulations of BOLD response amplitudes related to the spatial location of the to-be-ignored stream. Presenting the to-be-ignored speaker to the right was associated with increased BOLD responses in the left planum temporale, the anterior and midcingulate cortex, the medial portion of the supplementary motor area, the insula, the left middle frontal gyrus, the right precentral gyrus, the right intraparietal sulcus, parts of primary and secondary visual cortex, as well as in lobules V, VI, and VII of the cerebellum. Presenting the to-be-ignored speaker to the left led to increased response amplitudes in the right anterior planum temporale. Figure 4 depicts the activation clusters observed during block onset and selective listening phases.

Relating BOLD activity to speech envelope reconstruction accuracies

Our data show considerable differences in the tracking of the to-be-attended speech stream during the selective listening task, both on the interindividual level across subjects and on the intraindividual level across trials. The main goal of the fMRI data analysis was to relate these differences to differences in BOLD activation levels during selective listening phases.

A group-level regression analysis was used to identify significant relationships between the mean speech envelope reconstruction accuracy r_z(att) and individual mean BOLD amplitudes during onset- and sustained listening phases of the two-speaker blocks. During sustained selective listening phases, we found a significant positive effect in the right temporoparietal junction [$p < 0.05$, FWE corrected on the cluster-level; activation peak coordinates $(x, y, z) = (50, -48, 26)$]. BOLD amplitudes in this region increased with an increasing mean envelope reconstruction accuracy. Figure 5*A* depicts the location of the cluster as well as β estimates extracted from the effect peak as a function of the individual mean r_z(att). No brain region showed a negative relationship between speech envelope tracking accuracy and BOLD amplitude during selective listening blocks. In contrast to the sustained selective listening phases, we observed no positive or negative relationship between r_z(att) and onset-related BOLD response amplitudes. In addition, differences in BOLD response amplitudes related to the spatial position of the to-be-ignored stream were not linked to r_z(att) either during block onset or during sustained selective listening.

Complementing the group-level regression analysis, we also investigated the relationship between single-trial BOLD amplitudes and trial-by-trial fluctuations of speech envelope reconstruction accuracy r_z(att). This analysis did not, however, yield any statistically significant positive or negative results. Given the small number of trials and the potentially interfering effect of the MRI environment on single-trial speech envelope reconstruction accuracies, we decided nevertheless to further inspect qualitative results of this analysis on a lower uncorrected significance level. As shown in Figure 6, successively lowering the statistical threshold indicates a positive relationship between the single-trial speech envelope reconstruction accuracy and BOLD responses in large parts of the language network, including the superior temporal sulcus, Broca's area, and adjacent parts of the middle and inferior frontal gyri. Low envelope reconstructions accuracies were associated with a more scattered pattern, mainly affecting the white matter, ventricles, visual cortex, and cerebellum. The latter observation suggests that low single-trial envelope reconstruction accuracies may to some extent result from subject motion during data acquisition.

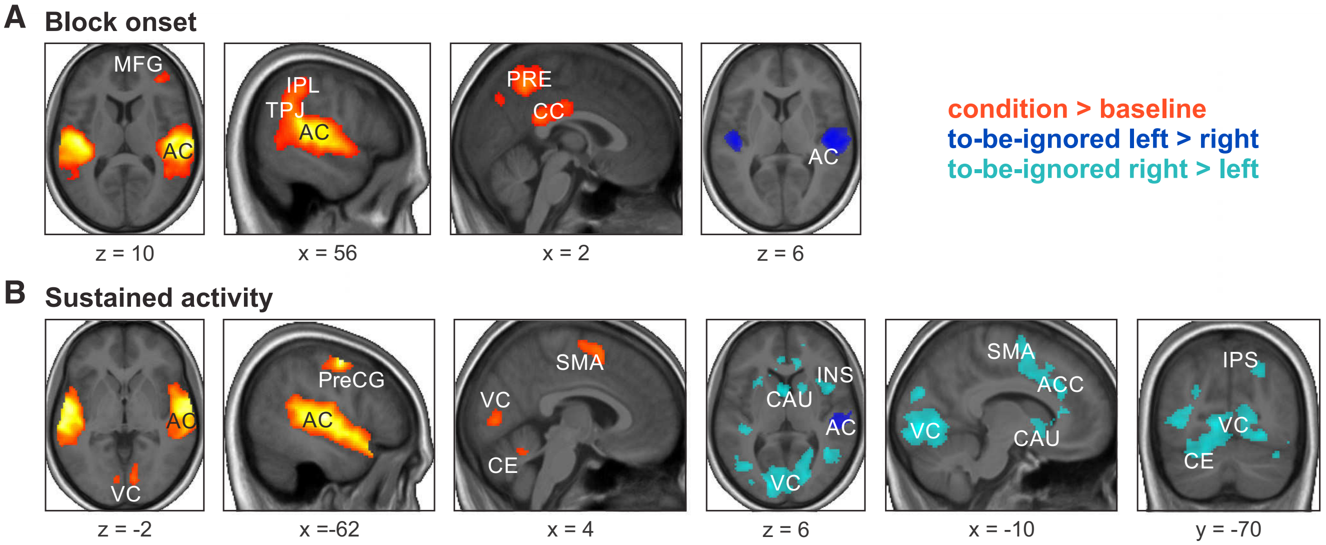

Figure 4. Onset–related and sustained BOLD activation patterns during selective listening. Shown are activation patterns for the contrasts two-speaker onset > baseline and two-speaker block > baseline as well as parametric modulations of BOLD response amplitudes related to the spatial position of the to-be-ignored stream in both contrasts. Activations are shown at a statistical threshold of $p < 0.05$, FWE corrected for multiple comparisons on the cluster-level. ***A***, Brain regions activated at block onset include large parts of the auditory cortex (AC), the right temporoparietal junction (TPJ), the right inferior parietal lobe (IPL), the middle frontal gyri (MFG), the precuneus (PRE), and the mid- and posterior cingulate cortex (CC). ***B***, Sustained activity during selective listening was found in the auditory cortex, the supplementary motor area (SMA), the precentral gyrus (PreCG), in parts of the visual cortex (VC), and in the cerebellum (CE). Spatial modulations of BOLD response amplitude were observed in auditory cortex, the insula (INS), the caudate nucleus (CAU), the anterior cingulate cortex (ACC), the supplementary motor area, the right intraparietal sulcus (IPS), the visual cortex, and the cerebellum.

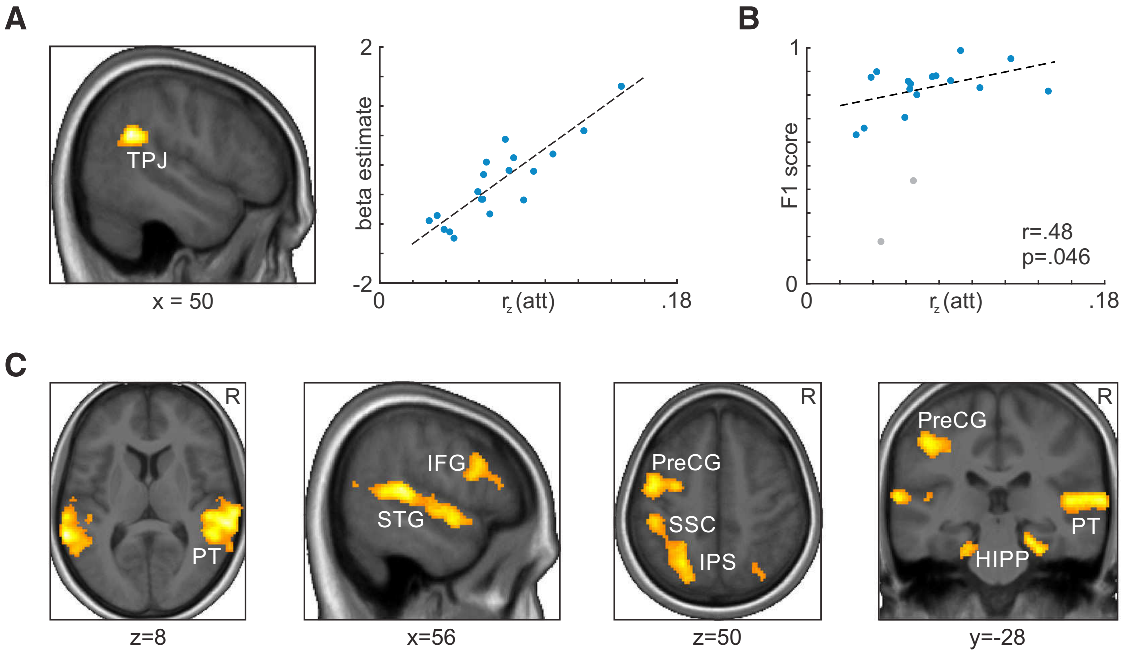

Figure 5. ***A***, BOLD amplitudes in the right temporoparietal junction (TPJ) were positively correlated with mean speech envelope reconstruction accuracies obtained for the to-be-attended stream during selective listening. The relationship between β estimates for the contrast two-speaker block > baseline extracted from the effect peak and mean r_z(att) scores is depicted on the right. ***B***, Mean r_z scores obtained for the to-be-attended speaker were positively correlated to selective listening performance obtained in a secondary behavioral task. Two outlier data points, which were excluded from the statistical analysis, are shown in gray. ***C***, During selective listening, the right TPJ showed significant task-related functional connectivity to the planum temporale (PT), the superior temporal gyrus (STG), the intraparietal sulcus (IPS), the left somatosensory cortex (SSC), the precentral gyrus (PreCG), right inferior frontal gyrus (IFG), and the hippocampus (HIPP). For both fMRI analyses, effects are depicted at a statistical threshold of $p < 0.05$, FWE corrected on the cluster level.

Functional connectivity of the right temporoparietal junction during selective listening

Sustained BOLD activity in the right temporoparietal junction was found to be related to individual differences in the cortical tracking of the to-be-attended speech stream, suggesting that this region interacts with auditory sensory areas during selective listening. To investigate this question, we subsequently studied the functional connectivity of the right temporoparietal junction during two-speaker blocks. To assess specifically task-related functional connectivity, we subtracted the individual resting-state connectivity of the right temporoparietal junction from the pattern obtained during task performance. The resulting task-

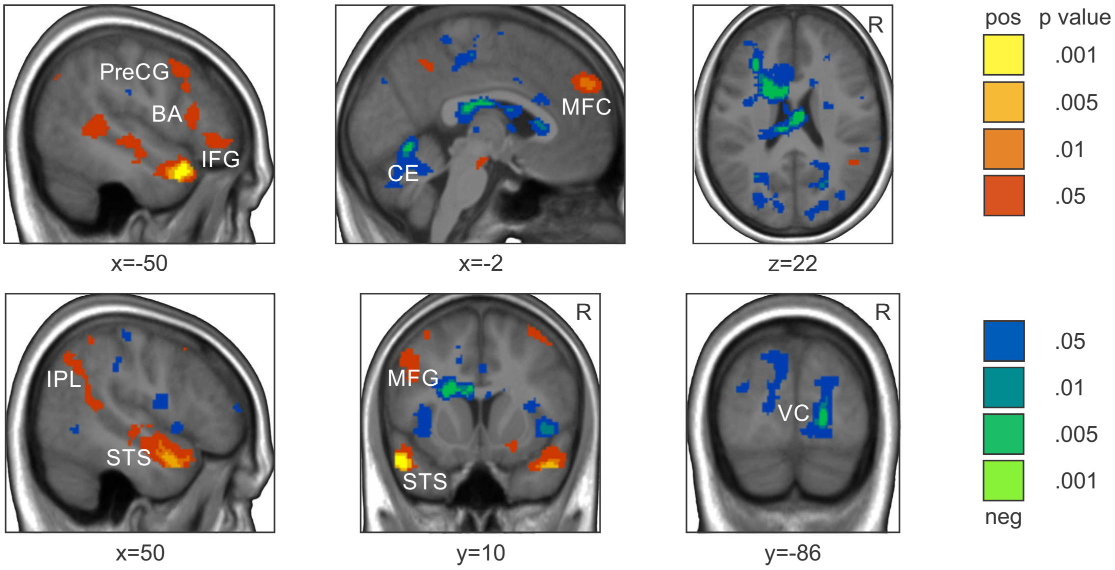

Figure 6. Trial-to-trial relationship between BOLD response amplitude and envelope reconstruction accuracy. No brain region showed a significant positive or negative relationship between single-trial BOLD amplitudes and the speech envelope reconstruction accuracy r_z(att). Lowering the statistical threshold however revealed a network of speech-related brain areas showing a positive relationship between both measures. Regions encompass the superior temporal sulcus (STS), Broca's area (BA), the middle (MFG) and inferior frontal gyri (IFG), the precentral gyrus (PreCG), the inferior parietal lobe (IPL), and the medial frontal cortex (MFC). Negative relationships between BOLD amplitude and r_z(att) tended to occur in the white matter, ventricles, visual cortex (VC), and cerebellum (CE). The latter finding suggests that low single-trial envelope reconstruction accuracies may to some extent result from subject motion during data acquisition. The figure depicts positive and negative relationships between single-trial fMRI response amplitudes and speech envelope reconstruction accuracy for the to-be-attended stream at different statistical thresholds. All voxel-level thresholds are uncorrected; minimum cluster size was set to 50 voxels.

related connectivity network is depicted in Figure 5*C*. During selective listening, the right temporoparietal junction shows significant functional connectivity to large parts of the anterior and posterior secondary auditory cortex, including the planum temporale, the planum polare, and parts of the superior temporal gyrus. Further connections encompass the right but mostly left superior parietal lobe and the intraparietal sulcus, parts of left and right posterior inferior frontal gyrus, the left somatosensory cortex and left supplementary motor area, the left and right hippocampus and parahippocampus, as well as lobules V, VI, and VII of the cerebellum.

We tested whether the connectivity strength between the right temporoparietal junction and other regions during selective listening is modulated by the speech envelope reconstruction accuracy r_z(att). This analysis revealed no significant findings.

Behavioral data

During the EEG-fMRI experiment, participants had to answer yes/no questions on the content of the to-be-attended story. On average, 80.7 ± 14.3% of the questions were answered correctly (total range: 50.0–94.4%). The average score obtained in the delayed recall task of the attended story's content after finishing the experiment was 9.5 ± 4.1 of 20 items (range: 4–17). The scores of both tasks were positively correlated across subjects (Pearson's $r = 0.57$, $p = 0.035$, FDR corrected).

After finishing the measurements, participants rated the task difficulty and the distraction by the to-be-ignored speech stream and the scanner background noise. Most of the subjects could follow the to-be-attended story reasonably well (rating: 6.7 ± 2.1 (M ± SD); range: 3–10; scale: 1 = not at all, 10 = very good), although both the to-be-ignored stream (rating: 3.3 ± 2.2; range: 1–10; scale: 1 = not at all, 10 = very strong) and the scanner background noise (rating: 4.9 ± 2.6) posed a considerable source of distraction. On average, the scanner noise was rated to be more distracting than the to-be-ignored speech stream (Wilcoxon signed-rank test, $p = 0.046$, FDR corrected).

Complementing the measures acquired during the selective speech tracking task, individual data on general spatial selective listening abilities were obtained using a behavioral deviance detection task in which participants had to detect deviant tones in one ear while ignoring similar deviants occurring in the other ear. Task blocks without contralateral auditory input served as the baseline condition. The response accuracy in this task was significantly reduced in the two-stream condition (F_1 score = 0.83 ± 0.10) compared with baseline (F_1 score = 0.94 ± 0.07; $t_{(15)} = -4.74$, $p = 0.002$, FDR corrected), but generally stayed on a high level. F_1 scores obtained in the two-stream condition were not correlated with the number of correctly answered questions on the contents of the to-be-attended story (Pearson's $r = 0.14$; $p = 0.589$, FDR corrected).

We tested for a relationship between the individual envelope reconstruction accuracy for the to-be-attended stream and selective listening performance. There was no relationship between r_z(att) and the percentage of correctly answered questions in the selective listening task (Pearson's $r = 0.20$; $p = 0.244$, FDR corrected). In contrast, we observed a significant positive correlation between mean r_z(att) values and F_1 scores in the two-stream condition of the behavioral deviance detection task, suggesting that envelope reconstruction scores were positively related to individual spatial selective listening abilities (Pearson's $r = 0.48$; $p = 0.046$, FDR corrected). The relationship between F_1 scores and r_z(att) is depicted in Figure 5*B*.

Discussion

Here, we combined an EEG-based speech envelope reconstruction approach with simultaneously acquired fMRI data to study the relationship between attentional modulations in the cortical tracking of an attended speech stream and neural activity in frontoparietal brain areas. The only brain region showing a significant relationship between BOLD response amplitudes during selective listening and the individual speech envelope reconstruction accuracies for the to-be-attended speech stream was the right temporoparietal junction, a core region of the ventral attention network. Complementing this finding, the temporoparietal junction showed task-specific functional connectivity to secondary auditory cortex and brain areas related to attention control during selective listening. Envelope reconstruction accuracies for the to-be-attended speech were further found to be positively correlated with performance in a behavioral dichotic target detection task, supporting the view that differences in speech tracking are related to individual selective listening abilities.

Speech envelope reconstruction accuracy and auditory selective attention

The applied stimulus reconstruction approach is based on a linear decoding method in which the envelope of a speech signal is reconstructed from the simultaneously recorded EEG data (Crosse et al., 2016b). Similar methods have been applied successfully to decode information in different sensory domains and using various electrophysiological recording techniques (Lalor et al., 2006; Ding and Simon, 2012b; Crosse et al., 2015; Crosse et al., 2016a). Although the approach works best using high-density EEG, it is robust to significant electrode reduction and has even been demonstrated with EEG montages restricted to the listener's ear (Mirkovic et al., 2015; Mirkovic et al., 2016; Fiedler et al., 2017). Complementing these findings, we here show that the EEG stimulus reconstruction approach can be used with data obtained in an adverse fMRI environment. This enabled us to study the relationship between attentional modulations in the tracking of speech and sustained frontoparietal brain activity during selective listening.

We could not establish a statistically significant relationship between trial-to-trial fluctuations in envelope reconstruction accuracy and single-trial BOLD responses. However, qualitatively, r_z(att) values tended to be positively related to BOLD fluctuations in the anterior superior temporal sulcus and, to a lesser extent, in adjacent frontal brain areas involved in speech processing, indicating that neural ensembles within these regions track the to-be-attended speech stream. This observation is consistent with recent MEG and electrocorticography findings, also reporting speech tracking in superior temporal and lateral frontal structures (Zion Golumbic et al., 2013; Vander Ghinst et al., 2016). Zion Golumbic et al. (2013) further demonstrated that, whereas early auditory sensory areas track both attended and ignored speech streams in multispeaker situations, the tracking gets more selective in later processing stages.

Stimulus reconstruction accuracies have been linked successfully to behavioral task performance in auditory, visual, and audiovisual settings. (Crosse et al., 2015; O'Sullivan et al., 2015; Crosse et al., 2016a; Presacco et al., 2016). In particular, O'Sullivan and colleagues (2015) reported a positive relationship between single-trial speech envelope reconstruction accuracy for the attended stream during selective listening and trial outcome. Complementing these findings, we here observed a positive correlation between the mean speech envelope reconstruction accuracy for the to-be-attended stream and individual selective listening abilities. However, we did not establish a direct relationship between speech envelope reconstruction accuracies and selective listening performance in the same experiment, but rather relied on results obtained in a secondary behavioral task. The number of correctly answered questions on the to-be-attended speech stream during the experiment was not correlated with individual differences in the tracking of this stream. We speculate that the comparatively small number of yes/no questions asked in our study, which concentrated on the last portion of the selective listening blocks only, cannot reliably capture individual differences in selective listening performance.

Frontoparietal brain activity during selective listening

Previous fMRI studies on auditory selective attention reported activations in core regions of the dorsal attention network, particularly within the intraparietal sulcus but also in the frontal eye fields and the middle frontal gyri, but to some extent also in the ventral attention network, including the right temporoparietal junction, the insula, and the inferior frontal gyrus (Hill and Miller, 2010; Kong et al., 2014; Seydell-Greenwald et al., 2014; Michalka et al., 2015). In some disagreement with this, we here found no sustained BOLD activations during selective listening in these regions. It cannot be completely ruled out that BOLD responses in frontoparietal brain regions observed in previous experiments were partially misattributed to selective listening. In most previous experiments, selective listening epochs were significantly shorter than in our study, making it difficult to dissociate onset-related and sustained activations related to selective listening. In addition, some studies applied memory-related or deviance detection tasks to control for attention allocation, which themselves trigger neural activity within frontoparietal regions (Huang et al., 2012; Huang et al., 2013; Puschmann et al., 2016; Albouy et al., 2017). Conversely, we here applied a continuous task design lacking periods of rest. Brain regions showing similar activation levels in all task conditions are unlikely to be identified in such designs. Indeed, direct contrasts among single-speaker, two-speaker, and question blocks revealed no differences within the frontoparietal attention network (data not reported), suggesting that frontoparietal brain regions involved in controlling attention show similar sustained activation levels in all conditions.

Although our data show no increased activation levels within the dorsal or ventral attention networks during selective listening per se, BOLD responses in the right intraparietal sulcus were modulated significantly by the spatial location (left vs right) of the to-be-ignored stream. This result is consistent with previous electrophysiological studies providing strong experimental evidence for a role of the intraparietal sulcus in controlling spatial auditory attention (Kerlin et al., 2010; Ahveninen et al., 2013; Huang et al., 2014). We suggest that this region is involved orienting spatial attention away from the to-be-ignored speaker to facilitate the tracking of the centrally presented speech stream.

Role of the temporoparietal junction

The right temporoparietal junction was the only brain region in which BOLD response amplitudes were significantly related to interindividual differences in speech envelope tracking accuracies for the to-be-attended stream. Subjects with high mean envelope reconstruction accuracies had consistently higher activation levels in this region than participants with a generally poor cortical tracking of the attended speech stream. Single-trial response amplitudes in the right temporoparietal junction and trial-to-trial fluctuations in speech envelope reconstruction accuracy were, however, not robustly linked during the experiment. This suggests that activations in this region, unlike BOLD responses along the superior temporal sulcus and inferior frontal cortex, are not directly related to the actual tracking of the speech envelope. Instead, we propose that the right temporoparietal junction is involved in controlling attention during selective listening, allowing for a better overall maintenance of selective attention on the to-be-attended stream. Therefore, brain regions functionally connected to the right temporoparietal junction included not only the secondary auditory cortex but also parts of the ventral and the dorsal attention network, particularly the superior parietal lobe and the intraparietal sulcus.

The right temporoparietal region is a core structure of the ventral attention network, which is classically associated with target detection and stimulus-driven reorienting of attention to salient sensory events (Corbetta et al., 2008; Vossel et al., 2014). Recent MEG studies demonstrated a relationship between right temporoparietal activity and the success of voluntary spatial attention switches between streams, suggesting that this region is also involved in controlling voluntary orienting of auditory spatial attention (Ahveninen et al., 2013; Larson and Lee, 2014). In this view, right temporoparietal junction activity during sustained selective listening could be related to a spatial reorienting after lapses in which attention has been dragged to the to-be-ignored stream. Increased activation levels may thus be related to efficient and fast orienting processes during selective listening, leading to a better overall tracking of the to-be-attended stream. However, right temporoparietal activity during the onset of the to-be-ignored stream, likely to be related to an initial orientation of attention to the to-be-attended stream, did not correlate with speech tracking abilities.

Conclusion

We demonstrate that the stimulus reconstruction method can be applied reliably to EEG data acquired in the adverse environment of an MRI scanner. We used this method to study the relationship between the individual differences in the cortical tracking of an attended speech stream in a "cocktail party" situation and BOLD responses in frontoparietal brain regions involved in controlling auditory attention. Our results suggest that the right temporoparietal junction, a core region of the ventral attention network, essentially contributes to a successful cortical tracking of the attended speech stream. Given the known role of this region in orienting attention and attention switching, we speculate that the right temporoparietal junction may particularly contribute to an efficient reorienting of attention after short lapses of attention, allowing for an overall better tracking of the to-be-attended stream. However, additional research on the temporal dynamics of neural activity in this region during sustained selective listening is needed to clarify its role in controlling attention in "cocktail party" situations.

References

Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S, Lin FH, Raij T, Sams M, Vasios CE, Belliveau JW (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proc Natl Acad Sci U S A 108:4182–4187. CrossRef Medline

Ahveninen J, Huang S, Belliveau JW, Chang WT, Hämäläinen M (2013) Dynamic oscillatory processes governing cued orienting and allocation of auditory attention. J Cogn Neurosci 25:1926–1943. CrossRef Medline

Albouy P, Weiss A, Baillet S, Zatorre RJ (2017) Selective entrainment of theta oscillations in the dorsal stream causally enhances auditory working memory performance. Neuron 94:193–206.e5. CrossRef Medline

Anderson S, Kraus N (2010) Sensory-cognitive interaction in the neural encoding of speech in noise: a review. J Am Acad Audiol 21:575–585. CrossRef Medline

Ashburner J (2007) A fast diffeomorphic image registration algorithm. Neuroimage 38:95–113. CrossRef Medline

Bidelman GM, Dexter L (2015) Bilinguals at the "cocktail party": dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners' speech-in-noise recognition deficits. Brain Lang 143:32–41. CrossRef Medline

Bidet-Caulet A, Fischer C, Besle J, Aguera PE, Giard MH, Bertrand O (2007) Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. J Neurosci 27:9252–9261. CrossRef Medline

Bronkhorst AW (2015) The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten Percept Psychophys 77:1465–1487. CrossRef Medline

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

Coffey EBJ, Mogilever NB, Zatorre RJ (2017) Speech-in-noise perception in musicians: a review. Hear Res 352:49–69. CrossRef Medline

Corbetta M, Patel G, Shulman GL (2008) The reorienting system of the human brain: from environment to theory of mind. Neuron 58:306–324. CrossRef Medline

Crosse MJ, Butler JS, Lalor EC (2015) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J Neurosci 35:14195–14204. CrossRef Medline

Crosse MJ, Di Liberto GM, Lalor EC (2016a) Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J Neurosci 36:9888–9895. CrossRef Medline

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016b) The Multivariate Temporal Response Function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604. Medline

Delorme A, Makeig S (2004) EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21. CrossRef Medline

Ding N, Simon JZ (2012a) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89. CrossRef Medline

Ding N, Simon JZ (2012b) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109:11854–11859. CrossRef Medline

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735. CrossRef Medline

Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J (2017) Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng 14:036020. CrossRef Medline

Gomes H, Duff M, Ramos M, Molholm S, Foxe JJ, Halperin J (2012) Auditory selective attention and processing in children with attention-deficit/hyperactivity disorder. Clin Neurophysiol 123:293–302. CrossRef Medline

Green JJ, Doesburg SM, Ward LM, McDonald JJ (2011) Electrical neuroimaging of voluntary audiospatial attention: evidence for a supramodal attention control network. J Neurosci 31:3560–3564. CrossRef Medline

Hill KT, Miller LM (2010) Auditory attentional control and selection during cocktail party listening. Cereb Cortex 20:583–590. CrossRef Medline

Huang S, Belliveau JW, Tengshe C, Ahveninen J (2012) Brain networks of novelty-driven involuntary and cued voluntary auditory attention shifting. PLoS One 7:e44062. CrossRef Medline

Huang S, Seidman LJ, Rossi S, Ahveninen J (2013) Distinct cortical networks activated by auditory attention and working memory load. Neuroimage 83:1098–1108. CrossRef Medline

Huang S, Chang WT, Belliveau JW, Hämäläinen M, Ahveninen J (2014) Lateralized parietotemporal oscillatory phase synchronization during auditory selective attention. Neuroimage 86:461–469. CrossRef Medline

Kerlin JR, Shahin AJ, Miller LM (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". J Neurosci 30:620–628. CrossRef Medline

Kim H (2014) Involvement of the dorsal and ventral attention networks in oddball stimulus processing: a meta-analysis. Hum Brain Mapp 35:2265–2284. CrossRef Medline

Kong L, Michalka SW, Rosen ML, Sheremata SL, Swisher JD, Shinn-Cunningham BG, Somers DC (2014) Auditory spatial attention representations in the human cerebral cortex. Cereb Cortex 24:773–784. CrossRef Medline

Lalor EC, Pearlmutter BA, Reilly RB, McDarby G, Foxe JJ (2006) The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage 32:1549–1561. CrossRef Medline

Larson E, Lee AK (2014) Switching auditory attention using spatial and non-spatial features recruits different cortical networks. Neuroimage 84:681–687. CrossRef Medline

Lee AK, Rajaram S, Xia J, Bharadwaj H, Larson E, Hämäläinen MS, Shinn-Cunningham BG (2012) Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. Front Neurosci 6:190. CrossRef Medline

Lee AK, Larson E, Maddox RK, Shinn-Cunningham BG (2014) Using neuroimaging to understand the cortical mechanisms of auditory selective attention. Hear Res 307:111–120. CrossRef Medline

McDermott JH (2009) The cocktail party problem. Curr Biol 19:R1024–R1027. CrossRef Medline

Mechelli A, Price CJ, Friston K, Ashburner J (2005) Voxel-based morphometry of the human brain: methods and applications. Curr Med Imaging Rev 1:105–113. CrossRef

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236. CrossRef Medline

Michalka SW, Kong L, Rosen ML, Shinn-Cunningham BG, Somers DC (2015) Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. Neuron 87:882–892. CrossRef Medline

Mirkovic B, Debener S, Jaeger M, De Vos M (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng 12:046007. CrossRef Medline

Mirkovic B, Bleichner MG, De Vos M, Debener S (2016) Target speaker detection with concealed EEG around the ear. Front Neurosci 10:349. CrossRef Medline

Niazy RK, Beckmann CF, Iannetti GD, Brady JM, Smith SM (2005) Removal of FMRI environment artifacts from EEG data using optimal basis sets. Neuroimage 28:720–737. CrossRef Medline

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 25:1697–1706. CrossRef Medline

Peelle JE, Cusack R, Henson RN (2012) Adjusting for global effects in voxel-based morphometry: gray matter decline in normal aging. Neuroimage 60:1503–1516. CrossRef Medline

Petersen EB, Wöstmann M, Obleser J, Lunner T (2017) Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J Neurophysiol 117:18–27. CrossRef Medline

Power AJ, Lalor EC, Reilly RB (2011) Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. Cereb Cortex 21:1223–1230. CrossRef Medline

Presacco A, Simon JZ, Anderson S (2016) Effect of informational content of noise on speech representation in the aging midbrain and cortex. J Neurophysiol 116:2356–2367. CrossRef Medline

Puschmann S, Huster RJ, Thiel CM (2016) Mapping the spatiotemporal dynamics of processing task-relevant and task-irrelevant sound feature changes using concurrent EEG-fMRI. Hum Brain Mapp 37:3400–3416. CrossRef Medline

Rönnberg J, Lunner T, Zekveld A, Sörqvist P, Danielsson H, Lyxell B, Dahlström O, Signoret C, Stenfelt S, Pichora-Fuller MK, Rudner M (2013) The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. Front Syst Neurosci 7:31. CrossRef Medline

Ross B, Hillyard SA, Picton TW (2010) Temporal dynamics of selective attention during dichotic listening. Cereb Cortex 20:1360–1371. CrossRef Medline

Seydell-Greenwald A, Greenberg AS, Rauschecker JP (2014) Are you listening? Brain activation associated with sustained nonspatial auditory attention in the presence and absence of stimulation. Hum Brain Mapp 35:2233–2252. CrossRef Medline

Simon JZ (2015) The encoding of auditory objects in auditory cortex: insights from magnetoencephalography. Int J Psychophysiol 95:184–190. CrossRef Medline

Sussman E, Winkler I, Wang W (2003) MMN and attention: Competition for devience detection. Psychophysiology 40:430–435. CrossRef Medline

Teder W, Kujala T, Näätänen R (1993) Selection of speech messages in free-field listening. Neuroreport 5:307–309. CrossRef Medline

Van Rijsbergen CJ (1979) Information retrieval, Ed 2. Newton, MA: Butterworth.

Vander Ghinst M, Bourguignon M, Op de Beeck M, Wens V, Marty B, Hassid S, Choufani G, Jousmäki V, Hari R, Van Bogaert P, Goldman S, De Tiège X (2016) Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. J Neurosci 36:1596–1606. CrossRef Medline

Vossel S, Geng JJ, Fink GR (2014) Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. Neuroscientist 20:150–159. CrossRef Medline

Whitfield-Gabrieli S, Nieto-Castanon A (2012) Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect 2:125–141. CrossRef Medline

Xiang J, Simon J, Elhilali M (2010) Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. J Neurosci 30:12084–12093. CrossRef Medline

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 77:980–991. CrossRef Medline