Received: 1 August 2021 | Accepted: 10 October 2021

DOI: 10.1111/ejn.15504

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

Selective auditory attention within naturalistic scenes modulates reactivity to speech sounds

Hanna Renvall[1,2,3] | **Jaeho Seol**[1,2] | **Riku Tuominen**[1,2] | **Bettina Sorger**[4] | **Lars Riecke**[4] | **Riitta Salmelin**[1,2]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland

[2]Aalto NeuroImaging, Aalto University, Espoo, Finland

[3]BioMag Laboratory, HUS Diagnostic Center, Helsinki University Hospital, University of Helsinki and Aalto University School of Science, Helsinki, Finland

[4]Department of Cognitive Neuroscience, Maastricht University, Maastricht, The Netherlands

Correspondence
Hanna Renvall, Department of Neuroscience and Biomedical Engineering, Aalto University, P.O. Box 12200, FI-00076 Aalto, Espoo, Finland.
Email: hanna.renvall@aalto.fi

Funding information
Päivikki ja Sakari Sohlbergin säätiö; Suomen Kulttuurirahasto; Academy of Finland, Grant/Award Numbers: 127401, 277655, 292552, 315553, 321460

Edited by: John Foxe

Abstract

Rapid recognition and categorization of sounds are essential for humans and animals alike, both for understanding and reacting to our surroundings and for daily communication and social interaction. For humans, perception of speech sounds is of crucial importance. In real life, this task is complicated by the presence of a multitude of meaningful non-speech sounds. The present behavioural, magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) study was set out to address how attention to speech versus attention to natural non-speech sounds within complex auditory scenes influences cortical processing. The stimuli were superimpositions of spoken words and environmental sounds, with parametric variation of the speech-to-environmental sound intensity ratio. The participants' task was to detect a repetition in either the speech or the environmental sound. We found that specifically when participants attended to speech within the superimposed stimuli, higher speech-to-environmental sound ratios resulted in shorter sustained MEG responses and stronger BOLD fMRI signals especially in the left supratemporal auditory cortex and in improved behavioural performance. No such effects of speech-to-environmental sound ratio were observed when participants attended to the environmental sound part within the exact same stimuli. These findings suggest stronger saliency of speech compared with other meaningful sounds during processing of natural auditory scenes, likely linked to speech-specific top-down and bottom-up mechanisms activated during speech perception that are needed for tracking speech in real-life-like auditory environments.

Abbreviations: ANOVA, analysis of variance; BOLD, blood-oxygen-level dependent; dSPM, dynamical statistical parametric map; ECD, equivalent current dipole; fMRI, functional magnetic resonance imaging; FWHM, full width at half maximum; GLM, general linear model; ISI, interstimulus interval; MEG, magnetoencephalography; MNE, minimum norm estimate; MR, magnetic resonance; PT, planum temporale; RMS, root-mean-square value; SEM, standard error of the mean; SER, speech-to-environmental sound intensity ratio; SNR, signal-to-noise ratio; SOA, stimulus-onset-asynchrony; VOI, volume of interest.

KEYWORDS
auditory system, environmental sounds, functional magnetic resonance imaging, magnetoencephalography, speech processing

1 | INTRODUCTION

Humans can efficiently attend and react to behaviourally relevant perceptual features such as speech in their natural environment. The process is likely determined by an interplay between bottom-up feature analysis and top-down goal-directed attention to relevant sound attributes (Bregman, 1990). Auditory selective attention has been extensively studied using simplistic sounds (see, e.g., Picton & Hillyard, 1974; Näätänen & Michie, 1979; Alho, 1992; Lee et al., 2014), but the neural mechanisms may be different for the analysis of multiple and more complex concurrent sound streams—like speech amid other real-life sounds—given the greater perceptual load (Lavie, 1995). For segregating natural auditory objects from the auditory background, relevant spatial, temporal and spectral cues appear to be processed in an attention-dependent manner at the cortical locations generally involved in sound recognition (Hausfeld, Riecke, & Formisano, 2018; Renvall et al., 2016). Selectively attending to target sounds is accompanied by stronger auditory cortical steady-state responses to these sounds, compared with attending to background sounds, which may boost the perception of the target sound (e.g., Elhilali et al., 2009). Attention to speech among musical melodies increases blood-oxygenation-level dependent (BOLD) signals in the left frontal and temporal cortices (Yoncheva et al., 2010). Recent studies investigating concurrent speech sound streams have demonstrated that whereas early auditory cortical responses appear to represent both attended and ignored speech (Puvvada & Simon, 2017), the higher-order auditory cortices specifically track the spectral and temporal features of the attended speech signals (Ding & Simon, 2012; Kerlin et al., 2010; Mesgarani & Chang, 2012; Vander Ghinst et al., 2016) while suppressing those of the unattended speech (Mesgarani & Chang, 2012; Puvvada & Simon, 2017). Furthermore, the tracking of on-going speech is enhanced with increasing speech intelligibility (Peelle et al., 2013).

These results leave open the question of whether such processes are unique for speech or generalize to other natural sounds outside the language domain. Speech-specific processing in the superior temporal cortex is generally supported by results showing sensitivity to phoneme-related spectrotemporal features (Chang et al., 2010; Mesgarani et al., 2014) and to the temporal structure of speech compared with acoustically similar non-speech sounds (Overath et al., 2015). Speech specificity is often studied by comparing brain responses with speech and acoustically well-matched non-speech stimuli (Overath et al., 2015; Peelle et al., 2013; Scott et al., 2000). With an alternative paradigm that made use of the natural acoustical variability within speech sounds and a machine-learning approach, we recently demonstrated that magnetoencephalographic (MEG) auditory responses accurately follow the temporal unfolding of spoken words presented in isolation (Nora et al., 2020). Notably, a similar time-locking mechanism was not evident for processing of environmental sounds, not even for non-speech vocal sounds with spectrotemporal modulations comparable to spoken words and conveying the same meaning (e.g., coughing and laughter). The result suggested that the acoustic-phonetic content of speech sounds is tracked in a special manner compared with other natural sounds (Nora et al., 2020).

The present study was set out to address the proposed speech selectivity when meaningful auditory stimuli are heard in naturalistic auditory environments. Similarly to words, environmental sounds have semantic content and complex spectrotemporal patterns (Dick et al., 2007; Gygi et al., 2004). Environmental sounds can prime semantically related words (van Petten & Rheinfelder, 1995), and processing of both speech and environmental sounds is modulated by contextual cues (Ballas & Howard, 1987). When speech or environmental sounds are embedded in noise, cortical processing of the sounds varies both in a stimulus-specific and noise intensity-dependent manner: Sustained, left superior temporal activation starting at ~400 ms after stimulus onset, as measured with MEG, is modulated by the signal-to-noise ratio (SNR) for speech sounds but not environmental sounds, suggestive of stronger reactivity for speech than for other sounds in the auditory cortex (Renvall et al., 2012). However, given the differences in the spectral structure between the speech and environmental sounds in that study, part of the observed stimulus specificity might have originated from differential vulnerability to the applied noise.

Here, participants listened to auditory scenes consisting of superimposed speech and environmental sounds with varying speech-to-environmental sound intensity ratios (SERs), and attended to either speech or environmental sounds within the same set of stimuli. This experimental design allowed direct comparison of the two attentional states (focus on speech vs. environmental sounds) at three SERs (+18, 0 and −18 dB) as use of the same acoustic input for both conditions eliminated the confounding effect of acoustical differences between speech and environmental sounds. We predicted that, in accordance with earlier recordings (Chang et al., 2010; Hausfeld, Riecke, Valente, & Formisano, 2018; Mesgarani & Chang, 2012; Puvvada & Simon, 2017), the cortical activation pattern would primarily reflect the attended component of the complex auditory stimulus, especially during hierarchically later processing stages. Specifically, taken the notable time locking of auditory cortical responses to speech sounds (Nora et al., 2020), we hypothesized that attention to speech sounds, in particular, would modulate their cortical processing in an SER-related manner, compared with other complex sounds especially in the left planum temporale (PT), which has been suggested to be particularly involved in analysis of spectrotemporally complex sounds (Griffiths & Warren, 2002) and showed sensitivity to speech SNR in our earlier study (Renvall et al., 2012).

2 | MATERIALS AND METHODS

2.1 | Subjects

We studied, with informed consent, 11 Finnish-speaking adults (mean ± standard error of the mean [SEM] age 28 ± 1 years; 5 females, 6 males; all right-handed). None of them reported a history of hearing or neurological impairments. All subjects participated in both neuroimaging experiments (MEG and functional magnetic resonance imaging [fMRI]). The study had a prior approval of the Aalto University Research Ethics Committee, and it conforms to the Declaration of Helsinki.

The original datasets are not publicly available due to restrictions placed by the Aalto Research Ethics Committee. The data that support the findings of this study are available from the corresponding author with permission of the Aalto Research Ethics Committee.

A purely behavioural experiment employing the same design as in the functional neuroimaging experiments was conducted on eight additional subjects who did not participate in the MEG or fMRI experiments.

Our sample size was relatively small according to current neuroimaging standards, and the current pandemic situation hinders us from increasing it at this point. However, our main MEG results were very robust, and it is unlikely that increasing the sample size would affect them.

2.2 | Experimental stimuli

The stimuli consisted of 48 superimpositions of spoken words and environmental sounds (mean [±standard deviation] duration 870 ± 100 ms). The 48 spoken words were consonant-initial, eight-letter common Finnish nouns, pronounced by eight speakers (four males and four females; six words/speaker), and they were recorded in an acoustically shielded room. The 48 environmental sounds were collected from the internet and comprised, for example, sounds of tool use, animal cries, traffic sounds and nonlinguistic human sounds (e.g., laughing and crying). In the combined sounds, the speech and environmental sound excerpts were not semantically associated to each other (e.g., speech sound 'tomaatti' [tomato] and sound of baby crying; 'kapteeni' [captain] and guitar sound).

All individual sounds were first low-pass filtered at 8 kHz. The speech and environmental sounds were then superimposed with SERs of +18, 0 and −18 dB. These particular SERs were selected on the basis of our previous study (Renvall et al., 2012), in which they were found to produce the largest effects on cortical activity. The amplitudes of the different combined stimuli were adjusted so that all stimuli had the same overall root-mean-square value (RMS). At all three SERs, both sound excerpts within the stimuli were clearly distinguishable (see behavioural outcome in Section 3). The stimuli were sampled at 41 kHz (16-bit, mono), and they had rise and fall times of 20 ms. Figure 1a illustrates the spectrograms and sound waveforms for one original speech sound ('kaupunki', i.e., 'town') and one environmental sound (*alarm clock*) and for their superimposition at 0 dB.

Although the auditory stimulation was kept constant between the two *Focus of Attention* conditions, we also aimed at controlling for the possibility that the superimposed stimuli could contain spectrotemporal features attracting attention more to the linguistic than environmental sound attributes or vice versa. To this end, more detailed acoustical analyses of the original sounds (perceived loudness) and of the superimposed sounds used in the experiment (power distribution across stimulus frequencies, temporal modulations within sounds, correlation between spectral frequency channels) were conducted, and they are described in Supporting Information.

2.3 | Experimental design

During the neuroimaging experiments, the sounds were presented in trains of four. The subjects were instructed to attend to either speech or environmental sound

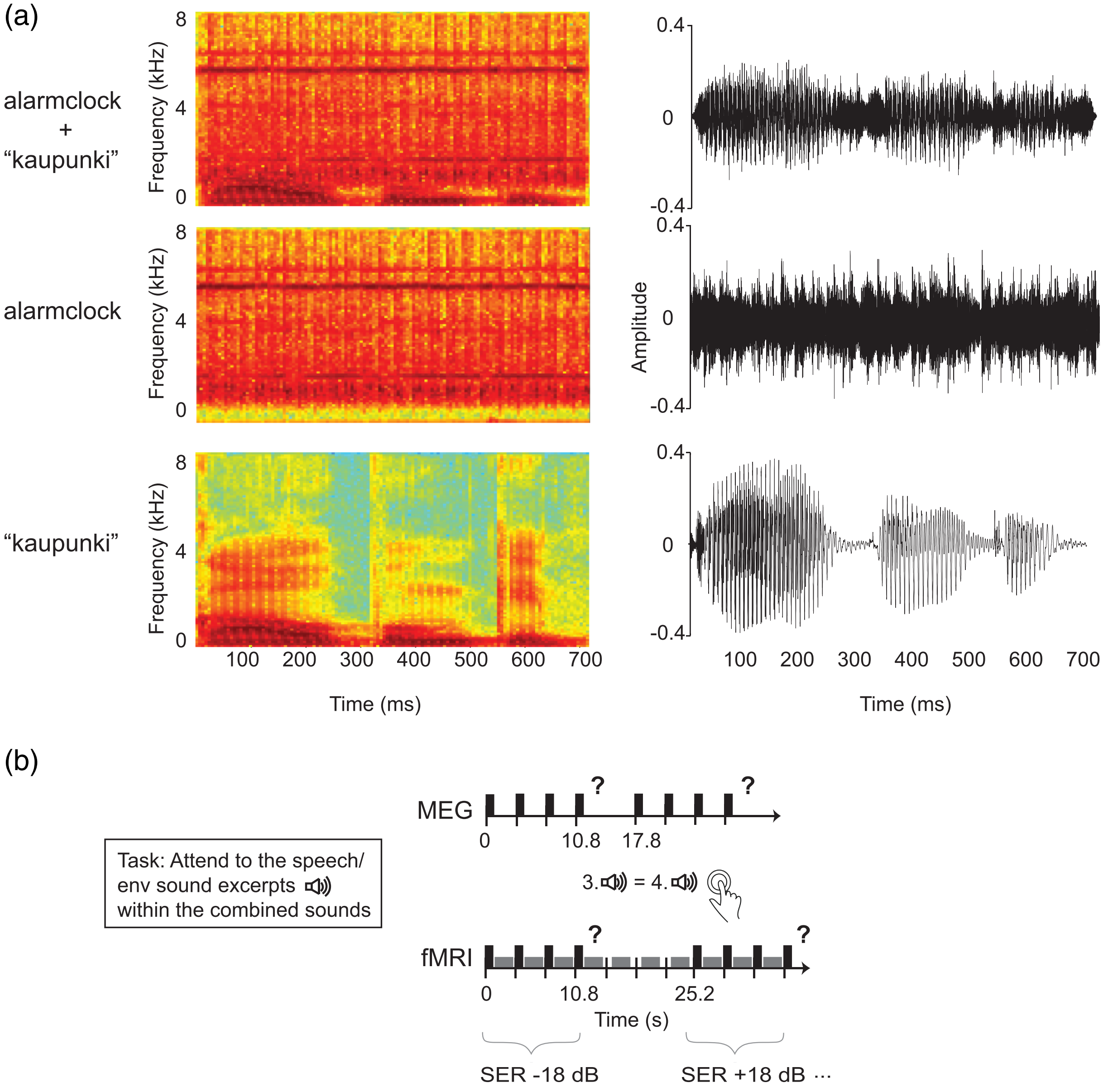

FIGURE 1 An example stimulus and the experimental design. (a) The spectrograms and waveforms for an example stimulus (sound of an alarm clock combined with the speech sound 'kaupunki', i.e., 'town'; top row) and for the original sound excerpts (middle and bottom rows). (b) Schematic presentation of the trials in the magnetoencephalography (MEG) (above) and functional magnetic resonance imaging (fMRI) (below) experiments (black = auditory stimuli, grey = fMRI scanning time). After the last (4th) stimulus within each trial, the subject decided whether the 3rd and 4th stimuli contained the same speech/environmental sound excerpt, depending on which line of sound excerpts the subject was attending to. Stimulus trains with different speech-to-environmental sound intensity ratios (SERs) (+18, 0 and −18 dB) were presented in a random order, and the same condition did not repeat immediately

excerpts within the sounds and to respond with a finger lift if the attended excerpts in the last two stimuli in a train were the same. For this task, in addition to the 48 experimental sounds, 16 new combinations from the original speech and environmental sounds were created to serve as catch trials. These sounds consisted of the same attended speech/environmental excerpt as the preceding stimulus on a given trial, but the non-attended sound excerpt was different. Cortical and behavioural responses to catch trials were excluded from the analyses.

In both experiments, the superimposed speech/environmental sounds were presented in trains of four with a stimulus-onset-asynchrony (SOA) of 3600 ms. The particular SER of the superimposed sounds remained the same within a given stimulus train and varied between the trains. Stimulus trains with different SERs were presented in a random order, and the same condition did not repeat immediately.

In the beginning of each experimental MEG/fMRI run, each comprising 40 stimulus trains, the subject was instructed to attend to either speech or environmental sound excerpts within the stimuli and respond with a finger lift if the attended excerpts in the last two stimuli within a train were the same (occurring in 10% of the stimulus trains). The response hand was alternated across subjects, and the catch trials with motor responses were discarded from both MEG and fMRI analyses.

The experimental paradigm was designed to be well-suited for both MEG and fMRI recordings. Within each research modality, the measurement duration was adjusted on the basis of the assumed SNR, based on previous comparable studies (e.g., Renvall et al., 2012), while keeping the number of times each stimulus was presented the same in both the MEG and fMRI recordings. In the MEG experiment, the 48 combined sounds were presented at each SER four times (twice in each attentional condition), and each stimulus train was followed by a rest interval of 7 s. In the fMRI experiment, a subset of 24 superimposed sounds (duration 865 ± 90 ms) was presented, at each SER four times (twice in each attentional condition); here, each stimulus train was followed by a 14.4-s rest interval. The order of runs in which participants attended to speech or environmental sounds, as well as the order of fMRI and MEG experiments, was counterbalanced across subjects. During the MEG and fMRI experiments, the sounds were delivered to the subjects binaurally at a similar, comfortable listening level through plastic tubes and earpieces. The experimental design is schematically described in Figure 1b. For assuring proper vigilance for both experiments (MEG and fMRI), they were conducted on separate days.

2.4 | Behavioural data: Acquisition and analysis

A purely behavioural experiment employing the same 2 (*Focus of Attention*) × 3 (*SER*) design as in the functional neuroimaging experiments was conducted on eight subjects who did not participate in the MEG or fMRI experiments. They listened to pairs of stimuli in two separate runs. In each stimulus pair, one sound was a superimposed sound used also in the neuroimaging experiment, and the other one was a new combination from the 48 original speech and environmental stimuli. The subject's task was to listen to the sound pairs and indicate with a button press as soon as possible whether the attended excerpts (speech or environmental sounds, in different runs) in the two sounds were the same. Altogether 99 sound pairs (three SERs, 33 sound pairs in each SER) were presented during both runs, with an SOA of 2 s and an inter-pair interval of ~3 s. The repetition-detection performance was taken as a measure for how well listeners could disentangle the target sound from the superimposed sounds.

During the neuroimaging experiments, the behavioural responses were too scarce for statistical inference. However, the hit rate of behavioural responses across SERs was 84 ± 9% in the 'attention to speech' task and 95 ± 8% in the 'attention to environmental sound' task (pooled across the MEG and fMRI experiments), indicating that the subjects followed properly the task instructions.

2.5 | MEG data: Acquisition and analysis

Auditory evoked fields were recorded in a magnetically shielded room while the subject was seated with the head covered by the MEG helmet. A 306-channel Vectorview neuromagnetometer (Elekta Neuromag, Helsinki, Finland) was used, which contains 102 identical triple sensors, comprising two orthogonal planar first-order gradiometers and one magnetometer, each of them coupled to a Superconducting QUantum Interference Device. Four head-position-indicator coils were attached to the scalp, and their positions were measured with a 3D digitizer; the head-coordinate frame was anchored to the two periauricular points and the nasion. The head position with respect to the MEG sensor array was determined by briefly feeding current to the marker coils before the actual measurement.

The MEG signals were band-pass filtered at 0.03–330 Hz, digitized at 1000 Hz and averaged from 300 ms before the stimulus onset to 1500 ms after it, setting as baseline the 300-ms interval immediately preceding the stimulus onset. The averaged signals were digitally low-pass filtered at 40 Hz. The horizontal and vertical electro-oculograms were recorded to identify and discard data contaminated by eye blinks and movements. The responses from the 2nd to the 4th sound in the stimulus trains were averaged; the response to the first stimulus in each train was omitted from the analysis due to the longer preceding interstimulus interval (ISI). A minimum of 70 artefact-free responses were collected per condition.

For source-space analysis, the head was modelled as a homogeneous sphere. The model parameters were optimized for the intracranial space based on individual MR images that were available for all subjects. The responses were analysed by first segregating the recorded sensor-level signals into separable cortical-level spatiotemporal components, by means of guided neural current modelling (equivalent current dipole [ECD]; Hämäläinen et al., 1993), separately for each subject. The model parameters of an ECD represent the location, orientation and strength of the net current in an activated brain area. Only ECDs explaining more than 85% of the local field variance during the response peaks in a subset of 16–20 gradiometer channels were included in the final model. Based on this criterion, 2–5 spatiotemporal components were selected in the participants' individual models. The location of neural activity, within the limits of MEG's spatial resolution (Hansen et al., 2010), was similar across the different experimental conditions. Therefore, to maximize the goodness-of-fit of the model and to diminish uncertainty due to variability of model accuracies, for each participant, the components were identified in the condition with the strongest signals, thus best SNR. Thereafter, it was verified that the obtained ECDs explained well the responses also in the other experimental conditions. The location of the components explaining the field patterns at around 100 and 400 ms was found to be very similar, in agreement with previous studies (Bonte et al., 2006; Helenius et al., 2002; Renvall et al., 2012; Uusvuori et al., 2008). To prevent interactions between these ECDs when all relevant ECDs were brought into the same model, 400-ms component was used to model both responses. In all subjects, the component explained well also the 100-ms deflections, resulting in responses that exceeded at least 4 times the standard deviation of the noise level calculated from the 300-ms prestimulus baseline. The same component also accounted for the current flow of opposite direction at ∼200 ms.

In the auditory modality, ECD models have been shown to align well with distributed modelling approaches (Vartiainen et al., 2009). For verifying the spatial distribution of activity obtained with ECD modelling, the cortical generators were additionally visualized with a distributed source model, using minimum norm estimates (MNEs; Hämäläinen & Ilmoniemi, 1994) with the MNE Suite software package (Gramfort et al., 2014). MNE implements the L2 MNE of the source distribution; that is, it determines the current distribution that explains the measurements and has the smallest L2-norm. MNE analysis results in distributed models of the cortical activation, but it does not provide information of the actual shape or extent of the activated area. For MNE analysis, the cortical surface of each participant was reconstructed from the corresponding MR images with the Freesurfer software (Dale & Sereno, 1993; Fischl et al., 1999). Each hemisphere was covered with ∼5000 potential source locations. Currents oriented normal to the cortical surface were favoured by weighting the transverse currents by a factor of .3 (Lin et al., 2006), and depth weighting was used to reduce the bias towards superficial sources. Noise-normalized MNEs (dynamical statistical parametric maps, dSPMs) were calculated over the whole cortical area to estimate the SNR in each potential source location (Dale et al., 2000). A noise covariance matrix was estimated from the 300-ms prestimulus baseline periods in the data. For group-level visualization, the MNEs of each individual participant were first normalized to the maximum value of that participant and subsequently morphed, with spatial smoothing, to a common (one participant's) brain. MNEs were calculated from 300 ms before the stimulus onset to 1500 ms after it in all stimulus conditions but illustrated here (see Section 3) for an exemplary time window of 700–800 ms at the extreme SERs of +18 and −18 dB.

To confirm that the responses were not contaminated by the decision signal (on the 4th sound), we additionally analysed the responses only to the 2nd and 3rd sounds in a trial. Due to the noisier signals (fewer responses available), the responses were analysed at the sensor level only: The response latencies and amplitudes were measured from the vector sum $\sqrt{\left(\frac{\partial Bz}{\partial x}\right)^2 + \left(\frac{\partial Bz}{\partial y}\right)^2}$ of the MEG channel pair showing the maximum signal. In signal strength comparisons, the vector sums simplify the analysis when the orientation and/or strength of the neural current changes as a function of time, while the source location remains essentially the same.

A repeated-measures analysis of variance (ANOVA) with *Focus of attention* and *SER* as within-subject factors was used for the statistical analysis of main effects on the behavioural responses and *Hemisphere*, *Focus of attention* and *SER* as within-subject factors for cortical source strengths and response durations (based on the ECD modelling). The results were Bonferroni-corrected for multiple comparisons, and Greenhouse–Geisser correction was applied if sphericity assumption was violated.

2.6 | fMRI data: Acquisition and analysis

To minimize a possible masking effect of acoustic scanner noise, the sounds were presented during 1600-ms silent periods between consecutive 2000-ms fMRI scans; the sounds started 400 ms after the beginning of silence. The imaging was performed with a MAGNETOM Skyra 3-T scanner (Siemens Healthcare) with a 20-channel head coil. In each subject, two runs of 327 volumes were acquired with a gradient-echo echo planar imaging sequence (field of view = 200 × 200 mm^2, time of repetition = 3600 ms, time to echo = 30 ms, flip angle = 75°, 36 interleaved slices with a thickness of 3 mm and no gap in between, number of excitations = 1, acquisition matrix size = 64 × 64). Structural MRI images were obtained from each subject with a standard spoiled-gradient-echo sequence after the last functional run.

The functional and anatomical images were analysed with BrainVoyager QX (Brain Innovation, Maastricht, the Netherlands). Preprocessing consisted of slice scan-time correction, linear-trend removal, temporal high-pass filtering (cut-off 5 cycles per run) and 3D motion correction. Spatial smoothing (full width at half maximum [FWHM]: 4 mm) was applied to the fMRI data. Functional slices were co-registered to the anatomical data, and both data sets were normalized to Talairach space (Talairach & Tournoux, 1988).

In the analysis, each stimulus train of four was considered a block. To obtain an overall estimate of the activated brain areas, the fMRI time series were first analysed on a whole-brain, voxel-by-voxel basis with multisubject random-effect general linear models (GLMs) separately for the three 'attention-to-speech' and 'attention-to-environmental sounds' conditions. The data were subsequently analysed with repeated measures ANOVA, with *Focus of attention* and *SER* as within-subjects factors. The obtained cortical maps were then thresholded using a spatial cluster-size threshold method to control for multiple comparisons (Goebel et al., 2006) and to facilitate direct comparison of the activated areas between the present results and our earlier related findings (Renvall et al., 2012). Similarly to the statistical approach of Goebel et al. (2006), a voxel-level threshold was first set to $t_{10} = 3.4$ ($p < .007$, uncorrected; attention to spoken words/environmental sounds relative to the rest). Subsequently, the maps were corrected based on their spatial smoothness and a Monte Carlo simulation (1000 iterations) that estimates the cluster-level false positives and finds the minimum cluster size that yields an overall false-positive rate of $\alpha = .05$. For display, the corrected maps were superimposed on one participant's Talairach-transformed anatomical data.

Finally, to compare the results with those of our previous study (Renvall et al., 2012), we restricted the analysis to non-smoothed data and the volume of interest (VOI) that had showed sensitivity to speech SNR in that earlier study. This region concentrated on the left-hemispheric PT. Voxels within the region were pooled and analysed with the GLM described above. The resulting individual parameter estimates (β values) were then submitted to a 2nd-level (random-effect) analysis using the same ANOVA as described above.

3 | RESULTS

3.1 | Behavioural results outside the MEG/fMRI experiment

Subjects were faster in recognizing repetition of environmental sounds than repetition of speech sounds across the three SERs (mean ± SEM response time 687 ± 24 ms for environmental sounds vs. 930 ± 79 ms for speech, $F(1,7) = 9.6$, $p < .04$; hit rate 97 ± 1% vs. 88 ± 4%, $F(1,7) = 6.3$, $p = .08$; see Figure 2a). The generally longer reaction times to speech sounds likely reflect that the participants needed to hear a longer part of speech than environmental sound before being able to decide on its similarity to the previously presented sound.

For both response time and hit rate, there was a significant interaction between *Focus of attention* and *SER* (reaction time $F(2,14) = 14.1$, $p < .001$; hit rate $F(2,14) = 45.0$, $p < .001$). Subsequent analysis revealed a significant effect of *SER* when subjects were attending to speech (reaction time $F(2,14) = 14.1$; $p < .002$; hit rate $F(1.02,7.2) = 25.9$, Greenhouse–Geisser corrected $p = .002$): the subjects made more mistakes and responded more slowly with decreasing SER. In contrast, when attending to environmental sounds, SER had no significant influence on reaction time or hit rate ($p > .85$). In a post hoc analysis, the reaction times and hit rates were found to differ between *Focus of attention* to speech and environmental sounds at an SER of −18 dB but not at +18 or 0 dB (pair-wise t tests, $p < .01$). In sum, participants' performance was generally better for environmental vs. speech sounds, and this effect was strongest at the lowest SER of −18 dB, due to a drop in performance when attending to speech at this SER.

3.2 | MEG results: Cortical sources

The strongest MEG signals were observed bilaterally over the temporal cortices (Figure 3). In agreement with previous studies (cf. Hari, 1990), responses at ~100 ms (N100m) were explained by ECDs in the left and right supratemporal cortex, around PT, in all subjects, and the same sources also explained well the subsequent sustained responses that corresponded to the N400m responses observed in earlier MEG studies of spoken word processing (Bonte et al., 2006; Helenius et al., 2002; Renvall et al., 2012; Uusvuori et al., 2008). In addition, in the left hemisphere of 6/11 subjects and right hemisphere of 8/11 subjects, the measured activity suggested the existence of another current source, peaking at around 250 ms, with more variable location and current orientation (see Figure 3), also in agreement with earlier

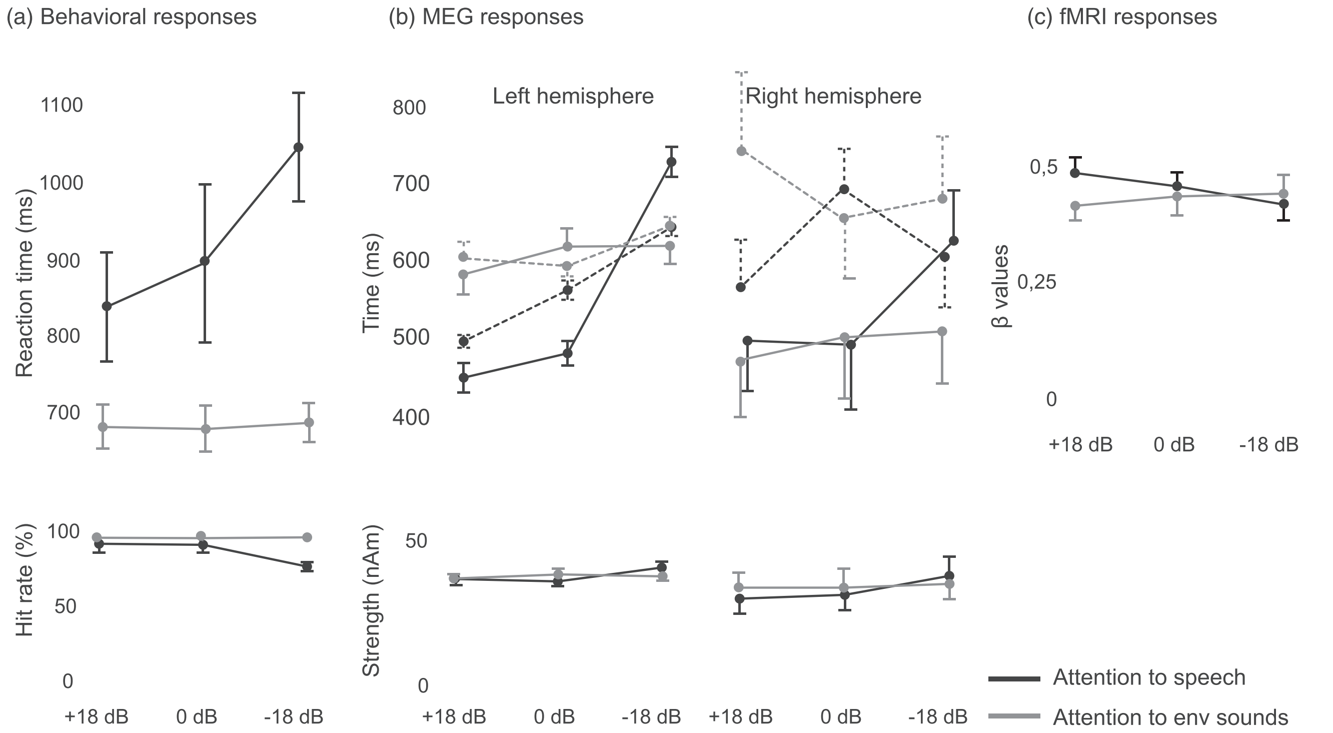

FIGURE 2 Behavioural and functional neuroimaging results. (a) Reaction times (above) and hit rates (below) in the behavioural experiment. (b) Duration (solid line) and peak time (dotted line) of the magnetoencephalography (MEG) sustained responses, above; peak signal strength, below. Left-hemisphere data on the left, right-hemisphere data on the right. Note that the response duration is measured as the difference between the time points of 50% of the maximum activation, on the rising and falling slope of the response. (c) General linear model (GLM) β values within the speech signal-to-noise ratio (SNR)-sensitive region (obtained from Renvall et al., 2012). The vertical bars indicate standard error of the mean (SEM). The asterisks mark significant interactions and differences between attentional conditions (*$p < .05$, **$p < .01$, ***$p < .005$)

observations (Bonte et al., 2006; Renvall et al., 2012; Uusvuori et al., 2008).

The transient N100m responses were of similar magnitude and duration in all stimulus and attentional conditions. The same observation was made for the 250-ms responses (see Figure 3).

3.3 | Effect of SER and *Focus of attention* on the sustained response

Prominent sustained responses, with maximum amplitude of at least four times the standard deviation of the prestimulus noise level, were detected in all subjects in the left hemisphere. This long-lasting sustained activation (measured as the difference between the time points of 50% of the maximum activation, on the rising and falling slopes of the response) differed across conditions (see Figures 2b and 3). There was a significant main effect of *SER* ($F(1,10) = 13.9$, $p < .01$; linear trend for contrasts): The response duration was shorter for higher SER. Furthermore, significant interaction of *SER* and *Focus of attention* was observed ($F(1,10) = 13.5$, $p < .01$; linear trend for contrasts): When attending to speech, the response duration decreased with increasing *SER*. Post hoc *analyses* demonstrated this to be true especially in the left hemisphere (effect of *SER*, $F(2,20) = 10.7$, $p = .002$, see Figures 2b and 3a for all SERs, and Figure 3b for +18- and −18-dB SERs) where the responses decreased in duration at the higher SER in all subjects (range 64–622 ms, mean ± SEM 278 ± 16 ms; see Figure 3a inserts for the interindividual variability of the effect). In contrast, SER did not influence response duration when attending to environmental sounds ($F(2,20) = .11$, $p = .9$).

In the right hemisphere, the responses were more variable between individuals in both *Focus-of-attention* conditions, and post hoc *analyses* revealed no significant interaction between *SER* and *Focus of attention* ($F(1,10) = 2.4$, $p = .30$). However, effect of *SER* remained significant ($F(1,10) = 13.1$, $p < .01$) in the right hemisphere. The peak amplitudes did not differ between *SER* nor *Focus-of-attention* conditions in either hemisphere.

The effect of SER on response duration in the left hemisphere was thus specific to the attended speech sounds, in agreement with the prediction. In line with

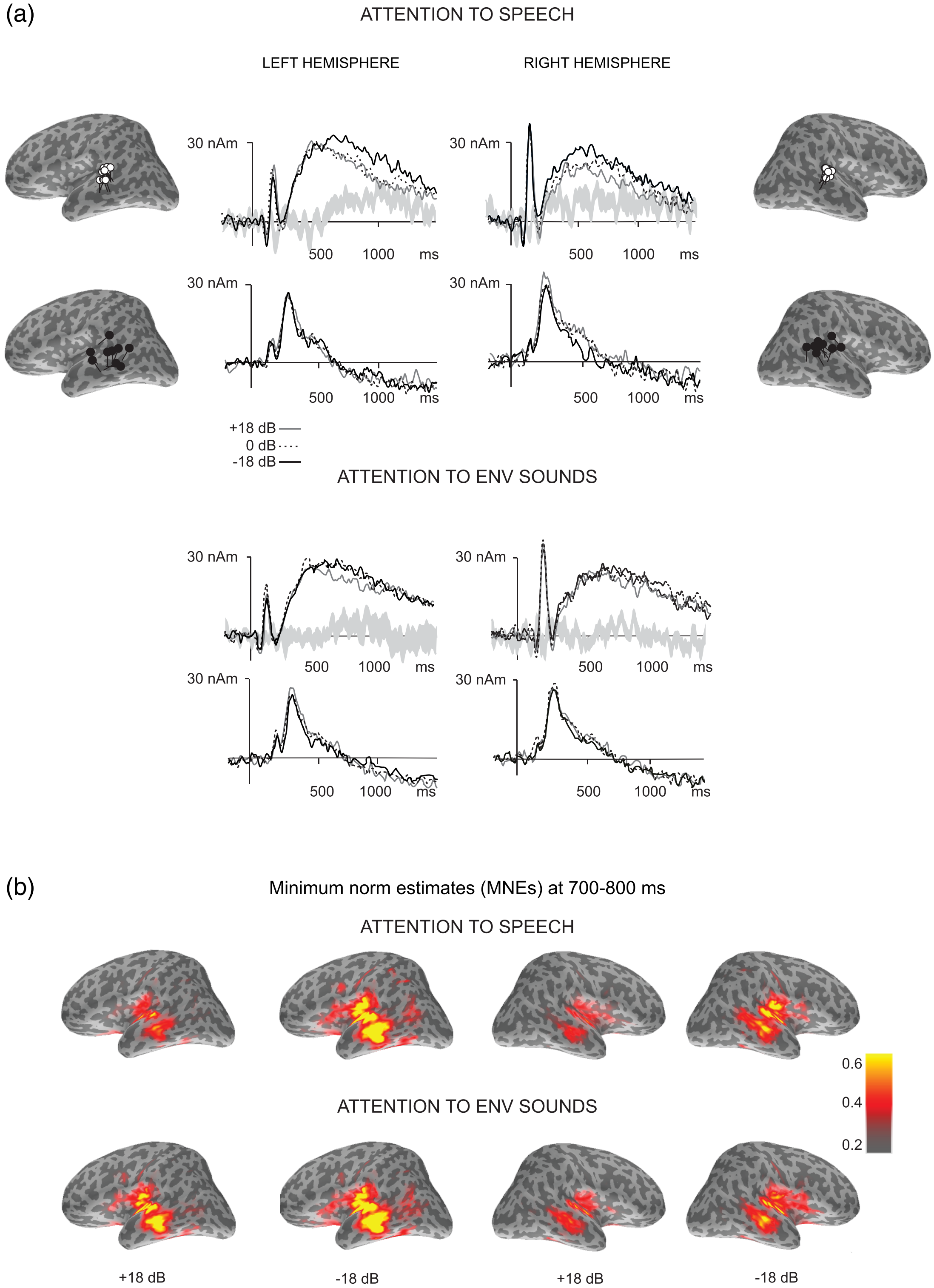

FIGURE 3 Magnetoencephalography (MEG) group-level results for both *Focus-of-attention* conditions. (a) Locations and orientations of individual equivalent current dipoles (ECDs) used to model the 100-ms and sustained responses (white dots, black tails), and 250-ms responses (black dots), superimposed on one subject's cortical surface and the corresponding averaged ECD time courses from −250 to 1500 ms with respect to the stimulus onset at all speech-to-environmental sound intensity ratios (SERs) (−18, 0 and +18 dB). The grey inserts under the ECDs that were used to model the 100-ms and sustained responses depict the difference curve calculated between SERs −18 and +18 dB over individual subjects (average ± standard error of the mean [SEM]). Left hemisphere on the left and right hemisphere on the right. (b) Average minimum norm estimate (MNE) dynamical statistical parametric map (dSPM) distributions, morphed to one subject's brain and normalized, in an exemplary time window of 700–800 ms after stimulus onset for SERs +18 and −18 dB in both *Focus-of-attention* conditions

this result, the response duration in the left hemisphere was shorter at the SERs of +18 and 0 dB when attending to speech than environmental sounds ($p < .05$, pooled across conditions +18 and 0 dB), whereas no such difference was observed between conditions at −18-dB SER ($p = .16$).

In line with the ECD analysis, MNEs (Figure 3b, depicted here at an exemplary time window of 700–800 ms after the stimulus onset) illustrate diminished neuronal activity at SER +18 dB compared with SER −18 dB when the subjects were attending to speech. No such difference related to SER is seen when attending to environmental sounds.

For confirming that the results were not contaminated by the decision signal related to the presentation of the 4th stimulus in a trial, we analysed the responses to only 2nd–3rd stimuli (see Section 2). The results were similar to those obtained for the 2nd–4th sounds: The duration of the sustained MEG response in the left hemisphere decreased significantly from −18- to +18-dB SER when attending to speech (865 ± 30 vs. 541 ± 46 ms, $p < .0003$), whereas no similar change was observed when attending to environmental sounds (820 ± 73 vs. 789 ± 74 ms, $p = .55$). When attending to speech at +18-dB SER, the response duration was significantly shorter than when attending to the environmental part of the same exact stimuli ($p < .005$).

The subjects listened to the exact same stimuli in the two *Focus-of-attention* conditions. However, we additionally controlled for the possibility that acoustical variability within the stimuli would have contributed to how the subjects attended to their speech/non-speech content. We divided the MEG responses on the basis of the closeness of their environmental sound parts to human speech sounds (human non-speech sounds, animal sounds and 'other sounds' that included sounds, e.g., from traffic, tools and household machinery; see Supporting Information). The duration of the sustained MEG response was similar for all stimulus subgroups ($p = .23$), speaking against an effect of acoustical variability on the observed responses.

3.4 | fMRI results

The experimental sounds evoked widespread BOLD activity in bilateral temporal regions, covering Heschl's gyri and PT (see Figure 4a). Further whole-brain (voxel-by-voxel) analysis revealed no significant effect of *Focus of attention* or *SER*. The analysis was subsequently restricted to the VOI that showed sensitivity to speech SNR in our previous fMRI study (Renvall et al., 2012; Figure 4b): left superior temporal cortex in the vicinity of Heschl's gyrus and PT. Within this VOI, the β values from the GLM analysis (Figure 3c) showed a statistically significant interaction of *Focus-of-attention* and *SER* ($F(2,20) = 3.7$, $p = .04$): There was a tendency of β values to decrease with decreasing SER when attending to speech (test for linear trend $F(1,10) = 4.4$, $p = .06$). No such trend was observed when participants attended to environmental sounds ($F(1,10) = .6$, $p = .46$).

4 | DISCUSSION

We investigated cortical processing of speech embedded in short naturalistic auditory scenes where both the relative intensity of speech and environmental sounds, as well as the focus of subjects' attention between the sound excerpts, were varied. The paradigm enabled us to study the cortical mechanisms related to attending to speech vs. environmental sounds while keeping the acoustical stimulation unchanged between attention conditions. When subjects attended to speech sounds within the combined stimuli, activity especially in the left temporal cortex varied with the relative intensity of the speech and environmental sound excerpts: When the speech in the stimuli could be readily recognized (i.e., at higher SER), sustained MEG responses at ~500–1200 ms after stimulus onset were shorter in duration, and BOLD fMRI responses in the left PT were stronger. The behavioural results underscored the effect of SER specifically on speech processing: The higher the SER was, the faster and more correct the subjects were in disentangling the speech sound from the combined stimuli. Remarkably, no similar effects, either neural or behavioural, were detected when the subjects were attending to the environmental sound excerpts within the exact same sounds.

Importantly for the interpretation of our results, the participants listened to the exact same stimuli in both attentional conditions. The results are unlikely to depend on a generally different level of task difficulty between the speech and environmental sound excerpts within the stimuli, as the cortical responses in the two attentional conditions differed most strongly at SERs (+18 and 0 dB) at which the behavioural responses were most comparable. Furthermore, none of our statistical analyses on the acoustical features within the experimental sounds (perceived loudness, power distribution across stimulus frequencies, temporal modulations and correlation between spectral frequency channels) suggested features that could have driven the focus of attention differently between sounds in the two conditions. Indeed, the smallest acoustical differences occurred between the experimental stimuli at the highest SERs of +18- and 0-dB SERs (see Supporting Information), which, in turn,

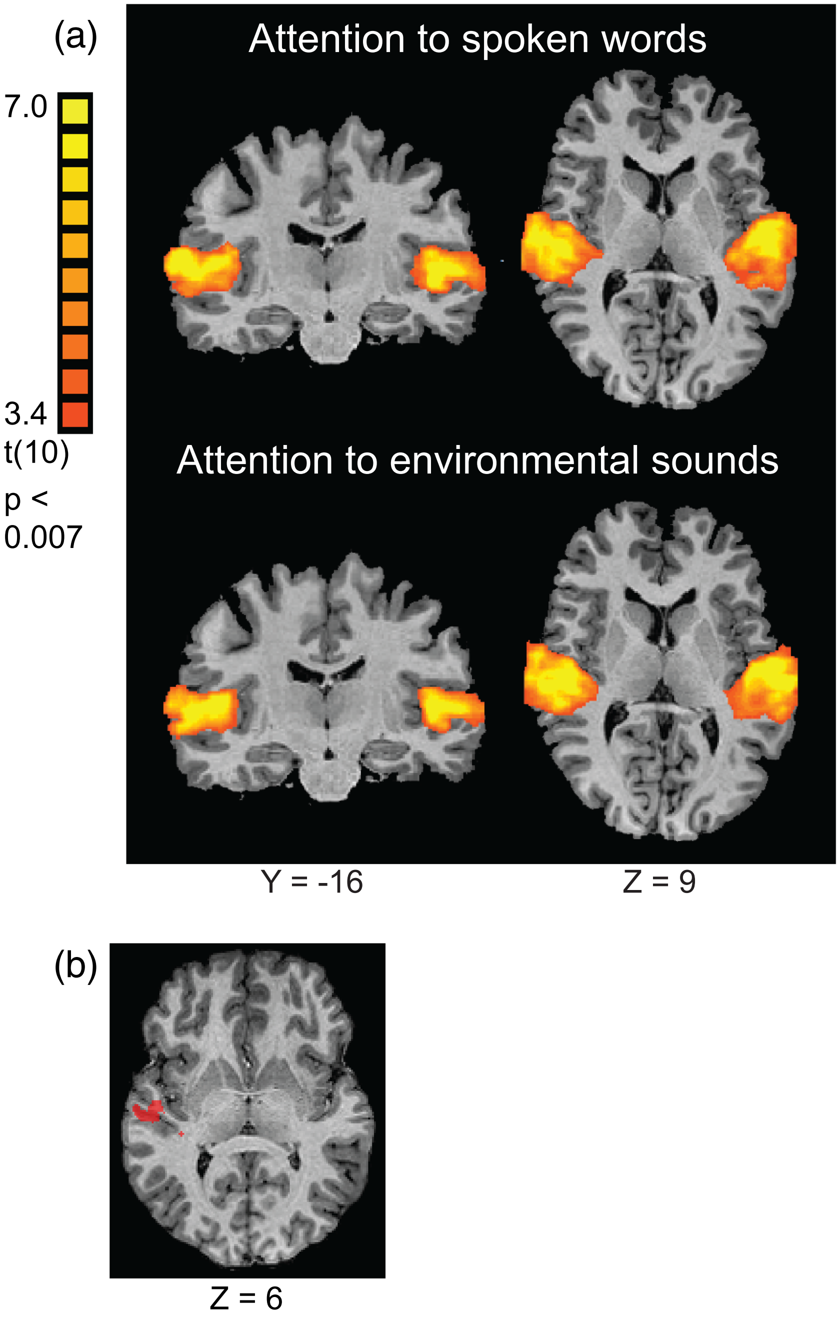

FIGURE 4 Functional magnetic resonance imaging (fMRI) group-level results. (a) Task vs. rest. Brain regions activated when the subjects were attending to speech sound excerpts (top) and environmental sound excerpts (bottom) within the sounds averaged across sound intensity ratios (SERs) compared with rest (whole-brain multisubject random-effect general linear model analyses). (b) Visualization of the speech signal-to-noise ratio (SNR)-sensitive region within the left hemisphere (according to Renvall et al., 2012). The β values within the particular region are presented in Figure 2c

showed the strongest cortical effects. Our interpretation is further supported by our previous study (Renvall et al., 2012) in which similar speech and environmental sounds to present study were used but embedded in increasing level of noise. In that study, the recognition accuracy of environmental sounds got worse with increasing noise, that is, with increasing task difficulty, whereas no similar drop in recognition was observed for speech sounds. Still, the MEG responses to speech sounds showed very similar reactivity as a function of SNR as in the present study: The better the SNR (Renvall et al., 2012) or stronger the SER (present study), the shorter the duration of the sustained MEG response.

We suggest that our results speak for speech specificity in the left temporal cortex in natural listening situations, reflected here as stronger fMRI activity and especially as the markedly shorter-lasting MEG responses for the more salient speech stimuli, to be discussed in the following.

4.1 | Cortical processing of speech differs from processing of other sounds

Cortical processing of speech consists of both bottom-up perceptual processing of the acoustical cues within the speech sounds and of top-down modulation that depends on the speech context, semantic content, listener's expectations and possible visual input (see, e.g., Rönnberg et al., 2008, Davis & Johnsrude, 2007). In noisy auditory conditions, cortical processing of speech is affected by the bandwidth and relative intensity of the noise (Ding & Simon, 2012; Renvall et al., 2012; Seither-Preisler et al., 2003), but in a non-linear manner that possibly supports, for example, perception of prosody regardless of the acoustic background (Ding & Simon, 2012). The relative contribution of top-down processing is likely to be emphasized in natural auditory scenes where speech sounds are intermingled with other sounds of the environment. Thus, our observed attention effects may reflect top-down modulations related to the selection of behaviourally relevant features among the auditory environment.

In primate auditory cortex, it has been demonstrated that early perceptual processes likely depend on strictly hierarchical connections between auditory areas (e.g., Hackett & Kaas, 1998), while broadly distributed top-down connections from, for example, parabelt areas to the primary core area, without corresponding feedforward projections (de la Mothe et al., 2006; Hackett et al., 2014), probably support top-down modulatory activity. Functional, task-dependent changes in the spectrotemporal receptive fields have, indeed, been observed in the primary auditory cortex of ferrets, relating to the animal's ability to adapt to changing auditory demands (for a review, see Fritz et al., 2005). Similarly in humans, speech processing has been suggested to reflect a functional hierarchy, such that the hierarchically early auditory areas are more involved in the acoustical analysis, whereas processing of, for example, speech intelligibility relies more on higher-order auditory areas (Davis & Johnsrude, 2003). Support for task-related dynamical shaping of auditory neural representations in humans comes from electrocorticographic studies (Chang et al., 2010; Mesgarani & Chang, 2012; Zion Golumbic et al., 2013) demonstrating that neuronal activity specifically tracks the attended speech signal among multiple overlapping speech sources.

We have recently shown that MEG responses to speech sounds reflect the sounds' spectrotemporal characteristics in a time-locked manner, whereas similar time locking is not observed for environmental sounds or human non-speech sounds with comparable spectrotemporal modulations (Nora et al., 2020). The results speak for neural activation specifically tracking the speech sound spectrogram, possibly essential for encoding relevant acoustic–phonetic features during speech processing. Such time locking may be especially vulnerable to disturbances caused, for example, by overlapping sound sources, whereas analysis of environmental sounds, processed over longer time chunks, would be more resistant to simultaneous sound sources. In line with this interpretation, the magnitude of phase locking of auditory cortical responses to on-going speech has been demonstrated to decrease with diminishing speech intelligibility (Peelle et al., 2013). Our present non-invasive recordings are in line also with related electrocorticographic findings (Chang et al., 2010; Mesgarani & Chang, 2012; Zion Golumbic et al., 2013) and extend them by suggesting specific processing of speech sounds among other natural sounds, not only during early acoustic encoding (Nora et al., 2020) but also during real-life-like auditory processing, presumably via accentuated top-down modulatory activity suggested by the current results.

4.2 | MEG and fMRI suggest sensitivity to speech saliency

Several earlier fMRI studies have demonstrated speech-specific reactivity, especially in the left temporal lobe (Benson et al., 2001, 2006; Binder et al., 2000; Davis & Johnsrude, 2003; Vouloumanos et al., 2001). For example, areas in the left superior temporal sulcus have been shown to be activated by intelligible speech (Scott et al., 2000), syllables (Liebenthal et al., 2005), vowels (Obleser et al., 2006) and to reflect sensitivity to speech SNR (Renvall et al., 2012). Alternatively, speech perception has been suggested to emerge from integrated activation of areas processing both non-speech and speech sounds (Price et al., 2005). Our results bring together both views and speak for specific sensitivity to speech saliency predominantly in the left supratemporal cortex.

Speech sounds are presumably the most socially crucial auditory signals, and their successful encoding requires mapping of the acoustic cues to the appropriate linguistic categories both rapidly and in a speaker-invariant manner (Liberman et al., 1967; for a review, see Kleinschmidt & Jaeger, 2015). Such a complex task is likely to depend on specificity at different processing levels. To our knowledge, the selective processing of speech and environmental sounds has not been earlier compared under matched acoustic conditions. Attention is known to accentuate the processing of speech sounds in effortful listening conditions, especially in the bilateral temporal areas (Wild et al., 2012). The late timing (>500 ms after stimulus onset) of the observed attention effect and its location in the left PT are in line with earlier studies on auditory selective attention (Hall et al., 2000; Ross et al., 2010). The increased speech recognizability manifested here in MEG as temporal sharpening of the stimulus-locked sustained responses and in fMRI as a stronger BOLD response. The relationship between MEG evoked responses and BOLD fMRI remains poorly understood (but see Hall et al., 2014). Under monotonic auditory stimulation, BOLD responses follow more closely the N100m than later sustained MEG responses (Gutschalk et al., 2010). Similarly, our previous results demonstrated better correlation of BOLD responses with the 100-ms MEG than the sustained MEG responses to speech sounds embedded in noise (Renvall et al., 2012). It is possible that also here the evoked MEG and BOLD responses in the left temporal cortex reflect partly different speech-specific neuronal processes: Whereas the MEG responses may highlight direct tracking of the acoustic content of the speech signal, disturbed by the overlapping noise and accentuated by stimulus-specific attention, the BOLD response may be more related to the actual saliency of the stimulus.

4.3 | Sensitivity to speech is reflected in the response timing regardless of the focus of attention

The sustained MEG responses were generally of shorter duration when the SER was higher, irrespective of whether the subjects were attending to speech or environmental sounds in the stimuli. Analysis of speech sounds has long been considered a rather automatic process (Näätänen et al., 2001, 2007; von Kriegstein et al., 2003), with a mandatory semantic access despite attempts to ignore the sounds (for a review, see, e.g., Bowers et al., 2009). Indeed, when attending to one speech stream among others, the ignored speech is also represented in the neural responses, especially in the hierarchically early auditory areas (Ding & Simon, 2012; Hausfeld, Riecke, Valente, & Formisano, 2018; Puvvada & Simon, 2017; Zion Golumbic et al., 2013). Our finding of the faster buildup of cortical response in the left hemisphere with the increasing SER may reflect such sensitivity to speech regardless of the attentional demands. However, future studies should also address the relative contribution of different task and acoustic demands with respect to the observed stimulus specificity, for example, using tasks that specifically target acoustic vs. semantic aspects of the stimuli.

5 | CONCLUSIONS

Overall, our results provide evidence for SER-related attentional modulation of the auditory cortical activity during natural auditory stimulation, and the effect appears to be specific to attending to speech. The observed modulation of the sensory cortical representations is likely to be related, on the one hand, to time-locked cortical tracking of speech sounds and, on the other hand, to top-down mechanisms for selecting behaviourally relevant objects from the auditory background.

ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland (grant numbers 292552 and 315553 to RS and grant numbers 127401, 277655 and 321460 to HR), Suomen Kulttuurirahasto and Päivikki ja Sakari Sohlbergin säätiö.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

HR and RS designed the study. HR, JS and RT collected the neuroimaging and behavioural data. HR, JS, RT, BS and LR performed the data analyses. HR, LR and RS wrote the manuscript, with contribution from all the authors.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.15504.

DATA AVAILABILITY STATEMENT

The original datasets are not publicly available due to restrictions placed by the Aalto Research Ethics Committee. The data that support the findings of this study are available from the corresponding author with permission of the Aalto Research Ethics Committee.

ORCID

Hanna Renvall https://orcid.org/0000-0001-7589-7826

REFERENCES

Alho, K. (1992). Selective attention in auditory processing as reflected by event-related brain potentials. *Psychophysiology*, *29*, 247–263. https://doi.org/10.1111/j.1469-8986.1992.tb01695.x

Ballas, J. A., & Howard, J. H. Jr. (1987). Interpreting the language of environmental sounds. *Environment and Behavior*, *19*, 91–114. https://doi.org/10.1177/0013916587191005

Benson, R. R., Richardson, M., Whalen, D. H., & Lai, S. (2006). Phonetic processing areas revealed by sinewave speech and acoustically similar nonspeech. *NeuroImage*, *31*, 342–353. https://doi.org/10.1016/j.neuroimage.2005.11.029

Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., & Liberman, A. M. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and Language*, *78*, 364–396. https://doi.org/10.1006/brln.2001.2484

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, *10*, 512–528. https://doi.org/10.1093/cercor/10.5.512

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex*, *16*, 115–123. https://doi.org/10.1093/cercor/bhi091

Bowers, J. S., Davis, C. J., Mattys, S. L., Damian, M. F., & Hanley, D. (2009). The activation of embedded words in spoken word identification is robust but constrained: Evidence from the picture-word interference paradigm. *Journal of Experimental Psychology. Human Perception and Performance*, *35*, 1585–1597. https://doi.org/10.1037/a0015870

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press. https://doi.org/10.7551/mitpress/1486.001.0001

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*, 1428–1432. https://doi.org/10.1038/nn.2641

Dale, A. M., Liu, A. K., Fischl, B. R., Bruckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*, 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Dale, A. M., & Sereno, M. I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, *5*, 162–176. https://doi.org/10.1162/jocn.1993.5.2.162

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience*, *23*, 3423–3431. https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*, 132–147. https://doi.org/10.1016/j.heares.2007.01.014

de la Mothe, L. A., Blumell, S., Kajikawa, Y., & Hackett, T. A. (2006). Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. *The Journal of Comparative Neurology*, *496*, 27–71. https://doi.org/10.1002/cne.20923

Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D'Amico, S., Wilson, S., Bates, E., & Pizzamiglio, L. (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: Evidence from functional magnetic resonance imaging and lesion data. *Journal of Cognitive Neuroscience*, *19*, 799–816. https://doi.org/10.1162/jocn.2007.19.5.799

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 of the United States of America*, *109*, 11854–11859. https://doi.org/10.1073/pnas.1205381109

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, *7*, e1000129. https://doi.org/10.1371/journal.pbio.1000129

Fischl, B., Sereno, M., & Dale, A. (1999). Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*, 195–207. https://doi.org/10.1006/nimg.1998.0396

Fritz, J., Elhilali, M., & Shamma, S. (2005). Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hearing Res*, *206*, 159–176. https://doi.org/10.1016/j.heares.2005.01.015

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From singlesubject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, *27*, 392–401. https://doi.org/10.1002/hbm.20249

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, *25*, 348–353. https://doi.org/10.1016/S0166-2236(02)02191-4

Gutschalk, A., Hämäläinen, M. S., & Melcher, J. R. (2010). BOLD responses in human auditory cortex are more closely related to transient MEG responses than to sustained ones. *Journal of Neurophysiology*, *103*, 2015–2026. https://doi.org/10.1152/jn.01005.2009

Gygi, B., Kidd, G. R., & Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America*, *115*, 1252–1265. https://doi.org/10.1121/1.1635840

Hackett, T. A., de la Mothe, L. A., Camalier, C. R., Falchier, A., Lakatos, P., Kajikawa, Y., & Schroeder, C. E. (2014). Feedforward and feedback projections of caudal belt and parabelt areas of auditory cortex: Refining the hierarchical model. *Frontiers in Neuroscience*, *8*, 72. https://doi.org/10.3389/fnins.2014.00072

Hackett, T. A., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditorycortex in macaque monkeys. *The Journal of Comparative Neurology*, *394*, 475–495. https://doi.org/10.1002/(SICI)1096-9861(19980518)394:4<475::AID-CNE6>3.0.CO;2-Z

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Summerfield, A. Q., Palmer, A. R., Elliott, M. R., & Bowtell, R. W. (2000). Modulation and task effects in auditory processing measured using fMRI. *Human Brain Mapping*, *10*, 107–119. https://doi.org/10.1002/1097-0193(200007)10:3<107::AID-HBM20>3.0.CO;2-8

Hall, E. L., Robson, S. E., Morris, P. G., & Brookes, M. J. (2014). The relationship between MEG and fMRI. *NeuroImage*, *102*, 80–91. https://doi.org/10.1016/j.neuroimage.2013.11.005

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—Theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*, 413–497. https://doi.org/10.1103/RevModPhys.65.413

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*, 35–42. https://doi.org/10.1007/BF02512476

Hansen, P. C., Kringelbach, M. L., & Salmelin, R. (2010). *MEG—An Introduction to Methods*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195307238.001.0001

Hari, R. (1990). The neuromagnetic method in the study of the human auditory cortex. In F. Grandori, M. Hoke, & G. L. Romani (Eds.), *Auditory evoked magnetic fields and electric potentials* (pp. 222–282). Karger.

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

Helenius, P., Salmelin, R., Service, E., Connolly, J., Leinonen, S., & Lyytinen, H. (2002). Cortical activation during spoken-word segmentation in nonreading-impaired and dyslexic adults. *The Journal of Neuroscience*, *22*, 2936–2944. https://doi.org/10.1523/JNEUROSCI.22-07-02936.2002

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience*, *30*, 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148–203. https://doi.org/10.1037/a0038695

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468. https://doi.org/10.1037/0096-1523.21.3.451

Lee, A. K., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111–120. https://doi.org/10.1016/j.heares.2013.06.010

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461. https://doi.org/10.1037/h0020279

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*, 1621–1631. https://doi.org/10.1093/cercor/bhi040

Lin, F.-H., Witzel, T., Ahlfors, S., Stufflebeam, S., Belliveau, J., & Hämäläinen, M. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *NeuroImage*, *31*, 160–171. https://doi.org/10.1016/j.neuroimage.2005.11.054

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

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*, 1006–1010. https://doi.org/10.1126/science.1245994

Näätänen, R., & Michie, P. T. (1979). Early selective-attention effects on the evoked potential: A critical review and reinterpretation. *Biological Psychology*, *8*, 81–136. https://doi.org/10.1016/0301-0511(79)90053-X

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*, 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). "Primitive intelligence" in the auditory cortex. *Trends in Neurosciences*, *24*, 283–288. https://doi.org/10.1016/S0166-2236(00)01790-2

Nora A, Faisal A, Seol J, Renvall H, Formisano E, & Salmelin R (2020). Dynamic time-locking mechanism in the cortical representation of spoken words. ENEURO.0475-19.2020

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., & Rauschecker, J. P. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping*, *27*, 562–571. https://doi.org/10.1002/hbm.20201

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*, 903–911. https://doi.org/10.1038/nn.4021

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

Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II. Effects of attention. *Electroencephalography and Clinical Neurophysiology*, *36*, 191–199. https://doi.org/10.1016/0013-4694(74)90156-4

Price, C., Thierry, G., & Griffiths, T. (2005). Speech-specific auditory processing: Where is it? *Trends in Cognitive Science*, *9*, 271–276.

Puvvada, K. C., & Simon, J. Z. (2017). Cortical representations of speech in a multitalker auditory scene. *The Journal of Neuroscience*, *37*, 9189–9196. https://doi.org/10.1523/JNEUROSCI.0938-17.2017

Renvall, H., Formisano, E., Parviainen, T., Bonte, M., Vihla, M., & Salmelin, R. (2012). Parametric merging of MEG and fMRI reveals spatiotemporal differences in cortical processing of words and environmental sounds in background noise. *Cerebral Cortex*, *22*, 132–143. https://doi.org/10.1093/cercor/bhr095

Renvall, H., Staeren, N., Barz, C. S., Ley, A., & Formisano, E. (2016). Attention modulates the auditory cortical processing of spatial and category cues in naturalistic auditory scenes. *Frontiers in Neuroscience*, *10*, 254. https://doi.org/10.3389/fnins.2016.00254

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*, S99–S105. https://doi.org/10.1080/14992020802301167

Ross, B., Hillyard, S. A., & Picton, T. W. (2010). Temporal dynamics of selective attention during dichotic listening. *Cerebral Cortex*, *20*, 1360–1371. https://doi.org/10.1093/cercor/bhp201

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*, 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Seither-Preisler, A., Krumbholz, K., & Lütkenhöner, B. (2003). Sensitivity of the neuromagnetic N100m deflection to spectral bandwidth: A function of the auditory periphery? *Audiology & Neuro-Otology*, *8*, 322–337. https://doi.org/10.1159/000073517

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: An approach to cerebral imaging*. Thieme Medical Publishers.

Uusvuori, J., Parviainen, T., Inkinen, M., & Salmelin, R. (2008). Spatiotemporal interaction between sound form and meaning during spoken word perception. *Cerebral Cortex*, *18*, 456–466. https://doi.org/10.1093/cercor/bhm076

van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, *33*, 485–508. https://doi.org/10.1016/0028-3932(94)00133-a

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. *The Journal of Neuroscience*, *36*, 1596–1606. https://doi.org/10.1523/JNEUROSCI.1730-15.2016

Vartiainen, J., Parviainen, T., & Salmelin, R. (2009). Spatiotemporal convergence of semantic processing in reading and speech perception. *The Journal of Neuroscience*, *29*, 9271–9280. https://doi.org/10.1523/JNEUROSCI.5860-08.2009

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research. Cognitive Brain Research*, *17*, 48–55. https://doi.org/10.1016/S0926-6410(03)00079-X

Vouloumanos, A., Kiehl, K. A., Werker, J. F., & Liddle, P. F. (2001). Detection of sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and non-speech. *Journal of Cognitive Neuroscience*, *13*, 994–1005. https://doi.org/10.1162/089892901753165890

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *The Journal of Neuroscience*, *32*, 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Yoncheva, Y. N., Zevin, J. D., Maurer, U., & McCandliss, B. D. (2010). Auditory selective attention to speech modulates activity in the visual word form area. *Cerebral Cortex*, *20*, 622–632. https://doi.org/10.1093/cercor/bhp129

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*, 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Renvall, H., Seol, J., Tuominen, R., Sorger, B., Riecke, L., & Salmelin, R. (2021). Selective auditory attention within naturalistic scenes modulates reactivity to speech sounds. *European Journal of Neuroscience*, *54*(10), 7626–7641. https://doi.org/10.1111/ejn.15504