PLOS One

RESEARCH ARTICLE

Perceiving speech from a familiar speaker engages the person identity network

Gaël Cordero[1]*, **Jazmin R. Paredes-Paredes**[1], **Katharina von Kriegstein**[2], **Begoña Díaz**[1]

1 Department of Psychology, Faculty of Medicine and Health Sciences, Universitat Internacional de Catalunya, Barcelona, Spain, **2** Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

* gcordero@uic.es

Citation: Cordero G, Paredes-Paredes JR, von Kriegstein K, Díaz B (2025) Perceiving speech from a familiar speaker engages the person identity network. PLoS One 20(5): e0322927. https://doi.org/10.1371/journal.pone.0322927

Editor: Patrick Bruns, University of Hamburg, GERMANY

Received: November 25, 2024

Accepted: March 31, 2025

Published: May 14, 2025

Data availability statement: The data necessary to replicate the results of this study, as well as the stimuli and scripts that controlled stimuli presentation, are publicly available in Open Science Framework https://osf.io/z2dsu/?view_only=ca5b224056404582abdf4bc0b09d782a

Abstract

Numerous studies show that speaker familiarity influences speech perception. Here, we investigated the brain regions and their changes in functional connectivity involved in the use of person-specific information during speech perception. We employed functional magnetic resonance imaging to study changes in functional connectivity and Blood-Oxygenation-Level-Dependent (BOLD) responses associated with speaker familiarity in human adults while they performed a speech perception task. Twenty-seven right-handed participants performed the speech task before and after being familiarized with the voice and numerous autobiographical details of one of the speakers featured in the task. We found that speech perception from a familiar speaker was associated with BOLD activity changes in regions of the person identity network: the right temporal pole, a voice-sensitive region, and the right supramarginal gyrus, a region sensitive to speaker-specific aspects of speech sound productions. A speech-sensitive region located in the left superior temporal gyrus also exhibited sensitivity to speaker familiarity during speech perception. Lastly, speaker familiarity increased connectivity strength between the right temporal pole and the right superior frontal gyrus, a region associated with verbal working memory. Our findings unveil that speaker familiarity engages the person identity network during speech perception, extending the neural basis of speech processing beyond the canonical language network.

Introduction

Each human voice is acoustically unique, a feature which allows us to recognize familiar speakers, but which adds computational complexity to speech perception, i.e., the process by which speech sounds are decoded from the continuous speech signal and identified. Voice and linguistic information are intertwined in the speech signal to an extent that the acoustic cues that identify speech sounds (i.e.,

Funding: This work was supported by grants from the Ministerio de Ciencia e Innovación (PID2019- 106924GA-I00 and PID2022-137368NB-I00) awarded to BD. GC was supported by a doctoral fellowship from the Universitat Internacional de Catalunya (2021) and a scientific exchange grant awarded by the European Molecular Biology Organization (9778). A grant financed by the Agència de Gestió d'Ajuts Universitaris i de Recerca (2021 SGR 00625) also supported this work. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

phonemes) vary across speakers [1]. For instance, the percept of ambiguous vowels changes in accordance with the acoustic properties of voices [2,3]. Prior experience with a voice increases the robustness of the neural representation of speech and facilitates speech intelligibility [4–6]. Remarkably, the brain can harness voice cues with minimal exposure. In experimental contexts, the effects of voice priors on speech perception and recognition have been observed after just a few minutes or as little as two sentences of exposure to the speaker's voice [7–10]. The use of voice priors in speech perception is in line with proposals of human perception which argue that the central nervous system forms stable percepts of highly variable stimuli, including phonemes, by exploiting its cumulative knowledge [11,12].

Interactions between voice and speech processes are unexpected when considering the different brain regions involved. Voice-sensitive regions (i.e., responsive to speaker information embedded in the speech signal) are primarily right-lateralized and are functional constituents of the person identity network [13,14]. This network includes regions that exhibit unimodal sensitivity, such as the voice-sensitive anterior temporal lobe and superior temporal sulcus, as well as regions that have been proposed to integrate information associated with person identities, such as the supramarginal and angular gyri [13–22]. Speech perception predominantly engages the left superior temporal sulcus and gyrus [23], which in turn are part of the language network. This network also includes the bilateral left inferior and middle frontal gyri, and middle temporal gyrus [23–26]. Previous neuroimaging studies have identified two potential neurofunctional mechanisms which might support the use of voice priors during speech perception. Firstly, studies have reported an increase in interhemispheric functional connectivity between right voice-sensitive regions and left speech-sensitive regions when recognizing speech from multiple speakers as opposed to recognizing speech from a single speaker [27–29]. Secondly, several studies have identified an overlap between the neural substrates of voice perception and recognition and speech perception; areas along the temporal cortices and right temporoparietal junction are sensitive to both voice and phonetic information [5,21,22,29–31]. Most of the studies that have found evidence in favor of either of these two neurofunctional mechanisms investigated how physical properties of the speaker's voice, such as pitch or vocal tract length, modulated brain responses during the performance of speech tasks. To the best of our knowledge, only one study has investigated how the brain exploits voice priors during a speech recognition task. Holmes and Johnsrude (2021) studied the brain responses associated with speech recognition from familiar and unfamiliar speakers in two listening conditions: in the presence of competing speech and in the absence of competing speech. Responses in left posterior temporal regions, not including the primary auditory cortex, displayed greater similarity between listening conditions for the familiar speakers as compared to the unfamiliar ones. Holmes and Johnsrude interpreted these findings as an indication that speech representations are more resistant to competing speech when the target speaker is familiar. Their finding suggests that voice familiarity interacts with goal-driven attention to facilitate speech recognition in noisy environments [5]. In the present study we investigate whether functionally overlapping regions are also engaged when perceiving speech from a familiar speaker in the absence of competing speech, and therefore when attention to voice properties is not task relevant.

We hypothesized that perceiving speech from a familiar speaker, relative to an unfamiliar speaker, would engage the neural mechanisms proposed to support the use of voice-specific information during speech perception: interhemispheric connectivity [27–29] and functionally overlapping regions [5,21,22,29–31]. Twenty-seven adults were included in this functional magnetic resonance imaging (fMRI) study. We measured Blood-Oxygen-Level-Dependent (BOLD) activity during a speech perception task with auditory disyllabic non-words before and after familiarizing participants with one of the speakers featured in the task (Fig 1). An independent functional localizer was employed to define voice-sensitive and speech-sensitive Regions of Interest (ROI), allowing us to test for changes in connectivity and BOLD activity in these specific regions, as observed in previous studies [5,21,22,27–31]. Drawing from the two neurofunctional mechanisms that have been proposed to support the interaction between voice and speech processes, we expected that the voice sensitive region would exhibit an increase in interhemispheric connectivity in association with conducting the speech perception

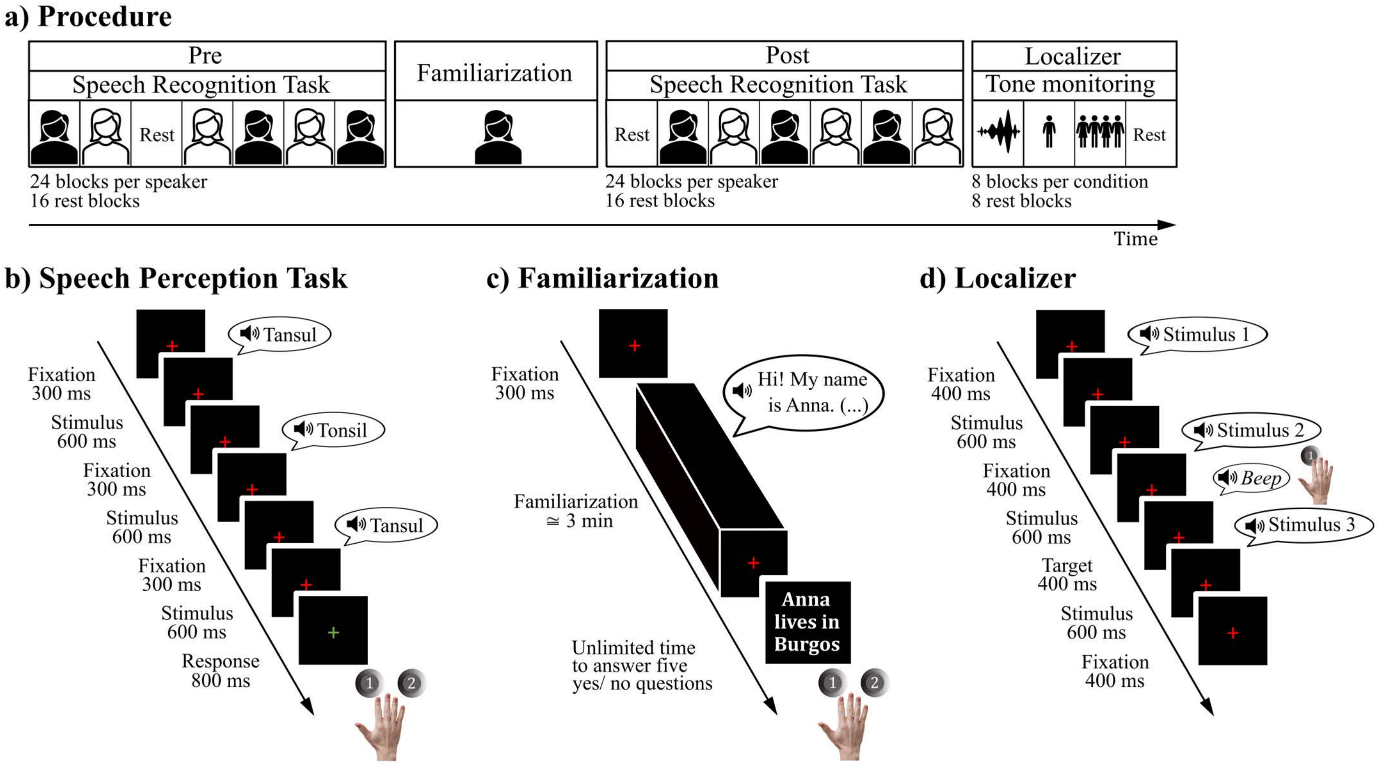

Fig 1. Procedure and trial design of the different tasks that the participants conducted in the scanner. (a) Shows the four runs that composed the procedure. The first three runs (i.e., Pre, Speaker Familiarization, and Post) corresponded to the experiment while the fourth run corresponded to the Independent Functional Localizer for voice and speech-sensitive regions. (b) Depicts the trial structure of the speech perception task that participants conducted in the Pre and Post runs. Each block was composed of four trials. Trials followed an ABX design; participants had to respond if the third stimulus of each trial was a repetition of either the first or second stimulus presented during that trial. (c) Portrays the run in which participants were familiarized with one of the speakers. It began with the presentation of a recording of one of the speakers featured in the speech perception task. This recording was a first-person narration which contained autobiographical information of a fictional character. When this recording finalized, five statements concerning the identity of the fictional character were presented. Participants had to judge whether these affirmations were true or false. Note that, despite the text of the Familiarization appearing in the English language in the figure, the Familiarization was conducted in the native language of the participants, Spanish. (d) Portrays the three conditions that composed the Independent Functional Localizer for voice and speech-sensitive regions: spectrally rotated speech, single-speaker, and multi-speaker. During the Independent Functional Localizer participants were instructed to press a button when a pure tone was presented.

https://doi.org/10.1371/journal.pone.0322927.g001

task with the familiar speaker, relative to the unfamiliar speaker. Furthermore, we predicted that speaker familiarity, relative to speaker unfamiliarity, would be associated with greater responses in regions in which voice and speech processes exhibit neural overlap, such as the right posterior superior temporal gyrus [21] or left posterior middle temporal gyrus [30].

Methods

Participants

A total of 31 graduate and undergraduate students were recruited for this study between the 20th of January 2022 and the 24th of January 2024. Technical errors related to the MRI-compatible headphones and the response box rendered the fMRI data of three participants unsuitable for analysis. Moreover, the behavioral data of one participant suggested that random responses had been provided throughout the speech perception task (Pre accuracy: 48.95%; Post accuracy: 48.43%). These four participants were excluded from all analyses. The final sample consisted of 27 adults (15 female) whose mean age was 21 years, ranging from 18 to 26. This sample size is similar than that of recent studies which have investigated interactions between voice and speech processes by means of fMRI [5,21]. All participants were right-handed according to the Oldfield handedness questionnaire [32]. Participants were native Spanish speakers who did not have substantial musical training, since musicianship has been associated with enhanced voice processing abilities [33]. Substantial musical training was defined as meeting a minimum of 2 of the 3 following criteria; (1) the onset of musical training having occurred prior to the age of 12 years, (2) having partaken in musical training for a minimum of 5 years, and (3) being part of a musical group or ensemble, either currently or in the past (criteria as in [34]). None of the participants had a history of auditory, neurological, psychiatric, language or learning disorders, and they had normal or corrected-to-normal vision. The Ethics Committee of the Medical Faculty and Health Sciences of the Universitat Internacional de Catalunya (Spain) approved the procedures (Study: PSI-2020-05; version: 17/11/20). All participants provided their written informed consent and were monetarily rewarded for their time (10€/ hour).

Stimuli

For the experiment, two sets of 40 disyllabic non-words were created. One set was built with Spanish phonemes and the other with Arab phonemes (each set was designed by a native speaker of the respective language). The purpose of the two language datasets will be reported elsewhere. Two adult (19 and 24 years) female Arab-Spanish native bilinguals recorded the stimuli, amounting to a total of 160 non-words. The two speakers also recorded a text describing autobiographical details of a fictional character in the first-person, henceforth referred to as the Familiarization, which had a duration of 2 minutes 45 seconds (see Appendix 1 for the original text in Spanish and an English translation). Previous studies have shown that directing the attention of listeners to identity characteristics of a speaker leads to voice-specific characteristics being learnt [21,35,36] as well as improving speech intelligibility [9]. Thus, we designed the Familiarization with the intent of familiarizing participants with both the identity and with voice-specific characteristics of the familiar speaker. Five written statements concerning the contents of the Familiarization were also created (see S1 Table in Appendix 1 for the original statements in Spanish and an English translation).

Three native Spanish listeners (including two of the authors) considered that the identity of the speakers was prone to confusion on account of the speakers having relatively similar-sounding voices. To remedy this, two versions of each speaker -a tone higher and a tone lower than the original voice- were created by changing the pitch of the original recordings ("Change Pitch", Audacity v. 3.0.2, Audacity Team). The manipulated versions were employed in the experiment. Manipulation of the pitch has been successfully used to create different perceptual versions of a single voice [29,37,38]. The original, unmodified stimuli were not used.

A third set of 40 disyllabic non-words with Spanish phonemes was assembled for an independent functional localizer of voice and speech-sensitive areas. Four native Spanish speakers, different from the speakers from the experiment, recorded the non-words. As with the stimuli from the experiment, we increased the acoustic diversity of the voices by creating a tone higher and a tone lower version of each voice ("Change Pitch" function, Audacity v. 3.0.2, Audacity Team). All 12 voices were used in the localizer. A copy of the resulting 480 stimuli (40 non-words x 12 voices) were spectrally rotated, a process which preserves the spectral complexity of speech while removing all linguistic information [39], by means of a custom script (Blesser3, Version 3.1., downloaded from: https://www.phon.ucl.ac.uk/resource/software-other.php) in MATLAB (Version R2022b, MathWorks, Inc., Natick, MA USA). By employing different stimuli in the independent functional localizer from those employed in the experimental procedure, we sought to ensure that the results obtained from the analysis of our experimental data are specific to the experimental conditions and not confounded by the characteristics of the stimuli employed during the experimental procedure [40]. Multiple speakers were recorded and a copy of the stimuli was spectrally rotated due to the characteristics of our independent functional localizer protocols, which followed previous studies [39,41]. For further details regarding the independent functional localizer protocols, please see the Intependent functional localizer section).

Stimuli recordings were conducted using an Audio-Technica AT2020 microphone, a Marantz Professional Sound Shield Live vocal reflection filter, and with the software Audacity (Version 3.0.0., Audacity Team) in a sound-attenuated room. A noise reduction procedure was applied to all audio clips with Audacity software. The volume of all audio clips was normalized to 75 dB, a fade was applied to the first and last 50 ms of all clips, and the duration of all non-words was equalized to 600 ms with the PRAAT Vocal Toolkit [42] ("stretch" method available in the "Change Duration" function).

Procedure

The fMRI procedure (Fig 1, panel a) consisted of two parts: the experiment and an independent functional localizer to identify the location of the voice and speech-sensitive regions in our group of participants. To avoid our results being influenced by cognitive effort, we designed all tasks with the intent of high accuracy being easily attainable in all conditions. Participants were familiarized with the experimental procedure before entering the MRI-machine. They performed 8 trials of the experimental task with one of the speakers employed in the independent functional localizer. Both the experiment and the independent functional localizer were performed using a custom Presentation script (Version 22.1, Neurobehavioral Systems, Inc., Berkeley, CA). Audiovisual stimuli presentation was controlled using a VisuaStim Digital system (Resonance Technology Inc., Northridge, CA). Auditory stimuli were delivered through MRI-compatible Serene Sound headphones (Resonance Technology Inc., Northridge, CA) which offer an attenuation of 30 dB of scanner noise. Visual stimuli were presented via MRI-compatible goggles also manufactured by Resonance Technology Inc.

Experiment. The experimental paradigm aimed at evaluating the effect of the factor speaker familiarity (i.e., Familiar (Fam) and Unfamiliar (Unfam)) on speech perception. During two runs (i.e., Pre and Post) participants performed a speech perception task which followed an ABX design. Three non-words were sequentially presented in each trial. The first and second non-words differed, while the third was a repetition of either the first or the second. Participants were tasked with responding if the third non-word was a repetition of the first or of the second non-word presented in the trial. Throughout the experiment, non-words were enunciated by two speakers (i.e., Fam and Unfam), with only one speaker per trial and block (Fig 1, panel b). The two runs, Pre and Post, followed a block design with the same procedures. Thus, the experiment contained four condition blocks: Fam/Pre, Unfam/Pre, Fam/Post, and Unfam/Post. Between the Pre and the Post, the Familiarization of one of the two speakers was presented (Fig 1, panel c). Despite our use of the Fam – Unfam nomenclature in the Pre run, it is crucial to note that during the Pre run, participants had no prior experience with neither speaker. Only in the Post run was a familiar speaker presented, since the Familiarization occurred after the Pre run. However, for the sake of simplicity, we will employ the nomenclature Fam/ Unfam to distinguish speakers regardless of run. In other words, we refer to the speaker with whom participants were familiarized with as the Fam speaker in both the Pre and Post run, despite participants not being familiar with said speaker during the Pre run.

Each run (i.e., Pre and Post) contained 24 blocks per speaker (i.e., Fam and Unfam). Each block was composed of 4 trials. A trial began with the presentation of a red cross. After a 300 ms delay, three auditory non-words were presented with an inter-stimulus interval (ISI) of 300 ms. The first and second non-words were different from one another, and the third non-word was a repetition of either the first or the second. Participants performed an ABX task: they had to indicate which of the two initial non-words was repeated in the third place by pushing with their right index or middle finger one of two buttons on a response box. Participants were instructed to press the button positioned under their index finger to indicate that the repeated non-word was the first non-word presented during the trial, while pressing the button assigned to their middle finger indicated that the repeated non-word was a repetition of the second. Response time started from the beginning of the third non-word until 800 ms after its offset, as indicated by the change of the red fixation cross to green. Each block had a duration of 14 seconds and each trial of 3 and a half seconds. Sixteen rest blocks, which also had a duration of 14 seconds, were also presented. During the rest blocks, no auditory stimulation was presented, and a red fixation cross appeared on the screen. The 4 block conditions (i.e., Pre/Fam, Pre/Unfam, Post/Fam, and Post/Unfam) were pseudo-randomly presented, with the restriction of not allowing the consecutive presentation of more than five blocks of the same condition. The selection of the stimuli that composed each condition block was restricted to avoid repeating non-words between trials of a single block. Each non-word was enunciated by one of the two speakers in each run to diminish perceptual learning effects related to the use of the same speaker - non-word pair in the two runs. The presentation of the rest blocks was also pseudo-random, with two rest blocks never being presented consecutively.

Between the Pre and Post runs, participants were informed that one of the two speakers was going to introduce themselves and that they would later be tested on autobiographical details associated with the identity of the speaker. Participants then heard the Familiarization, a ca. 3 minute speech sample narrated in the first-person that contained extensive autobiographical information, from one of the two speakers employed in the Pre and Post runs. Participants were informed prior to the familiarization that they would be tested on the autobiographic details associated with the identity of the speaker, encouraging them to attend the presentation. The Familiarization was followed by 5 visually presented statements concerning the autobiographical information of the fictional character. Participants had to judge whether these statements were true and provided their answers without the pressure of a time limit by pushing one of two buttons with either their right index or middle finger. Pressing the button positioned under their index finger indicated that the participant judged the affirmation to be true, while pressing the middle finger button indicated that they considered the affirmation to be false. The speaker that acted as the familiar speaker and the version (i.e., high or low pitch) was counterbalanced across participants to ensure that potential differences between the familiar and unfamiliar speakers were not stimuli driven. The unfamiliar speaker had the opposite pitch dimension (either high or low) than the familiar speaker to ensure discriminability. The duration of the experimental procedure was approximately 35 minutes (Pre and Post = 15 minutes each; Familiarization = circa 5 minutes).

Independent functional localizer. The independent functional localizer for voice and speech-sensitive regions was administered after the experiment (Fig 1, panel a). The localizer followed a block design with three different conditions: single-speaker, multi-speaker, and spectrally rotated speech (Fig 1, panel d). Non-words were presented in the single-speaker and multi-speaker condition blocks, while spectrally rotated speech was presented in the equally named condition blocks. Throughout the localizer, participants conducted a pure-tone detection task.

The independent functional localizer was composed of 8 blocks per condition. All condition blocks had a duration of 14 seconds and consisted in the presentation of 14 auditory stimuli with an ISI of 400 ms while a red fixation cross was presented. The design of the three condition blocks is based on previous studies [39,41]. In the single-speaker blocks, participants heard 14 non-words enunciated by a single voice. The voice employed in each single-speaker block was randomly selected and the non-words were pseudo-randomly determined to avoid repeating the same non-word more than once per block. In the multi-speaker blocks, participants listened to one non-word enunciated by 14 voices. For each block, the non-word was selected randomly, and the voices were pseudo-randomly ordered to avoid consecutive presentations of the same voice. In the spectrally rotated speech blocks, participants listened to 14 spectrally rotated non-words enunciated by a single voice. The speaker was randomly determined for each block and the spectrally rotated non-words were pseudo-randomly selected to avoid within-block repetitions.

Twice per condition block, a 200 ms pure tone (1300 Hz) was presented following a 200 ms delay after the last syllable of a block. The blocks in which the pure tone was presented were pseudo-randomly determined to avoid the presentation of the tone in adjacent blocks. Participants indicated that they had heard the pure tone via button press with their right index finger. Responses were considered correct if delivered up to 1 and a half seconds after the onset of the pure tone. Eight rest blocks of the same duration as the condition blocks, i.e., 14 seconds, were also presented. No auditory stimulus was presented during the rest blocks, solely a red fixation cross. The presentation of condition and rest blocks was pseudo-randomized to avoid the consecutive presentation of two blocks of the same condition or of two rest blocks. The duration of the independent functional localizer was 7m 30s.

Data acquisition. Functional images and structural T1-weighted images were acquired on a Siemens 3T Magnetom Prisma Fit MR scanner with a 20-channel head coil (Siemens Healthcare, Erlangen, Germany) at the Hospital Clínic in Barcelona (Spain). Adjustable padding was placed on both sides of the participants' heads to stabilize head position. For the functional images, we used a Multi-Band Echo-Planar Imaging (MB-EPI) sequence (slice thickness = 2.6 mm, number of slices = 42, order of acquisition = interleaved, multiband acceleration factor = 6, TR = 875 ms, TE = 30 ms, flip angle = 65°, FOV = 250 mm, voxel size = 2.6 mm^3; phase encoding direction: anterior-posterior). We employed a slice tilt of -30° to reduce signal losses in the temporal lobes [43,44], where previous studies have localized voice and speech-sensitive regions [13,14,24–26,45]. A total of 1024 EPI volumes were acquired during both the Pre and the Post runs (192 volumes per condition and 128 rest volumes per run). A total of 512 volumes (128 volumes per condition and 128 rest volumes) were acquired during the independent functional localizer. Additionally, a B0 field-map was acquired before the first functional run (Short TE = 4.92 ms, Long TE = 7.38 ms). Lastly, we acquired the T1-weighted structural images by using a high-resolution GRAPPA EPI sequence (slice thickness = 0.80 mm, number of slices = 208, GRAPPA factor = 2, TR = 2400 ms, TE = 2.22 ms, FOV = 256 mm).

Data analysis

All participants included in the data analysis: (i) achieved a response accuracy above 75% in the speech perception task in both runs of the experimental procedure, (ii) provided correct answers to a minimum of four of the five Familiarization questions, and (iii) provided an answer to a minimum of five of the six presentations of the pure tone during the independent functional localizer. Behavioral response accuracy in the experimental task was analyzed with a generalized linear mixed effects model fitted in R (Version 4.1.0, R Core Team, 2017) with the addition of RStudio (Version 2022.2.2.485, RStudio Team) and the lme4 package [46]. We conducted a trial-by-trial analysis in which response accuracy was modelled as the dependent variable, Run (2 levels: Pre and Post), Speaker (2 levels: Fam and Unfam), and the interaction of Run and Speaker as fixed effects, and participant ID as a random effect with random intercept. We also conducted trial-by-trial analysis of the reaction time data with a linear mixed effects model which had the same structure as the generalized mixed effects model employed to analyze the accuracy data. Behavioral effects were considered significant at $p < .05$. The fMRI data was preprocessed and analyzed with SPM 12 (Wellcome Centre for Human Neuroimaging, London, UK; implemented in MATLAB Version R2022b, MathWorks, Inc., Natick, MA USA) with the addition of the CONN toolbox [47,48].

Preprocessing of the fMRI data. For each participant, the preprocessing pipeline began with realigning and unwarping the fMRI data. During unwarping, the B0 field-map was used for susceptibility distortion correction. Preprocessing continued with the identification of potential outlier scans from the global BOLD signal and the estimation of motion parameters of each subject during image acquisition. Following standard procedures [47,48], acquisitions which exhibited framewise displacement above 0.9 mm or global BOLD signal changes above 5 standard deviation were flagged as potential outliers and were considered as confounding effects during the denoising procedure conducted prior to the functional connectivity analyses. Functional images were then co-registered to the participant's anatomical image and normalized to Montreal Neurological Institute (MNI) standard stereotactic space. Functional data was smoothed with an 8mm full width half maximum Gaussian kernel.

For the functional connectivity analyses, the preprocessed functional data was denoised using the CONN Toolbox. Confounding effects to the BOLD signal (noise components from cerebral white matter and cerebrospinal fluid, estimated motion parameters, identified outlier scans, and constant task effects) were used as temporal covariates and removed from the BOLD functional data by linear regression. The BOLD time series was then bandpass filtered between 0.008 and 0.09 Hz to reduce the effect of low-frequency drifts and high-frequency physiological noise [49].

Functional connectivity analysis. We sought to identify changes in functional connectivity associated with speaker familiarity during speech perception. Hence, our seed region for all connectivity analysis was the voice-sensitive Region of Interest (ROI) (see Independent functional localizer section) which is known to respond to the experimental manipulation conducted here, i.e., voice familiarity [13,16,17,45,50–56]. We used the CONN Toolbox to perform both seed-to-whole brain and ROI-to-ROI analyses using the implemented generalized Psychophysiological Interaction (gPPI) procedure [57,58]. In the seed-to-whole brain analysis, target voxels were all voxels in the brain, except for those contained in the seed region. For the ROI-to-ROI analysis, target voxels were solely those voxels included in the speech-sensitive ROI (see Independent functional localizer section). To conduct gPPI, we first extracted an averaged BOLD time-course of the seed region and used it as a physiological regressor. For the first-level analysis, we generated a PPI regressor for each condition (i.e., Pre/Fam; Pre/Unfam; Post/Fam; Post/Unfam) by calculating the element-by-element product between psychological and physiological regressors. We then computed how strongly the time course of the seed region correlated with the PPI regressor of a target voxel. This pair-wise computation was made for every possible seed-target pair to measure task-dependent changes in functional connectivity for each participant. These results were converted to z-scores using the Fisher's z-transformation before calculating group-level averaged functional connectivity scores. An interaction contrast [(Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam)] was computed to investigate if the functional connectivity of the Voice-Sensitive ROI (VSR) is modulated by speaker familiarity during speech perception. To discard the possibility that differences in functional connectivity were caused by behavioral performance differences, we performed a second analysis with the interaction contrast in which we included as control covariates the accuracy and reaction time associated with the speech perception task. For the accuracy covariate, the accuracy percentage participants attained in each experimental condition were transformed into rationalized arcsine units to increase the data's suitability for statistical analysis [59]. The reaction time covariate was included in milliseconds. The accuracy and reaction time covariates of each participant were both calculated following the interaction contrast computed for the neuroimaging data analysis: (Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam). Moreover, we conducted exploratory analyses to investigate if the behavioral covariates exhibited a significant correlation with modulations in functional connectivity. For completeness, main effect contrasts of Run [(Post/Fam + Post/Unfam) – (Pre/Fam + Pre/Unfam)] and of Speaker [(Post/Fam + Pre/Fam) – (Post/Unfam + Pre/Unfam)] were also conducted.

Activity analysis. The activity analysis aimed to investigate changes in brain responses associated with speaker familiarity during speech perception. Activity analysis was conducted on the preprocessed data prior to denoising. Statistical parametric maps were generated for each participant by modeling the evoked hemodynamic response of each block type separately (i.e., Pre/Fam; Pre/Unfam; Post/Fam; Post/Unfam) as boxcar functions convolved with a synthetic hemodynamic response function using the general linear model approach [60]. Following a similar approach as in the functional connectivity analyses, we conducted an interaction contrast as well as contrasts modelling both main effects (i.e., Run and Speaker) at the first level. At the second level, one-sample t-tests across the first-level contrasts images of all participants were used. To ascertain that the results of our main analysis (i.e., the interaction contrast) were not due to differences in participant's performance in the speech perception task, we conducted a second analysis in which the accuracy and reaction time scores of participants were included as control covariates (see Functional connectivity analysis section for details on the calculation of the behavioral scores included as covariates). As in the connectivity analyses, exploratory analyses were conducted to investigate if the behavioral covariates exhibited a significant correlation with modulations in BOLD activity.

Statistical thresholds. For the connectivity analysis, results were considered significant at a voxel-height (cluster-forming) threshold of $p < .001$ and a cluster size threshold at $p < .05$ corrected for False Discovery Rate (FDR), in accordance with the recommended thresholds for cluster-level inferences based on Random Field Theory detailed in the handbook of the CONN toolbox [47]. The Harvard-Oxford cortical atlas, as implemented in the CONN toolbox, was employed to identify the regions in which significant results were obtained. Activity analyses effects were considered significant if they were present at p < .05 Family Wise Error (FWE) corrected at the peak level and a minimum cluster size of 10 voxels at the whole-brain level, or at p < .05 FWE corrected at the peak level for the ROI and Bonferroni corrected for the two ROIs (p < .025 FWE corrected). Regions in which significant activity results were obtained were labelled using the Neuromorphometrics atlas implemented in SPM12.

Definition of regions of interest (ROIs). We used the independent functional localizer to define group-based ROIs. For each participant, we computed a statistical parametric map by modelling the evoked hemodynamic response for the three conditions separately (single speaker, multi-speaker, and spectrally rotated speaker) as boxcar functions convolved with a synthetic hemodynamic response function using a general linear model approach [60]. To define the VSR we followed an approach similar to that employed by Belin & Zatorre (2003) [41]. We contrasted the BOLD response elicited by the multi-speaker condition (i.e., varied voice information; constant phoneme information) against the single speaker condition (i.e., constant voice information; varied phonetic information) at the first-level and employed a one-sample t-test across the first-level contrast images of all participants at the second-level. To define the Speech-Sensitive ROI (SSR) we employed an approach presented by Scott and colleagues (2000) [39]; the BOLD responses elicited during the single speaker condition (i.e., constant voice information; varied phoneme information) were contrasted against the perception of acoustic stimuli with comparable temporal and spectral complexity as speech (i.e., spectrally rotated speech) at the first-level and a one-sample t-test across the first-level contrast images of all participants at the second-level.

The ROIs were defined as all contiguous voxels responsive at $p < 0.05$ uncorrected located in an anatomical position in line with the literature (S2 and S3 Tables in Appendix 2 for a comparison between the ROIs defined here with clusters reported by previous studies). Anatomical position was determined using the Neuromorphometrics atlas implemented in SPM12. A cluster situated in the right temporal pole which extended into the right anterior superior temporal sulcus was defined as the Voice-Sensitive Region (VSR, MNI coordinates: 42, 14, -23; k = 24 voxels) (S1 Fig, panel A in Appendix 2). The Speech-Sensitive Region (SSR) was defined as a cluster situated in the left posterior superior temporal gyrus (MNI coordinates: -57, -31, 10; k = 33 voxels) (S1 Fig, panel B in Appendix 2). The ROIs were created and exported to the functional image space using the Marsbar Toolbox [61].

Results

Behavioral results

Participants exhibited high accuracy in the Speech Perception Task in all the conditions that composed the experimental procedure (see Table 1). A generalized linear mixed effects model was used to examine the effect that Run (2 levels: Pre and Post), Speaker (2 levels: Fam and Unfam), and the interaction of these two factors had on the probability that participants would deliver a correct response on a trial-by-trial basis. The delivered responses (coded as correct: 1; incorrect or miss: 0) were modelled as the dependent variable, run (Pre and Post), Speaker (Fam and Unfam), and their interaction as fixed effects, and participant ID as a random effect. The estimates of the fixed effects suggested that neither Run, Speaker, nor their interaction had a significant effect on the probability of participants providing a correct response (Run: $\beta = -0.15$, SE = 0.13, z = −1.12, p = .26; Speaker: $\beta = 0$, SE = 0.14, z = 0, p = 1; Run x Speaker: $\beta = -0.027$, SE = −0.19,

Table 1. Performance in the speech perception task.

Condition	Mean	SD	Range
Accuracy (%)			
Pre/Fam	95.23	5.08	78.12–100
Pre/Unfam	95.20	3.76	85.41–100
Post/Fam	95.90	4.66	83.33–100
Post/Unfam	96.05	3.47	88.54–100
Reaction time (ms)			
Pre/Fam	768.88	132.75	410.47–988.85
Pre/Unfam	767.04	139.00	383.34–1033.05
Post/Fam	770.20	106.57	606.17–993.00
Post/Unfam	762.21	110.58	563.13–1000.73

https://doi.org/10.1371/journal.pone.0322927.t001

z = −13, p = .89). Similarly, no significant effect was obtained for Run, Speaker, nor their interaction in the RT model (Run: $\beta = -3.58$, SE = 57.97, t = −0.06, p = .95; Speaker: $\beta = -43.18$, SE = 57.90, t = −0.74, p = 0.45; Run x Speaker: $\beta = 32.96$, SE = 81.95, t = 0.40, p = .68). The homogenous performance across experimental conditions attained by the participants suggests that any modulations in brain functional connectivity or activity associated with perceiving speech from the familiar speaker are unlikely to be caused by differences in cognitive effort between the experimental conditions.

Functional connectivity results

Conducting seed-to-whole brain analysis with the interaction contrast [i.e., (Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam)] revealed a significant effect in the connectivity of the VSR with the right Superior Frontal Gyrus (SFG; see Table 2 and Fig 2; for a comparison between the coordinates of this cluster and those of previous similar studies, see S4 Table in Appendix 4). This result cannot be explained by difficulty differences in the conditions of the speech perception task. Analysis of the behavioral data showed that task difficulty was comparable across conditions (see Behavioral results section). Furthermore, a second analysis in which the accuracy scores and RT data were added as second-level control covariates still showed significant modulations of the functional connectivity between the VSR and the rSFG (t(24) = 5.86, p = .010 FDR corrected). Exploratory analysis conducted to investigate if the behavioral covariables correlated with changes in

Table 2. Functional connectivity results of the seed-to-whole brain analyses.

Contrast [a]	Brain region [b]	Peak MNI coordinates			Size *p*-FDR	*k*
		x	*y*	*z*		
Interaction contrast: (Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam)	Right Superior Frontal Gyrus	24	04	54	.004	76
Main effect of Run: (Post/Fam + Post/Unfam) – (Pre/Fam + Pre/Unfam)	Right Supramarginal Gyrus (posterior division)	38	−34	28	.007	80
	Left Central Opercular Cortex	−50	−6	6	.010	64
	Right Frontal Pole	16	28	38	.017	52
Main effect of Speaker: (Pre/Fam + Post/Fam) – (Pre/Unfam + Post/Unfam)	Left Superior Frontal Gyrus	−18	18	66	.013	77

k, cluster size.

[a] Seed region in all contrasts was an independently localized voice-sensitive region (peak MNI coordinate: 42, 14, -23; right temporal pole).

[b] Cluster labels obtained from the Harvard-Oxford cortical atlas as implemented in the CONN toolbox.

https://doi.org/10.1371/journal.pone.0322927.t002

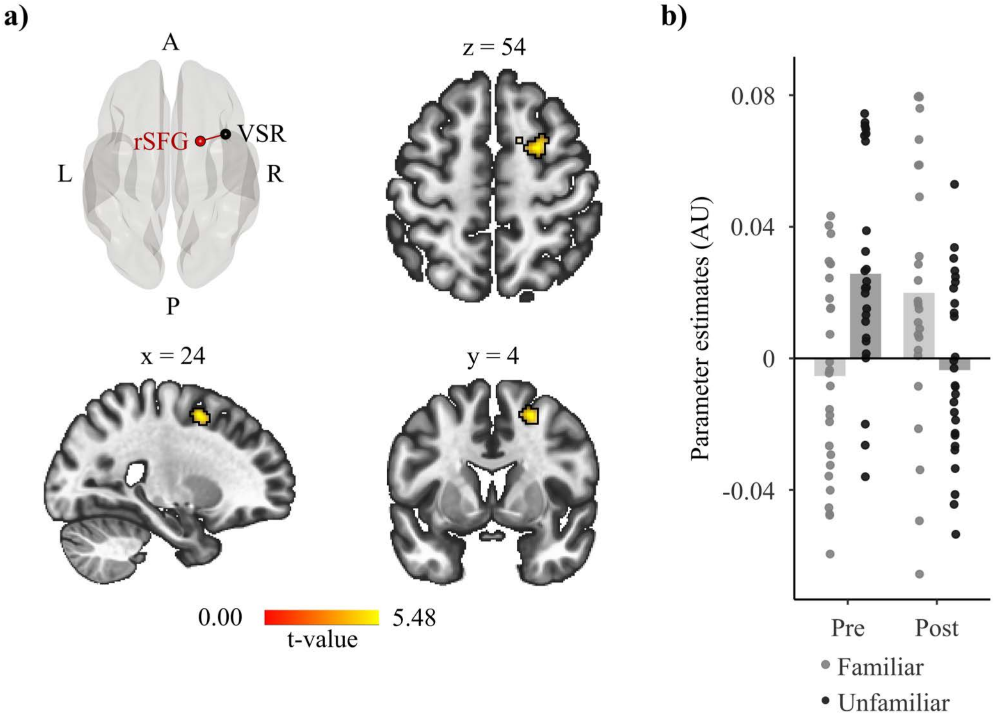

Fig 2. Functional connectivity results of the interaction contrast. (a) Speech perception from a familiar speaker led to a change in functional connectivity strength between the independently localized Voice-Sensitive Region (VSR), employed as seed region, and the right Superior Frontal Gyrus (rSFG; t = 5.48; p = .004 FDR-corrected). Glass brain shows the approximate location of the voxel that exhibited the strongest change in functional connectivity strength of the rSFG and VSR. Functional connectivity results were considered significant at a voxel-height threshold of p < .001 and cluster-wise threshold of p < .05 FDR-corrected. A = anterior. R = right, P = posterior, L = left. (b) Bar-plots (AU: Arbitrary Units) display the functional connectivity strength between the VSR and rSFG in each condition. Participant-specific values for each condition are represented by the scatterplots overlayed with the respective conditions. Values plotted in bar plots were extracted with the REX toolbox [62] as implemented in the CONN toolbox.

https://doi.org/10.1371/journal.pone.0322927.g002

functional connectivity strength in the interaction contrast did not reveal significant effects. The contrast modelling the main effect of Run [i.e., (Post/Fam + Post/Unfam) – (Pre/Fam + Pre/Unfam)] unveiled an increase in functional connectivity between the VSR and both the right Supramarginal Gyrus (rSMG) and the left Central Opercular Cortex. Furthermore, the main effect of Run also revealed a decrease in connectivity strength between the VSR and the right FP (see Table 2 and S3 Fig Appendix 3). The contrast that modelled a main effect of Speaker [i.e., (Post/Fam + Pre/Fam) – (Post/Unfam + Pre/Unfam)] revealed an increase in connectivity strength between the VSR and the left SFG (see Table 2 and S4 Fig in Appendix 3).

The ROI-to-ROI analyses revealed no significant effects between the VSR and the SSR for neither the interaction contrast (T(26) = 1.19; *p*-FDR = .30) nor for the contrasts modelling the main effects of Run (T(26) =.70; *p*-FDR = .66) and Speaker (T(26) =.13; *p*-FDR = .96).

Activity results

Performing the activity analysis at the whole brain level with the interaction contrast [i.e., (Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam)] revealed a significant effect in the rSMG (see Table 3 and Fig 3; for a comparison between these peak coordinates and those of previous studies, see S5 Table in Appendix 4). Inspection of the first-level results revealed that 18 participants exhibited the interaction for Run and Speaker in the cluster that was significant at the second level. The sensitivity exhibited by the rSMG to speaker familiarity cannot be explained by differential degrees of difficulty in

Table 3. Activity results at the whole brain level.

Contrast	Brain region[a]	Peak MNI coordinates			t	Peak *p*-FWE	*k*
		x	*y*	*z*			
Interaction contrast: (Post/Fam – Pre/Fam) – (Post/Unfam – Pre/Unfam)	Right Supramarginal Gyrus	54	−37	31	6.90	.004	19
Main effect of Run: (Post/Fam + Post/Unfam) – (Pre/Fam + Pre/Unfam)	Left Inferior Frontal Gyrus	−45	5	19	6.92	.004	16
	Right Inferior Frontal Gyrus	39	8	25	6.83	.005	20
	Right Superior Parietal Lobule	21	−67	43	6.54	.010	10
Main effect of Speaker: (Pre/Fam + Post/Fam) – (Pre/Unfam + Post/Unfam)						n.s.	

n.s., non-significant.

[a]Regions labelled in accordance with the Neuromorphometrics atlas implemented in SPM12

https://doi.org/10.1371/journal.pone.0322927.t003

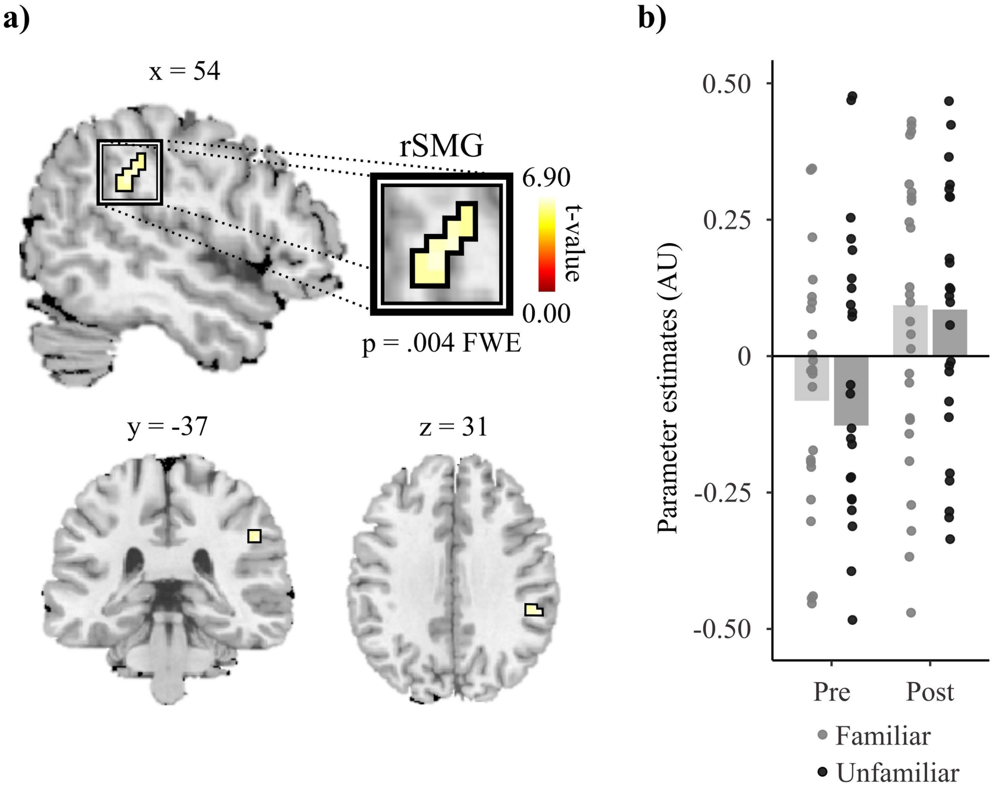

Fig 3. Activity results at the whole brain level. (a) The BOLD activity recorded from the right Supramarginal Gyrus (rSMG) exhibited a significant interaction for speaker familiarity and run (t = 6.90; p = .004 FWE corrected at the whole brain level). (b) Bar plots show the parameter estimates obtained from the peak voxel in each condition. Scatterplots overlayed with the respective conditions show participant-specific values. At the group level, speaker familiarity led to BOLD activity in this region increasing 0.17 AU in Post relative to Pre (i.e., Post/Fam – Pre/Fam)".

https://doi.org/10.1371/journal.pone.0322927.g003

the conditions of the speech perception task; accuracy and reaction time was comparable across conditions (see section 3.1.). Furthermore, a second analysis which included these behavioral measures as second-level control covariates confirmed that the rSMG is sensitive to speaker familiarity (t(24) = 6.86; p = .007 FWE corrected). The main effect analysis of Run [i.e., (Post/Fam + Post/Unfam) – (Pre/Fam + Pre/Unfam)] revealed greater activity in two clusters which bilaterally covered the Inferior Frontal Gyri (IFGs), and a third cluster located in the right SPL (see Table 3 and S7 Fig in Appendix 3).

The analysis modelling the main effect of speaker [i.e., (Post/Fam + Pre/Fam) – (Post/Unfam + Pre/Unfam)] did not reveal any significant modulation in activity.

Employing the two independently localized ROIs for small volume correction in the interaction contrast revealed significant effects (see Fig 4) in both the VSR (MNI: 51 8–23; t = 4.73; p = .001 FWE corrected, and Bonferroni corrected for the

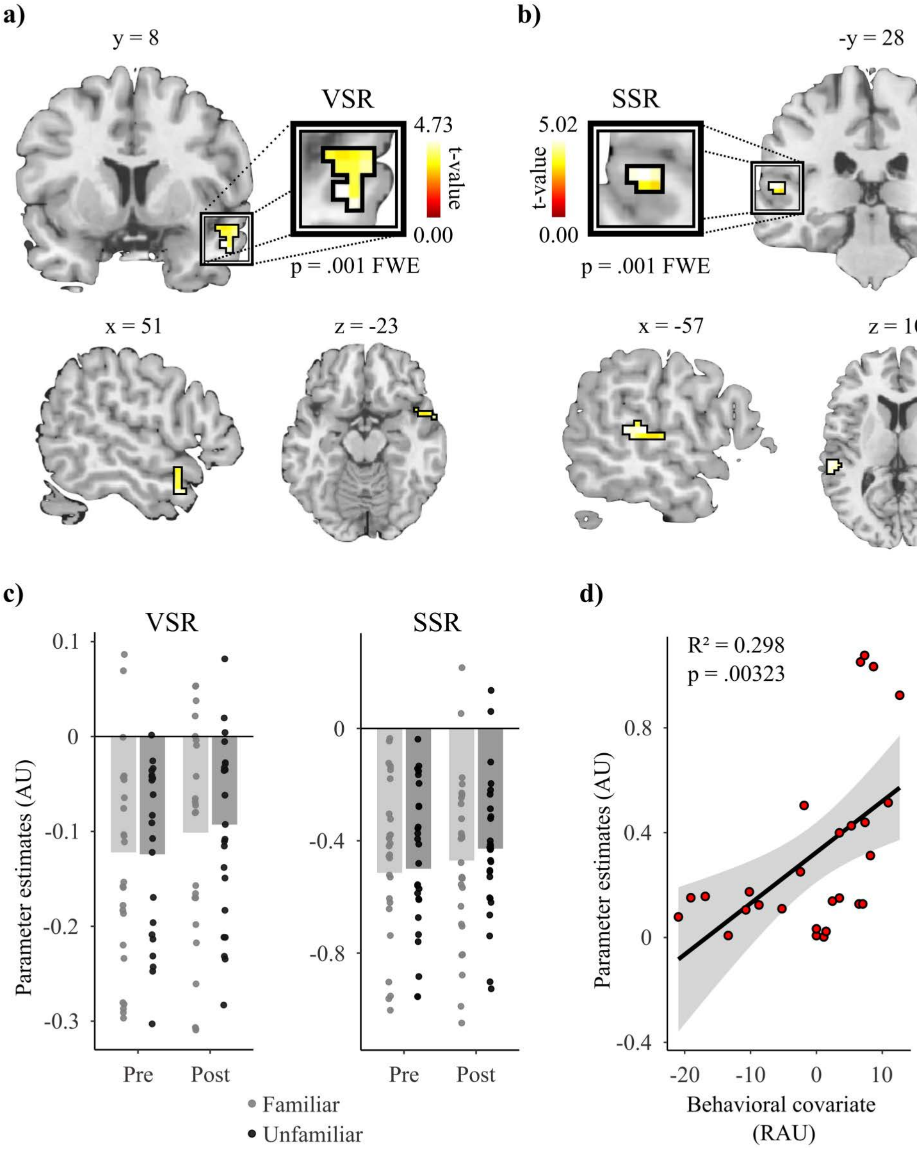

Fig 4. Activity results of the interaction contrast in the independently localized regions of interest. Small volume correction of the interaction contrast revealed significant interactions in both independently localized regions of interest: the Voice-Sensitive Region (VSR), located in the right temporal pole (a), and in the Speech-Sensitive Region (SSR), located in the left posterior superior temporal sulcus (b). Both results were significant (VSR: t = 4.73, p = .001, FWE corrected; SSR: t = 5.02; p = .001, FWE corrected) at p < .05 FWE corrected at the peak level for the ROI and Bonferroni corrected for the two ROIs (i.e., p < .025 FWE corrected). Bar plots (c) show the parameter estimate obtained from the peak voxel of each ROI in each condition at the group level. Scatterplots overlayed with the bar plots show participant-specific values for the respective conditions. At the group level, speaker familiarity (i.e., Post/Fam – Pre/Fam) increased the BOLD activity in the VSR and the SSR by .02 and .04 AU, respectively. (d) Exploratory analysis revealed an association between the accuracy participants attained in the speech recognition task (RAU: Rationalized Arcsine Units) and the interaction contrast parameter estimate extracted from the peal voxel of the SSR. This association was marginally significant after Bonferroni correction for the two ROIs (t = 3.36; p = .025 FWE corrected).

https://doi.org/10.1371/journal.pone.0322927.g004

two ROIs) and in the SSR (MNI: -57–28 10; t = 5.02; p = .001 FWE corrected and Bonferroni corrected for the two ROIs). Examination of the first-level results revealed that 27 and 24 participants exhibited the interaction effect in the VSR and SSR, respectively. The sensitivity of these two regions to speaker familiarity was confirmed by a second analysis in which the accuracy and RT scores of participants were included in the interaction contrast as second-level control covariables: a significant interaction remained present in both ROIs (VSR: t(24) = 4.60; p = .001 FWE corrected and Bonferroni corrected; SSR: t(24) = 5.80, p = .001 FWE corrected and Bonferroni corrected for the two ROIs). Lastly, exploratory analysis revealed a marginally significant association between the accuracy covariate and the interaction effect in the SSR (t = 3.36, p = .025 FWE and Bonferroni corrected).

Discussion

There are three main findings from our study on the influence of speaker familiarity on speech perception. Firstly, speaker familiarity was associated with changes in BOLD activity in regions of the person identity network: the right Temporal Pole (rTP), which we independently localized as a voice sensitive region, and the right Supramarginal Gyrus (rSMG). Secondly, the posterior left Superior Temporal Gyrus (lSTG), a region which we localized as speech-sensitive, also exhibited sensitivity to speaker-familiarity. Thirdly, speaker familiarity led to an increase in connectivity strength between the rTP and the right Superior Frontal Gyrus (rSFG). These findings cannot be explained by difficulty differences in the conditions of the speech perception task. The accuracy and reaction time of participants was comparable across conditions and all results remained significant when analyses included these two behavioral measures as second-level control covariates. Our results align with our hypothesis of speech perception from familiar speakers engaging regions which exhibit a functional overlap between voice and speech processes. However, contrary to our predictions, perceiving speech from a familiar speaker did not lead to an increase in interhemispheric connectivity. Our findings suggest that speech perception from a familiar speaker engages regions of the person identity network, speech-selective regions which are additionally sensitive to speaker familiarity, and a right-lateralized network which, we propose, supports the harnessing of voice priors during speech perception for the encoding and maintenance of speech information in the verbal memory system.

One of our main findings is the sensitivity to speaker familiarity during speech perception exhibited by the rTP and the rSMG. Both regions have been associated with contributing to person identity recognition (rTP: 51,56; rSMG: [15,63]). However, research suggests that the specific functional contributions of these two regions to the process of speaker recognition during speech perception differ. The rTP has been proposed as being the region where person identity information is stored, allowing for the recognition of a speaker from the perception of their voice [13,16,17,41,64]. Here, the task participants conducted while their BOLD activity was recorded did not require speaker recognition. However, research has shown that speaker information is processed during speech perception despite voice properties not being task relevant [65–67]. Therefore, the sensitivity to speaker familiarity exhibited by the rTP in the present study could be attributed to the role of this region in recognizing the voice of the familiarized speaker. Regarding the rSMG, in addition to its implication in person identity recognition [15,63], research suggests that it is also involved in phoneme perception. Vaden and collaborators (2010) found that the BOLD activity of the rSMG is subject to fMRI phoneme repetition-suppression effects, which led them to propose that this region is sensitive to phoneme predictability [68]. Furthermore, evidence suggests that the rSMG supports voice information influencing phoneme perception. Myers & Theodore (2017) found that the rSMG responded to speakers producing phonemes that aligned with the listener's prior experience of the talker's speech. These results were interpreted as indicating the rSMG contributes to the speech perceptual system accommodating the idiosyncratic ways in which different talkers produce their phonemes [21,22]. Building on these proposals, we suggest that the sensitivity to speaker familiarity exhibited by the rSMG in the current study may reflect a similar role: monitoring whether the familiar speaker produces the expected speaker-specific variations. This monitorization would allow the speech perceptual system to detect when expectations differ from percepts. Our proposal aligns with claims of the perceptual system harnessing all available information to optimize speech perception [11,12,69–73].

Our second main finding is the sensitivity to speaker familiarity exhibited by the lSTG, a speech-sensitive region [23–26]. This finding is in line with previous studies that have reported that the lSTG is also sensitive to speaker-specific information [5,29,31]. Holmes and Johnsrude (2021) investigated speech recognition from both familiar and unfamiliar speakers under two listening conditions: one with competing speech and one without. BOLD response patterns in the posterior lSTG exhibited greater similarity between listening conditions when the target speaker was familiar, compared to when the target speaker was unfamiliar. The between condition similarity in response patterns for familiar speakers positively correlated with the benefit participants exhibited in speech comprehension from the familiar speakers. The authors interpreted their findings as an indication that top-down attention mechanisms exhibited greater engagement when processing speech from familiar speakers as compared to unfamiliar speakers in the presence of competing speech. They proposed that this increased engagement led to more robust phoneme representations in the posterior lSTG [5]. Here, we found that the BOLD activity of the posterior lSTG exhibits differential engagement during speech perception from speakers that differ in familiarity. Furthermore, our analysis revealed a marginally significant positive association between the benefit in speech perception associated with speaker familiarity and the BOLD activity interaction in the lSTG. Our findings align with the proposal by Holmes and Johnsrude (2021) of speaker familiarity leading to enhanced phoneme representations in this speech-sensitive region. However, the task design employed in the present study requires us to propose an alternative cognitive mechanism to the top-down attentional modulation proposed by Holmes & Johnsrude (2021). Our speech perception task did not feature competing speech. Therefore, directing attention to the voice properties of speakers was not task-relevant. We suggest that the enhanced phoneme representations associated with speaker familiarity in the lSTG is due to this region also encoding speaker-specific characteristics, as proposed by previous studies [29,31]. This interpretation might seem in conflict with our independent functional localization of the posterior lSTG as a speech-sensitive region. However, what we suggest is that voice information, in addition to phoneme information, is also represented in the posterior lSTG. While left lateralized temporal regions are firmly established as favoring the encodement of linguistic information [23–26], a growing body of research suggests that left lateralized temporal regions are also implicated in vocal identity processing [29,30,74–76]. This redundancy in voice information representation would allow for the robust processing of speaker-specific phonetic variations despite functional disruption of voice-sensitive regions, as a recent study has shown [77].

Our third finding corresponds to an increase in functional connectivity strength in response to speaker familiarity between the independently defined voice-sensitive region and the rSFG. The recruitment of the rSFG might be associated with the involvement of this region in verbal working memory, i.e., the cognitive ability that allows us to retain and mentally manipulate linguistic information [78–81]. The speech perception task, conducted with both the unfamiliar and familiar speaker, required the engagement of verbal working memory; participants had to retain the presented non-words to identify which non-word was later repeated. The observed increase in connectivity strength between the voice-sensitive region and rSFG in association with the familiar speaker suggests that familiarity with the voice of the familiar speaker contributed to retaining in working memory the non-words produced by said speaker during the speech perception task. Previous studies show that familiarity with stimuli improves performance in working memory tasks [82–86]. For instance, familiar faces are easier to remember than unfamiliar faces [84]. While in the present study no behavioral benefit was observed for the familiar speaker, probably due to a ceiling effect observed in all conditions of the speech perception task, previous studies show that working memory automatically recruits prior knowledge to enhance its functioning. The increase in functional connectivity between the voice sensitive area and the SFG might reflect the recruitment of voice priors for use in verbal working memory.

The three main findings of the study partially support our hypothesis. We hypothesized that perceiving speech from a familiar speaker would engage similar neurofunctional mechanisms as those observed in previous studies that have investigated the processing of voice characteristics during speech perception. Two such mechanisms have been proposed: i) interhemispheric functional connectivity between right voice-sensitive regions and left speech-sensitive regions

[27–29] and ii) overlap between the neural substrates supporting both processes [5,21,22,29–31]. Our findings reveal that recognizing speech from a familiar speaker engages the second of these two mechanisms; regions which are sensitive to both voice and speech information (i.e., rSMG, and lSTG). Regarding why we did not find evidence for the other proposed mechanism; studies that have reported increases in interhemispheric connectivity during speech perception featured multi-speaker conditions [27–29], which our design did not include. These differences in design and results suggests that the engagement of interhemispheric connectivity to support speech perception is associated with scenarios which feature multiple, unfamiliar, and rapidly changing speakers. This type of multi-speaker situations are presumably more perceptually demanding relative to the presentation of speakers in isolation, as in the present study.

In addition to the findings involving interactions between speaker conditions before and after the Familiarization, our analyses revealed results associated with a main effect of Run. These results include an interhemispheric increase in connectivity strength between the rTP and the left central operculum, a region involved in auditory and linguistic processing [87], and an increase in BOLD activity in three regions involved in person identity recognition: the bilateral Inferior Frontal Gyri (IFG) and the right Superior Parietal Lobule (SPL) (for a review, see 13). The bilateral IFG are involved in voice-identity recognition [13,88–90] while both the right IFG and SPL are multimodal regions which participate in recognizing person identities that have been recently learnt in laboratory experiments [13,56,91–94]. The observed main effect of Run suggests that participants learnt voice-specific characteristics of both the familiar and the unfamiliar speakers, regardless of the availability of autobiographic information associated with the identity of each speaker. This interpretation is in line with previous neuroimaging studies that showed that greater familiarity with a voice leads to these regions of the person identity network exhibiting greater BOLD responses [52,56,90,95]. During the Pre run, participants heard each speaker enunciate non-words for approximately 3 minutes. These 3 minutes of exposure may have been enough for participants to recognize the speakers featured in the Post run as being the speakers of the Pre run, as previous studies have shown that voices can be accurately recognized after as little as a couple of sentences of exposure [9,96,97]. We suggest that the main effect of Run reflects general familiarity with the two voices featured in the task, whereas the interaction effects reflect richer, speaker-specific knowledge influencing speech perception. In support of the distinct nature of the main effects of run and the interaction effects, behavioral research has shown that the exposure required for mere voice recognition is considerably less than the exposure required for voice-related processes to influence speech perception [9]. However, it should be noted that the results of the main effect of Run could reflect task-learning effects not associated with voice information. It remains for future research to investigate whether initial exposure to a speaker activates regions such as the bilateral IFGs, with greater exposure subsequently engaging regions sensitive to both voice and speech information, as observed in the present study.

The present study reveals the engagement of regions pertaining to the person identity network, beyond voice-specific areas, during the performance of a speech perception task. However, our design does not allow us to fully attribute functional specificity to the reported changes in BOLD activity and functional connectivity as a function of speaker familiarity. While we recorded the BOLD signal of participants as they performed the speech perception task, the findings could reflect voice familiarity or voice-related processes. The absence of an experimental condition in which speech perception was not performed prevents us from unequivocally determining the nature of the connectivity changes captured in the present study. However, a recent high powered (n = 218) fMRI study found that the neural underpinnings that support the general perception of human vocalizations (i.e., speech, laughter, sighing, crying, coughing, onomatopoeias…) include numerous regions that did not exhibit a significant interaction effect in the current study, such as the bilateral inferior frontal and precentral gyri, the amygdala, and the thalamus [98]. Similarly, previous studies that have investigated functional connectivity changes associated with voice-specific processes such as recognition of familiar voices [56,64], voice monitoring [99], and passive exposure to voices [100] have found modulations between regions distinct from the ones we observed, mostly constrained to the temporal lobes and inferior frontal gyri. Jointly, these studies suggest that our interaction results are not solely attributable to voice familiarity or other voice-related processes.

Conclusions

Our findings reveal the engagement of the person identity network during speech perception, extending the neural underpinnings of speech processing beyond the canonical language network. Additionally, we show that one of the neurofunctional mechanisms proposed by previous studies as underpinning the interaction between speech and voice processing, i.e., neural overlap for voice and speech processes, is engaged when perceiving speech from a familiar speaker. These findings contribute to understanding the brain mechanisms that support the use of voice priors during speech perception.

Supporting information

S1 Appendix. Original Familiarization text and translation. File additionally contains the statements employed to ensure participant attended to the Familiarization.
(DOCX)

S2 Appendix. ROI figures and comparison of ROI coordinates with previous studies.
(DOCX)

S3 Appendix. Main effect results' figures and interaction results' figures in representative participants.
(DOCX)

S4 Appendix. Comparison of the coordinates of results obtained at the whole-brain level with previous studies.
(DOCX)

Author contributions

Conceptualization: Gaël Cordero, Begoña Díaz.

Data curation: Gaël Cordero.

Formal analysis: Gaël Cordero, Katharina von Kriegstein, Begoña Díaz.

Funding acquisition: Gaël Cordero, Begoña Díaz.

Investigation: Gaël Cordero, Jazmin R. Paredes-Paredes.

Methodology: Gaël Cordero, Katharina von Kriegstein, Begoña Díaz.

Project administration: Begoña Díaz.

Resources: Begoña Díaz.

Software: Gaël Cordero.

Supervision: Katharina von Kriegstein, Begoña Díaz.

Visualization: Gaël Cordero.

Writing – original draft: Gaël Cordero, Begoña Díaz.

Writing – review & editing: Gaël Cordero, Jazmin R. Paredes-Paredes, Katharina von Kriegstein, Begoña Díaz.

References

1. Peterson GE, Barney HL. Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America. 1952;24(2):175–84. https://doi.org/10.1121/1.1906875
2. Ladefoged P, Broadbent DE. Information Conveyed by Vowels. The Journal of the Acoustical Society of America. 1957;29(1):98–104. https://doi.org/10.1121/1.1908694
3. Magnuson JS, Nusbaum HC, Akahane-Yamada R, Saltzman D. Talker familiarity and the accommodation of talker variability. Atten Percept Psychophys. 2021;83(4):1842–60. https://doi.org/10.3758/s13414-020-02203-y PMID: 33398658

4. Bradlow AR, Pisoni DB. Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. J Acoust Soc Am. 1999;106(4 Pt 1):2074–85. https://doi.org/10.1121/1.427952 PMID: 10530030

5. Holmes E, Johnsrude IS. Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar. Neuroimage. 2021;237:118107. https://doi.org/10.1016/j.neuroimage.2021.118107 PMID: 33933598

6. Johnsrude IS, Mackey A, Hakyemez H, Alexander E, Trang HP, Carlyon RP. Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci. 2013;24(10):1995–2004. https://doi.org/10.1177/0956797613482467 PMID: 23985575

7. Bradlow AR, Bent T. Perceptual adaptation to non-native speech. Cognition. 2008;106(2):707–29. https://doi.org/10.1016/j.cognition.2007.04.005 PMID: 17532315

8. Clarke CM, Garrett MF. Rapid adaptation to foreign-accented English. J Acoust Soc Am. 2004;116(6):3647–58. https://doi.org/10.1121/1.1815131 PMID: 15658715

9. Holmes E, To G, Johnsrude IS. How Long Does It Take for a Voice to Become Familiar? Speech Intelligibility and Voice Recognition Are Differentially Sensitive to Voice Training. Psychol Sci. 2021;32(6):903–15. https://doi.org/10.1177/0956797621991137 PMID: 33979256

10. Trude AM, Brown-Schmidt S. Talker-specific perceptual adaptation during online speech perception. Language and Cognitive Processes. 2012;27(7–8):979–1001. https://doi.org/10.1080/01690965.2011.597153

11. Gilbert CD, Sigman M. Brain states: top-down influences in sensory processing. Neuron. 2007;54(5):677–96. https://doi.org/10.1016/j.neuron.2007.05.019 PMID: 17553419

12. Kleinschmidt DF. Structure in talker variability: How much is there and how much can it help?. Lang Cogn Neurosci. 2019;34(1):43–68. https://doi.org/10.1080/23273798.2018.1500698 PMID: 30619905

13. Blank H, Wieland N, von Kriegstein K. Person recognition and the brain: merging evidence from patients and healthy individuals. Neurosci Biobehav Rev. 2014;47:717–34. https://doi.org/10.1016/j.neubiorev.2014.10.022 PMID: 25451765

14. Maguinness C, Roswandowitz C, von Kriegstein K. Understanding the mechanisms of familiar voice-identity recognition in the human brain. Neuropsychologia. 2018;116(Pt B):179–93. https://doi.org/10.1016/j.neuropsychologia.2018.03.039 PMID: 29614253

15. Awwad Shiekh Hasan B, Valdes-Sosa M, Gross J, Belin P. "Hearing faces and seeing voices": Amodal coding of person identity in the human brain. Sci Rep. 2016;6:37494. https://doi.org/10.1038/srep37494 PMID: 27881866

16. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000;403(6767):309–12. https://doi.org/10.1038/35002078 PMID: 10659849

17. Belin P, Zatorre RJ, Ahad P. Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res. 2002;13(1):17–26. https://doi.org/10.1016/s0926-6410(01)00084-2 PMID: 11867247

18. Carter RM, Huettel SA. A nexus model of the temporal-parietal junction. Trends Cogn Sci. 2013;17(7):328–36. https://doi.org/10.1016/j.tics.2013.05.007 PMID: 23790322

19. Hölig C, Föcker J, Best A, Röder B, Büchel C. Activation in the angular gyrus and in the pSTS is modulated by face primes during voice recognition. Hum Brain Mapp. 2017;38(5):2553–65. https://doi.org/10.1002/hbm.23540 PMID: 28218433

20. Lattner S, Meyer ME, Friederici AD. Voice perception: Sex, pitch, and the right hemisphere. Hum Brain Mapp. 2005;24(1):11–20. https://doi.org/10.1002/hbm.20065 PMID: 15593269

21. Luthra S, Magnuson JS, Myers EB. Right Posterior Temporal Cortex Supports Integration of Phonetic and Talker Information. Neurobiol Lang (Camb). 2023;4(1):145–77. https://doi.org/10.1162/nol_a_00091 PMID: 37229142

22. Myers EB, Theodore RM. Voice-sensitive brain networks encode talker-specific phonetic detail. Brain Lang. 2017;165:33–44. https://doi.org/10.1016/j.bandl.2016.11.001 PMID: 27898342

23. Fedorenko E, Ivanova AA, Regev TI. The language network as a natural kind within the broader landscape of the human brain. Nat Rev Neurosci. 2024;25(5):289–312. https://doi.org/10.1038/s41583-024-00802-4 PMID: 38609551

24. Friederici AD, Gierhan SME. The language network. Curr Opin Neurobiol. 2013;23(2):250–4. https://doi.org/10.1016/j.conb.2012.10.002 PMID: 23146876

25. Hodgson VJ, Lambon Ralph MA, Jackson RL. Multiple dimensions underlying the functional organization of the language network. Neuroimage. 2021;241:118444. https://doi.org/10.1016/j.neuroimage.2021.118444 PMID: 34343627

26. Turkeltaub PE, Coslett HB. Localization of sublexical speech perception components. Brain Lang. 2010;114(1):1–15. https://doi.org/10.1016/j.bandl.2010.03.008 PMID: 20413149

27. Deng Z, Chandrasekaran B, Wang S, Wong PCM. Training-induced brain activation and functional connectivity differentiate multi-talker and single-talker speech training. Neurobiol Learn Mem. 2018;151:1–9. https://doi.org/10.1016/j.nlm.2018.03.009 PMID: 29535043

28. Kreitewolf J, Gaudrain E, von Kriegstein K. A neural mechanism for recognizing speech spoken by different speakers. Neuroimage. 2014;91:375–85. https://doi.org/10.1016/j.neuroimage.2014.01.005 PMID: 24434677

29. von Kriegstein K, Smith DRR, Patterson RD, Kiebel SJ, Griffiths TD. How the human brain recognizes speech in the context of changing speakers. J Neurosci. 2010;30(2):629–38. https://doi.org/10.1523/JNEUROSCI.2742-09.2010 PMID: 20071527

30. Chandrasekaran B, Chan AHD, Wong PCM. Neural processing of what and who information in speech. J Cogn Neurosci. 2011;23(10):2690–700. https://doi.org/10.1162/jocn.2011.21631 PMID: 21268667

31. Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 2008;322(5903):970–3. https://doi.org/10.1126/science.1164318 PMID: 18988858

32. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971;9(1):97–113. https://doi.org/10.1016/0028-3932(71)90067-4 PMID: 5146491

33. Xie X, Myers E. The impact of musical training and tone language experience on talker identification. J Acoust Soc Am. 2015;137(1):419–32. https://doi.org/10.1121/1.4904699 PMID: 25618071

34. Kaganovich N, Kim J, Herring C, Schumaker J, Macpherson M, Weber-Fox C. Musicians show general enhancement of complex sound encoding and better inhibition of irrelevant auditory change in music: an ERP study. Eur J Neurosci. 2013;37(8):1295–307. https://doi.org/10.1111/ejn.12110 PMID: 23301775

35. Goldinger SD. Words and voices: episodic traces in spoken word identification and recognition memory. J Exp Psychol Learn Mem Cogn. 1996;22(5):1166–83. https://doi.org/10.1037//0278-7393.22.5.1166 PMID: 8926483

36. Theodore RM, Blumstein SE, Luthra S. Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. Atten Percept Psychophys. 2015;77(5):1674–84. https://doi.org/10.3758/s13414-015-0854-0 PMID: 25824889

37. Díaz B, Hintz F, Kiebel SJ, von Kriegstein K. Dysfunction of the auditory thalamus in developmental dyslexia. Proc Natl Acad Sci U S A. 2012;109(34):13841–6. https://doi.org/10.1073/pnas.1119828109 PMID: 22869724

38. Kreitewolf J, Friederici AD, von Kriegstein K. Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. Neuroimage. 2014;102 Pt 2:332–44. https://doi.org/10.1016/j.neuroimage.2014.07.038 PMID: 25087482

39. Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 2000;123 Pt 12(Pt 12):2400–6. https://doi.org/10.1093/brain/123.12.2400 PMID: 11099443

40. Saxe R, Brett M, Kanwisher N. Divide and conquer: a defense of functional localizers. Neuroimage. 2006;30(4):1088–96; discussion 1097-9. https://doi.org/10.1016/j.neuroimage.2005.12.062 PMID: 16635578

41. Belin P, Zatorre RJ. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 2003;14(16):2105–9. https://doi.org/10.1097/00001756-200311140-00019 PMID: 14600506

42. Corretge R. Praat Vocal Toolkit. 2012. Available: http://www.praatvocaltoolkit.com

43. Devlin JT, Russell RP, Davis MH, Price CJ, Wilson J, Moss HE, et al. Susceptibility-induced loss of signal: comparing PET and fMRI on a semantic task. Neuroimage. 2000;11(6 Pt 1):589–600. https://doi.org/10.1006/nimg.2000.0595 PMID: 10860788

44. Weiskopf N, Hutton C, Josephs O, Deichmann R. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage. 2006;33(2):493–504. https://doi.org/10.1016/j.neuroimage.2006.07.029 PMID: 16959495

45. Stevenage SV. Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. Neuropsychologia. 2018;116(Pt B):162–78. https://doi.org/10.1016/j.neuropsychologia.2017.07.005 PMID: 28694095

46. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Usinglme4. J Stat Soft. 2015;67(1). https://doi.org/10.18637/jss.v067.i01

47. Nieto-Castanon A. Handbook of functional connectivity Magnetic Resonance Imaging methods in CONN. 2020. Available: https://www.researchgate.net/publication/339460691_Handbook_of_functional_connectivity_Magnetic_Resonance_Imaging_methods_in_CONN.

48. Whitfield-Gabrieli S, Nieto-Castanon A. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect. 2012;2(3):125–41. https://doi.org/10.1089/brain.2012.0073 PMID: 22642651

49. Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med. 1995;34(4):537–41. https://doi.org/10.1002/mrm.1910340409 PMID: 8524021

50. Andics A, McQueen JM, Petersson KM, Gál V, Rudas G, Vidnyánszky Z. Neural mechanisms for voice recognition. Neuroimage. 2010;52(4):1528–40. https://doi.org/10.1016/j.neuroimage.2010.05.048 PMID: 20553895

51. Imaizumi S, Mori K, Kiritani S, Kawashima R, Sugiura M, Fukuda H, et al. Vocal identification of speaker and emotion activates different brain regions. Neuroreport. 1997;8(12):2809–12. https://doi.org/10.1097/00001756-199708180-00031 PMID: 9295122

52. Latinus M, Crabbe F, Belin P. Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex. 2011;21(12):2820–8. https://doi.org/10.1093/cercor/bhr077 PMID: 21531779

53. Nakamura K, Kawashima R, Sugiura M, Kato T, Nakamura A, Hatano K, et al. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia. 2001;39(10):1047–54. https://doi.org/10.1016/s0028-3932(01)00037-9 PMID: 11440757

54. Pisoni A, Sperandeo PR, Romero Lauro LJ, Papagno C. The Role of the Left and Right Anterior Temporal Poles in People Naming and Recognition. Neuroscience. 2020;440:175–85. https://doi.org/10.1016/j.neuroscience.2020.05.040 PMID: 32497758

55. von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res. 2003;17(1):48–55. https://doi.org/10.1016/s0926-6410(03)00079-x PMID: 12763191

56. von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L. Interaction of face and voice areas during speaker recognition. J Cogn Neurosci. 2005;17(3):367–76. https://doi.org/10.1162/0898929053279577 PMID: 15813998

57. Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage. 1997;6(3):218–29. https://doi.org/10.1006/nimg.1997.0291 PMID: 9344826

58. McLaren DG, Ries ML, Xu G, Johnson SC. A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage. 2012;61(4):1277–86. https://doi.org/10.1016/j.neuroimage.2012.03.068 PMID: 22484411

59. Studebaker GA. A "rationalized" arcsine transform. J Speech Hear Res. 1985;28(3):455–62. https://doi.org/10.1044/jshr.2803.455 PMID: 4046587

60. Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD. Statistical Parametric Mapping The Analysis of Functional Brain Images. Academic Press/ Elsevier: Amsterdam; 2007.

61. Brett M, Anton JL, Valabregue R, Poline JB. Region of Interest Analysis Using an SPM Toolbox. NeuroImage. 2002;16(2):769–1198. https://doi.org/10.1016/s1053-8119(02)90013-3

62. Duff EP, Cunnington R, Egan GF. REX: response exploration for neuroimaging datasets. Neuroinformatics. 2007;5(4):223–34. https://doi.org/10.1007/s12021-007-9001-y PMID: 17985253

63. Bestelmeyer PEG, Mühl C. Neural dissociation of the acoustic and cognitive representation of voice identity. Neuroimage. 2022;263:119647. https://doi.org/10.1016/j.neuroimage.2022.119647 PMID: 36162634

64. Kriegstein KV, Giraud A-L. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage. 2004;22(2):948–55. https://doi.org/10.1016/j.neuroimage.2004.02.020 PMID: 15193626

65. Grey S, van Hell JG. Foreign-accented speaker identity affects neural correlates of language comprehension. Journal of Neurolinguistics. 2017;42:93–108. https://doi.org/10.1016/j.jneuroling.2016.12.001

66. Ma Y, Yu K, Yin S, Li L, Li P, Wang R. Attention Modulates the Role of Speakers' Voice Identity and Linguistic Information in Spoken Word Processing: Evidence From Event-Related Potentials. J Speech Lang Hear Res. 2023;66(5):1678–93. https://doi.org/10.1044/2023_JSLHR-22-00420 PMID: 37071787

67. Van Berkum JJA, van den Brink D, Tesink CMJY, Kos M, Hagoort P. The neural integration of speaker and message. J Cogn Neurosci. 2008;20(4):580–91. https://doi.org/10.1162/jocn.2008.20054 PMID: 18052777

68. Vaden KI Jr, Muftuler LT, Hickok G. Phonological repetition-suppression in bilateral superior temporal sulci. Neuroimage. 2010;49(1):1018–23. https://doi.org/10.1016/j.neuroimage.2009.07.063 PMID: 19651222

69. Blank H, Davis MH. Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS Biol. 2016;14(11):e1002577. https://doi.org/10.1371/journal.pbio.1002577 PMID: 27846209

70. Cope TE, Sohoglu E, Sedley W, Patterson K, Jones PS, Wiggins J, et al. Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat Commun. 2017;8(1):2154. https://doi.org/10.1038/s41467-017-01958-7 PMID: 29255275

71. Eisenhauer S, Fiebach CJ, Gagl B. Context-Based Facilitation in Visual Word Recognition: Evidence for Visual and Lexical But Not Pre-Lexical Contributions. eNeuro. 2019;6(2):ENEURO.0321-18.2019. https://doi.org/10.1523/ENEURO.0321-18.2019 PMID: 31072907

72. MILLER GA, HEISE GA, LICHTEN W. The intelligibility of speech as a function of the context of the test materials. J Exp Psychol. 1951;41(5):329–35. https://doi.org/10.1037/h0062491 PMID: 14861384

73. van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A. 2005;102(4):1181–6. https://doi.org/10.1073/pnas.0408949102 PMID: 15647358

74. Perrachione TK, Pierrehumbert JB, Wong PCM. Differential neural contributions to native- and foreign-language talker identification. J Exp Psychol Hum Percept Perform. 2009;35(6):1950–60. https://doi.org/10.1037/a0015869 PMID: 19968445

75. Roswandowitz C, Kappes C, Obrig H, von Kriegstein K. Obligatory and facultative brain regions for voice-identity recognition. Brain. 2018;141(1):234–47. https://doi.org/10.1093/brain/awx313 PMID: 29228111

76. Salvata C, Blumstein SE, Myers EB. Speaker Invariance for Phonetic Information: an fMRI Investigation. Lang Cogn Process. 2012;27(2):210–30. https://doi.org/10.1080/01690965.2011.594372 PMID: 23264714

77. Luthra S, Mechtenberg H, Giorio C, Theodore RM, Magnuson JS, Myers EB. Using TMS to evaluate a causal role for right posterior temporal cortex in talker-specific phonetic processing. Brain Lang. 2023;240:105264. https://doi.org/10.1016/j.bandl.2023.105264 PMID: 37087863

78. Emch M, von Bastian CC, Koch K. Neural Correlates of Verbal Working Memory: An fMRI Meta-Analysis. Front Hum Neurosci. 2019;13:180. https://doi.org/10.3389/fnhum.2019.00180 PMID: 31244625

79. Marvel CL, Desmond JE. The contributions of cerebro-cerebellar circuitry to executive verbal working memory. Cortex. 2010;46(7):880–95. https://doi.org/10.1016/j.cortex.2009.08.017 PMID: 19811779

80. Strand F, Forssberg H, Klingberg T, Norrelgen F. Phonological working memory with auditory presentation of pseudo-words -- an event related fMRI Study. Brain Res. 2008;1212:48–54. https://doi.org/10.1016/j.brainres.2008.02.097 PMID: 18442810

81. Sweet LH, Paskavitz JF, Haley AP, Gunstad JJ, Mulligan RC, Nyalakanti PK, et al. Imaging phonological similarity effects on verbal working memory. Neuropsychologia. 2008;46(4):1114–23. https://doi.org/10.1016/j.neuropsychologia.2007.10.022 PMID: 18155074

82. Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychol Sci. 2004;15(2):106–11. https://doi.org/10.1111/j.0963-7214.2004.01502006.x PMID: 14738517

83. Brady TF, Konkle T, Alvarez GA. Compression in visual working memory: using statistical regularities to form more efficient memory representations. J Exp Psychol Gen. 2009;138(4):487–502. https://doi.org/10.1037/a0016797 PMID: 19883132

84. Jackson MC, Raymond JE. Familiarity enhances visual working memory for faces. J Exp Psychol Hum Percept Perform. 2008;34(3):556–68. https://doi.org/10.1037/0096-1523.34.3.556 PMID: 18505323

85. O'Donnell RE, Clement A, Brockmole JR. Semantic and functional relationships among objects increase the capacity of visual working memory. J Exp Psychol Learn Mem Cogn. 2018;44(7):1151–8. https://doi.org/10.1037/xlm0000508 PMID: 29648871

86. Starr A, Srinivasan M, Bunge SA. Semantic knowledge influences visual working memory in adults and children. PLoS One. 2020;15(11):e0241110. https://doi.org/10.1371/journal.pone.0241110 PMID: 33175852

87. Mălîia M-D, Donos C, Barborica A, Popa I, Ciurea J, Cinatti S, et al. Functional mapping and effective connectivity of the human operculum. Cortex. 2018;109:303–21. https://doi.org/10.1016/j.cortex.2018.08.024 PMID: 30414541

88. Rämä P, Poremba A, Sala JB, Yee L, Malloy M, Mishkin M, et al. Dissociable functional cortical topographies for working memory maintenance of voice identity and location. Cereb Cortex. 2004;14(7):768–80. https://doi.org/10.1093/cercor/bhh037 PMID: 15084491

89. Relander K, Rämä P. Separate neural processes for retrieval of voice identity and word content in working memory. Brain Res. 2009;1252:143–51. https://doi.org/10.1016/j.brainres.2008.11.050 PMID: 19063872

90. Shah NJ, Marshall JC, Zafiris O, Schwab A, Zilles K, Markowitsch HJ, et al. The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. Brain. 2001;124(Pt 4):804–15. https://doi.org/10.1093/brain/124.4.804 PMID: 11287379

91. Andreasen NC, O'Leary DS, Arndt S, Cizadlo T, Hurtig R, Rezai K, et al. Neural substrates of facial recognition. J Neuropsychiatry Clin Neurosci. 1996;8(2):139–46. https://doi.org/10.1176/jnp.8.2.139 PMID: 9081548

92. Haxby JV, Ungerleider LG, Horwitz B, Maisog JM, Rapoport SI, Grady CL. Face encoding and recognition in the human brain. Proc Natl Acad Sci U S A. 1996;93(2):922–7. https://doi.org/10.1073/pnas.93.2.922 PMID: 8570661

93. Leveroni CL, Seidenberg M, Mayer AR, Mead LA, Binder JR, Rao SM. Neural systems underlying the recognition of familiar and newly learned faces. J Neurosci. 2000;20(2):878–86. https://doi.org/10.1523/JNEUROSCI.20-02-00878.2000 PMID: 10632617

94. Maurage P, Joassin F, Pesenti M, Grandin C, Heeren A, Philippot P, et al. The neural network sustaining crossmodal integration is impaired in alcohol-dependence: an fMRI study. Cortex. 2013;49(6):1610–26. https://doi.org/10.1016/j.cortex.2012.04.012 PMID: 22658706

95. Bethmann A, Scheich H, Brechmann A. The temporal lobes differentiate between the voices of famous and unknown people: an event-related fMRI study on speaker recognition. PLoS One. 2012;7(10):e47626. https://doi.org/10.1371/journal.pone.0047626 PMID: 23112826

96. Zäske R, Awwad Shiekh Hasan B, Belin P. It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. Cortex. 2017;94: 100–12. https://doi.org/10.1016/j.cortex.2017.06.005

97. Zäske R, Volberg G, Kovács G, Schweinberger SR. Electrophysiological correlates of voice learning and recognition. J Neurosci. 2014;34(33):10821–31. https://doi.org/10.1523/JNEUROSCI.0581-14.2014 PMID: 25122885

98. Pernet CR, McAleer P, Latinus M, Gorgolewski KJ, Charest I, Bestelmeyer PEG, et al. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage. 2015;119:164–74. https://doi.org/10.1016/j.neuroimage.2015.06.050 PMID: 26116964

99. Zhang Y, Ding Y, Huang J, Zhou W, Ling Z, Hong B, et al. Hierarchical cortical networks of "voice patches" for processing voices in human brain. Proc Natl Acad Sci U S A. 2021;118(52):e2113887118. https://doi.org/10.1073/pnas.2113887118 PMID: 34930846

100. Aglieri V, Chaminade T, Takerkart S, Belin P. Functional connectivity within the voice perception network and its behavioural relevance. Neuroimage. 2018;183:356–65. https://doi.org/10.1016/j.neuroimage.2018.08.011 PMID: 30099078