8-15-2021

Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar

Emma Holmes
The University of Western Ontario

Ingrid S. Johnsrude
The University of Western Ontario, ijohnsru@uwo.ca

Citation of this paper:
Emma Holmes, Ingrid S. Johnsrude, Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar, NeuroImage, Volume 237, 2021, 118107, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2021.118107. (https://www.sciencedirect.com/science/article/pii/S1053811921003840)

NeuroImage 237 (2021) 118107

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar

Emma Holmes[a,*], Ingrid S. Johnsrude[a,b]

[a] The Brain and Mind Institute, University of Western Ontario, London, Ontario, N6A 3K7, Canada
[b] School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario, London, N6G 1H1, Canada

ARTICLE INFO

Keywords:
Speech
Voice
Familiarity
Attention
Auditory cortex
fMRI

ABSTRACT

When speech is masked by competing sound, people are better at understanding what is said if the talker is familiar compared to unfamiliar. The benefit is robust, but how does processing of familiar voices facilitate intelligibility? We combined high-resolution fMRI with representational similarity analysis to quantify the difference in distributed activity between clear and masked speech. We demonstrate that brain representations of spoken sentences are less affected by a competing sentence when they are spoken by a friend or partner than by someone unfamiliar—effectively, showing a cortical signal-to-noise ratio (SNR) enhancement for familiar voices. This effect correlated with the familiar-voice intelligibility benefit. We functionally parcellated auditory cortex, and found that the most prominent familiar-voice advantage was manifest along the posterior superior and middle temporal gyri. Overall, our results demonstrate that experience-driven improvements in intelligibility are associated with enhanced multivariate pattern activity in posterior temporal cortex.

Introduction

Speech can be difficult to understand when other conversations take place at the same time. Being familiar with a conversational partner is associated with better speech intelligibility when a competing talker is present (Nygaard et al. 1994; Nygaard and Pisoni 1998; Yonan and Sommers 2000; Levi et al. 2011; Johnsrude et al. 2013; Kreitewolf et al. 2017; Holmes et al. 2018; Domingo et al. 2020). This familiar-voice benefit is substantial—participants report 10–20% more sentences correctly when they are spoken by their friend or spouse than when they are spoken by someone unfamiliar, and this cannot be explained by different acoustics of familiar and unfamiliar voices since, in a subset of these studies, familiar and unfamiliar voices were identical over the group (Johnsrude et al. 2013; Kreitewolf et al. 2017; Holmes et al. 2018; Domingo et al. 2020). Despite this large and consistent benefit to intelligibility, the neural mechanisms by which familiarity improves intelligibility are currently unknown.

Previous functional imaging studies have typically manipulated intelligibility by changing speech acoustics or lexical predictability. Studies manipulating speech acoustics have demonstrated that better speech intelligibility is associated with greater activity around the superior temporal sulcus (Scott 2000; Wild, Yusuf, et al. 2012; STS; Kyong et al. 2014) and superior temporal gyrus (STG; Davis et al. 2011; Evans et al. 2016). In these studies, however, it is difficult to disentangle effects of acoustics from differences in intelligibility. A study manipulating lexical predictability (Wild, Davis, et al. 2012) measured responses to degraded speech when it was preceded by a visual word prime: speech was rated as clearer when the word prime matched the spoken word than when it was different. The improvement in speech clarity for speech preceded by matching word primes was associated with greater activity in bilateral STS and left STG, including cytoarchitectonically defined primary auditory cortex. These findings are consistent with the idea that more intelligible speech is associated with greater activity along the superior temporal lobe, including primary auditory cortex.

Recent neuroimaging analyses have moved beyond simple activation maps to characterise the multivariate pattern of activity within a brain area, which improves sensitivity to distributed activity (Mur et al. 2009; Haxby 2012). For example, Representational Similarity Analysis (RSA; Kriegeskorte et al. 2008; Diedrichsen and Kriegeskorte 2017) quantifies the difference between conditions as the 'distance' in representational space between their associated multivariate activities. These multivariate approaches can detect between-condition differences in the pattern of activity across voxels, even when average activity is the same. This approach has been used in previous studies to cluster stimuli into categories based on their associated patterns of brain activity; however, here, we use RSA in a novel way—to quantify the difference in distributed activity between clear and degraded speech. In this way, the RSA distance reflects the difference in distributed activity evoked by

* Corresponding author at: Emma Holmes. Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London, WC1N 3BG, United Kingdom.
E-mail address: emma.holmes@ucl.ac.uk (E. Holmes).

https://doi.org/10.1016/j.neuroimage.2021.118107.
Received 14 October 2020; Received in revised form 19 April 2021; Accepted 25 April 2021
Available online 30 April 2021.

speech that is presented as clear and degraded; in other words, reflecting the extent to which brain activity is affected by the speech degradation (i.e., how 'robust' brain activity is to degradation). Given that familiarity with a talker improves intelligibility in noise—in other words, making the intelligibility of speech in noise more similar to that of speech in quiet—we hypothesised that we could identify areas sensitive to intelligibility (controlling for acoustics) by comparing activation patterns for familiar compared to unfamiliar voices. We reasoned that regions exhibiting more similar (i.e., more robust) multivariate activity for speech presented alone and the same speech in noise when the talker is familiar, compared to unfamiliar, are sensitive specifically to intelligibility. This allowed us to ask whether familiarity-driven intelligibility enhancements are evident as early as primary auditory cortex (Wild, Davis, et al. 2012; Holmes et al. 2021), in non-primary auditory cortex (Davis and Johnsrude 2003; Adank 2012; Alain et al. 2018), or in higher areas such as the inferior frontal gyrus (Davis and Johnsrude 2003; Wild, Yusuf, et al. 2012; Alain et al. 2018).

We used ultra-high field fMRI (7 Tesla), combined with RSA, to measure activity that was elicited by sentences that were presented alone and by the same sentences that were presented simultaneously with a competing sentence spoken by a different talker. Comparing the multivariate activity in these two conditions revealed the extent to which the pattern of brain activity was disrupted by a competing (unfamiliar) talker. We compared conditions in which participants listened to speech spoken by a familiar talker (their friend or partner) with speech spoken by unfamiliar takers, who were the friends and partners of other participants. Thus, familiar and unfamiliar stimuli were acoustically matched across the group.

Materials and methods

Participants

We recruited 27 participants (9 male, 22 right-handed), who had taken part in a previous behavioural experiment on voice familiarity, and who had a friend or partner who had been recorded speaking a list of sentences. Participants were 19–68 years old (median = 22 years, inter-quartile range = 6), were native Canadian English speakers, and had average pure-tone audiometric thresholds better than 20 dB HL in each ear (measured at four octave frequencies between 0.5 and 4 kHz). They had known their friends and partners (8 male, 10 romantic partners) for .6–35.6 years (median = 3.1 years, inter-quartile range = 5.1) and reported speaking to them 3–84 hours per week (median = 29 hours, inter-quartile range = 21). The experiment was cleared by Western University's Health Sciences Research Ethics Board. Informed consent was obtained from all participants.

Design

First, participants completed an adaptive behavioural task to determine the target-to-masker ratio (TMR) for reporting 40% of sentences correctly when both talkers were unfamiliar. During the subsequent scanning session, all stimuli were presented at the adapted TMR—which ensured that the intelligibility level of the baseline (unfamiliar) condition was equivalent for all participants.

During the scanning session, we presented 6 experimental conditions in a 3 × 2 factorial design. Target sentences were either spoken by a familiar ("Familiar") or by one of two unfamiliar ("Unfam-1" and "Unfam-2") talkers. The unfamiliar talkers in the scanning session were different than those presented in the pre-scan behavioural task, to prevent participants becoming overly familiar with particular unfamiliar voices. During the scanning session, target sentences were either presented alone ("Alone") or in the presence of a competing sentence ("Masked"). Masking talkers were always unfamiliar and different from the target talker. In addition, we included silent trials that contained no acoustic stimuli.

Finally, we conducted a post-scan behavioural task to measure the intelligibility of the materials heard in the three Masked conditions in the scanner, which provided an independent measure of the familiar-voice benefit to intelligibility for each participant. Sentences from the three conditions (Familiar Masked; Unfam-1 Masked; Unfam-2 Masked) were presented in a randomized order.

Apparatus

The pre- and post-scan behavioural sessions were conducted in a quiet room. Acoustic stimuli were presented through a Steinberg Media Technologies UR22 sound card and were delivered binaurally through Grado Labs SR225 headphones. Participants viewed visual stimuli on the monitor of a Lenovo ThinkPad P50 20EN laptop and responded using a mouse.

While participants were in the MRI scanner, acoustic stimuli were presented through the same Steinberg Media Technologies UR22 sound card, which was connected to a stereo amplifier (PYLE PRO PCA1 for 22 participants, PYLE PRO PCAU22 for 5 participants). Acoustic stimuli were delivered binaurally through Sensimetrics insert earphones (Model S14 for 22 participants, Model S15 for 5 participants) and were presented at a comfortable listening level that was the same for all participants. Visual stimuli were projected onto a screen at one end of the magnet bore, which participants viewed through a mirror attached to the head coil.

Stimuli

Acoustic stimuli were spoken sentences that had been recorded by each participant's friend or spouse in a previous experiment. Sentences were from the Boston University Gerald (BUG) corpus (Kidd et al. 2008), which follow the structure: "<Name><verb><number><adjective><noun>". In the sub-set of sentences used in the experiment, there were two names ('Bob' and 'Pat'), eight verbs, eight numbers, eight adjectives, and eight nouns (displayed in Fig. 1). An example is "Bob brought three red flowers".

Sentences were recorded using a Sennheiser e845-S microphone connected to a Steinberg Media Technologies UR22 sound card. The recordings were conducted in a single-walled sound-attenuating booth (Eckel Industries of Canada, Ltd.; Model CL-13 LP MR). The sentences had an average duration of 2.5 seconds ($s = 0.3$). The levels of the digital recordings of the sentences were normalised to the same root mean square (RMS) power.

During the experiment, each participant heard sentences spoken by their familiar partner and sentences spoken by eight unfamiliar talkers, who were the partners of other participants in the experiment. For each participant, unfamiliar talkers were selected to be the same sex and roughly the same age as the participant's familiar partner (they also necessarily had a similar accent because we only recruited participants who were native speakers of Canadian English). Sentences spoken by six of the unfamiliar talkers were presented in the pre-scan behavioural adaptive test, and sentences spoken by the other two unfamiliar talkers were presented in the scanning session and post-scan behavioural test: this was to ensure that the unfamiliar talkers from the pre-scan behavioural were not familiar by the start of the scan.

We planned to present each voice to one participant (i.e., their partner) as a familiar talker and to two other participants as an unfamiliar talker. However, this was not possible because the partners of 8 people did not participate in this experiment. Thus, 8 voices were presented as unfamiliar but never as familiar, 10 voices were presented only once as familiar and once as unfamiliar, and 3 voices were only presented as familiar. In total, we used 36 different talkers. Thus, across the group, familiar and unfamiliar conditions were acoustically similar.

Fig. 1. Schematic of the response screen used for the tasks conducted outside the scanner (i.e., pre-scan and post-scan behavioural tasks).

Procedure

Pre-scan behavioural. To determine the target-to-masker ratio (TMR) for reporting 40% (chance = 0.02%) of sentences correctly, we used a weighted up-down procedure (Kaernbach, 1991). On each trial, participants heard two sentences from the BUG matrix spoken simultaneously by two different unfamiliar talkers of the same sex. The relative levels of the two sentences were determined by the TMR (in decibels) for each trial. They identified the four remaining words of the sentence that began with a particular target name ("Bob" or "Pat"), by clicking buttons on a screen (Fig. 1). The words in the masker sentence were always different to the words in the target sentence. We adapted the TMR in 3 separate, but interleaved, runs—which each contained a different pair of unfamiliar talkers. Each run stopped after 12 reversals and we calculated thresholds for each run as the median of the last 5 reversals. For each participant, we calculated the median of the thresholds across the three runs: this TMR value was used during the MRI session.

Functional MRI. During the MRI session, we presented 12 functional runs, each containing 25 trials (300 trials total) and lasting 3.33 minutes. We presented 48 trials in each of the six experimental conditions, as well as 12 silent trials. All 7 trial types were interleaved in a pseudorandom order, with the constraint that each run included 1 silent trial and 4 trials from each of the six experimental conditions (sentence content was selected randomly for each condition, without replacement, from the set of 48 sentences).

In three of the conditions, participants heard 48 sentences from the BUG matrix (Kidd et al. 2008), which were either spoken by their familiar ("Familiar Alone") or by one of their two unfamiliar ("Unfam-1 Alone" and "Unfam-2 Alone") talkers. In the other three conditions ("Familiar Masked", "Unfam-1 Masked" and "Unfam-2 Masked"), participants heard the same sentences spoken by the same three talkers, but they were presented simultaneously with a different sentence from the BUG matrix that was spoken by one of the two unfamiliar talkers. The sentences that were used as maskers were from the same set of 48 sentences that were used as targets. The onsets of the target and masker sentences were identical. For the two conditions in which one of the unfamiliar talkers was presented as the target, the masker sentence was spoken by the other unfamiliar talker. In the Familiar Masked condition, the Unfam-1 and Unfam-2 talkers were each presented as the masker talker on half of the trials. The words in the masker sentence were always different from those in the target sentence. We chose to use the same 48 sentences in all conditions so that the materials were linguistically matched across all conditions and the target stimuli were identical in the Alone and Masked conditions. We used different sentences for every trial within every condition, so that the same sentences were not presented too frequently, which could evoke repetition suppression; using a set of 48 different sentences meant that we did not need to repeat the same sentences on consecutive trials.

Fig. 2 illustrates the trial structure. We modified the task so it was more amenable to responses inside the MRI scanner. On each trial, the target sentence was the one that began with a particular name word ('Bob' or 'Pat'). Half of target sentences began with 'Bob' and the other half began with 'Pat'. Acoustic stimuli were positioned such that the middle of the target sentence occurred 4 seconds before the beginning of the first volume collection of a pair of volumes (see section below: 'MRI data acquisition'); this is a conventional design for auditory functional imaging (Hall et al. 1999; Schwarzbauer et al. 2006; Perrachione and Ghosh 2013). Thus, sentence onset was jittered across trials. At the beginning of each trial, the target name word was displayed visually on the screen (even when the target sentence was presented alone). The name word was presented on the screen for 300 ms at the beginning of each trial, then a fixation cross was presented for 3700 ms. Four seconds after the trial began, participants saw a probe sentence written on the screen. They were asked to indicate whether the probe sentence was the same as the target sentence they heard spoken. They held a button box in one hand and pressed one button if the probe sentence was the same and a different button if the probe sentence was different. The name word in the probe sentence was always the same as the target name. On half

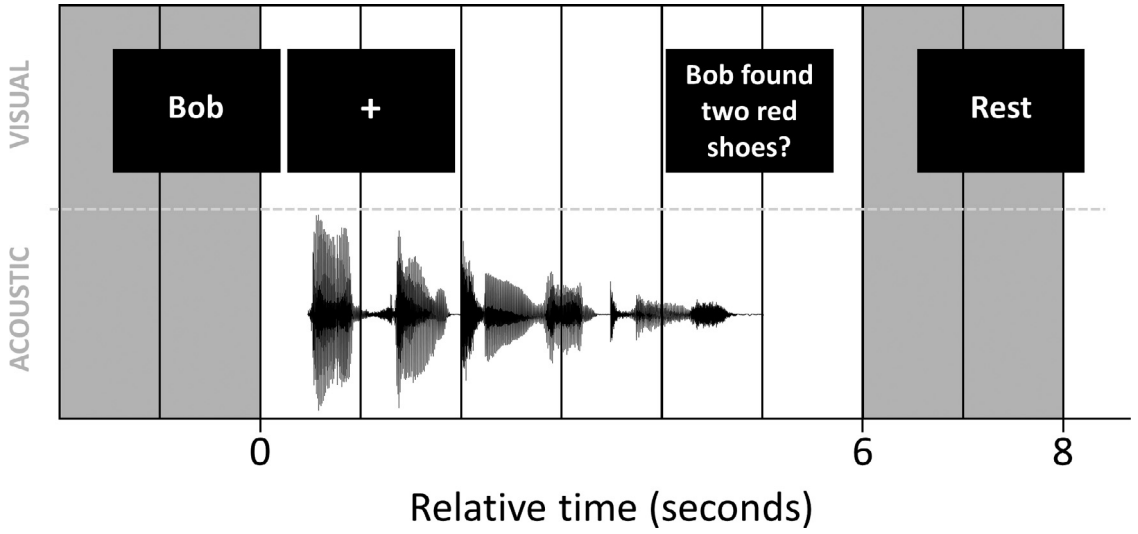

Fig. 2. Schematic of trial structure for the functional runs of the MRI session. An example trial is displayed, with the visual stimuli on the upper row, and the acoustic stimulus on the lower row. Each grey bar indicates one fMRI volume acquisition. White bars indicate 'silent' scans without volume acquisition. The acoustic stimulus is always presented during the 'silent' scans. The cue ("Bob") for the example trial is presented during the volume acquisitions for the previous scan, and the cue ("Rest") for the next (Silence) trial is presented during the volume acquisitions at the end of the trial.

of trials, the other four words were also the same. On the other half of trials, one of the four words was different. On Alone trials, the different word was selected randomly from the other words in the BUG corpus. On Masked trials, the different word was from the masker sentence. The placement of the incorrect word in the sentence (i.e., 2nd, 3rd, 4th, or 5th word) was counterbalanced across trials.

For the 12 silent trials, the visual cue word was "Rest", and no acoustic stimuli were presented.

Immediately before the scanning session, participants completed a practice, which contained 14 trials with the same (fixed) TMR that was used in the MRI session. The practice was conducted in a quiet room with the same equipment as the pre-scan adaptive task. The trial structure was identical to the functional runs of the scanning session. Participants responded using two keys on the laptop.

Post-scan behavioural. Finally, participants completed a behavioural task outside the scanner. We presented three conditions in which there was always a competing masker: Familiar Masked, Unfam-1 Masked, and Unfam-2 Masked. The trials were identical to those presented in the MRI session, but they were presented in a different (pseudorandomly interleaved) order. The post-scan behavioural was divided into two halves: In one half, target sentences began with the name word 'Bob', and in the other, target sentences began with the name word 'Pat'. The order of the name words was counterbalanced across participants. The structure of each trial was identical to the pre-scan adaptive part: participants identified the four remaining words from the target sentence by clicking buttons on a screen (Fig. 1). Participants completed 144 trials (48 in each of the three conditions), with a short break every 24 trials.

MRI data acquisition

MRI was conducted on a 7.0 Tesla Siemens MAGNETOM scanner at Robarts Research Institute, Western University (London, Ontario, Canada) with a 32-channel receive coil. At the beginning of the session, we acquired a whole-brain T1-weighted anatomical image for each participant with the following parameters: MP2RAGE; voxel size = 0.75 mm isotropic; 208 slices; PAT GRAPPA of factor 3; anterior-to-posterior phase encoding, time-to-repeat (TR) = 6000 ms, echo time (TE) = 2.83 ms.

T2*-weighted functional images were acquired using echo-planar imaging (EPI), with: voxel size = 1.75 mm isotropic; 63 slices; multiband acceleration of factor 3 with interleaved slices; field of view of 208 mm; TR = 1000 ms; echo spacing = 0.45 ms; PAT GRAPPA of factor 3; posterior-to-anterior phase encoding; bandwidth = 2778 Hz/Px. Acquisition was transverse oblique, angled away from the eyes, and in most cases covered the whole brain. (If the brain was too large for the field of view, slice positioning excluded the very top of the superior parietal lobule.) We used interleaved silent steady state (ISSS) imaging (Schwarzbauer et al. 2006): Each trial contained 7 'silent' scans (radio frequency pulses without volume acquisition) followed by 2 scans with volume acquisition (Fig. 2). Acoustic stimuli were presented during the silent period between volume acquisitions. We collected 52 volumes from each participant (2 per trial) in each of the 12 runs. The first two 'dummy' scans were presented immediately prior to the first trial of each run and were excluded from the analyses. We collected field maps immediately after the functional runs (short TE = 4.08 ms, long TE = 5.1 ms).

Analyses

For the analyses, we collapsed across the conditions in which unfamiliar voices were presented as targets (i.e., "Unfam-1 Alone" and "Unfam-2 Alone"; "Unfam-1 Masked" and "Unfam-2 Masked"). For all of the analyses, the number of participants (N) was 27.

Behavioural data

We calculated sensitivity (d') for target recognition performance during the MRI session using loglinear correction (Hautus 1995), and chance d' of 0.3. For the post-scan behavioural, we calculated the percentage of sentences in which participants reported all four words (after the name word) correctly. The data met the assumptions for normality, as assessed by non-significant Shapiro-Wilk and Kolmogorov-Smirnov tests, and by visual inspection of box plots and Q-Q plots. We used Pearson's product moment correlation coefficients to compare d' in the MRI session with percent correct in the post-scan behavioural session.

MRI data preprocessing and GLM

MRI data were preprocessed using SPM12 (Wellcome Centre for Human Neuroimaging, London, UK). Each participant's functional images (EPIs) were unwarped using their field maps and were realigned to the first image of the run. The functional and anatomical images were coregistered to the mean EPI, then normalised to the standard SPM12 tem-

plate (avg305T1). For RSA analyses, we took the mean of the two adjacent volumes for each trial (which were always the two volumes at the end of the trial, after the sentences had finished; see Fig. 2), to improve the signal-to-noise ratio. For the univariate analyses, we took the same average after applying spatial smoothing, to ensure the data met the assumptions of Gaussian random field theory for multiple comparisons correction (Worsley et al. 1992). For spatial smoothing, we used a Gaussian kernel with a full-width at half-maximum of 12 mm.

We analysed the results from each participant at the first level using a General Linear (convolution) Model that uses least squares to estimate all parameters simultaneously: we included 18 regressors of no interest, which included the 6 motion realignment parameters (3 directions and 3 rotations) and 12 regressors corresponding to each run. We applied no high-pass filtering, because of the long time period between volume acquisitions. Serial correlations were accounted for using the default autoregressive model in SPM12.

RSA

For RSA, we entered the unsmoothed images into the first level analysis. We extracted the betas from each participant that corresponded to each of the experimental conditions: this produced beta images with one value per voxel. The region of interest (ROI) was defined using the Neurosynth database: We used a meta-analysis of all studies (N = 81; 'association test') that included the term 'Speech Perception' and used this to mask the imaging data. We analysed the 'distance' between the betas for pairs of conditions using MATLAB 2017b. In other words, for each pair of conditions we asked: how (dis)similar is the distribution of beta values across voxels? We focussed on pairs of conditions in which the same sentences were spoken by the same talker, but in the presence or absence of a competing masker; for example, "Familiar Alone" compared with "Familiar Masked". We did this so that each distance reflects only the effect of the masker (which was present in the Familiar Masked condition and absent in the Familiar Alone condition) and not the effect of the target voice (which was the same in both conditions). Thus, within each subject, comparisons between Familiar and Unfamiliar distances are not affected by the acoustics of the voices. For the unfamiliar condition, we averaged the distances across the two unfamiliar voices for each participant. We performed the analyses once using correlations as the distance metric and once using Euclidean distances, and we obtained the same pattern of results using both methods. We, therefore, primarily report results using correlation distances, which were defined as 1 minus the Pearson's correlation coefficient. As a post-hoc analysis, we also repeated the analysis with the SPM t-maps (each condition contrasted against silent trials) rather than the beta values. For completeness—and to demonstrate the robustness of our results to the specific analysis method chosen—we show the results of all of these analyses in the Results section. At the group level, we compared distances for Familiar and Unfamiliar conditions using Wilcoxon signed rank tests for repeated samples.

For each participant, we extracted the distances between the Alone and Masked stimuli in the Familiar and Unfamiliar conditions and used the difference (within the entire Speech Perception ROI) as an index of the Familiar-Unfamiliar RSA difference. We refer to this as the RSA interaction. We then used a Spearman's correlation, across participants, to examine the relationship between the Familiar-Unfamiliar RSA interaction and the behavioural benefit to intelligibility that each participant obtained from their familiar voice (which was not normally distributed). We calculated this behavioural benefit from the post-scan behavioural test, as the difference between percent correct in the Familiar Masked condition and the Unfamiliar Masked conditions. As the demographic data violated assumptions of normality, we used Spearman's correlations to examine the relationships between the Familiar-Unfamiliar RSA interaction and the number of years participants had known each other or the number of hours they reported speaking to their friend or partner each week. We compared Spearman's correlations using a one-tailed test according to Eid et al. (2017). As a post-hoc analysis—to rule out differences in fundamental frequency and acoustic correlates of vocal tract length between each participant's familiar and unfamiliar voices as an explanation for the RSA interaction—we also used Spearman's correlations to examine the relationships between the Familiar-Unfamiliar RSA interaction and these acoustic attributes. For every participant, we calculated the average fundamental frequency and formant ratio (i.e., the second formant frequency divided by the first formant frequency) for sentences spoken by each of the three voices they heard during the experiment (which were extracted using Praat; Boersma and Weenink 2003); we then calculated the difference in these attributes between the participant's familiar voice and each of the two unfamiliar voices. The average difference across the two unfamiliar voices was used as an indication of the Familiar-Unfamiliar fundamental frequency difference and the Familiar-Unfamiliar formant spacing difference for each participant.

For analyses in which we use a primary auditory cortex ROI, we applied a bilateral mask of Te1.0 from the SPM Anatomy Toolbox (Eickhoff et al. 2005).

Searchlight RSA

We used the RSA toolbox (Nili et al. 2014) for the searchlight RSA analysis. We searched within the 'Speech Perception' Neurosynth ROI for areas that were particularly sensitive to the difference in distances between Familiar and Unfamiliar conditions. We defined an expected dissimilarity matrix (visualised in Fig. 3) based on the 6 conditions, which each contained 48 trials. The matrix contained a smaller value (0.5) for the Familiar Alone with Familiar Masked cells, than for the Unfamiliar Alone with Unfamiliar Masked cells (1.0). The remaining cells in the matrix were of no interest for this analysis and were, therefore, excluded. We used the correlation distance metric on the betas at the individual subject level. To compare the expected dissimilarity matrix with the data (6 distance measures per participant, corresponding to the cells in the expected dissimilarity matrix displayed in Fig. 3), we used Spearman's correlations, to identify areas showing greater dissimilarity for unfamiliar than familiar conditions (irrespective of the absolute values in the expected dissimilarity matrix—which were set to 0.5 and 1.0). As a post-hoc analysis, we used Kendall's tau-b instead of Spearman's correlations and obtained identical results. Our searchlight area was spherical with a radius of 15 mm. We judged that the size of this searchlight area would provide an acceptable trade-off between statistical power (which increases as the searchlight radius increases, because the number of data points increases, and means that patterns with a larger spatial extent are able to be detected) and spatial specificity (which decreases as the searchlight radius increases). Searchlight areas at the edge of the ROI—whose spherical area spanned voxels outside the ROI—were included with as many voxels were inside the ROI; in other words, these tests were conducted on fewer voxels. We assessed the significance of the correlation statistics for every searchlight area at the group level using t-tests, with false discovery rate (FDR) correction for the number of searchlight areas within the Speech Perception ROI.

Univariate analyses

For the univariate analyses, we entered the spatially smoothed images into the first level analyses, where we applied our contrasts of interest: the main effect of Familiarity (Familiar or Unfamiliar), the main effect of Masker (Alone or Masked), and the interactions. We also included the same 18 regressors of no interest that we included in the GLM for the RSA analyses. We analysed the resulting contrast images at the group level using one-sample t-tests. All contrasts were corrected for family-wise error (FWE; Worsley et al. 1992).

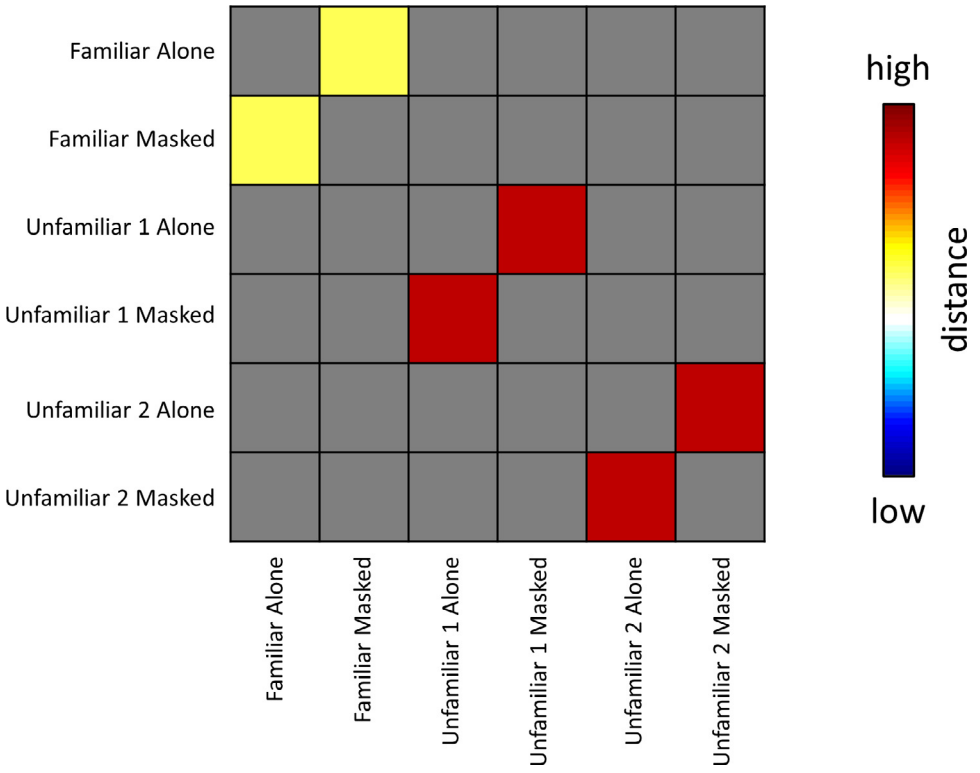

Fig. 3. Hypothesis representational dissimilarity matrix for the searchlight RSA analysis. The matrix contained a smaller distance value for the Familiar Alone with Familiar Masked cells, than for the Unfamiliar Alone with Unfamiliar Masked cells. The remaining cells in the matrix were of no interest for this analysis and were, therefore, excluded (grey cells in the figure).

Effects of interest

In all of the analyses, we were most interested in how the effect of masker (whether a target was masked by a competing unfamiliar voice or presented alone) depended on the familiarity of the target voice. Identical target stimuli (sentences and voices) were used in both Masked and Alone conditions, so the difference reflects the degree to which speech perception is affected by the presence of a masking sentence. Masking sentences were identical for Familiar and Unfamiliar conditions (always one of two unfamiliar voices, different from the target voice), and Familiar and Unfamiliar target voices were largely counterbalanced across participants (see Materials and Methods section for details). Thus, any difference in processing of familiar, compared to unfamiliar, voices when a masker is present cannot be explained by acoustics. In contrast, the main effect of Masker could be attributable to a variety of factors, including acoustic differences between the Alone and Masked conditions—and was therefore not of interest.

Data availability

The data generated during this study are available at the Open Science Framework (https://osf.io/bd6vr/).

Results

Replication of familiar-voice benefit to intelligibility

In the MRI system, participants performed well in the Familiar Alone (mean = 92.5%, S.E. = 1.9) and Unfamiliar Alone (mean = 91.4%, S.E. = 2.0) conditions. They performed less well in the Familiar Masked condition (mean = 69.5%, S.E. = 2.4) and most poorly in the Unfamiliar Masked condition (mean = 61.0%, S.E. = 1.9). Performance was better than chance (50%) in all four conditions [$t(26) > 5.75$, $p < .001$, $g_s > 2.15$].

A 2 × 2 ANOVA (factors: Familiarity and Masker) confirmed that sentences in familiar voices were more intelligible than sentences in unfamiliar voices [main effect of Familiarity: F(1, 26) = 16.29, $p < .001$, $\omega_p^2 = .35$], and sentences presented alone were more intelligible than masked sentences [main effect of Masker: F(1, 26) = 270.60, p < .001, $\omega_p^2 = .91$; see Fig. 4]. A significant Familiarity-Masker interaction [F(1, 26) = 6.99, $p = .014$, $\omega_p^2 = .18$] indicated better sensitivity (d') for familiar than unfamiliar voices when a masker was present [paired-sample t-test: $t(26) = 4.12$, $p < .001$, $d_z = .79$], but not when the target sentence was presented alone [$t(26) = 1.53$, $p = .14$, $d_z = .29$], which is probably due to a ceiling effect when only one sentence was presented.

Post-scan intelligibility testing revealed better performance for Familiar Masked (mean = 66.3%, S.E. = 4.0) than Unfamiliar Masked (mean = 46.9%, S.E. = 3.6) targets [$t(26) = 4.75$, $p < .001$, $d_z = .91$]. Performance in both conditions was significantly above chance (.004%) [$t(26) > 12.80$, $p < .001$, $g_s > 4.78$]. Across participants, post-scan intelligibility correlated with d′ in the scanning session, for both Familiar Masked [$r = .68$, $p < .001$; 95% CI = .39–.84] and Unfamiliar Masked [$r = .58$, $p = .001$; 95% CI = .26–.79] materials.

Pattern of activity is more robust for familiar voices

We targeted our analyses to brain regions known to be important for speech perception. We identified an ROI using a term-based meta-analysis in Neurosynth: the ROI was based on 81 studies using the search term "speech perception", which produced a 7217-voxel ROI that included superior and middle temporal gyri (and sulci) bilaterally, as well as left inferior temporal gyrus, left IFG and insula, left superior frontal gyrus, left precentral gyrus, right postcentral gyrus, and bilateral cerebellum (see Fig. 5).

Within the ROI, we used RSA to test the dissimilarity of multivariate representations between conditions that contained identical target sentences: we compared the Familiar Masked with the Familiar Alone condition, and the Unfamiliar Masked with the Unfamiliar Alone condition. Dissimilarities (correlation distances) were small overall (Fig. 6A), but were greater for sentences in unfamiliar voices (median = .0096; interquartile range [IQR] = .0022) than for sentences in a familiar voice (median = .0094; IQR = .0015) ($W = 290$, $p = .015$, $Z = 2.43$). Thus,

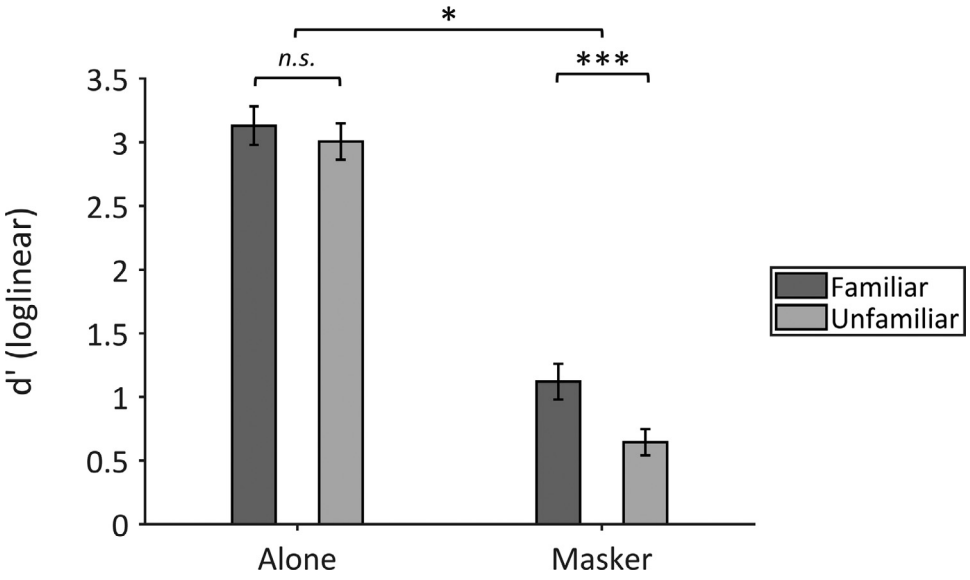

Fig. 4. Behavioural sensitivity (d′ with loglinear correction; $N = 27$) during the functional runs of the MRI session. Error bars display ±1 standard error of the mean. [*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; *n.s.* not significant]

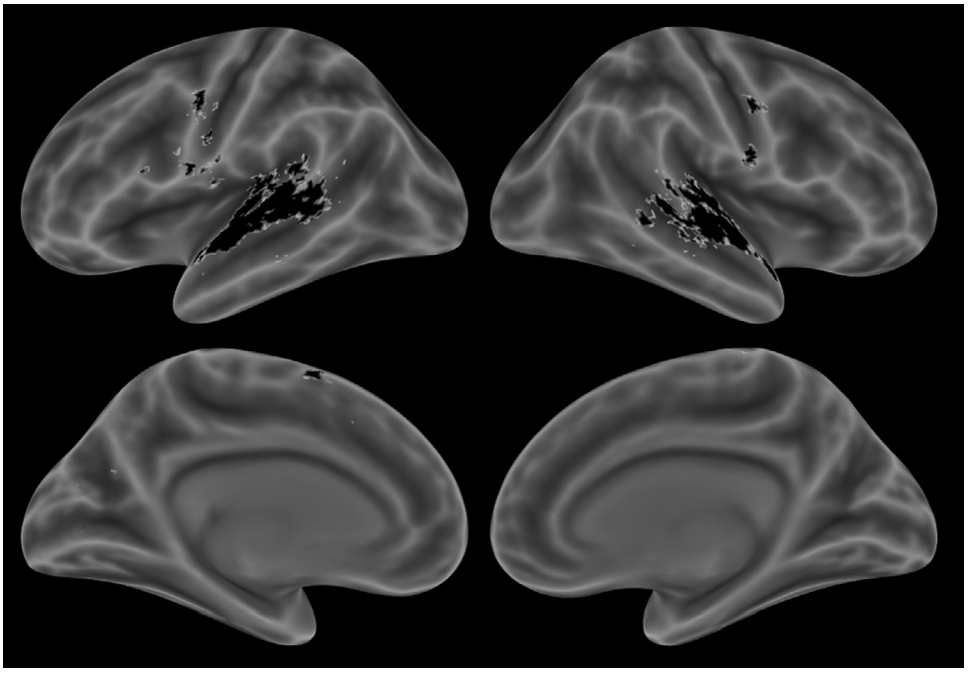

Fig. 5. Speech Perception mask from the Neurosynth database (generated from 81 studies with the 'association test' method), displayed on an inflated cortical surface. The left hemisphere is on the left side of the image. The mask contained 7217 voxels, which are indicated in black. All analyses were conducted in volumetric space, and are displayed on the cortical surface for visualisation only.

in this large ROI, the representation of speech is less influenced by a masker if the voice is familiar.

We replicated this result using different RSA methods (see Fig. 7; corresponding statistics are shown in the figure legend).

RSA interaction correlates with intelligibility benefit

We then examined whether the magnitude of the RSA interaction just described—the difference in Alone-Masked dissimilarity for Familiar and Unfamiliar voices in the Speech Perception ROI—correlated with the intelligibility benefit in individual participants. Fig. 6B shows the significant correlation between behavioural performance (in the post-scan intelligibility test) and the RSA interaction, across participants [$r_s = .51$, $p = .007$; 95% CI = .16–.74].

The RSA interaction did not correlate with the number of years participants had known their friend or partner [$r_s = -.05$, $p = .81$; 95% CI = -.38–.00] or the number of hours per week they spoke to them [$r_s = -.17$, $p = .39$; 95% CI = -.51–.00]; both of these correlations were significantly smaller than the correlation with behavioural performance [$z > 1.79$, $p < 0.037$]. In addition, the RSA interaction did not correlate with the difference in fundamental frequency [$r_s = .05$, $p = .81$; 95% CI = -.35–.41] or formant spacing [$r_s = -.18$, $p = .36$; 95% CI = -.53–.21] between the familiar and unfamiliar voices for each participant; both of these correlations were significantly smaller than the correlation with behavioural performance [$z > 1.73$, $p < 0.042$].

RSA interaction is most prominent in posterior STG, MTG and PT

We used searchlight RSA to find the brain areas within the Speech Perception ROI that were most sensitive to the RSA interaction. Fig. 6C–D shows the results of this analysis, thresholded at $p < .05$ FDR within the Speech Perception ROI (7217 voxels). 728 of the searchlight volumes (15 mm diameter) were significant. The centres of significant volumes

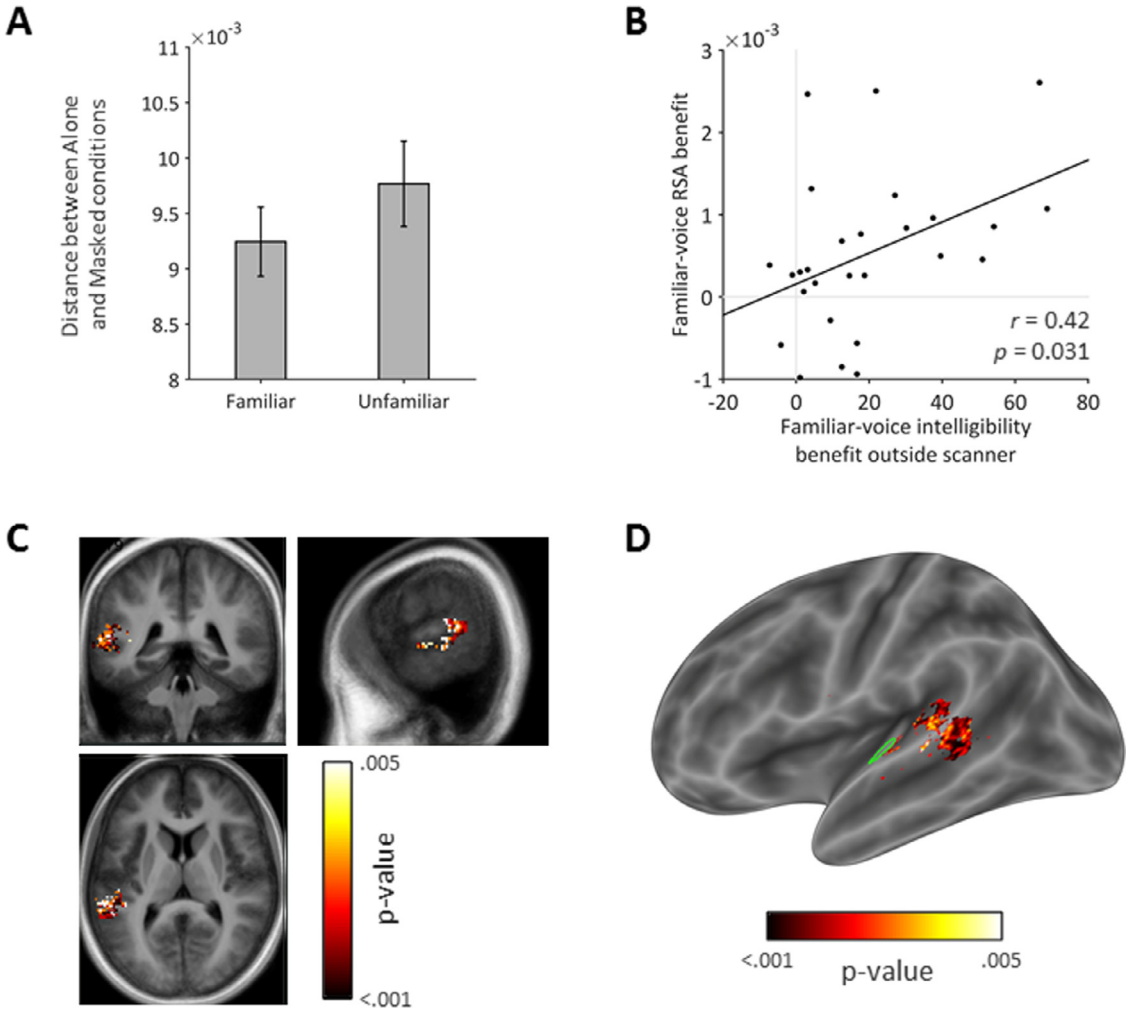

Fig. 6. Functional MRI results (N = 27). (A) Results from the Representational Similarity Analysis (RSA) in the Speech Perception ROI. The y-axis shows the correlation distance metric (1 - Pearson's correlation coefficient) between the Alone and Masked conditions, plotted separately for conditions in which the target sentence was Familiar or Unfamiliar. Error bars display ±1 standard error of the mean. (B) Correlation between the familiar-voice benefit to intelligibility (i.e., difference in percent correct between the Familiar Masked and Unfamiliar Masked conditions) measured in the post-scan behavioural task and the familiar-voice RSA benefit (i.e., the RSA interaction between Familiarity and Masker) in each participant. Each dot represents one participant. (C) Areas identified in the searchlight RSA (i.e., $p < 0.05$ FDR at the group level within the Speech Perception ROI; corresponding to $p < 0.005$ uncorrected, as plotted), displayed on sections from the average structural image (27 participants). Following neurological convention, the left hemisphere is on the left side of the image. (D) Results from the searchlight RSA displayed on an inflated cortical surface (left hemisphere only), plotted using BSPMVIEW (Spunt 2016). These results are the same as those plotted in panel C (which were conducted in volumetric space), but are visualised differently so that all significant results can be viewed in a single image. The area outlined in green indicates left Te1.0. In panels C and D, the colour bar indicates the uncorrected p-values, which were all $p < .05$ after applying false discovery rate (FDR) correction.

were located in left posterior STG and MTG, and left planum temporale (PT).

As a post-hoc exploratory analysis, to check for effects outside of the ROI, we conducted a whole-brain searchlight analysis. No searchlight volumes were significant after FDR correction.

No evidence of RSA interaction in primary auditory cortex

To check if there was evidence for familiar-voice effects in primary auditory cortex, we used two complementary approaches.

First, we used a primary auditory cortex ROI (Te1.0: Morosan et al. 2001; 409 voxels) to test whether this region—as a whole—showed different RSA distances between the Alone and Masked conditions for Familiar compared to Unfamiliar voices; in other words, whether there was evidence for an RSA interaction. Using the same method that we used for the speech perception ROI (in the section above: "Pattern of activity is more robust for familiar voices"). we found no significant difference in correlation distances between Familiar (median = .0094; IQR = .0025) and Unfamiliar (median = .0101; IQR = .0023) conditions ($W = 255$, $p = .11$, $Z = 1.59$).

Second, we checked whether the significant searchlight volumes within the Speech Perception ROI (from the section above: "RSA interaction is most prominent in posterior STG, MTG and PT") overlapped with primary auditory cortex. We compared the centres of significant RSA volumes displayed in Fig. 6C with the primary auditory cortex ROI. Centres of auditory cortex volumes were posterior and/or inferior to area Te1.0 (Fig. 6D), implying that significant interactions between Familiarity and Masker occur outside primary auditory cortex.

No evidence for difference in regions for familiar versus unfamiliar voices

For completeness, we also analysed the data using a standard univariate approach, using a threshold of $p < .05$ FWE. No voxels were significant at this threshold (either in a whole brain analysis or within the Speech Perception ROI) for the main effect of Familiarity or for the interaction between Familiarity (Familiar or Unfamiliar) and Masker (Alone

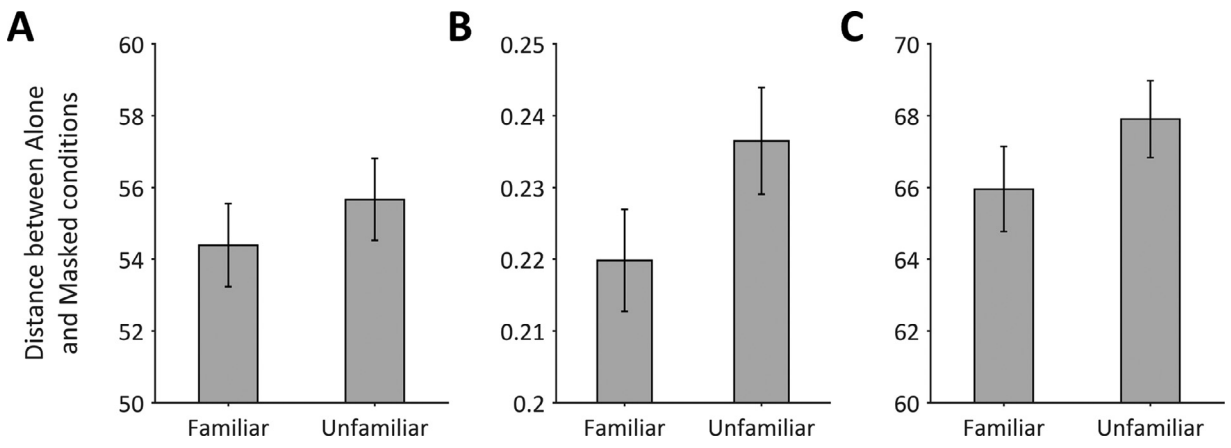

Fig. 7. Comparison of results from the Representational Similarity Analysis (RSA) in the Speech Perception ROI, using different methods. In all panels, the y-axis shows the distance metric between the Alone and Masked conditions, plotted separately for conditions in which the target sentence was Familiar or Unfamiliar. Error bars display ±1 standard error of the mean. (A) Euclidean distance calculated on the beta values differed between Familiar and Unfamiliar conditions ($S = 96$, $p = .025$, $Z = 2.23$). (B) Correlation distance (1 - Pearson's correlation coefficient) calculated on the SPM t-maps (each condition contrasted against silent trials) differed between Familiar and Unfamiliar conditions ($S = 33$, $p = .00018$, $Z = 3.75$). (C) Euclidean distance calculated on the SPM t-maps differed between Familiar and Unfamiliar conditions ($S = 71$, $p = .0046$, $Z = 2.83$). For comparison, the correlation distance calculated on the beta values are displayed in Figure 6A, and the corresponding statistics are reported in the Results section.

Table 1
Results from the univariate contrast between the Alone and Masked conditions. Statistical analyses were conducted at the group level using one-sample t-tests, and were thresholded at p = .05 after correcting for family-wise error (FWE). Peak locations were labelled using the Harvard-Oxford atlas based on the MNI co-ordinates. L: Left; R: Right.

Contrast	Peak location	t	p_{FWE}	MNI co-ordinates (mm)		
				x	y	z
Masked > Alone	Planum Temporale (L)	-13.24	< .001	-55	-21	4
	Supramarginal Gyrus (L posterior)	-10.19	< .001	-55	-42	11
	Middle Frontal Gyrus (L)	-13.11	< .001	-42	17	32
	Middle Frontal Gyrus (L)	-8.32	< .001	-40	3	54
	Inferior Frontal Gyrus, pars opercularis (L)	-8.03	.001	-50	19	11
	Superior Parietal Lobule (L)	-12.44	< .001	-31	-57	46
	Planum Temporale (R)	-10.21	< .001	62	-19	7
	Superior Frontal Gyrus (L)	-10.13	< .001	-5	12	58
	Paracingulate Gyrus (R)	-6.06	.031	10	28	35
	Angular Gyrus (R)	-9.63	< .001	35	-55	46
	Middle Frontal Gyrus (R)	-9.21	< .001	47	31	33
	Frontal Operculum Cortex (R)	-8.18	< .001	35	24	4
	Cerebral White Matter (R)	-7.31	.002	31	50	4
	Caudate (R)	-7.07	.004	12	10	7
	Location not in atlas	-6.95	.005	-38	-62	-33
	Cerebral White Matter (L)	-6.36	.017	-29	42	4
	Caudate (L)	-6.06	.031	-12	12	4
	Middle Frontal Gyrus (R)	-5.97	.037	35	10	63
	Cerebral White Matter (L)	-5.96	.038	-12	7	4
	Cerebral White Matter (R)	-5.95	.039	24	54	-5
	Cerebral White Matter (L)	-5.84	.049	-14	3	4
Alone > Masked	Paracingulate Gyrus (L)	8.97	< .001	-5	54	0
	Frontal Pole	8.24	< .001	0	59	23
	Hippocampus (L)	8.35	< .001	-29	-31	-12
	Temporal Pole (R)	8.05	.001	52	7	-30
	Supramarginal Gyrus (R anterior)	6.73	.008	55	-29	32
	Cerebral White Matter (L)	6.49	.013	-14	-52	28
	Frontal Pole (L)	6.32	.018	-16	45	49
	Frontal Pole (R)	5.97	.038	10	54	44
	Subcallosal Cortex (L)	5.95	.039	-3	7	-9
	Frontal Pole (R)	5.91	.042	9	55	40
	Frontal Pole (R)	5.91	.042	14	52	47
	Subcallosal Cortex (L)	5.85	.048	-2	10	-9

or Masked). We found a number of significant regions for the main effect of Masker (see Table 1 and Fig. 8), possibly reflecting differences in acoustics, or in processes contributing to intelligibility when a masker is present. Peaks for the contrast Masked > Alone were largely confined to the region of the Speech Perception ROI, whereas peaks for the contrast Alone > Masked were almost entirely outside this ROI.

Discussion

Representations of spoken sentences in left-temporal regions are less affected by competing speech when they are spoken by someone familiar. In other words, familiar voices that are presented with a competing sentence have a higher cortical signal-to-noise ratio (SNR) than unfa-

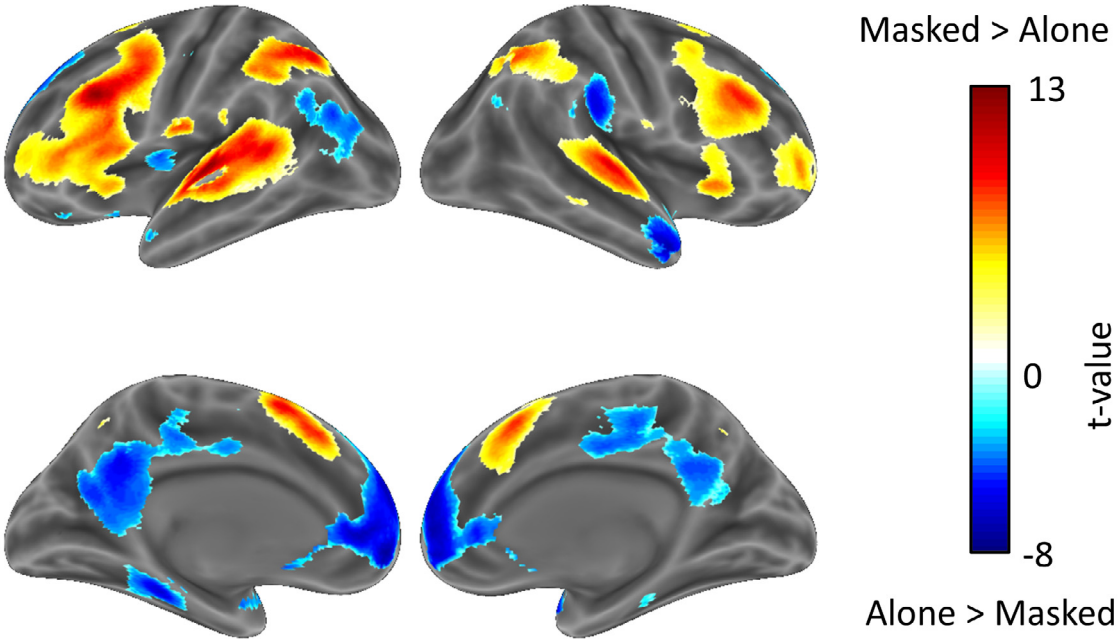

Fig. 8. Results from the univariate contrast between the Alone and Masked conditions, displayed on an inflated cortical surface. The left hemisphere is on the left side of the image. Coloured regions indicate voxels that survived a threshold of $p < .05$ after correcting for family-wise error (FWE). Warm colours indicate greater activity in the Masked than Alone conditions, and cool colours indicate greater activity in the Alone than Masked conditions. For statistics, see Table 1.

miliar voices that are presented with a competing sentence. The extent to which familiar voices elicited more robust multivariate patterns than unfamiliar voices correlated with the benefit to intelligibility that individuals obtained from the same familiar voice, and this correlation was significantly stronger than the correlation with the degree of familiarity (the number of years participants had known their friend or partner, and the number of hours they reported talking to them); neither measure of the degree of familiarity had a significant relationship to the multivariate effect. Thus, based on these measures, multivariate BOLD activity in speech-sensitive brain areas seems to index the intelligibility benefit that people gain from a familiar voice in the presence of a competing talker, rather than familiarity *per se*. Experience-driven changes in the *intelligibility* of familiar voices appears to be reflected in the representations of these voices in the left posterior STG and MTG, and in left PT. These regions are anatomically situated at intermediate stages of processing in auditory cortex, rather than primary cortex or at higher levels of the processing hierarchy such as IFG (Kaas et al. 1999; Scott and Johnsrude 2003; Peelle et al. 2010; Medalla and Barbas 2014).

We accounted for acoustic differences between familiar and unfamiliar voices in two ways. First, within each subject, we calculated distances between conditions in which the same target voice spoke the same sentences, but the masker differed. These distance values therefore remove responses specific to a target voice (e.g., related to its acoustics) and retain the effect of the masker. Each condition also contained exactly the same 48 target sentences, so these distance values remove responses specific to sentence content too. Second, voices were counterbalanced across the group such that unfamiliar talkers were familiar to other participants. Thus, the familiar-voice advantage is due to familiarity with a friend or partner's voice, rather than differences in voice acoustics between familiar and unfamiliar talkers.

Previous studies have identified sensitivity in left posterior STG and MTG to intelligibility by manipulating speech acoustics (Davis and Johnsrude 2003; Davis et al. 2011; Wild, Yusuf, et al. 2012; Evans et al. 2016) and the predictability of speech materials (Sohoglu et al. 2012; Wild, Davis, et al. 2012). Here, we demonstrate sensitivity in STG to intelligibility, using materials that are acoustically and linguistically identical across conditions. These results cannot be explained by acoustic factors, such as fundamental frequency, vocal tract length, accent, intonation, or other acoustic properties that differ between voices. The results reflect differences in the extent to which processing of a target sentence is affected by the presence of a competing sentence when the target sentence is spoken by a familiar compared to an unfamiliar person; this difference in processing may be associated with better top-down attention when a familiar voice is the target, leading to better intelligibility, which could arise because familiar voices are processed more efficiently than are unfamiliar voices (Holmes & Johnsrude, 2020).

Behavioural studies demonstrate that familiar voices are not simply more intelligible because they are more salient than unfamiliar voices (Johnsrude et al. 2013; Domingo et al. 2020; Holmes and Johnsrude 2020), even when familiar voices are presented less often as targets than unfamiliar voices (Holmes and Johnsrude 2020) (like in the current study in which each voice identity was presented as a target equally often, but the ratio of familiar to unfamiliar targets was 1:2). Therefore, the results obtained here are unlikely due to differential attentional salience of familiar and unfamiliar voices. In addition, any effects of processing familiar voices that occur in both the Familiar Alone and Familiar Masked conditions (which were interleaved) cannot explain our results, because the RSA analysis measured the difference between these conditions.

Our results demonstrate that the representation of spoken-sentence information in left posterior temporal regions is more resistant to interference by competing speech if the target talker is familiar. Our results can be thought of as reflecting better cortical SNR for familiar than unfamiliar voices. Cognitively, this could be underpinned by processes that are related to a reduction in informational masking (Wang et al. 2019; see Holmes and Johnsrude 2020), such as better segregation (Holmes et al. 2021) of speech in a familiar voice from masker sound, or better predictions about the low-level acoustic form of speech (Wild, Davis, et al. 2012) for familiar than unfamiliar voices.

Acoustically, the familiar-voice benefit to intelligibility relies critically on representations of the fundamental frequency and vocal tract length of the familiar talker (Holmes et al. 2018), so these are potential candidates for enhanced representation; the activity we observed could potentially reflect better representation of the pitch

(Griffiths et al. 1998; Gander et al. 2019) or other vocal characteristics for familiar than unfamiliar voices. From a neural perspective, increases in neuronal gain (Rabinowitz et al. 2011) of frequency channels corresponding to the frequencies of an attended voice (Rutten et al. 2019) may operate more efficiently for familiar than unfamiliar voices.

Bilateral STG and MTG have been shown to respond more to vocal than non-vocal sounds, and they have been previously labelled as 'temporal voice areas' (Belin et al. 2011; Bethmann and Brechmann 2014; Pernet et al. 2015; Agus et al. 2017). The area of STS that we found to be most sensitive to the familiar-unfamiliar voice difference is more posterior than the anterior and mid temporal voice areas reported in some studies (Belin et al. 2011; Pernet et al. 2015; Agus et al. 2017), but overlaps with posterior temporal voice areas reported in others (Warren et al. 2006; Birkett et al. 2007; Bethmann et al. 2012; Bethmann and Brechmann 2014; Pernet et al. 2015). Our finding that left STG is sensitive to the difference between familiar and unfamiliar voices suggests that these areas are also sensitive to the familiarity of voices. Previous imaging studies that compared familiar and unfamiliar voices have either used tasks that asked participants to judge voice familiarity (Birkett et al. 2007; Bethmann et al. 2012), or had participants passively listen to stimuli while speaker identity varied across conditions (Warren et al. 2006). In contrast, participants in this study were asked to focus on the intelligibility of spoken sentences in familiar and unfamiliar voices.

We found no significant differences between familiar and unfamiliar voices in the univariate analysis, consistent with the idea that speech spoken by familiar and unfamiliar people is processed in similar regions of the brain. The auditory face model (Belin et al. 2004, 2011) proposes that speech information and vocal identity are analysed in different areas of the brain: this idea is consistent with evidence that brain activity differs depending on whether the task is one of intelligibility or voice recognition (Von Kriegstein et al. 2003; Kriegstein and Giraud 2004; Bonte et al. 2014), Here, the task was to discriminate the content of speech (i.e., the words that were spoken), rather than to recognise the voice. The auditory face model does not explain how familiar-voice information affects speech intelligibility. Instead, our work builds upon evidence from a behavioural study showing that people use familiar-voice information in different ways when the goal is to understand the words spoken by someone familiar than when the goal is to recognise someone's identity from their voice (Holmes et al. 2018). Our RSA results cannot be explained by voice identification or recognition, because these processes would occur in both the Alone and Masked conditions and would, therefore, not be present in the RSA interaction between Familiarity and Masker. Our RSA results suggest that, in contrast to abstractionist accounts of speech perception (Lavner et al. 2001), in which talker-specific characteristics are stripped from the signal before the linguistic information is processed, information about a familiar talker is combined in the brain with information about the speech content, resulting in a more noise-resistant representations of (talker-specific) speech. This is more consistent with episodic accounts of speech processing (Goldinger 1998; Lachs et al. 2003), which posit that long-term representations of voice characteristics also participate in processes of lexical access and word recognition.

In this study, we chose to focus on regions known to be sensitive to speech perception, as we hypothesised this is where we would find areas that are sensitive to the familiar-voice benefit to intelligibility. Our ROI included several stages of auditory processing: primary auditory cortex, later stages of processing in auditory cortex, and higher areas outside of auditory cortex including IFG, left superior frontal gyrus, left precentral gyrus, and right postcentral gyrus (Kaas et al. 1999; Scott and Johnsrude 2003; Peelle et al. 2010; Medalla and Barbas 2014). We found no evidence that representations in primary auditory cortex or in areas at higher stages of processing in frontal cortex reflected the familiar-voice benefit to intelligibility. While we found significant searchlight volumes centred in left posterior STG and MTG and PT, this may in fact be an overestimate of the number of significant volumes—and the real region of sensitivity may be smaller than shown—given that the searchlight volumes overlapped considerably and are, therefore, spatially correlated.

Given that manipulating visual word primes to enhance intelligibility led to univariate activity in broadly similar regions of the brain that we found to be maximally sensitive to the familiar-voice benefit to intelligibility (Sohoglu et al. 2012; Wild, Davis, et al. 2012), similar mechanisms may underlie both effects. Such a result would suggest that these regions are not necessarily voice-specific, but are representing the brain's "best guess" at the linguistic content—reflecting the integration of signal content with content constructed through intelligibility-enhancing processes that involve context, predictability, and familiar-voice cues. The RSA methods used here will be helpful for exploring these possibilities in the future.

Conclusions

Overall, the current study demonstrates that posterior temporal cortex represents information about target speech more robustly in the presence of competing speech when the target talker is a friend or partner, compared to someone unfamiliar. Furthermore, the relative robustness of the representations for a familiar, compared to an unfamiliar, target talker correlates with the intelligibility benefit that participants gain from that familiar voice. Whether these posterior temporal regions are representing voice-specific speech information, or a more general, reconstructed 'best guess' at the identity of a masked speech signal, remains to be determined. This is a first step in establishing the neurobiological organization supporting the intelligibility benefit obtained when speech is in a familiar compared to unfamiliar voice. This benefit is large, and may be of substantial importance in everyday life, particularly for those with hearing impairment.

Author Contributions

EH: Writing. ISJ: Writing.

Declaration of interests

The authors declare no competing interests.

Credit authorship contribution statement

Emma Holmes: Conceptualization, Investigation, Data curtion, Software, Formal analysis, Visualization. **Ingrid S. Johnsrude:** Conceptualization, Funding acquisition.

Acknowledgements

This work was supported by funding from the Canadian Institutes of Health Research (CIHR; Operating Grant: MOP 133450), the Natural Sciences and Engineering Research Council of Canada (NSERC; Discovery Grant: 327429-2012), and Western University's Canada First Research Excellence Fund BrainsCAN initiative. We thank Joe Gati for his help piloting scanning parameters.

References

Adank, P., 2012. The neural bases of difficult speech comprehension and speech production: two Activation Likelihood Estimation (ALE) meta-analyses. Brain Lang. 122, 42–54.

Agus TR, Paquette S, Suied C, Pressnitzer D, Belin P. 2017. Voice selectivity in the temporal voice area despite matched low- level acoustic cues. 1–7.

Alain, C, Du, Y, Bernstein, LJ, Barten, T, Banai, K., 2018. Listening under difficult conditions: an activation likelihood estimation meta-analysis. Hum. Brain. Mapp. 39, 2695–2709.

Belin, P, Bestelmeyer, PEG, Latinus, M, Watson, R., 2011. Understanding voice perception. Br. J. Psychol. 102, 711–725.

Belin, P, Fecteau, S, Bédard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends. Cogn. Sci. 8, 129–135.

Bethmann, A, Brechmann, A., 2014. On the definition and interpretation of voice selective activation in the temporal cortex. Front. Hum. Neurosci. 8, 1–14.

Bethmann, A, Scheich, H, Brechmann, A., 2012. The temporal lobes differentiate between the voices of famous and unknown people: an event-related fMRI study on speaker recognition. PLoS One 7.
Birkett, PB, Hunter, MD, Parks, RW, Farrow, TF, Lowe, H, Wilkinson, ID, Woodruff, PW., 2007. Voice familiarity engages auditory cortex. Neuroreport 18, 1375–1378.
Boersma P, Weenink D. 2003. Praat: Doing phonetics by computer.
Bonte, M, Hausfeld, L, Scharke, W, Valente, G, Formisano, E., 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci. 34, 4548–4557.
Davis, MH, Ford, MA, Kherif, F, Johnsrude, IS., 2011. Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. J. Cogn. Neurosci. 23, 3914–3932.
Davis, MH, Johnsrude, IS., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.
Diedrichsen, J, Kriegeskorte, N., 2017. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput. Biol 13, e1005508.
Domingo, Y, Holmes, E, Johnsrude, IS., 2020. The benefit to speech intelligibility of hearing a familiar voice. J. Exp. Psychol. Appl. 26, 236–247.
Eickhoff, SB, Stephan, KE, Mohlberg, H, Grefkes, C, Fink, GR, Amunts, K, Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25, 1325–1335.
Eid, M, Gollwitzer, M, Schmitt, M., 2017. Statistik und forschungsmethoden.
Evans, S, McGettigan, C, Agnew, ZK, Rosen, S, Scott, SK., 2016. Getting the cocktail party started: masking effects in speech perception. J. Cogn. Neurosci. 28, 483–500.
Gander, PE, Kumar, S, Sedley, W, Nourski, KV, Oya, H, Kovach, CK, Kawasaki, H, Kikuchi, Y, Patterson, RD, Howard, MA, Griffiths, TD., 2019. Direct electrophysiological mapping of human pitch-related processing in auditory cortex. Neuroimage, 116076.
Goldinger, SD., 1998. Echoes of echoes? An episode theory of lexical access. Psychol. Rev. 105, 251–279.
Griffiths, TD, Büchel, C, Frackowiak, RSJ, Patterson, RD., 1998. Analysis of temporal structure in sound by the human brain. Nat. Neurosci. 1, 422–427.
Hall, DA, Haggard, MP, Akeroyd, MA, Palmer, AR, Summerfield, AQ, Elliott, MR, Gurney, EM, Bowtell, RW., 1999. Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7, 213–223.
Hautus, MJ., 1995. Corrections for extreme proportions and their biasing effects on estimated values of d′. Behav. Res. Method. Instrument. Comput. 27, 46–51.
Haxby, JV, 2012. Multivariate pattern analysis of fMRI: the early beginnings. Neuroimage 62, 852–855.
Holmes, E, Domingo, Y, Johnsrude, IS., 2018. Familiar voices are more intelligible, even if they are not recognized as familiar. Psychol. Sci. 29, 1575–1583.
Holmes, E, Johnsrude, IS., 2020. Speech spoken by familiar people is more resistant to interference by linguistically similar speech. J. Exp. Psychol. Learn Mem. Cogn. 46, 1465–1476.
Holmes, E, Zeidman, P, Friston, K, Griffiths, T., 2021. Difficulties with speech-in-noise perception related to fundamental grouping processes in auditory cortex. Cereb. Cortex.
Johnsrude, IS, Mackey, A, Hakyemez, H, Alexander, E, Trang, HP, Carlyon, RP, 2013. Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol. Sci. 24, 1995–2004.
Kaas, JH, Hackett, TA, Tramo, MJ., 1999. Auditory processing in primate cerebral cortex. Curr Opin Neurobiol 9, 164–170.
Kidd, G, Best, V, Mason, CR., 2008. Listening to every other word: examining the strength of linkage variables in forming streams of speech. J. Acoust. Soc. Am. 124, 3793–3802.
Kreitewolf, J, Mathias, SR, von Kriegstein, K., 2017. Implicit talker training improves comprehension of auditory speech in noise. Front. Psychol. 8, 1584.
Kriegeskorte, N, Mur, M, Bandettini, PA., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.
Kriegstein, K V, Giraud, A-L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948–955.
Kyong, JS, Scott, SK, Rosen, S, Howe, TB, Agnew, ZK, McGettigan, C., 2014. Exploring the roles of spectral detail and intonation contour in speech intelligibility: an fMRI study. J Cogn Neurosci 26 1748–1743.
Lachs L, McMichael K, Pisoni DB. 2003. Speech perception and implicit memory: evidence for detailed episodic encoding of phonetic events. In: Rethinking Implicit Memory. p. 215–235.
Lavner, Y, Rosenhouse, J, Gath, I., 2001. The prototype model in speaker identification by human listeners. Int. J. Speech Technol. 4, 63–74.
Levi, S V, Winters, SJ, Pisoni, DB., 2011. Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? J. Acoust. Soc. Am. 130, 4053–4062.
Medalla, M, Barbas, H., 2014. Specialized prefrontal "auditory fields": organization of primate prefrontal-temporal pathways. Front. Neurosci..
Morosan, P, Rademacher, J, Schleicher, A, Schormann, T, Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13, 684–701.
Mur, M, Bandettini, P a, Kriegeskorte, N., 2009. Revealing representational content with pattern-information fMRI–an introductory guide. Soc. Cogn. Affect. Neurosci. 4, 101–109.
Nili, H, Wingfield, C, Walther, A, Su, L, Marslen-Wilson, W, Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10, e1003553.
Nygaard, LC, Pisoni, DB., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 355–376.
Nygaard, LC, Sommers, MS, Pisoni, DB., 1994. Speech perception as a talker-contingent process. Psychol. Sci. 5, 42–46.
Peelle, JE, IS, Johnsrude, Davis, MH., 2010. Hierarchical processing for speech in human auditory cortex and beyond. Front. Hum. Neurosci. 4.
Pernet, CR, McAleer, P, Latinus, M, Gorgolewski, KJ, Charest, I, Bestelmeyer, PEG, Watson, RH, Fleming, D, Crabbe, F, Valdes-Sosa, M, Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 119, 164–174.
Perrachione, TK, Ghosh, SS., 2013. Optimized design and analysis of sparse-sampling fMRI experiments. Front. Neurosci. 7, 1–18.
Rabinowitz, NC, Willmore, BDB, Schnupp, JWH, King, AJ., 2011. Contrast gain control in auditory cortex. Neuron 70, 1178–1191.
Rutten, S, Santoro, R, Hervais-Adelman, AG, Formisano, E, Golestani, N., 2019. Cortical encoding of speech enhances task-relevant acoustic information. Nat Hum Behav 3.
Schwarzbauer, C, Davis, MH, Rodd, JM, Johnsrude, IS., 2006. Interleaved silent steady state (ISSS) imaging: a new sparse imaging method applied to auditory fMRI. Neuroimage 29, 774–782.
Scott, SK., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.
Scott, SK, Johnsrude, IS., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci..
Sohoglu, E, Peelle, JE, Carlyon, RP, Davis, MH., 2012. Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. 32, 8443–8453.
Spunt B. 2016. BSPMVIEW.
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. Cogn. Brain Res. 17, 48–55.
Wang, Y, Zhang, J, Zou, J, Luo, H, Ding, N., 2019. Prior knowledge guides speech segregation in human auditory cortex. Cereb. Cortex 29, 1561–1571.
Warren, JD, Scott, SK, Price, CJ, Griffiths, TD., 2006. Human brain mechanisms for the early analysis of voices. Neuroimage 31, 1389–1397.
Wild, CJ, Davis, MH, Johnsrude, IS., 2012. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage 60, 1490–1502.
Wild, CJ, Yusuf, A, Wilson, DE, Peelle, JE, Davis, MH, Johnsrude, IS., 2012. Effortful listening: the processing of degraded speech depends critically on attention. J. Neurosci. 32, 14010–14021.
Worsley, KJ, Evans, AC, Marrett, S, Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab 12, 900–918.
Yonan, CA, Sommers, MS., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol. Aging 15, 88–99.