Functional connectivity within the voice perception network and its behavioural relevance

Virginia Aglieri, Thierry Chaminade, Sylvain Takerkart, Pascal C Belin

NeuroImage 183 (2018) 356–365

Contents lists available at ScienceDirect

NeuroImage

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

Functional connectivity within the voice perception network and its behavioural relevance

Virginia Aglieri [a,*], Thierry Chaminade [a,b], Sylvain Takerkart [a,b], Pascal Belin [a,b,c]

[a] Institut des Neurosciences de la Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France
[b] Institute of Language, Communication and the Brain, Marseille, France
[c] International Laboratories for Brain, Music and Sound, Department of Psychology, Université de Montréal, McGill University, Montreal, QC, Canada

ARTICLE INFO

Keywords:
fMRI
Functional connectivity
Voice perception
Auditory cortex
Individual differences
Voice recognition

ABSTRACT

Recognizing who is speaking is a cognitive ability characterized by considerable individual differences, which could relate to the inter-individual variability observed in voice-elicited BOLD activity. Since voice perception is sustained by a complex brain network involving temporal voice areas (TVAs) and, even if less consistently, extra-temporal regions such as frontal cortices, functional connectivity (FC) during an fMRI voice localizer (passive listening of voices *vs* non-voices) has been computed within twelve temporal and frontal voice-sensitive regions ("voice patches") individually defined for each subject (N = 90) to account for inter-individual variability. Results revealed that voice patches were positively co-activated during voice listening and that they were characterized by different FC pattern depending on the location (anterior/posterior) and the hemisphere. Importantly, FC between right frontal and temporal voice patches was behaviorally relevant: FC significantly increased with voice recognition abilities as measured in a voice recognition test performed outside the scanner. Hence, this study highlights the importance of frontal regions in voice perception and it supports the idea that looking at FC between stimulus-specific and higher-order frontal regions can help understanding individual differences in processing social stimuli such as voices.

1. Introduction

Perceiving socially relevant stimuli such as human faces and voices is fundamental for social interactions and it is carried out in an automatic fashion for most of the population; however, this task requires different processing stages at the neural level. According to the "auditory face" model of cerebral voice processing (Belin et al., 2011, 2004; Blank et al., 2014), low-level visual and acoustic cues are first processed in subcortical structures and in primary visual/auditory areas; then, a finer structural analysis allows faces and voices to be detected and matched to internal templates; eventually, different types of higher-level information such as speech, emotion and speaker identity are processed in functionally dissociable, but interacting brain regions. Hence, thinking of face or voice perception as processes sustained by a stimulus-specific area does not account for this complex analysis and it is nowadays an outdated view. For instance, it is increasingly accepted that face perception involves a network of areas such as stimulus-specific regions, part of the "core face perception system" (e.g. face fusiform area, FFA), and regions which are not stimulus-specific but that still display significant sensitivity to faces, constituting an "extended face perception system" (e.g. anterior temporal lobe and prefrontal areas) (Castello et al., 2017; Haxby et al., 2000; Ishai, 2008; Tsao and Livingstone, 2008). In a similar way, voice perception can be associated to core regions of the voice perception network, located in the upper bank of the superior temporal gyrus/sulcus (Temporal Voice Areas, TVAs; Belin et al., 2000). Nevertheless, other areas have been found to show small but significant voice-specific activation when large number of subjects are included, namely several prefrontal regions and subcortical structures including the amygdalae, which could hence constitute the "extended" portion of the voice perception network (Pernet et al., 2015).

The study of structural and functional connectivity (FC) of the face perception network allowed clarifying why regions of the core and extended network are often co-activated (Fairhall and Ishai, 2007; Pyles et al., 2013; Turk-Browne et al., 2010). In the voice perception domain,

Abbreviations: FC, functional connectivity; STS, superior temporal sulcus; STG, superior temporal gyrus; IFG, inferior frontal gyrus; ROI, region of interest; TVA, temporal voice area; FVA, frontal voice area.

* Corresponding author.
E-mail addresses: virginia.aglieri@univ-amu.fr, vi.aglieri@gmail.com (V. Aglieri).

https://doi.org/10.1016/j.neuroimage.2018.08.011
Received 5 June 2018; Received in revised form 13 July 2018; Accepted 8 August 2018
Available online 9 August 2018

FC within the voice perception network remains quite under investigated and the available studies focused on clinical populations such as schizophrenia (Mou et al., 2013), on the relationship between face and voice recognition (von Kriegstein et al., 2005), and on speech perception (Osnes et al., 2011) and production (Flagmeier et al., 2014). Von Kriegstein and Giraud (2004) looked at FC during speaker recognition, finding that right TVA and ipsilateral parietal and prefrontal cortices were functionally coupled during such task. However, recognizing who is speaking is a complex process that goes beyond the understanding of the FC between core and extended voice areas during simple voice perception.

Hence, the first aim of the present study was to characterize FC within the voice perception network by looking at FC between defined regions of interest (ROI-to-ROI FC) during a functional magnetic resonance imaging (fMRI) task of passive voice listening. In particular, we identified three new bilateral voice-sensitive areas in prefrontal cortex, which we termed the "Frontal Voice Areas" (FVAs). As well, we defined three "voice patches" (named TVAs) distributed bilaterally along the antero-posterior axis of STS and STG, casting the results obtained by Pernet et al. (2015) through a cluster analysis of individual voice-selective peaks in a large ($n > 200$) cohort. Then, we investigated FC between the core temporal voice perception network (TVAs) and the extended frontal network (FVAs), as well as FC differences between right and left hemispheres and between anterior, middle and posterior TVAs.

Second, the study of FC could also reveal individual differences in person recognition abilities, as it has been previously observed in the domain of face perception (Avidan et al., 2014; Zhu et al., 2011). Even if less widely investigated than in the domain of face perception, there exists a considerable inter-individual variability in voice perception as well. For instance, the distribution of performances obtained at a simple voice recognition task (the Glasgow Voice Memory Test – GVMT; Aglieri et al. (2016)) spans from close to chance-level in subjects potentially affected by developmental phonagnosia (deficit in recognizing voice notwithstanding intact brain structures) to perfect performance in "super recognizers" of voices. This behavioural variability could have its neural correlates: Pernet et al. (2015) also observed high inter-individual variability in voice-specific BOLD activity (e.g. subjects with minimal activation within TVAs and subjects with distributed activation outside the TVAs). In favour of the existence of a relationship between individual differences at the neural and behavioural level, Watson et al. (2012) found that right TVA activation during auditory stimulation was modulated by voice recognition performance. However, this study did not find any significant correlation between voice recognition scores and voice-specific BOLD activity. This supports the hypothesis that since perception of social stimuli needs the concomitant activation of a distributed network, individual differences in person recognition abilities could become evident only at the level of structural (Thomas et al., 2009) and functional connectivity (Avidan et al., 2014; Zhu et al., 2011). The second aim of this study was hence to understand if inter-individual variability in voice recognition abilities assessed by the GVMT could have been reflected by different FC profiles within the voice perception network.

2. Methods

2.1. Subjects

92 subjects (58 females, mean age ± SD = 26.67 ± 10.41, [18–66]) with no previous history of mental illness and that self-reported normal audition were recruited for both the fMRI session (Voice Localizer; Belin et al., 2000) and a short behavioural assessment of voice recognition abilities (the Glasgow Voice Memory Test; Aglieri et al. (2016)). The two experimental sessions were carried out in different days, the order of the session being irrelevant. Participants have been recruited among students and personal at the University of Glasgow; no restrictions on mother tongue or handedness were applied. Prior to testing sessions, they all provided written informed consent, following the guidelines of the declaration of Helsinki.

2.2. The Glasgow voice Memory Test

The Glasgow voice Memory Test (GVMT) lasts 5 min and requires to memorize 8 voices (vowel/a/; 4 females and 4 males) and to immediately recognize them among a sequence of 16 voices (8 old/8 new). The same procedure is repeated for bell sounds, allowing to look for dissociations between the two different auditory processes. This test has been proven to be a valid method for a preliminary assessment of voice perception deficits since a developmental phonagnosic subject (KH (Garrido et al., 2009); was found to be impaired in voices but not in bells recognition (Aglieri et al., 2016). In the current study, 39 participants carried out the online version of the test (http://experiments.psy.gla.ac.uk/experiments/assessment.php?id=1270) and 53 were instead assessed in laboratory conditions through a version running on Media Control Functions (DigiVox, Montreal, Canada). Percentage of correct responses (PC) was computed for voices (PC voices) and bells recognition (PC bells), as well as the difference between PC for voices and bells (PC voices – PC bells). A Wilcoxon-rank sum test revealed no significant differences between the behavioural scores obtained by the online and laboratory samples (all $p > 0.05$), confirming previous results (Aglieri et al., 2016).

2.3. Voice localizer

The voice localizer is a 1-run block design lasting 10 min and 20 s which requires to passively listen to 20 blocks of vocal and 20 blocks of non-vocal sounds with eyes closed (Belin et al., 2000). Each stimulation block lasts 8s and it contains different short stimuli separated by at most 400 ms of no stimulation (further details on blocks content and exact stimuli duration can be found in Pernet et al. (2015)). Stimulation blocks are interleaved with 8s-long interstimulus intervals (ISI) presented in pseudo-random order to allow for the relaxation of hemodynamic response. Stimuli order is random but constant across subjects. About 60% of the stimuli were recorded appositely for this task and the rest was downloaded from public databases available in year 2000 or taken from recordings of American English vowels (Hillenbrand et al., 1995). Vocal blocks contain heterogeneous vocalizations obtained from 47 speakers (7 babies, 12 adults, 23 children and 5 elderlies) producing speech sounds (e.g. words, syllables or sentence extracts) and non-speech sounds (e.g. laughs, sighs, cries, neutral sounds like coughs, and onomatopoeias). Non-vocal blocks are made up of both natural sounds (e.g. falls, sea waves, wind, animal calls) and artificial sounds (e.g. cars, glass, alarms, clocks, and instrumental musical pieces). All stimuli (16 bit, mono, 22, 050 Hz sampling rate) were normalized through Root Mean Square, together with a 1-kHz tone used for volume calibration. Stimuli were presented using Media Control Functions (DigiVox, Montreal, Canada) via electrostatic headphones (NordicNeuroLab, Norway; or Sensimetrics, USA) at a comfortable level (80–85 dB Sound Pressure Level).

2.4. fMRI acquisition

Participants underwent the voice localizer in a 3 T Siemens (Erlangen, Germany) Tim Trio scanner at the Centre for Cognitive Neuroimaging, University of Glasgow. A 32-channel head coil was used for 25 subjects while the rest of the subjects were scanned with a 12-channel coil. For acquisition, a single-shot gradient-echo echo-planar imaging (EPI) sequence was used with the following parameters: 32 slices per volume, interleaved slices order, voxel size $3 \times 3 \times 3.3\ mm^3$, acquisition matrix 70×70, flip angle = 77°, echo time (TE) = 30 ms. The repetition time (TR) was 2 s as well as the acquisition time (TA), resulting in quasi-continuous scanning noise. For each participant, 310 EPI volumes were acquired together with a high-resolution 3D T1-weighted sagittal scan (voxel size $1\ mm^3$ isotropic; acquisition matrix $256 \times 256 \times 192$).

2.5. Preprocessing of fMRI data

SPM12b MRI was used to analyse data (r6080 — Wellcome Department of Cognitive Neurology, University College London). First, six parameters accounting for subjects' motion have been estimated (affine transformation) and used to realign the 310 EPI volumes to their mean. T1 images have been then co-registered to the mean of realigned EPIs through normalized mutual information. Furthermore, T1 images were segmented into their native space tissue components (gray and white matter, and CSF); these tissues were in turn iteratively co-registered to obtain a DARTEL template, which accounts for structural variability across subjects. These two steps were performed through the segmentation algorithm included in SPM12. This procedure uses diffeomorphic registration to preserve cortical topology applying a membrane bending energy or Laplacian model (Ashburner, 2007). At this step, normalization parameters (flow fields) were created and then used to normalize co-registered T1 images, realigned EPI images and the 3 native tissues to the DARTEL template and finally to the MNI space (affine registration). In the normalization step, a Gaussian smoothing kernel of 1 mm has been applied to all images to avoid aliasing. EPI images were further smoothed using a 4 mm Gaussian kernel.

2.6. Group analysis

Normalized EPI images resampled at $1.5 \times 1.5 \times 1.5\,mm^3$ and smoothed as previously described were used to build a design matrix for each subject. The design matrix was made up of nine regressors: two for stimulation blocks (one vocal (V), one non-vocal (NV)), six realignment parameters and one constant. Vocal and non-vocal regressors were obtained by convolving boxcar functions representing the onset and duration of stimulation blocks with the canonical hemodynamic response function (HRF). Before model estimation, each voxel's time course was filtered through a high-pass filter at 128 s and auto-correlation was modelled using an auto-regressive matrix of order 1. After model estimation, two different contrasts were computed: 1) V *vs* NV, 2) V + NV *vs* baseline. The first contrast was used to look at group results while the second one (all sounds *vs* baseline) served as control for individual subjects' supra-threshold responses – e.g. a subject was excluded if no supra-threshold voxel reached significance. To look at group results, the images of the first level contrast (V *vs* NV) of all 92 subjects were inserted in a one-sample *t*-test to assess where BOLD signal change was significantly different from zero ($p < 0.05$, corrected for family-wise error at the voxel level). Significant clusters were then anatomically labelled according both to the SPM Anatomy toolbox (Eickhoff et al., 2005) and to the Harvard-Oxford atlas (Desikan et al., 2006) since the two atlases gave slightly different results.

2.7. Definition of seed regions at the individual level

In order to account for the inter-individual variability in the localization of voice responsive regions, we defined a systematic procedure that was used separately on each subject to create a seed region (either called region of interest - ROI) within each of the "voice patches" identified at the group level. The term "voice patch" has been introduced by Pernet et al. (2015) to indicate three sub-regions of bilateral STS/STG (anterior, middle and posterior) showing preferential activation for voices compared to other sounds as well high test-retest reliability. Here, the term voice patches is used to indicate both the three sub-regions along the STS/STG equivalent to those ones identified in Pernet et al. (2015), but specific to our group analysis (TVAs; Fig. 1), and three bilateral frontal regions that showed supra-threshold activation for the contrast of interest: the frontal voice areas (FVAs; Fig. 1), further detailed in the results section. Within each of these voice patches, we identified individual peaks of the V vs. NV contrast. This was done in a slightly different way in the TVAs and the FVAs. In the frontal lobe, since the three regions were clearly separated at the group level, we simply chose the strongest peak (highest z-value) obtained at the individual level within each of the three clusters (peak coordinates of these three clusters are reported in Table 3). Since the test-retest reliability of these frontal patches remains to be confirmed, the reproducibility of these patches was tested as following: for five subjects randomly selected from our sample, either the first or second half of the voice localizer blocks was modelled (10 vocal and 10 non-vocal blocks each). For each of these two sub-datasets, individual peaks for V vs. NV contrast were then detected among the six frontal search-zones obtained at the group-level V vs. NV contrast (N = 92). This analysis in the five selected subjects showed that there was at least one peak in each of the six frontal search-zones in both the two "new" datasets. However, the Euclidean difference between the first peak in the two datasets spanned between 1.50 and 21.48 mm ($M \pm SD = 7.33 \pm 5.04$).

In the temporal lobe, the group-level analysis resulted in one large cluster in each hemisphere. In order to detect the three voice patches previously described in Pernet et al. (2015) but specific to our sample and to use them as seeds for functional connectivity, instead of arbitrarily dividing the temporal cluster in an anterior, middle and posterior cluster and look for individual peaks within each of these search-zones, the first ten peaks within the entire temporal cluster resulting from group-level analysis were first identified for each subject; among these peaks, the nearest ones in Euclidean distance to the voice patches coordinates (three for each hemisphere) reported in Pernet et al. (2015) were selected as temporal voice patches, for each subject. Coordinates of temporal voice patches specific to our sample are reported in Table 3.

Two subjects that did not show any peak within at least one of the regions defined as search-zones for individual peaks were excluded from further analysis.

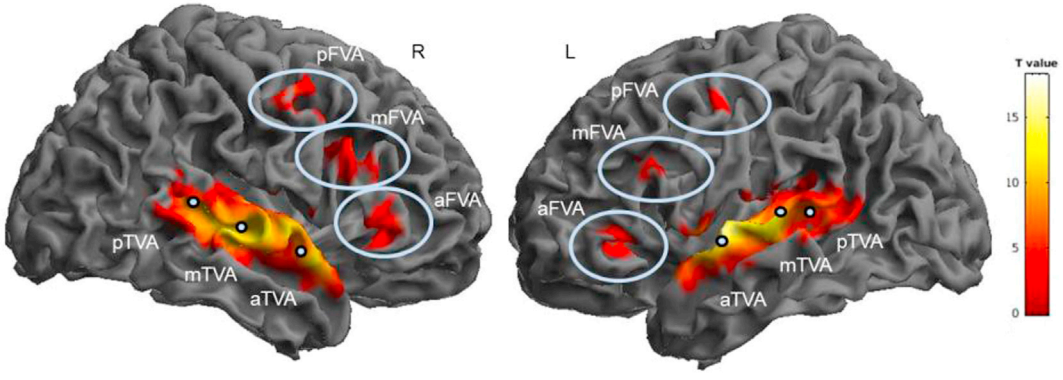

Fig. 1. Random effect analysis in 92 subjects. Suprathreshold clusters showing higher activation for the contrasts V *vs* NV (N = 92, $p < 0.05$ FWE voxel-level corrected, extent threshold = 0 mm^3). The black-light blue points represent approximate location of the six TVAs. The frontal clusters where individual FVAs were defined are instead in the light blue circles (a = anterior; m = middle; p = posterior).

This procedure yielded 12 peaks per subject (six FVAs and six TVAs). To define individual ROIs (or seeds), we selected the 100 voxels closest (in Euclidean distance) to the corresponding individual peak and located in the gray matter, as obtained from segmentation of normalized T1 image.

2.8. Functional connectivity analysis

Task-related ROI-to-ROI FC was analysed with the SPM toolbox CONN (Whitfield-Gabrieli and Nieto-Castanon, 2012). In doing so, 1st level SPM matrices of each subject (estimated on functional images pre-processed as previously described), T1 images, the segmentation of white matter, gray matter and CSF transformed in the DARTEL template and normalized to MNI space were first imported in CONN. The 12 individual subject's ROIs obtained as described above (six FVAs and six TVAs) were then selected as individual ROIs. Additionally, an aCompCorr denoising process (Behzadi et al., 2007) was applied to eliminate confounds of white matter, CSF, subject motion and effects of the task (voice, non-voice as well as rest). This procedure allows to remove the temporal time series of each confound from EPI images and to apply a band-pass filter ($0.01\ Hz < f < 0.10\ Hz$) to the residual time series. Adding regressors accounting for task effects in this preprocessing step allowed preventing that the main effects of the task drove the estimation of the correlations quantifying FC. By choosing the option "functional connectivity" (also referred to as "weighted GLM") offered in the CONN toolbox, each condition of interest was then described by a boxcar function and convolved with a canonical hemodynamic response function (HRF). For each subject, average time series weighted per condition were extracted across all voxels within each ROI and bivariate correlation coefficients between each pair of ROIs considered in isolation were computed; these coefficients were then Fisher-transformed at the group-level and a weight of 1 was assigned to vocal blocks and of −1 to non-vocal ones.

The scores obtained at the GVMT for voices, bells and the difference between percent correct for voices and bells were also added as 2nd level covariates, mean-centred to obtain a group mean of 0. To better characterize the TVAs, the differences in FC between anterior, middle and posterior TVAs were also investigated by first comparing bilateral seeds ("anterior-posterior gradient analysis"); if any result was present, FC was then computed for right and left seeds separately to find the hemisphere driving the effect. Possible differences in FC of left and right seeds of both FVAs and TVAs were investigated ("lateralization analysis"). Here, if some significant result was present, the effect of anterior, middle and posterior seed was tested separately.

To test the directional hypothesis that FC between frontal and temporal regions increased together with voice recognition abilities, as well as with the difference between voices and bells recognition, a one-sided (positive) hypothesis testing was applied to the correlation between behavioural and FC measures.

All group-level results were corrected for multiple comparisons (false discovery rate - FDR) at seed-level ($p < 0.05$) within the CONN toolbox. Seed-level correction is a parametric correction method offered in CONN that corrects for multiple comparisons arising from testing significance of FC between a seed and multiple target ROIs. In case of parametric correction, no family-wise error (FWE) correction is offered in the toolbox. All statistical tests, except the correlations between behavioural scores and ROI-to-ROI FC, were two-sided since they were not driven by any a-priori hypothesis.

3. Results

3.1. Glasgow voice Memory Test (GVMT)

The results obtained by the 90 subjects retained for FC analysis in recognition of voices, bells and the difference between these two scores are reported in Table 1. The Kolgomorov-Smirnoff test revealed that all score distributions violated normality. A Spearman correlation revealed that performance for voices and bells recognition were significantly correlated ($rho = 0.35$, $p < 0.001$).

3.2. Random-effect analysis

As suggested by Fig. 1, the second-level random-effects analysis of the contrast of interest (V > NV) resulted in an extended cluster of activation in bilateral STG/STS and in bilateral pre-frontal smaller clusters. As described above, the 12 ROIs (6 TVAs and 6 FVAs) were extracted from these clusters. However, other extra-temporal areas reached significance at p-FWE < 0.05 at the voxel level (Table 2).

3.3. ROI-to-ROI FC

3.3.1. Network

Voice-specific BOLD signal change in one ROI positively modulated the activity of the other ROIs; no negative modulation between the 12 ROIs was observed (Fig. 2 and Table 4). That is, the functional connectivity in the whole network of six TVAs and six FVAs increased for voices compared to non-vocal sounds. Five ROIs showed significantly positive FC values to all other ROIs: bilateral middle TVAs, bilateral posterior TVAs and the left posterior FVA. All voice patches had significant inter-hemispheric positive connections to homologue areas. Table 4 reports F-test values, resulting from the ROI-to-ROI FC analysis testing any effect between a seed and the matrix containing all target ROIs, and FC intensity, the sum of the significant *t*-test values assessing FC between a seed and a target. These two measures are informative of the FC strength of one seed relative to the whole network.

3.3.2. Gradient

Posterior TVAs showed significantly higher FC than anterior ones toward right posterior FVA (Fig. 3; $t\ (89) = 2.96$, *p*-FDR corrected two-sided < 0.05). This effect was not driven by left nor right TVAs.

3.3.3. Lateralization

When comparing left and right ROIs, the only significant result was found for the frontal regions: FC between the left anterior TVA and the FVAs was higher in the left compared to the right hemisphere (Fig. 3; $t\ (89) = 3.82$, *p*-FDR corrected two-sided < 0.001). This effect was driven by anterior FVAs (no significant differences were found when middle and posterior FVAs were selected separately).

3.3.4. Correlation with GVMT

There was a significant correlation between voice recognition percent correct and ROI-to-ROI FC between anterior and posterior right FVAs (Fig. 4; $t\ (89) = 2.83$, *p*-FDR-corrected (one-sided, positive) = 0.03), meaning that better scores for voice recognition were associated with higher FC between these two ROIs. When the difference between voices and bells scores was selected as 2nd level covariate, the significant correlation between anterior and posterior FVAs was still present ($t\ (89) = 2.88$, *p*-FDR corrected (two-sided) = 0.03, *p*-FDR corrected (one-sided, positive) = 0.015) but in addition this behavioural measure also correlated with FC between right posterior FVA and right posterior TVA ($t\ (89) = 2.89$, p-FDR corrected (two-sided) = 0.03, p-FDR corrected

Table 1
GVMT results. Range, means, standard deviations (SD) and 95% confidence intervals observed for the scores obtained in voices and bells recognition and for the difference between voices and bells recognition.

Score	Min	Max	Mean ± SD	95% CI	
Percent Correct voices (%)	37.5	100	78.12 ± 12.68	75.47	80.78
Percent Correct bells (%)	75	100	87.43 ± 9.39	85.46	89.40
Percent Correct voices – Percent Correct bells (%)	−37.5	0	−9.3 ± 12.71	−11.96	−6.64

Table 2
Random effect analysis results for the contrast V *vs* NV at $p < 0.05$ FWE voxel-level corrected, extent threshold = 10 mm^3. The 1st column of the table reports the MNI coordinates of local maxima (peaks separated by more than 8 mm); the 2nd column contains t-values estimating the contrast of interest (height threshold $t(1,91) = 5.27$, $p < 0.05$ FWE corrected); the 3rd column reports relative cluster size; the 4th column and 5th column contain the anatomical location of the clusters as labelled by the two different atlases.

Coordinates (x y z)	t	Cluster size	Labelling – Anatomy toolbox	Labelling – Harvard-Oxford atlas
−60 -8 -2	18.25	7772	Left STG	Left STG – anterior
60 -26 2	15.88	8085	Right STG	Right STG – posterior
−20 -9 -16	9.71	241	Left hippocampus	Left accumbens
48 18 24	7.69	1970	Right IFG triangularis	Right IFG opercularis
51 -2 48	8.84	452	Right precentral gyrus	Right precentral gyrus
−52 -8 48	8.08	168	Left postcentral gyrus	Left precentral gyrus
−39 27 -3	8	506	Left IFG orbitalis	Frontal orbital cortex
28 -2 -18	7.85	291	Right amygdala	Right accumbens
−48 16 21	7.49	491	Left IFG opercularis	Left IFG opercularis
24 0 8	7.46	150	Right putamen	Right pallidum
14 -14 9	7.26	168	Right prefrontal thalamus	Right caudate
−14 -26 -6	6.73	42	Left parietal thalamus	Left caudate
14 -27 -4	6.65	37	Right parietal thalamus	Right caudate
12 -3 10	6.21	72	Right temporal thalamus	Right caudate
−12 -4 10	5.99	49	Left prefrontal thalamus	Left caudate
−22 -6 6	5.78	16	Left pallidum	Left pallidum

Table 3
Voice patches localization. 1st column: name of the voice patch localized along the STG/STS (TVAs) and prefrontal regions (FVAs); 2nd column: MNI centre coordinates of the voice patches; 3rd column and 4th columns: anatomical labelling according to two different atlases.

Voice patch	Coordinates (x y z)	Labelling – Anatomy toolbox	Labelling – Harvard Oxford atlas
Left aTVA	−62 -4 0	STG	STG anterior division
Left mTVA	−66 -28 4	MTG	STG posterior division
Left pTVA	−58 -38 6	MTG	STG posterior division
Right aTVA	58 2 -8	Temporal pole	STG anterior division
Right mTVA	58 -20 -2	STG	STG posterior division
Right pTVA	50 -32 4	STG	STG posterior division
Left aFVA	−39 27 -3	IFG (orbitalis)	IFG (orbitalis)
Left mFVA	−48 16 21	IFG (opercularis)	IFG (opercularis)
Left pFVA	−52 -8 48	Postcentral gyrus	Precentral gyrus
Right aFVA	54 32 0	IFG (triangularis)	IFG (triangularis)
Right mFVA	48 18 24	IFG (triangularis)	IFG (opercularis)
Right pFVA	51 -2 48	Precentral gyrus	Precentral gyrus

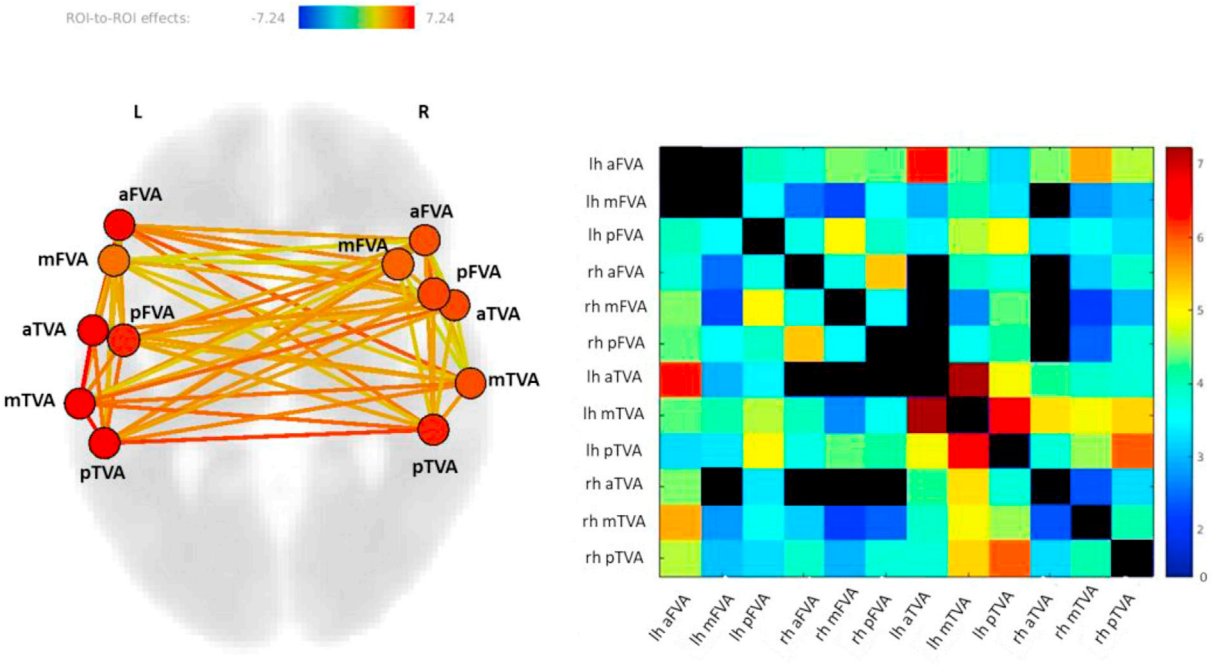

Fig. 2. Functional connectivity within the voice perception network. Axial view of the ROI-to-ROI FC (left; p-FDR seed-level corrected < 0.05) and connectivity matrix reporting t-values for the contrast V *vs* NV (right). The black squares in the connectivity matrix are used to visualize non-significant correlations at p-FDR seed-level corrected <0.05 or same-seed correlation. Note that both colorbars report *t*-test values between two seeds (DOF = 89), but the minimum value in the left colorbar is negative (−7.24) while in the connectivity matrix is 0. Lh = left hemisphere; rh = right hemisphere.

Table 4
ROI-to-ROI FC. F-test values (assessing any effect between a seed and the matrix containing all target ROIs), FC intensity (sum of significant *t*-test values expressing statistical significance of FC between two seeds) and number of ROIs showing significant FC with each ROI at p-FDR seed-level corrected < 0.05, ordered by F-values strength.

ROI	F (11, 79)	FC intensity	Number
left mTVA	8.41	52.7	11
left aTVA	8.25	36.82	8
left pTVA	6.55	49.86	11
right mTVA	5.24	39.38	11
right pTVA	4.98	43.87	11
right mFVA	4.95	31.12	9
left aFVA	4.8	45.23	10
right pFVA	4.74	34.57	9
rigth aTVA	4.38	26.6	7
left pFVA	4.15	43.28	11
right aFVA	3.81	35.97	10
left mFVA	2.75	28.04	9

(one-sided, positive) = 0.015) and with FC between right posterior FVA and left middle TVA (t (89) = 2.32, p-FDR corrected (one-sided, positive) = 0.04). ROI-to-ROI FC during V vs NV condition was not modulated in contrast by performance obtained at bells recognition (all p-FDR corrected > 0.05).

4. Discussion

4.1. Summary of results

The investigation of functional connectivity (FC) within the voice perception network, constituted by 3 frontal and 3 temporal seeds individually defined in each hemisphere based on group voice-specific activation (the so-called temporal and frontal "voice patches"), helped to better characterize the functional role of voice sensitive regions and their relation to behavior.

First, we showed that during voice perception there was a positive FC between the different seeds of the voice perception network; that is, FC between all pairs of regions of the voice network increases when hearing a vocal sound compared to a non-vocal sound. When comparing FC between the two hemispheres ("lateralization analysis"), we found that FVAs (in particular, anterior ones) and anterior TVA were more positively coupled in the left than right hemisphere. Looking instead at FC profile of the voice patches localized along the STS ("gradient analysis"), we found that bilateral posterior TVAs showed higher FC to right posterior FVA (located within the precentral gyrus) compared to anterior TVAs.

Importantly, we showed that ROI-to-ROI FC within the voice perception network can be behaviorally relevant: subjects better at voice recognition also had higher fronto-frontal FC (between posterior and anterior right FVAs) during perception of vocal compared to non-vocal sounds. The difference between voices and bells recognition resulted as well in a significant correlation with FC: subjects better at voices than bells recognition also showed higher fronto-frontal FC (again, between posterior and anterior right FVAs) but also fronto-temporal FC (between posterior right FVA and 1) left middle TVA and 2) right posterior TVA.

4.2. The extended system of voice perception: the frontal voice areas

Our study not only confirms previous observations on the involvement of extra-temporal regions (in particular, frontal ones) in voice perception processes in humans (Andics et al., 2010; Bestelmeyer et al., 2012; Charest et al., 2012; Fecteau et al., 2005; Latinus et al., 2011; Pernet et al., 2015; Zäske et al., 2017), but also shows that their activation positively covariates with voice-elicited BOLD response in temporal voice areas. Frontal cortices have previously been found to show activation related to socially salient (e.g. trustworthiness, gender) vocal cues (Bestelmeyer et al., 2012; Charest et al., 2012; Jones et al., 2015) and to process speaker identity, independently of verbal information (Latinus et al., 2011; Zäske et al., 2017). Furthermore, frontal regions such as the inferior frontal gyrus (IFG) are nowadays ascertained to be part of the face perception network (Duchaine and Yovel, 2015); notwithstanding, frontal regions are still considered to have a secondary role in voice perception as demonstrated by the fact that they are not consistently activated during voice-related tasks (Bonte et al., 2014). This could be due to the fact that there is a considerable anatomical and functional inter-subject variability in prefrontal regions (Juch et al., 2005), which can be overcome using large cohorts of subjects (Pernet et al., 2015) and, possibly, a preprocessing step involving normalization to a sample-specific template as performed in the present study. Importantly, the existence of individual peaks within the frontal search-zones even in new datasets created by splitting in half the voice localizer scans, points towards the reproducibility of the frontal patches in similar fMRI designs. However, further investigation is needed in order to clarify the anatomical variability of frontal patches across subjects.

The frontal voice-sensitive regions that we observed by contrasting vocal and non-vocal sounds were bilaterally located and occupied different regions of the frontal cortices, detailed here below. According to the Harvard-Oxford atlas, the anterior FVA cluster had slightly different locations in the two hemispheres: the left seed occupied the pars orbitalis of the IFG (BA 47), while the right one was located in the pars triangularis of the IFG (BA 45). As for the middle and posterior FVAs, their anatomical location was comparable in the two hemispheres (even if posterior FVA

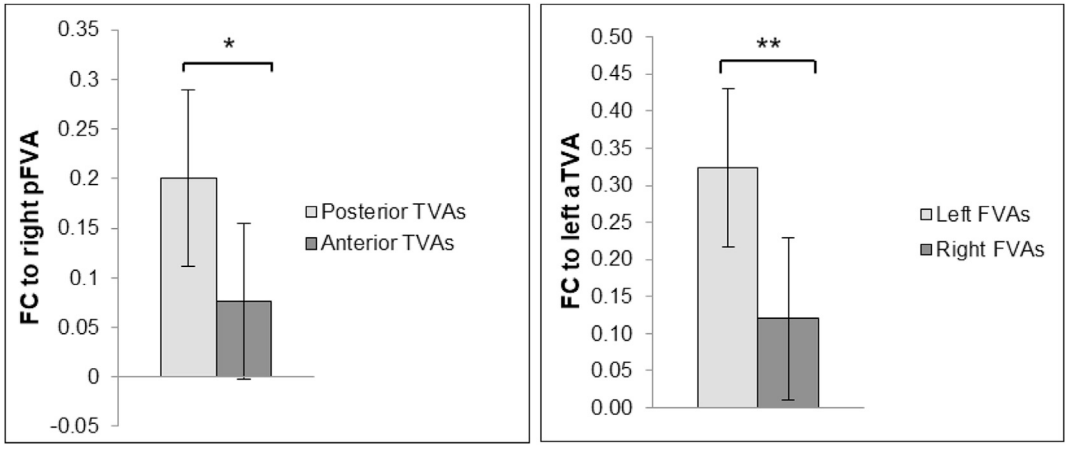

Fig. 3. Bar graphs representing differences in ROI-to-ROI FC between posterior and anterior TVAs (left graph) and between left and right FVAs (right graph). Error bars represent 95% confidence intervals (*p-FDR corrected (two-sided) < 0.05; **p-FDR corrected (two-sided) < 0.001).

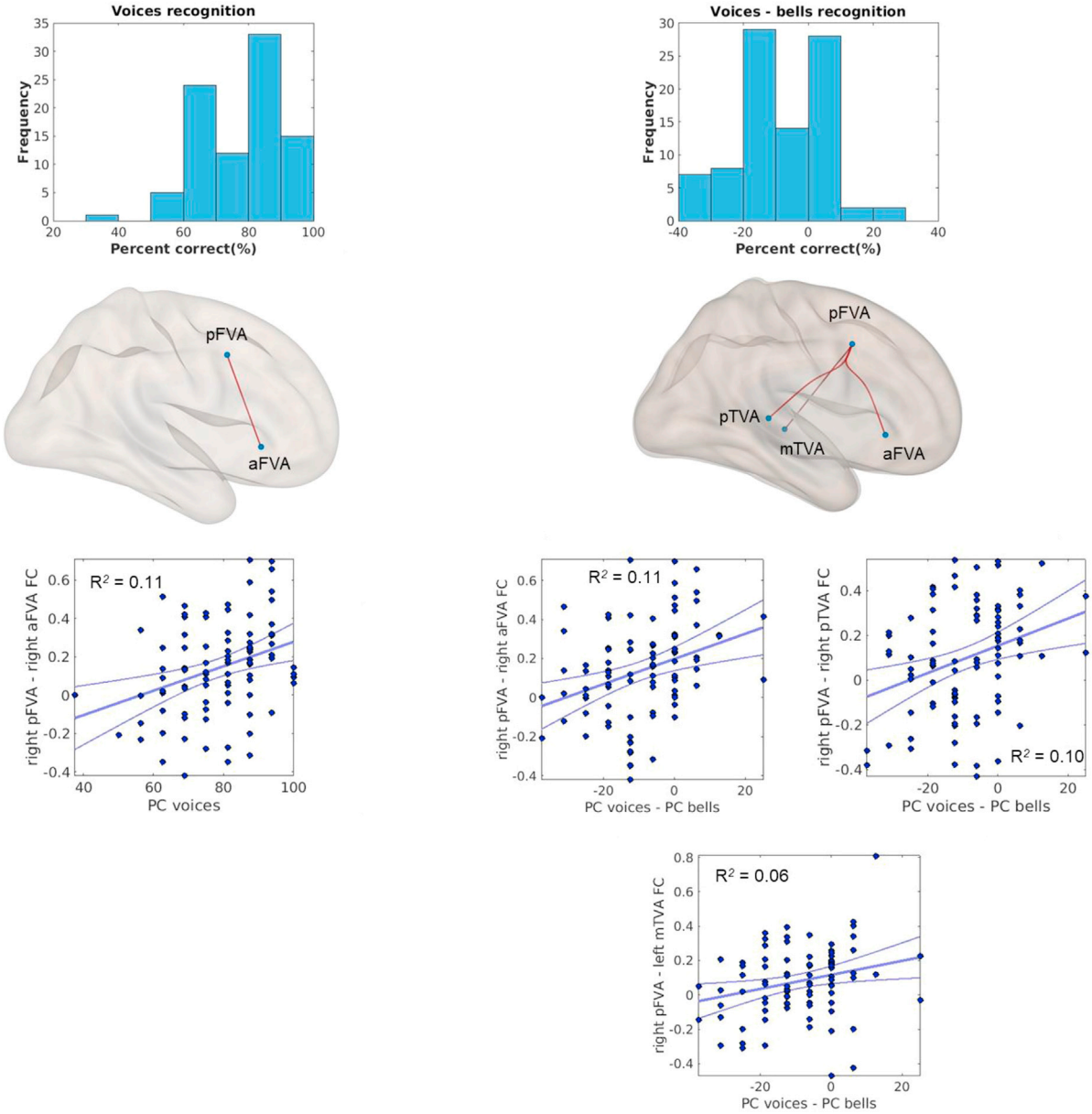

Fig. 4. Correlation between FC and voice recognition scores. 1st row: distribution of percent correct responses for voices (left) and of the difference between voices and bells recognition (right); 2nd row: illustration of the FC between regions showing significant correlation with voice recognition scores; 3rd/4th row: scatterplots of the significant correlations between ROI-to-ROI FC and the two different scores (R^2 = coefficient of determination of Pearson correlation coefficient).

was slightly more posterior in the left hemisphere), being middle FVAs in the IFG opercularis (BA 44) and posterior FVAs in the precentral gyrus (BA 6). In the left hemisphere, IFG has been most often associated to language processing (Hagoort, 2005), while the functional profile of its rightward counterpart remains much less understood. What is known is that the IFG portions (triangularis, orbitalis and opercularis) have distinct functional properties (Paulesu et al., 1997). A recent meta-analytic study concluded that the orbital IFG in the left hemisphere processes both semantic and emotional information (Belyk et al., 2017). The right IFG triangularis could have a role in carrying out analysis of prosody and pitch, hence functions tight to vocal perception (Rota et al., 2009; Zatorre et al., 1992). The pars opercularis of IFG (middle FVA) seems instead to belong to the dorsal stream of speech perception (Erickson et al., 2017) and furthermore, it has been associated to social cognition processes (Hamzei et al., 2016; Shamay-Tsoory et al., 2009). As for the precentral gyrus, previous evidences underscored its involvement in both speech perception and production (Cheung et al., 2016; Pulvermüller et al., 2006; Wilson et al., 2004). Therefore, these frontal regions that are mostly known for their role in linguistic processing could constitute the extended system of the voice perception network. The extended network of face perception includes areas not strictly visual and it is thought to process variable aspects of facial stimuli such as emotional expression and viewpoint (Haxby et al., 2000); likewise, it could be hypothesized that frontal areas encode those variable aspects of voice such as prosody and pitch. However, further research is needed to clarify their precise role in voice perception.

4.3. *Functional connectivity within the voice perception network*

The analysis of task-related FC (voice *vs* non-voice passive listening) revealed that all seeds showed positive bivariate correlation values within the voice perception network. Importantly, this increase in FC was not ascribed to a change in signal to noise ratio (SNR) because task-related activity was modelled together with ROIs time course, similarly

to what it is done in psychophysiological interaction (PPI; (Friston, 2011; O'Reilly et al., 2012). More specifically, only bilateral middle and posterior TVAs and posterior left FVA showed significantly positive FC to all other TVAs and FVAs. Anatomically, these TVAs are located in the posterior portion of the STS, a region considered to encode complex temporal cues of acoustical stimuli (von Kriegstein and Giraud, 2004), as well as a norm-based acoustical voice space (Andics et al., 2013). The observed positive FC between temporal voice patches located in the posterior STS and all other voice patches could then highlight the importance of the posterior STS as a functional hub within the voice perception network, in accordance with its structural and functional connectivity profile during speech perception (Saur et al., 2010). The posterior left FVA is located instead in the precentral gyrus, a region that we have seen to be activated in both speech production and perception and more precisely, near the larynx motor area (Brown et al., 2007). However, since our analysis gave no information on the directionality of functional connectivity, a top-down influence exerted by regions involved in voice perception on areas encoding larynx movements during passive listening can only be hypothesized.

The finding of FC between frontal regions (in particular, anterior ones) and anterior TVA stronger in the left than right hemisphere could be explained in terms of a correspondence between anatomical and functional connectivity (Saygin et al., 2012). This difference could in fact be ascribed to the structure of the arcuate fasciculus, a white matter association tract belonging to the superior longitudinal fasciculus which connects frontal, temporal and parietal regions; this structure is considered to have higher fiber density in the left than the right hemisphere (Glasser and Rilling, 2008; Nucifora et al., 2005; Vernooij et al., 2007) and to have a pivotal role in speech perception processes (Bernal and Ardila, 2009; Sarubbo et al., 2015).

The observation of different FC profiles of anterior and posterior TVAs contributes instead to the debate around the functional heterogeneity of the STS (and its sub-divisions), considered to be a multi-modal hub for socially relevant processes such as face perception (in particular eye gaze), integration of auditory and visual stimuli and action perception (Allison et al., 2000; Hein and Knight, 2008). Our finding is in line with the results reported in a meta-analysis on structural and functional connectivity of the STS (Erickson et al., 2017); according to these authors, anterior portions of the STS show in fact more co-activation with regions of the ventral auditory pathway (e.g. IFG orbitalis), while most posterior regions of the STS share more connections with dorsal stream areas (e.g. precentral gyrus, supplementary motor area and inferior parietal lobule) which have a major role in speech production (Saur et al., 2008). It is also worth mention that the posterior portion of the STS and the precentral gyrus are part of a fronto-temporal network supporting perception of multisensory social categories (Lahnakoski et al., 2012); hence, the finding of higher FC between posterior STS (compared to anterior STS) and right posterior FVA during voice perception could reflect the importance of these regions in perception of salient social stimuli such as voices.

4.4. Behavioural relevance of FC in voice perception

Our results show that voice recognition abilities as assessed through an off-line short unfamiliar voice recognition test (GVMT (Aglieri et al., 2016); correlated with fronto-frontal and fronto-temporal FC measured during voice *vs* non-voice perception. Even if this correlation explained only the 10% of the total variance, we believe it to be relevant because the relationship between individual differences in voice perception processes at the behavioural and neural levels remains under investigated, in particular when compared to the face perception domain (Wang et al., 2012; Zhu et al., 2011). To our knowledge, there are only few studies that investigated this relationship in the voice perception domain by looking at voice-elicited BOLD response (hence, disregarding FC) during different tasks. The first of these studies employed Positron Emission Tomography (PET) to look at regional cerebral blood flow associated to discrimination of familiar and unfamiliar speakers, finding that activity in right temporal and left frontal poles increased with better performance at familiar voices recognition (Nakamura et al., 2001). However, recognizing familiar voices is a complex process that does not allow separating mnemonic from acoustic perceptual processes (Kreiman and Sidtis, 2011). The second study observed a covariation of voice-related activity in middle/posterior left STS and voice identification performance during an identification task (Andics et al., 2010). Bonte et al. (2014) found a significant brain-behavior correlation by using a machine learning approach: here, speaker classification accuracy of voxels in the left posterior STG correlated with identification abilities of learned voices. It must be noted, however, that these two studies used an identification task performed on previously learned voices, not a simple perception task. When using a simple voice perception task, Watson et al. (2012) observed that voice recognition abilities correlated with BOLD response elicited by undifferentiated sounds but not with voice-specific activity. As such, voice-induced BOLD activity could be insufficient to predict voice recognition abilities. Rather, we here demonstrated that ROI-to-ROI FC within the voice perception network can be behaviorally relevant: FC between right posterior FVA and 1) right anterior FVA, 2) right posterior TVA, 3) left middle TVA increased with the difference between voices and bells recognition scores obtained in a short unfamiliar voice recognition test performed outside the scanner. This means that subjects better at voices than bells recognition also had stronger fronto-temporal FC or fronto-frontal FC during vocal compared to non-vocal perception.

According to our results, it seems that right posterior FVA and in particular its FC to bilateral posterior STS and right IFG can have an important role in predicting voice recognition abilities outside the scanner. As reported above, this region is involved in both speech perception and production (Cheung et al., 2016; Pulvermüller et al., 2006; Wilson et al., 2004) and importantly, in the right hemisphere it could have a specific role in encoding speakers' identity (Blank et al., 2014; Latinus et al., 2011). Its relevance in voice perception has also been confirmed by its preferential activation for prosody *vs* phoneme processing (Meyer et al., 2002; Sammler et al., 2015). Other studies highlighted its functional role in attentional mechanisms during auditory perception (Michalka et al., 2015) and working memory for tones (Koelsch et al., 2009). Furthermore, as reported above, this area, together with posterior STS, is part of a network sustaining undifferentiated social stimuli perception (Lahnakoski et al., 2012); hence, it could be hypothesized that the strength of FC between these two regions could be associated not only to individual differences in voice recognition abilities, as demonstrated by our results obtained during a voice perception task, but in social stimuli processing more generally.

Subjects with higher scores for voices than for bells recognition also had higher voice-specific FC between right posterior and anterior FVAs, this last region having a role in prosody perception (Rota et al., 2009). However, since the FC between these two frontal areas also showed positive correlation with voice recognition scores, we believe it to be less specific than the correlation with fronto-temporal FC. It is in fact possible that high scores in voice recognition are associated to high scores in bells recognition. Indeed, these scores were highly correlated, possibly mirroring common cognitive processes; it is hence only by dissociating them that we can get a specific measure of voice recognition abilities.

For interpreting these results, it is important to look at motor and sensorimotor accounts of speech perception, in particular at the action-perception integration model (Schomers and Pulvermüller, 2016). According to this model, speech perception involves a strong inter-communication of auditory and motor regions involved in speech production and phonological rehearsal, as confirmed by the observation of improved speech discrimination accuracy when seeing tongue movements (D'Ausilio et al., 2014). Therefore, it could be possible that FC between premotor cortex, in particular the region encoding for larynx and mouth movement, and regions such as IFG and STS, involved in auditory and phonological processing, could facilitate voice recognition

abilities by allowing multi-modal integration of the different information carried by human voices (e.g. mentally rehearsing vocal apparatus movements could fine-tune voice perception). In support of this explanation we can also cite cases of dyslexia (a language disorder in which phonological processing is impaired) characterized by impaired unfamiliar voice recognition (Perea et al., 2014; Stevenage, 2017).

Finally, this is the first study documenting a relationship between FC and behavioural performance in voice perception in healthy subjects; yet, a positive correlation between right fronto-temporal FC and voice recognition abilities has been previously observed in the clinical domain, namely in schizophrenic patients characterized by auditory hallucinations (Mou et al., 2013). The importance of looking at FC during voice perception is also supported by a recent study that found that fronto-temporal FC during mother's voice perception could predict future social communication skills as measured by a standardized scale (Abrams et al., 2016). Even if a systematic comparison between these studies and ours is not possible since we are in presence of different populations, these very similar findings can confirm the importance of looking at FC (in particular between regions of the core and extended voice perception network) when investigating individual differences in social processes such as voice perception.

Conflicts of interest

None.

Acknowledgements

Funding: This work was supported by grant AJE201214 from French Foundation for Medical Research, and grants ANR-16-CONV-0002 (Institute of Language, Communication and the Brain), ANR-11-LABX-0036 (Brain and Language Research Institute), BBSRC grants BB/E003958/1, BBJ003654/1 and BB/I022287/1, ESRC-MRC large grant RES-060-25-0010 and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

We would like to thank the two anonymous reviewers since their useful comments considerably contributed to improve the quality of this work.

References

Abrams, D.A., Chen, T., Odriozola, P., Cheng, K.M., Baker, A.E., Padmanabhan, A., Ryali, S., Kochalka, J., Feinstein, C., Menon, V., 2016. Neural circuits underlying mother's voice perception predict social communication abilities in children. Proc. Natl. Acad. Sci. U. S. A. 113, 6295–6300. https://doi.org/10.1073/pnas.1602948113.

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., Belin, P., 2016. The Glasgow Voice Memory Test: assessing the ability to memorize and recognize unfamiliar voices. Behav. Res. Meth. 1–14.

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cognit. Sci. 4, 267–278.

Andics, A., McQueen, J.M., Petersson, K.M., 2013. Mean-based neural coding of voices. Neuroimage 79, 351–360.

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage 52, 1528–1540.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007.

Avidan, G., Tanzer, M., Hadj-Bouziane, F., Liu, N., Ungerleider, L.G., Behrmann, M., 2014. Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. Cerebr. Cortex 24, 1565–1578.

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37, 90–101.

Belin, P., Bestelmeyer, P.E., 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 Cognit. Sci. 8, 129–135.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Belyk, M., Brown, S., Lim, J., Kotz, S.A., 2017. Convergence of semantics and emotional expression within the IFG pars orbitalis. Neuroimage 156, 240–248. https://doi.org/10.1016/j.neuroimage.2017.04.020.

Bernal, B., Ardila, A., 2009. The role of the arcuate fasciculus in conduction aphasia. Brain 132, 2309–2316. https://doi.org/10.1093/brain/awp206.

Bestelmeyer, P.E.G., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., Belin, P., 2012. Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. Cerebr. Cortex 22, 1263–1270. https://doi.org/10.1093/cercor/bhr204.

Blank, H., Wieland, N., von Kriegstein, K., 2014. Person recognition and the brain: merging evidence from patients and healthy individuals. Neurosci. Biobehav. Rev. 47, 717–734.

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.

Brown, S., Ngan, E., Liotti, M., 2007. A larynx area in the human motor cortex. Cerebr. Cortex 18, 837–845.

Castello, M.V. di O., Halchenko, Y.O., Guntupalli, J.S., Gors, J.D., Gobbini, M.I., 2017. The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. Sci. Rep. 7, 12237. https://doi.org/10.1038/s41598-017-12559-1.

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2012. Cerebral processing of voice gender studied using a continuous carryover fMRI design. Cerebr. Cortex 23, 958–966.

Cheung, C., Hamilton, L.S., Johnson, K., Chang, E.F., 2016. The auditory representation of speech sounds in human motor cortex. eLife Sci. 5, e12577. https://doi.org/10.7554/eLife.12577.

D'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J.J., Fadiga, L., 2014. Vision of tongue movements bias auditory speech perception. Neuropsychologia 63, 85–91.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Duchaine, B., Yovel, G., 2015. A revised neural framework for face processing. Annu. Rev. Vis. Sci. 1, 393–416. https://doi.org/10.1146/annurev-vision-082114-035518.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25, 1325–1335.

Erickson, L.C., Rauschecker, J.P., Turkeltaub, P.E., 2017. Meta-analytic connectivity modeling of the human superior temporal sulcus. Brain Struct. Funct. 222, 267–285.

Fairhall, S.L., Ishai, A., 2007. Effective connectivity within the distributed cortical network for face perception. Cerebr. Cortex 17, 2400–2406. https://doi.org/10.1093/cercor/bhl148.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Sensitivity to voice in human prefrontal cortex. J. Neurophysiol. 94, 2251–2254.

Flagmeier, S.G., Ray, K.L., Parkinson, A.L., Li, K., Vargas, R., Price, L.R., Laird, A.R., Larson, C.R., Robin, D.A., 2014. The neural changes in connectivity of the voice network during voice pitch perturbation. Brain Lang. 132, 7–13. https://doi.org/10.1016/j.bandl.2014.02.001.

Friston, K.J., 2011. Functional and effective connectivity: a review. Brain Connect. 1, 13–36.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: a selective deficit of vocal identity recognition. Neuropsychologia 47, 123–131.

Glasser, M.F., Rilling, J.K., 2008. DTI tractography of the human brain's language pathways. Cerebr. Cortex 18, 2471–2482.

Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends Cognit. Sci. 9, 416–423. https://doi.org/10.1016/j.tics.2005.07.004.

Hamzei, F., Vry, M.-S., Saur, D., Glauche, V., Hoeren, M., Mader, I., Weiller, C., Rijntjes, M., 2016. The dual-loop model and the human mirror neuron system: an exploratory combined fMRI and DTI study of the inferior frontal gyrus. Cerebr. Cortex 26, 2215–2224. https://doi.org/10.1093/cercor/bhv066.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cognit. Sci. 4, 223–233.

Hein, G., Knight, R.T., 2008. Superior temporal sulcus—it's my area: or is it? J. Cognit. Neurosci. 20, 2125–2136.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099–3111.

Ishai, A., 2008. Let's face it: it's a cortical network. Neuroimage 40, 415–419. https://doi.org/10.1016/j.neuroimage.2007.10.040.

Jones, A.B., Farrall, A.J., Belin, P., Pernet, C.R., 2015. Hemispheric association and dissociation of voice and speech information processing in stroke. Cortex 71, 232–239.

Juch, H., Zimine, I., Seghier, M.L., Lazeyras, F., Fasel, J.H.D., 2005. Anatomical variability of the lateral frontal lobe surface: implication for intersubject variability in language neuroimaging. Neuroimage 24, 504–514. https://doi.org/10.1016/j.neuroimage.2004.08.037.

Koelsch, S., Schulze, K., Sammler, D., Fritz, T., Müller, K., Gruber, O., 2009. Functional architecture of verbal and tonal working memory: an FMRI study. Hum. Brain Mapp. 30, 859–873. https://doi.org/10.1002/hbm.20550.

Kreiman, J., Sidtis, D., 2011. Foundations of Voice Studies: an Interdisciplinary Approach to Voice Production and Perception. John Wiley & Sons.

Lahnakoski, J.M., Glerean, E., Salmi, J., Jääskeläinen, I.P., Sams, M., Hari, R., Nummenmaa, L., 2012. Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. Front. Hum. Neurosci. 6. https://doi.org/10.3389/fnhum.2012.00233.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cerebr. Cortex 21, 2820–2828.

Meyer, M., Alter, K., Friederici, A.D., Lohmann, G., von Cramon, D.Y., 2002. FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. Hum. Brain Mapp. 17, 73–88.

Michalka, S.W., Kong, L., Rosen, M.L., Shinn-Cunningham, B.G., Somers, D.C., 2015. Short-term memory for space and time flexibly recruit complementary sensory-biased frontal lobe attention networks. Neuron 87, 882–892.

Mou, X., Bai, F., Xie, C., Shi, J., Yao, Z., Hao, G., Chen, N., Zhang, Z., 2013. Voice recognition and altered connectivity in schizophrenic patients with auditory hallucinations. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 44, 265–270. https://doi.org/10.1016/j.pnpbp.2013.03.006.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047–1054. https://doi.org/10.1016/S0028-3932(01)00037-9.

Nucifora, P.G., Verma, R., Melhem, E.R., Gur, R.E., Gur, R.C., 2005. Leftward asymmetry in relative fiber density of the arcuate fasciculus. Neuroreport 16, 791–794.

O'Reilly, J.X., Woolrich, M.W., Behrens, T.E., Smith, S.M., Johansen-Berg, H., 2012. Tools of the trade: psychophysiological interactions and functional connectivity. Soc. Cognit. Affect Neurosci. 7, 604–609.

Osnes, B., Hugdahl, K., Specht, K., 2011. Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. Neuroimage 54, 2437–2445. https://doi.org/10.1016/j.neuroimage.2010.09.078.

Paulesu, E., Goldacre, B., Scifo, P., Cappa, S.F., Gilardi, M.C., Castiglioni, I., Perani, D., Fazio, F., 1997. Functional heterogeneity of left inferior frontal cortex as revealed by fMRI. Neuroreport 8, 2011–2016.

Perea, M., Jiménez, M., Suárez-Coalla, P., Fernández, N., Viña, C., Cuetos, F., 2014. Ability for voice recognition is a marker for dyslexia in children. Exp. Psychol. 61, 480–487. https://doi.org/10.1027/1618-3169/a000265.

Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 164–174.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A. 103, 7865–7870. https://doi.org/10.1073/pnas.0509989103.

Pyles, J.A., Verstynen, T.D., Schneider, W., Tarr, M.J., 2013. Explicating the face perception network with white matter connectivity. PLoS One 8 e61611. https://doi.org/10.1371/journal.pone.0061611.

Rota, G., Sitaram, R., Veit, R., Erb, M., Weiskopf, N., Dogil, G., Birbaumer, N., 2009. Self-regulation of regional cortical activity using real-time fMRI: the right inferior frontal gyrus and linguistic processing. Hum. Brain Mapp. 30, 1605–1614. https://doi.org/10.1002/hbm.20621.

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P.E., Belin, P., 2015. Dorsal and ventral pathways for prosody. Curr. Biol. 25, 3079–3085.

Sarubbo, S., De Benedictis, A., Merler, S., Mandonnet, E., Balbi, S., Granieri, E., Duffau, H., 2015. Towards a functional atlas of human white matter. Hum. Brain Mapp. 36, 3117–3136.

Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., Weiller, C., 2008. Ventral and dorsal pathways for language. Proc. Natl. Acad. Sci. Unit. States Am. 105, 18035–18040. https://doi.org/10.1073/pnas.0805234105.

Saur, D., Schelter, B., Schnell, S., Kratochvil, D., Küpper, H., Kellmeyer, P., Kümmerer, D., Klöppel, S., Glauche, V., Lange, R., Mader, W., Feess, D., Timmer, J., Weiller, C., 2010. Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension. Neuroimage 49, 3187–3197. https://doi.org/10.1016/j.neuroimage.2009.11.009.

Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D., Saxe, R.R., 2012. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. Nat. Neurosci. 15, 321–327.

Schomers, M.R., Pulvermüller, F., 2016. Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. Front. Hum. Neurosci. 10. https://doi.org/10.3389/fnhum.2016.00435.

Shamay-Tsoory, S.G., Aharon-Peretz, J., Perry, D., 2009. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain 132, 617–627. https://doi.org/10.1093/brain/awn279.

Stevenage, S.V., 2017. Drawing a distinction between familiar and unfamiliar voice processing: a review of neuropsychological, clinical and empirical findings. Neuropsychologia. https://doi.org/10.1016/j.neuropsychologia.2017.07.005.

Thomas, C., Avidan, G., Humphreys, K., Jung, K., Gao, F., Behrmann, M., 2009. Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nat. Neurosci. 12, 29–31. https://doi.org/10.1038/nn.2224.

Tsao, D.Y., Livingstone, M.S., 2008. Mechanisms of face perception. Annu. Rev. Neurosci. 31, 411–437.

Turk-Browne, N.B., Norman-Haignere, S.V., McCarthy, G., 2010. Face-specific resting functional connectivity between the fusiform gyrus and posterior superior temporal sulcus. Front. Hum. Neurosci. 4. https://doi.org/10.3389/fnhum.2010.00176.

Vernooij, M.W., Smits, M., Wielopolski, P.A., Houston, G.C., Krestin, G.P., van der Lugt, A., 2007. Fiber density asymmetry of the arcuate fasciculus in relation to functional hemispheric language lateralization in both right-and left-handed healthy subjects: a combined fMRI and DTI study. Neuroimage 35, 1064–1076.

von 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. https://doi.org/10.1016/j.neuroimage.2004.02.020.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.-L., 2005. Interaction of face and voice areas during speaker recognition. J. Cognit. Neurosci. 17, 367–376.

Wang, L., Saalmann, Y.B., Pinsk, M.A., Arcaro, M.J., Kastner, S., 2012. Electrophysiological low-frequency coherence and cross-frequency coupling contribute to BOLD connectivity. Neuron 76, 1010–1020. https://doi.org/10.1016/j.neuron.2012.09.033.

Watson, R., Latinus, M., Bestelmeyer, P.E., Crabbe, F., Belin, P., 2012. Sound-induced activity in voice-sensitive cortex predicts voice memory ability. Front. Psychol. 3.

Whitfield-Gabrieli, S., Nieto-Castanon, A., 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect. 2, 125–141. https://doi.org/10.1089/brain.2012.0073.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7, 701.

Zäske, R., Hasan, B.A.S., Belin, P., 2017. It doesn't matter what you say: FMRI correlates of voice learning and recognition independent of speech content. Cortex 94, 100–112.

Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A., 1992. Lateralization of phonetic and pitch discrimination in speech processing. Science 256, 846–849.

Zhu, Q., Zhang, J., Luo, Y.L.L., Dilks, D.D., Liu, J., 2011. Resting-state neural activity across face-selective cortical regions is behaviorally relevant. J. Neurosci. 31, 10323–10330. https://doi.org/10.1523/JNEUROSCI.0873-11.2011.