Learning-Induced Changes in the Cerebral Processing of Voice Identity

Marianne Latinus, F. Crabbe, P. Belin

Learning-induced changes in the cerebral processing of voice identity.

Marianne Latinus, Frances Crabbe and Pascal Belin.

[1] Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, G12 8QB, Glasgow, UK.

[2] International Laboratories for Brain, Music and Sound (BRAMS), McGill University and Université de Montréal, Montreal, QC, Canada.

Correspondence should be addressed to M.L. (marianne.latinus@glasgow.ac.uk)

Abstract

Temporal voice areas showing a larger activity for vocal than non-vocal sounds have been identified along the superior temporal sulcus (STS); more voice-sensitive areas have been described in frontal and parietal lobes. Yet, the role of voice-sensitive regions in representing voice identity remains unclear. Using an fMR-adaptation design, we aimed at disentangling acoustic- from identity-based representations of voices. Sixteen participants were scanned while listening to pairs of voices drawn from morphed continua between two initially unfamiliar voices, before and after a voice learning phase. In a given pair, the first and second stimuli could be identical or acoustically different and, at the second session, perceptually similar or different. At both sessions, right mid STS/STG and superior Temporal Pole (sTP) showed sensitivity to acoustical changes. Critically, voice learning induced changes in the acoustical processing of voices in inferior frontal cortices (IFC). At the second session only, right IFC and left cingulate gyrus showed sensitivity to changes in perceived identity. The processing of voice identity appears to be subserved by a large network of brain areas ranging from the sTP, involved in an acoustic-based representation of unfamiliar voices, to areas along the convexity of the IFC for identity-related processing of familiar voices.

Keywords: fMR-adaptation, Temporal Voice Area, Inferior Frontal Cortex, Superior Temporal Cortex, Voice Morphing

Recognition of individuals is an important ability for humans; it has tremendous biological significance in terms of social interaction. Processing of paralinguistic information of voices is a means of speaker identification and of other auditory derived-semantic information such as age, emotional state or gender (Belin et al. 2004). Voices are efficiently used to recognise individuals (Nakamura et al. 2001; Schweinberger et al. 1997); this skill is present in all normal adult listeners (Kreiman 1997; Papcun et al. 1989) from birth (DeCasper and Fifer 1980) with a long evolutionary history (Belin 2006; Charrier et al. 2001; Marchant-Forde et al. 2002).

Temporal voice areas (TVA) showing greater activity for vocal than for non-vocal sounds have been identified along the upper bank of the superior temporal sulcus (STS, superior temporal cortex – STC) (Belin et al. 2000; von Kriegstein et al. 2003; von Kriegstein et al. 2005). Among the TVA, only the anterior part of the right STS showed a voice-preferential response even for non-speech vocal stimuli (Belin et al. 2002). Since then, more voice-sensitive responses have been found in frontal cortices, particularly the bilateral orbitofrontal cortex, and in parietal cortices (Fecteau et al. 2005; Stevens 2004; Von Kriegstein and Giraud 2004; von Kriegstein et al. 2005). The role of the TVA remains unclear as they have been shown to be involved in high-level auditory processing such as identity processing, particularly the temporal pole, or/and in low-level acoustic processing (Andics et al. 2010). Areas outside the TVA, notably the inferior frontal cortex (IFC), also showed sensitivity to voice familiarity (Von Kriegstein and Giraud 2004). Most of the above studies investigated voice recognition using either acoustically variable stimuli or different tasks making it difficult to disentangle acoustic processing, top-down modulation and identity processing *per se*. However, the processing of acoustic information and that of identity are two stages of a unique processing stream (Belin *et al.* 2004). It is important to tease them apart to identify brain regions truly involved in the processing of vocal identity. In an attempt to disentangle acoustic

processing from identity processing, Andics et al. (2010), using morphed stimuli of learned unfamiliar voices, reported acoustic sensitive regions within bilateral STS and prefrontal cortices (PFC) whereas voice identity processing involved bilateral STS, bilateral anterior temporal pole, left amygdala and left posterior STS (Andics et al. 2010). Interestingly, bilateral STS showed sensitivity to both acoustical and identity processing; this activation could be due to the stimuli included in the identity contrast that also differed acoustically (Andics et al. 2010).

Here, we attempt to provide evidence for separable neural substrates involved in acoustic and identity processing of vocal information. To that aim we investigated the nature of cerebral voice representation in healthy young adults by measuring changes in brain activity associated with learning new voice identities. We compared activity evoked by pairs of vocal stimuli, before and after voice learning, to separate acoustical- and identity-based cerebral processing using similar stimuli and tasks. We used an fMRI adaptation design: pairs of stimuli in which one stimulus property is repeated are known to induce a decrease of the blood oxygenation level-dependant (BOLD) signal in brain areas sensitive to that particular property (Grill-Spector and Malach 2001; Henson et al. 2003; Rotshtein et al. 2005). Subjects learned to recognise three unfamiliar voices and were scanned before and after learning the voices using an identical paradigm. We presented pairs of stimuli drawn from identity continua generated by morphing between all three possible couplings of the unfamiliar voices. Three types of pairs were presented: SAME pairs consisted of a repetition of the same voice stimulus; WITHIN pairs comprised two different voices taken on the same side of an identity continuum, such that at session 2 (but not session 1), subjects would perceive the two stimuli as being similar in terms of identity despite their being physically different; BETWEEN pairs consisted of two different voices taken on different sides of an identity continuum, such that at both sessions, subjects would perceive the two stimuli as being different both physically and in terms

of identity (Figure 1). The physical distance between the two stimuli of a pair was similar (30%) for WITHIN and BETWEEN pairs. We expected brain areas sensitive to acoustical properties of the stimuli to show a decreased BOLD signal for SAME compared to BETWEEN and WITHIN pairs, because the former was the only pair with no acoustical changes (contrast: WITHIN + BETWEEN vs. SAME); these areas are expected to be similar in both sessions irrespective of learning. However, we hypothesized that brain regions sensitive to vocal identity would show, at the second session, a decreased response for WITHIN and SAME pairs but not for BETWEEN pairs due to the crossing of the identity boundary in BETWEEN pairs (contrast: BETWEEN vs. SAME + WITHIN, at session2); at the first session, the voices were unfamiliar, therefore we did not expect regions showing a different sensitivity to WITHIN and BETWEEN pairs.

Materials and methods

Participants

Sixteen participants (7 males, 22.7 years +/- 0.88) from the under- and post-graduate population of Glasgow University with no native language restrictions took part in the experiment (11 were native English speakers from Scotland or England, one of them was studying French as an undergraduate; 2 were native German speakers; 1 was a native French speaker; the last two were native Polish and Portuguese speakers). All subjects reported normal audition. All subjects gave informed written consent, they were paid at a standard rate of £6 per hour; the study was approved by the ethics committee of Glasgow University.

Stimuli

Voice samples were drawn from a database of French-Canadian voices (Baumann and Belin 2008). Stimuli used in the experiments were sustained French vowels (/a/, /é/, /è/, /o/, and /U/; duration of 670 ms) from three male French Canadian speakers (voice A, B and C). There is growing evidence for an interaction between language and speaker recognition (Perrachione and Wong 2007). Using stimuli not drawn from the participants' native language could have interfered to some extent with identity processing, despite the limited linguistic content of our stimuli (vowels), probably contributing to a greater variability in subjects performance. All the stimuli used in the experiment were normalised for energy (RMS) using *Matlab* (*The MathWorks, Inc., Natick, Massachusetts, USA*). Voice stimuli were unfamiliar to the subjects prior to voice learning sessions (cf. below) during which they learned to associate each voice sample with one of 3 identities. Subjects were scanned prior to and after the voice learning sessions.

The 3 voices were morphed with each other using STRAIGHT (Kawahara et al. 1999) in *Matlab* in order to create 3 voice identity continua (5% to 95% in 15% steps) per vowel, i.e. 15 continua. STRAIGHT decomposes voice stimuli into five parameters (fundamental frequency – f0, frequency structure corresponding mostly to formant frequencies, time, spectro-temporal density and aperiodicity) that can be manipulated independently of one another. Landmarks to be put in correspondence across voices were manually identified in each stimulus time-frequency space; they corresponded to the four first frequency bands with the highest energy at the start and end of each original vowel. Morphed stimuli were then re-synthesised based on the linear (time and aperiodicity) and logarithmic (f0, the frequency structure and spectro-temporal density) interpolation of those time-frequency landmarks. Each stimulus of a continuum between voices A and B was generated using different values of a weight parameter X allowing the creation of a morphed stimulus containing X percent of

information of voice A and 100-X percent of information of voice B. Values of X between 0% and 100% correspond to morphed stimuli intermediate between A and B.

Voice learning

After the first scanning session, subjects were familiarised with 3 voice identities using the following procedure. Voice samples presented during the training sessions consisted of 2 stories (one in English, one in French), as well as isolated words and vowels; a training session lasted for about 20 min. Note that our subjects did not necessarily understand French, however, as the vowels were French vowels and the task was to pay attention to identity and not speech, we believe that using a French story helped our subjects learn the voices. Stimuli were presented binaurally at a level of 80 dB via headphones (*Beyerdynamic DT770*) using MCF software (*Digivox; Montreal, QC, Canada*) in a sound-proof cabin. A training session comprised 3 parts.

(1) In the first part, subjects carefully listened to the two stories and learned to associate a name presented on the computer screen with a particular voice. The 3 names used were: Phil, Ian and Dave.
(2) The second part consisted of a 3 alternative forced-choice (3-AFC) identification task on words and vowels (in French and English); feedback was provided on subjects' answer and if an incorrect response was given, the sound and the correct answer were presented again.
(3) The third part was a test phase in which only vowels were presented and subjects performed the 3-AFC without feedback.

Subjects did one training session per day until their performance at the final 3-AFC task was above 66% (discrimination threshold in a 3-AFC task (Kingdom and Prins 2010)). On average, training lasted 6.4 days (range: 3-10), and performance at the final 3-AFC task was

85% (chance level of 33%). One male subject did not reach the critical threshold of 66% in identification task even after 10 sessions; he was thus removed from all further analysis except for the voice localiser analysis.

After the first and last training sessions, subjects performed three 2-AFC identification tasks on stimuli drawn from the 3 different voice identity continua (A-B, A-C and B-C; Figure 1A). Results of the 2-AFC classification task performed at the end of the first and last learning sessions are presented in Figure 1A. A 2 (session: 1, 2) x 7 (morph levels: 5%, 20%...95%) repeated measures ANOVA (*SPSS 15.0*) showed that classification performance was significantly different between sessions for the different morph levels (session x morph levels: $F(78,6) = 12.63$, $p < 0.001$). Subjects performed better on the continua end-points at the last session than at the first session (Figure 1A); Morph50, i.e. the ambiguous stimuli, was perceived as having no learned identities at either session.

fMRI experiment

Stimuli and Design

Stimuli were presented binaurally through MRI-compatible headphones (*NNL – NordicNeuroLab, Bergen, Norway*) using MCF software (*Digivox, Montreal, QC, Canada*) at a loud but comfortable level of 80 dB SPL(C).

Three morphed stimuli were drawn from each of the voice identity continua: Morph5, Morph35 and Morph65 corresponding to morph levels 5%, 35% and 65%, respectively, based on preliminary data showing that Morph5 and Morph35, but not Morph65, are perceived as a similar identity after learning. Morph5 and Morph65 were equally distant from Morph35 in terms of physical distance (30% distance). Crucially, while morph stimuli were equally unknown before the learning phase, after learning, Morph 5 and Morph 35 were perceived as

the same learned identities and Morph35 and Morph65 were perceived as different identities (Figure 1A).

We then created pairs with those stimuli; the first stimulus was always Morph35, the second stimulus was either Morph5, Morph35 or Morph65. Thus, three pair types were used: WITHIN (Morph5-Morph35, 30% physical change), SAME (Morph35-Morph35, no difference), BETWEEN (Morph35-Morph65, 30% physical change) (Figure 1B). Crucially, WITHIN and BETWEEN pairs were characterized by similar amount of acoustical change (30%); however after learning, the amount of perceptual change was much larger for BETWEEN than for WITHIN pairs (Figure 1A). The symmetrical set of stimuli (Morph35, Morph65 and Morph95) was also extracted from the continuum and similar pairs were built with those stimuli (for clarity, we only use the labels Morph5, Morph35 and Morph 65 in the rest of the text, since the direction of a given continuum was arbitrary and balanced across stimuli).

An fMRI run consisted of presenting 15 different pairs of each condition (WITHIN, SAME and BETWEEN) as well as 15 null trials (i.e. silent trials with no stimulation) in a pseudo-random order different across subjects and sessions. Pairs drawn from the two sides of a continuum were presented in two different runs; however, continua corresponding to different identities and vowels were mixed within a run. In each run, there were 15 stimuli per condition i.e. each pair drawn from each continuum (15 continua: 3 identities times 5 vowels); there were thus no repetition of a pair within a run. Each run was repeated 4 times; run order was randomised across subjects and sessions.

Subjects performed a same/different discrimination task in the scanner. They listened to stimulus pairs and decided whether the two sounds were the same (SAME) or different (WITHIN or BETWEEN) based on acoustical, not identity, differences. We chose to use an acoustic-based task in order to control the information processed by the subjects across the two

fMRI sessions. Such a control would have been impossible using a 'direct' identity discrimination task since identities were not known at the first session, before voice learning.

Image Acquisition and Analysis

Functional images covering the whole brain (field of view: FOV: 210 mm, 32 slices, voxel size 3x3x3 mm) were acquired on a 3T Tim Trio Scanner (Siemens) using an echoplanar imaging (EPI) sequence (interleaved, TR: 3.5s, TE: 30ms, Flip Angle: 77°, matrix size: 70^2). The sequence used in the experiment was a "sequence with gaps" (TR: 3.5s; TA: 1.8s, see Figure 1C) so that the sound were presented on a silent background; 8 runs of 4 min (70 volumes) were acquired; 10 volumes were recorded with no stimulation at the end of a run to create a baseline. At the end of each fMRI session, high resolution T1-weighted images (anatomical scan) were obtained (FOV: 256mm, 192 slices, voxel size: 1x1x1 mm, Flip angle: 9°, TR: 1.9s, TE: 2.52ms, matrix size: 256^2). In order to allow region of interest analyses (ROI), voice-selective areas were localized using a "voice localizer" scan: 8s blocks of auditory stimuli containing either vocal or non-vocal sounds ((Belin et al. 2000) – available online: http://vnl.psy.gla.ac.uk/resources_main.php); the voice localiser (TR: 2s, TE: 30ms, flip angle: 77°, voxel size: 3^3, matrix size: 70^2) was acquired in either one of the fMRI sessions (i.e. before or after the learning phase).

Data were analysed using SPM5 software (Wellcome Department of Imaging Neuroscience, London, UK; http://www.fil.ion.ucl.ac.uk/spm). First, the anatomical scan was AC-PC centred; this correction was applied to all the EPI images. Functional images were then motion corrected; all scans were aligned to the first scan of the last run, and a mean image (average of all scans) was created. The within session anatomical scan was co-registered to the mean image and segmented. The anatomical scan and the functional images were then normalised to the Montréal Neurological Institute (MNI) template using the parameters issued from the segmentation keeping the voxel resolution of the original scan (1x1x1 and 3x3x3 respectively).

Functional images were then smoothed with a Gaussian function with a full-width at half maximum (FWHM) of 8x8x8 mm. The first two volumes of each session were not included in the analysis of the data to allow for stabilisation of the scanner. Individual contrast images were generated for each pair types (SAME, WITHIN, BETWEEN) defined as independent conditions. Voxel-based random effects analysis, i.e. group level statistics, were performed on the individual contrasts across the brain volumes using a factorial design (2 (session) x3 (pair types) ANOVA). Processing and 3D rendering of brain anatomy for display purposes were performed using BrainVisa (http://brainvisa.info, IFR49, Neurospin, Saclay, France) from a T1 weighted scan of an individual subject normalised into the MNI space. Identification of brain areas was done using the *aal* brain atlas (Tzourio-Mazoyer et al. 2002) via XjView 8 (http://www.alivelearn.net/xjview8/).

Results

Behavioural results

During scanning, subjects performed a same-different discrimination task on stimuli from each pair presented. A 2 (session) x 3 (pair types: WITHIN, SAME, BETWEEN) repeated measures ANOVA was run on percent correct and reaction time data. Performance was similar across sessions (first: 74.2%; second: 76.8%; $F(14,1) = 1.6$, $p = 0.226$). Performance were better for SAME (93%) than for WITHIN (68%) and BETWEEN (65%) regardless of session ($F(28,2) = 57.99$, $p < 0.001$ corrected for non-sphericity (Greenhouse-Geisser)); there was no significant interaction between pair types and session ($F(28,2) = 3.69$, $p = 0.06$ corrected for non-sphericity).

There was no effect of pair types ($F(28,2) = 0.54$, $p = 0.49$ corrected) or session ($F(14,1) = 1.4$, $p = 0.257$) on reaction times, nor any interaction between those factors ($F(28,2) = 1.31$, $p= 0.28$ corrected).

fMRI Results

Sensitivity to acoustical differences

Whole-brain sensitivity to acoustical changes was assessed by contrasting activity between pairs showing acoustical changes (WITHIN and BETWEEN) to pairs with no acoustical change (SAME). Analyses pooled across sessions revealed 6 clusters ($p<0.001$ uncorrected, cluster threshold of 10 voxels) of activation presenting a larger activity to different than same pairs (Table 1, Figure 2A). These included bilateral anterior STC (superior Temporal Pole), bilateral insulae/anterior inferior frontal cortices (aIFC) and, in the right hemisphere only, middle STC and areas along the convexity of the inferior frontal gyrus (posterior IFC/MFG in Table 1).

Separate analyses for sessions 1 and 2 showed that the right middle STC showed a greater response to different than same pairs at both fMRI sessions, overlapping with the TVA (Figure 2B, C). Right anterior STC/superior Temporal Pole (sTP) was activated at the first session whereas left anterior STC/sTP, bilateral aIFC and right posterior IFC/MFG (Middle Frontal Gyrus) showed sensitivity to acoustical change at the second session only, i.e. after the voice learning (blue regions in Figure 2B).

Sensitivity to identity changes

Identity changes in a pair differed across sessions. At the first session, identity changes were linked to physical changes as stimuli were not associated with any identity: thus WITHIN and BETWEEN pairs corresponded to comparable identity changes. In contrast, at the second session, Morph35 and Morph5 were perceived as a same identity whereas Morph35 and Morph65 were perceived as different identities (Fig 1A); thus BETWEEN pairs were characterized at the second session by a marked change in identity whereas both WITHIN and SAME pairs kept perceived identity unchanged (identity adaptation). Consequently, regions

underlying the perception of learned identities were investigated with the following contrast at the second session only: BETWEEN vs. (SAME + WITHIN), i.e., pairs with identity change vs. pairs without identity change.

Whole brain sensitivity ($p < 0.01$ uncorrected, extent voxels threshold of 10) to identity changes assessed by the above described contrast revealed 3 clusters of activation along the convexity of right IFC (posterior and anterior IFC, pIFC/precentral Gyrus), and the left cingulate gyrus (Table 2, Figure 3A). We examined the same contrast (BETWEEN > (SAME + WITHIN)) at the first session, prior to learning; this showed no significant differences at the same statistical threshold. Involvement of the right pIFC in processing voice identity was further confirmed by a conjunction analysis of (BETWEEN > WITHIN) and (BETWEEN > SAME) (Figure 3).

Temporal Voice Areas – ROI Analyses

The Temporal Voice Areas (TVA) identified by the independent voice localizer were located as expected along the upper bank of the STS; 3 clusters were identified surviving a threshold of 7.07 (threshold T value for a $p < 0.05$ Family-Wise Error (FWE)-corrected, see Table 3, Figure 4).

ROI analyses were performed using Marsbar in 8mm-radius spheres around the 3 TVA maxima (Table 3). They showed that bilateral STC in the TVA showed sensitivity to acoustical differences: greater activity for different (WITHIN and BETWEEN) than SAME pairs ($p < 0.001$ corrected, at both session and for each session independently) but not for identity change ($p = 0.67$, $p = 0.09$ for left and right STC respectively). Interestingly, ROI analyses of the superior part of the Temporal Pole (aSTC) showed an effect of session as sensitivity to acoustical differences was significant at the first session ($p = 0.028$ corrected) but not at the second session ($p = 0.09$).

Discussion

We aimed at disentangling brain regions involved in acoustical representations of voice from the ones involved in identity voice representations. To that purpose we scanned participants before and after voice learning in an fMRI adaptation design (Grill-Spector and Malach 2001; Rotshtein et al. 2005) using pairs of stimuli in which either acoustical properties or identity properties were repeated so that each will lead to adaptation (i.e. decreases in BOLD signal) in specific brain areas. fMR-adaptation measures decreases in BOLD responses for repeated stimulus presentation; it is unclear how adaptation at the cerebral level, as measured with fMRI, relates to the activity of the underpinning neuronal populations. Different models have been proposed to explain the potential neural mechanisms underlying adaptation: neural fatigue, sharpening, or facilitation (Grill-Spector et al. 2006). The exact underlying mechanisms are unclear yet adaptation paradigms are widely used in fMRI to investigate the cognitive processing of different stimuli in the brain (Aguirre 2007; Rotshtein et al. 2005).
Three main results emerged from this study: 1) the temporal lobe, and particularly the TVA (Belin et al. 2000), were involved in acoustical processing of voices regardless of familiarity with the voice, i.e. irrespective of the fMRI session; 2) after voice learning, the processing of acoustical information also involved bilateral inferior frontal cortex (IFC); 3) adaptation to voice identity is shown, only after voice learning, in regions along the right IFC.

After the first and last learning sessions, subjects performed a 2-AFC on different voice identity continua. Results showed a significant improvement in the classification of continua end-points confirming that subjects had learned the 3 voices and demonstrating that, after voice learning, Morph5 and Morph35 were indeed classified as the same identity while Morph35 and Morph65 were perceived as different identities.

Sensitivity to acoustical voice information in the Superior Temporal Cortex (STC).

Comparing activity between pairs showing acoustical changes (WITHIN and BETWEEN pairs) to pairs with no change (SAME pairs) allowed us to identify brain regions involved in the processing of acoustical information. Bilateral STC, with a more anterior location in the left hemisphere, showed sensitivity to acoustical representation of voices (Figure 2A). Region of interest (ROI) analyses in the TVA (Belin et al. 2000) confirmed the involvement of bilateral mid STC in acoustical processing of voices consistent with previous studies (Andics et al. 2010; Formisano et al. 2008; Von Kriegstein and Giraud 2004).

The anterior part of the STC, i.e. the superior part of the temporal pole (TP), is also part of the TVA. Emerging evidence demonstrates the involvement of the anterior part of the temporal lobe of both humans and macaques (Petkov et al. 2008) in an invariant representation of voice identity, regardless of voice familiarity (unfamiliar voices – (Belin and Zatorre 2003; Formisano et al. 2008; Imaizumi et al. 1997; von Kriegstein et al. 2003); familiar voices – (Andics et al. 2010; Nakamura et al. 2001). Voice discrimination, i.e. the processing of unfamiliar voices, is impaired, in phonagnosics, by damage to bilateral temporal lobes (Van Lancker and Kreiman 1987; Van Lancker et al. 1988). In the present study, ROI analyses of the TVA and whole brain analyses of the "acoustical" contrast (WITHIN+BETWEEN > SAME) showed that the right superior temporal pole (sTP) was involved, at the first session only, in the processing of acoustic information. It should be noted that, at the first session, acoustical changes in the stimuli (in WITHTIN and BETWEEN pairs) were associated with perceptual changes in the stimuli: voices were unfamiliar, thus, two voices that sounded different could be perceived as different identities. At the second session, however, because participants learned to recognize the voices, the perceived identity was independent of acoustical change in the stimuli constituting the pairs: same identity for WITHIN pairs, different for BETWEEN pairs. Moreover, a dramatic decrease in the BOLD response of the sTP was seen at the second session for all three pair types, showing that activity in the right

sTP is reduced for familiar voices. Hence, our findings suggest that right sTP is implicated in an acoustic-based representation of unfamiliar voices; as soon as voices are familiar, activation in the right sTP decreased. On the contrary, whole brain analysis of the same contrast revealed the activation of the left sTP after voice learning, suggesting a dissociation between left and right sTP (Table 1): while right sTP is involved in an acoustic-based representation of unfamiliar voices, the left sTP seems to be involved in an acoustic-based representation of familiar voices.

Our findings could seem inconsistent with some studies reporting involvement of the right temporal pole in an identity-based representation of familiar voices. The temporal pole has also been described as being a multimodal area sensitive to various information related to person recognition as it is activated by face/voice recognition (Sestieri et al. 2006; von Kriegstein et al. 2006) and by retrieval of episodic memory (Ellis et al. 1989; Gorno-Tempini et al. 1998). The discrepancy between our study and others showing sensitivity to voice familiarity in the temporal pole could reflect the multimodal aspects of these other studies in which familiar voices were consistently associated to a face either because they used a learning procedure involving a face/voice association (Andics et al. 2010), or because they used acquaintances' voices thus necessarily associated with a face (Nakamura et al. 2001). Moreover, the region described here differed from that described in the others studies (Andics et al. 2010; Nakamura et al. 2001; von Kriegstein et al. 2005), which is located in a more inferior, medial part of the temporal pole. Similarly this region has also been found to respond to identity processing of familiar faces (Rotshtein et al. 2005) confirming that the involvement of the inferior TP may reflects multimodal association between familiar faces and voices.

Our findings together with previous studies show that the temporal pole comprised at least two distinct regions, with distinct functional roles. The right inferior TP seems to be a multimodal area involved in a non verbal representation of person knowledge that responds to

familiar faces or voices, but the latter only after a face/voice association either due to lab training or to familiarity itself (Andics et al. 2010; Gorno-Tempini et al. 1998; Hailstone et al. 2010; Nakamura et al. 2001; Rotshtein et al. 2005; von Kriegstein et al. 2005). On the contrary, the right sTP appears to be involved in an acoustic-based representation of unfamiliar voices (Belin and Zatorre 2003; Imaizumi et al. 1997; von Kriegstein et al. 2003; Von Kriegstein and Giraud 2004). Further studies are needed to clarify the specific role of the different parts of the temporal pole.

Sensitivity to acoustical information in inferior frontal cortices after voice learning.

An unexpected result of this experiment was the effect of voice learning on the brain network involved in the acoustic processing of voices. While before voice learning only right mid STC and sTP showed sensitivity to acoustical changes in the stimuli, after voice learning, bilateral IFC and insulae were recruited in order to process acoustical information in the stimuli consistent with previous studies reporting acoustical sensitivity in IFC for learned voices (Andics et al. 2010; von Kriegstein et al. 2006). Von Kriegstein and Giraud (2006) showed that, whereas activity in the TVA increased slightly after a face or name/voice association, activity in right prefrontal regions dramatically increased after a learned association consistent with Andics et al. (2010). Activation of the IFC is also reported for unfamiliar voices with no learned association, however, in both studies, voices were previously presented to the subjects (Stevens 2004; Von Kriegstein and Giraud 2004). Our findings confirmed that recruitment of IFC for the processing of acoustical information occurred only for previously heard voices as it emerged at the second session only. Activation of the IFC only at the second session could suggest that, after the first presentation of vocal stimuli, a vocal acoustical imprint is build, and when heard a second time, recollection of this acoustical imprint occurs, i.e. previously heard voices are represented in an "acoustical voice space" (Andics et al. 2010).

Identity-based representation of voices involved the right inferior frontal cortex.

Representation of voice identity was assessed by contrasting BETWEEN pairs, i.e. pairs showing an identity change, to pairs showing no identity change (WITHIN and SAME pairs). Before voice learning, no areas showed sensitivity to voice identity suggesting that voice identity "categories" did not exist previous to learning, i.e. there was no unconscious grouping of Morph35 and Morph5 when the voices were unfamiliar. At the second session, after learning the voices, right frontal areas and the left cingulate gyrus showed sensitivity to voice identity. Part of the frontal areas sensitive to identity change in the stimuli overlapped with regions responding to acoustical changes in the stimuli (Figure 3). Involvement of frontal areas in processing voice identity has been previously described for familiar voices or learned voices (Andics *et al.* 2010; Von Kriegstein and Giraud 2004; von Kriegstein *et al.* 2006); yet, with learned voices, Andics et al. (2010) did not report voice identity processing in IFC. In Andics et al. (2010) the task during scanning was explicitly to recognise the learned voices, which was not the case here or in von Kreigstein & Giraud (2006) which could explain the discrepancy in the results reported between the three studies. Such differences are to be expected based on recent evidence showing an influence of tasks on the activity measured in response to vocal stimuli (Bonte, 2009). At the second session, posterior IFC showed a higher activity to BETWEEN pairs than SAME or WITHIN whereas its activity was similar across pairs at the first session; it is also the only region highlighted by the conjunction test. This designates posterior IFC as a potential candidate for being an area involved in the representation of voice identity, irrespective of acoustical information. This conclusion is drawn from a comparison of pre- and post-learning observations, using the same task and stimuli; therefore, it should not be sensitive to the choice of task but instead reflect familiar voice identity processing.

We did not find activation of the right parietal cortex (Van Lancker et al. 1988; von Kriegstein et al. 2006), posterior STS (Andics et al. 2010) and fusiform gyrus (von Kriegstein

et al. 2005) in the representation of voice identity. Right parietal cortex and posterior STS are multimodal areas receiving input from the different sensory systems (Calvert et al. 2000; Sestieri et al. 2006) while the fusiform gyrus is activated by visual stimuli, faces in particular (Kanwisher et al. 1997). Von Kriegstein & Giraud (2006) and Andics et al. (2010) used a paradigm requiring a face/voice association in order to achieve voice learning, whereas we used a voice/name association. The activity in those multimodal areas could be due to retrieving visual information associated to the learned voice, explaining why they are not present in our study; similarly, when using a name/voice association, von Kriegstein and Giraud (2006) do not report activation in the right parietal cortex and FG. Interestingly, the left cingulate gyrus showed sensitivity to identity changes at the second session only. Left cingulate gyrus has been described as part of the network responding to familiar voices (Von Kriegstein and Giraud 2004); it was however, not reported in Andics et al. (2010) study with learned voices. One explanation could be that the training used in the present study is more likely to be similar to natural voice learning as participants were exposed to different vocal items, stories, words, and thus have a general idea of the person's voice. This suggests that cingulate gyrus could be involved in storage, memorisation and retrieval of familiar voices regardless of whether these are associated with visual or episodic information.

Conclusion

The Temporal Voice Areas, mid STC, are involved in acoustical processing of voices regardless of voice familiarity. The right superior temporal pole seems to be involved in acoustic-based representation of *unfamiliar* voices; whereas, the left temporal pole and bilateral IFC seem to be involved in an acoustic-based representation of familiar voices. The processing of vocal identity seems to involve a network of areas, located along the convexity of the right Inferior Frontal Gyrus, functioning in a hierarchical manner. A particular role is given to the posterior IFC that showed a dramatic change in its response after voice learning.

Acknowledgment

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC, grant number BB/E003958/1), an Economic and Social Research Council/Medical Research Council grant (RES-060-25-0010) and the Royal Society (IJP 2008/R1).

References

Aguirre GK. 2007. Continuous carry-over designs for fMRI. Neuroimage. 35: 1480-1494.
Andics A, McQueen JM, Petersson KM, Gal V, Rudas G, Vidnyanszky Z. 2010. Neural mechanisms for voice recognition. Neuroimage. 52: 1528-1540.
Belin P. 2006. Voice processing in human and non-human primates. Philos Trans R Soc Lond B Biol Sci. 361: 2091-2107.
Belin P, Fecteau S, Bedard C. 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn Sci. 8: 129-135.
Belin P, Zatorre RJ. 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 14: 2105-2109.
Belin P, Zatorre RJ, Ahad P. 2002. Human temporal-lobe response to vocal sounds. Brain Research Cogn Brain Research. 13: 17-26.
Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature. 403: 309-312.
Bonte M, Valente G, Formisano E. 2009. Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. J Neurosci. 29: 1699-1706.
Calvert GA, Campbell R, Brammer MJ. 2000. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol. 10: 649-657.
Charrier I, Mathevon N, Jouventin P. 2001. Mother's voice recognition by seal pups. Nature. 412: 873.
DeCasper AJ, Fifer WP. 1980. Of human bonding: Newborns prefer their mothers' voices. Science. 208: 1174-1176.
Ellis AW, Young AW, Critchley EM. 1989. Loss of memory for people following temporal lobe damage. Brain. 112 (Pt 6): 1469-1483.
Fecteau S, Armony JL, Joanette Y, Belin P. 2005. Sensitivity to voice in human prefrontal cortex. J Neurophysiol. 94: 2251-2254.
Formisano E, De Martino F, Bonte M, Goebel R. 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 322: 970-973.
Gorno-Tempini ML, Price CJ, Josephs O, Vandenberghe R, Cappa SF, Kapur N, Frackowiak RS. 1998. The neural systems sustaining face and proper-name processing. Brain. 121 (Pt 11): 2103-2118.
Grill-Spector K, Henson R, Martin A. 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci. 10: 14-23.
Grill-Spector K, Malach R. 2001. fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychol. 107: 293-321.
Hailstone JC, Crutch SJ, Vestergaard MD, Patterson RD, Warren JD. 2010. Progressive associative phonagnosia: A neuropsychological analysis. Neuropsychologia. 48: 1104-1114.
Henson RN, Goshen-Gottstein Y, Ganel T, Otten LJ, Quayle A, Rugg MD. 2003. Electrophysiological and haemodynamic correlates of face perception, recognition and priming. Cereb Cortex. 13: 793-805.
Imaizumi S, Mori K, Kiritani S, Kawashima R, Sugiura M, Fukuda H, Itoh K, Kato T, Nakamura A, Hatano K, Kojima S, Nakamura K. 1997. Vocal identification of speaker and emotion activates different brain regions. Neuroreport. 8: 2809-2812.
Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neuroscience. 17: 4302-4311.
Kawahara H, Masuda-Katsuse I, Cheveigne Ad. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. . Speech Communication. 27: 187-207.
Kingdom FAA, Prins N. 2010. Psychophysics: A Practical Introduction London: Academic Press: an imprint of Elsevier. 297 p.
Kreiman J. 1997. Listening to voices: theory and practice in voice perception research. In: Johnson K, Mullenix J, eds. Talker Variability in Speech Research New York: Academic Press p 85-108.

Marchant-Forde JN, Marchant-Forde RM, Weary DM. 2002. Responses of dairy cows and calves to each other's vocalisations after early separation. Applied Animal Behaviour Science. 78: 19-28.
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.
Papcun G, Kreiman J, Davis A. 1989. Long-term memory for unfamiliar voices. J Acoust Soc Am. 85: 913-925.
Perrachione TK, Wong PC. 2007. Learning to recognize speakers of a non-native language: implications for the functional organization of human auditory cortex. Neuropsychologia. 45: 1899-1910.
Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. 2008. A voice region in the monkey brain. Nat Neurosci. 11: 367-374.
Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ. 2005. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat Neuroscience. 8: 107-113.
Schweinberger SR, Herholz A, Sommer W. 1997. Recognizing famous voices: influence of stimulus duration and different types of retrieval cues. J Speech Lang Hear Res. 40: 453-463.
Sestieri C, Di Matteo R, Ferretti A, Del Gratta C, Caulo M, Tartaro A, Olivetti Belardinelli M, Romani GL. 2006. "What" versus "where" in the audiovisual domain: an fMRI study. Neuroimage. 33: 672-680.
Stevens AA. 2004. Dissociating the cortical basis of memory for voices, words and tones. Brain Res Cogn Brain Res. 18: 162-171.
Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 15: 273-289.
Van Lancker D, Kreiman J. 1987. Voice discrimination and recognition are separate abilities. Neuropsychologia. 25: 829-834.
Van Lancker DR, Cummings JL, Kreiman J, Dobkin BH. 1988. Phonagnosia: a dissociation between familiar and unfamiliar voices. Cortex. 24: 195-209.
von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL. 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Research Cogn Brain Research. 17: 48-55.
Von Kriegstein K, Giraud AL. 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage. 22: 948-955.
von Kriegstein K, Giraud AL. 2006. Implicit multisensory associations influence voice recognition. PLoS Biol. 4: e326.
von Kriegstein K, Kleinschmidt A, Giraud AL. 2006. Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cereb Cortex. 16: 1314-1322.
von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL. 2005. Interaction of face and voice areas during speaker recognition. Journal of Cognitive Neuroscience. 17: 367-376.

Tables

Table 1: Main Effect of Physical Differences

	Coordinates - mm			T Values	P Values (uncorrected)	Cluster Size
	x	**y**	**z**			
(Within + Between) > Same						
Left						
aSTC/sTP	-51	3	-9	3.87	0.0001	22 **
Insula/aIFC	-36	21	0	3.75	< 0.0002	11 **
Right						
aIFC/Insula	33	27	0	4.59	<0.0001	82 **
Middle STG	60	-27	3	4.44	<0.0001	81
pIFC/MFC	54	18	30	4.04	<0.0001	58 **
aSTC/sTP	54	9	-9	3.74	< 0.0002	10 *

Whole brain analyses. Clusters of more than 10 voxels surviving a threshold of T > 3.11 (p < 0.001, uncorrected).

IFC – Inferior Frontal Cortex, STG – Superior Temporal Cortex, MFG – Middle Frontal Cortex, sTP – superior Temporal Pole. a – anterior, p – posterior.

** Session 1 only. ** Session 2 only.*

Table 2: Main Effect of Perceptual Differences

	Coordinates - mm			T Values	P Values (uncorrected)	Cluster Size
	x	**y**	**z**			
Between > (Within + Same) – session two only						
Left						
Cingulate Gyrus	-12	9	45	3.19	0.001	16*
Right						
pIFC	45	0	21	3.61	0.0002	22*
aIFC	45	33	9	3.28	0.001	28*
pIFC/Precentral G	39	6	36	2.74	0.004	19

Whole brain analysis, displays masked by all versus silence. Clusters of more than 10 voxels surviving a threshold of $T > 2.37$ ($p<0.01$). Maxima surviving at $p < 0.001$. Precentral G – Precentral Gyrus.*

Table 3: Temporal Voice Areas

Voice > Non-Voice	Coordinates			T values	P Values (FWE)	Cluster Size
	x	**y**	**z**			
Left						
Mid STC	-57	-19	-2	9.17	0.003	161
Right						
Mid STC	51	-34	4	8.91	0.005	168
Superior TP	57	5	-14	7.82	0.02	13

Whole brain analyses. Clusters surviving a threshold of T > 7.07 (FWE 0.05). Mid STC – middle Superior Temporal Cortex, TP – Temporal Pole.

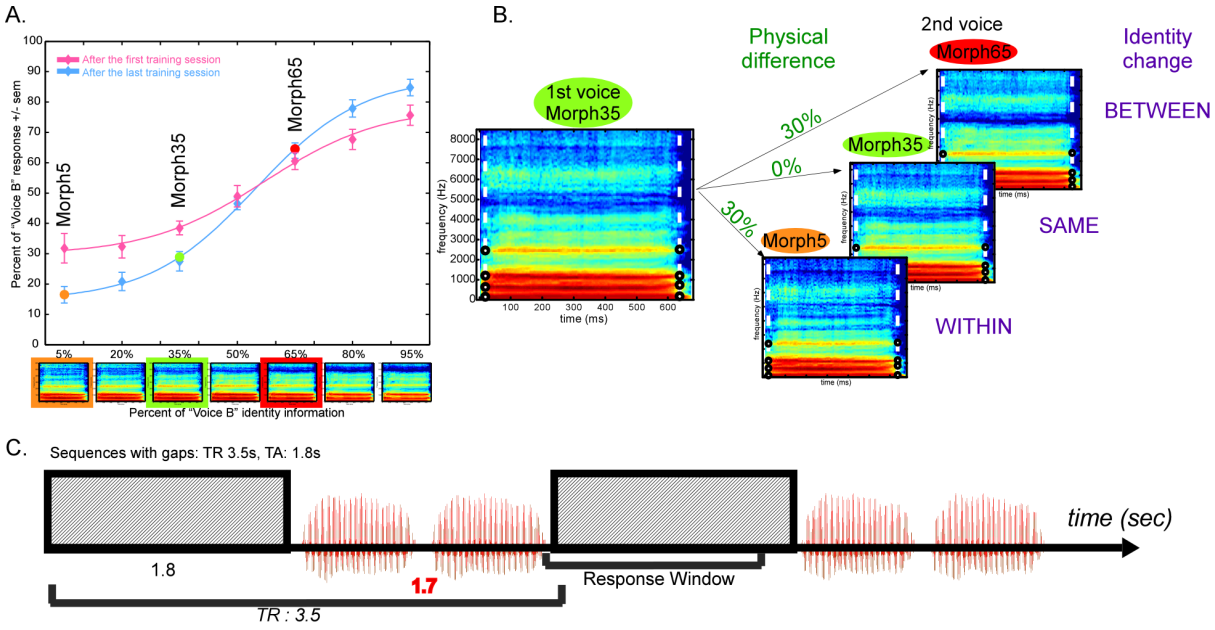

Figure 1. Stimuli and Design. A. Results of the 2-AFC task after the first (red) and last (blue) learning sessions. Stimuli used in the experiments are indicated by dots. B. Example of the stimuli and pair types used in the experiment. C. Illustration of the fMRI sequence used in the experiment.

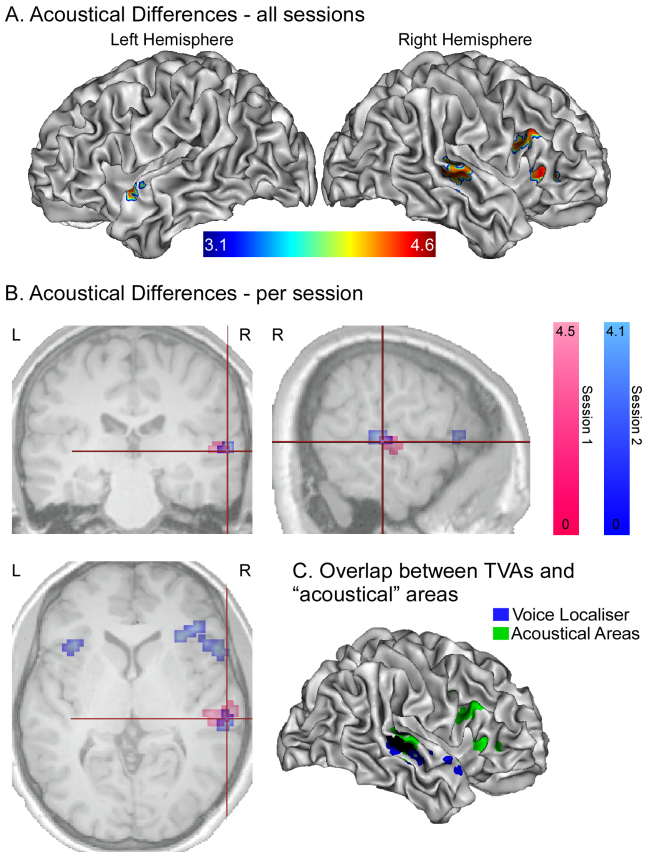

Figure 2. Sensitivity to acoustical changes. A. Areas highlighted by the WITHIN + BETWEEN greater than SAME contrasts pooled across sessions. Whole brain analysis, $p<0.001$ uncorrected; extent threshold of 10 voxels. B. Areas sensitive to acoustical changes per session. Note that only the right mid STS is activated at the first session where as at the second session, prefrontal areas are activated. C. Overlap between the acoustical areas (green) and the TVA (blue) in the right STS.

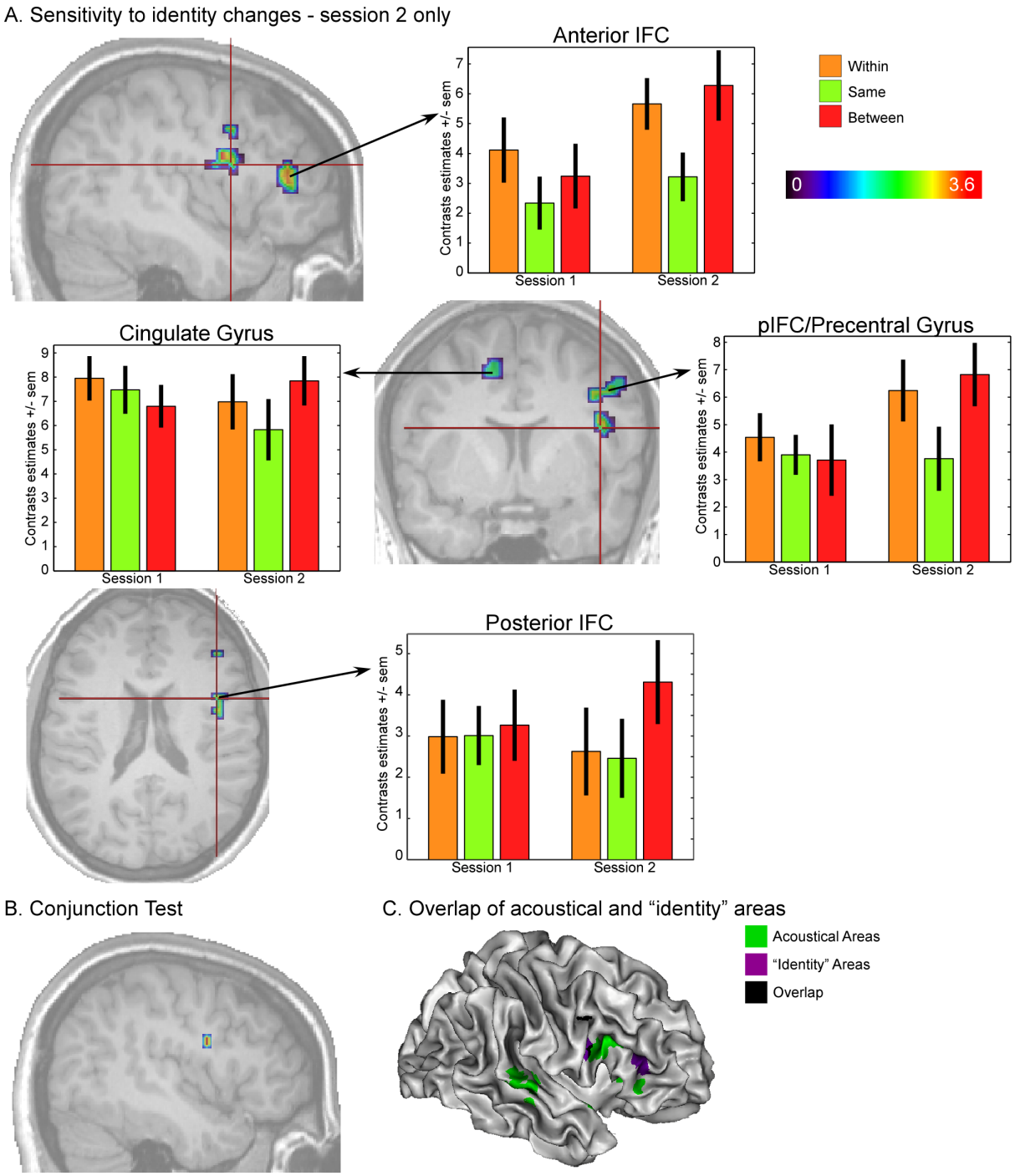

Figure 3. Sensitivity to identity changes after voice learning. A. Areas revealed by a whole brain analysis of the BETWEEN > SAME + WITHIN contrasts revealing identity-sensitive regions ($p<0.01$, extent threshold of 10 voxels). Bar graphs represent mean contrast estimates for each condition against baseline in an 8mm-radius sphere around the maximum of each cluster. B. Cluster revealed by the conjunction test of BETWEEN > WITHIN and BETWEEN > SAME. C. Overlap (black) between the acoustical areas (green) and the "identity" areas (purple).

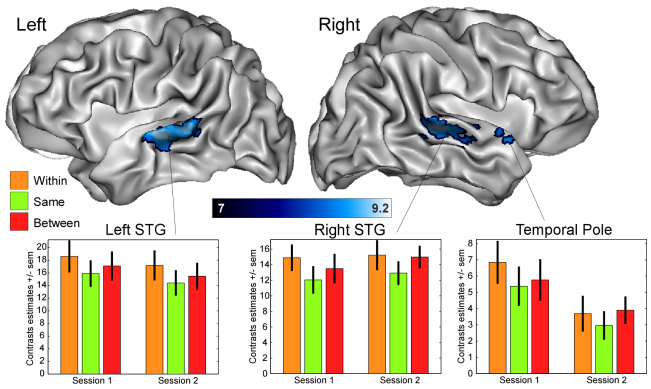

Figure 4. Temporal Voice Areas. In blue activity for voice greater than non voice ($p<0.05$, FWE corrected). Bar graphs represent the mean contrast estimates for each condition against baseline in an 8mm-radius sphere around the three maxima of the voice localiser.