Normal voice processing after posterior superior temporal sulcus lesion

Guo Jiahui[a*], Lúcia Garrido[b], Ran R. Liu[c], Tirta Susilo[d], Jason J. S. Barton[e] and Bradley Duchaine[a]

[a]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA; [b]Division of Psychology, Department of Life Sciences, Brunel University, Uxbridge UB8 3PH, United Kingdom; [c]Department of Medicine, division of neurology, Kingston General Hospital, Queen's University, Ontario, Canada; [d]School of Psychology, Victoria University of Wellington, Wellington, New Zealand; [e]Departments of Medicine (Neurology) & Opthalmology and Visual Sciences, University of British Columbia, Canada

[*] Corresponding author at: Dartmouth College, Hanover, Department of Psychological and Brain Sciences, NH, USA, 03755. Tel.: +1 603 277 0856.
E-mail address: Jiahui.Guo.GR@dartmouth.edu

Abstract

The right posterior superior temporal sulcus (pSTS) shows a strong response to voices, but the cognitive processes generating this response are unclear. One possibility is that this activity reflects basic voice processing. However, several fMRI and magnetoencephalography findings suggest instead that pSTS serves as an integrative hub that combines voice and face information. Here we investigate whether right pSTS contributes to basic voice processing by testing Faith, a patient whose right pSTS was resected, with eight behavioral tasks assessing voice identity perception and recognition, voice sex perception, and voice expression perception. Faith performed normally on all the tasks. Her normal performance indicates right pSTS is not necessary for intact voice recognition and suggests that pSTS activations to voices reflect higher-level processes.

Keywords: *voice perception, pSTS, patient study*

1. Introduction

The superior temporal sulcus (STS) extends anteriorly from the inferior parietal lobe along the entire temporal lobe and is one of the longest sulci in the brain. The STS plays a central role in processing social information, including the perception of faces (Haxby, Hoffman, & Gobbini, 2000), voices (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000), and biological motion (Yovel & O'Toole, 2016). In addition to representing social perceptual information, the STS integrates different social percepts to generate higher-level representations (Campanella & Belin, 2007; Frith & Frith, 2003; Yovel & O'Toole, 2016). A meta-analysis of more than 100 fMRI studies of the STS found motion processing, face processing, and audiovisual integration reliably activated the posterior STS (pSTS), speech processing activated the anterior STS, and theory-of-mind tasks led to activity along the entire STS (Hein & Knight, 2008). A recent study used localizers to identify areas in the STS showing selective responses to a variety of social stimuli (Deen, Koldewyn, Kanwisher, & Saxe, 2015). The results showed selectivity to theory-of-mind reasoning in the angular gyrus and surrounding sulci as well as the middle-to-anterior STS, biological motion in the most posterior region of STS, face processing in a more anterior region of pSTS with weaker face responses in middle-to-anterior STS, and a broad response to voices that peaked in middle STS.

The STS response to voices extended into pSTS (Deen et al., 2015), and a number of other studies have also suggested that the pSTS processes information about voices. In the first paper to identify voice-selective areas, Belin et al. (2000) reported three clusters that responded selectively to voices including one in pSTS (see also Watson, Latinus, Noguchi, et al., 2014). The pSTS, along with anterior STS, showed an elevated response when participants attended to vocal identity rather than the meaning of a sentence (von Kriegstein & Giraud, 2004). In an adaptation study, repetition suppression to vocal identity was found in pSTS and middle STS bilaterally (Andics et al., 2010). Notably, in a fair proportion of voice studies, effects are more pronounced in right STS than left STS (Belin, Zatorre, & Ahad, 2002; Belin et al., 2000; Gainotti, 2013), and pSTS's response to voices may also be stronger on the right than the left (Schall, Kiebel, Maess, & von Kriegstein, 2014).

What sort of cognitive computations do pSTS activations to voices reflect? One possibility is that pSTS carries out fundamental voice processing by representing the auditory properties of voices and categorizing vocal identity and characteristics like sex, expression, and age. These processes depend on voice-selective regions in more anterior regions of STS (Belin, Bestelmeyer, Latinus, & Watson, 2011; Bestelmeyer, Belin, & Grosbras, 2011), but pSTS may also play a role in them. Another possibility is that right pSTS voice activations are driven by higher-level voice processing such as computations integrating voice representations with information from faces and other types of social information (Belin et al., 2011; Campanella & Belin, 2007; Thurman, van Boxtel, Monti, Chiang, & Lu, 2016; Yovel & O'Toole, 2016). Consistent with this last account, pSTS shows cross-modal fMRI adaptation effects between faces and voices when facial expressions are similar to the preceding vocal expression (Watson, Latinus, Noguchi, et al., 2014). Such integration may be specific to people: a conjunction analysis of fMRI data indicated that right pSTS integrates face and voice information but not visual and auditory information about objects (Watson, Latinus, Charest, Crabbe, & Belin, 2014). A recent study found strong correlations between the strength of the preference of a voxel in pSTS for visual mouth movements and the magnitude of its auditory speech response, as well as its preference for vocal sounds (Zhu & Beauchamp, 2017). Magnetoencephalography (MEG) suggests that the right pSTS shows a stronger response to combined face-voice stimuli than the sum of unimodal face and voice components (Hagan et al., 2009). Finally, Deen et al. (2015) found that a pSTS region of interest identified with a dynamic face localizer showed comparable activation to voices and faces.

To clarify the role of the pSTS in voice processing suggested by fMRI studies, it would be helpful to have complementary data from a lesion study. Here, we assessed the role of right pSTS in voice processing with behavioral experiments in a patient, Faith, whose right pSTS was lost due to a tumor resection. The surgery left the more anterior regions of Faith's STS intact, including those containing the temporal voice areas (TVAs) in the middle and anterior STS. We tested Faith with eight behavioral tasks that tap a wide range of voice processing abilities including identity discrimination, identity memory, sex categorization, and expression categorization. Impaired performance with some or all of the tasks would support the hypothesis that right pSTS is involved in basic aspects of voice processing, while

intact performance would be more consistent with the hypothesis that the voice activations seen in the pSTS are reflections of higher-level voice processing.

2. Method & Results

2.1 Patient Case

Faith is a right-handed speech therapist, and English is her native language. In 2009 she had a right occipitotemporal resection to remove a tumor, and in July 2015, she had a second resection along the margin of the same location followed by proton radiation therapy. Following her first surgery, she noted severe face processing deficits. Her impairments affect many types of face processing, including perception of identity, expression, and gaze (Susilo, Wright, Tree, & Duchaine, 2015). Faith believes her ability to process voices remains normal. She completed the first eight tasks described below in April 2015 when she was 52-years-old, and did a final task (three-alternative expression test) in February 2016 when she was 53.

2.2 Faith's lesion and its overlap with voice-selective activations in normal participants

2.2.1 Anatomical scan

Faith was scanned on a 3.0-T Phillips MR scanner (Philips Medical Systems, WA, USA) with a SENSE (SENSitivity Encoding) 32-channel head coil. An anatomical volume was acquired using a high-resolution 3D magnetization-prepared rapid gradient-echo sequence (220 slices, field of view = 240 mm, acquisition matrix = 256 × 256, voxel size = 1 × 0.94 × 0.94 mm). This scan was skull stripped and then warped to Talairach space. The high-resolution MR images of Faith's brain (Figure 1A) show a lesion extending from the fusiform to the superior part of temporal lobe in the right hemisphere, encompassing a large part of her posterior superior temporal sulcus (pSTS). The estimated lesion size on the axial, coronal, and sagittal axes is 43 mm, 37 mm, and 37 mm, respectively.

2.2.2 Peak coordinates from five papers

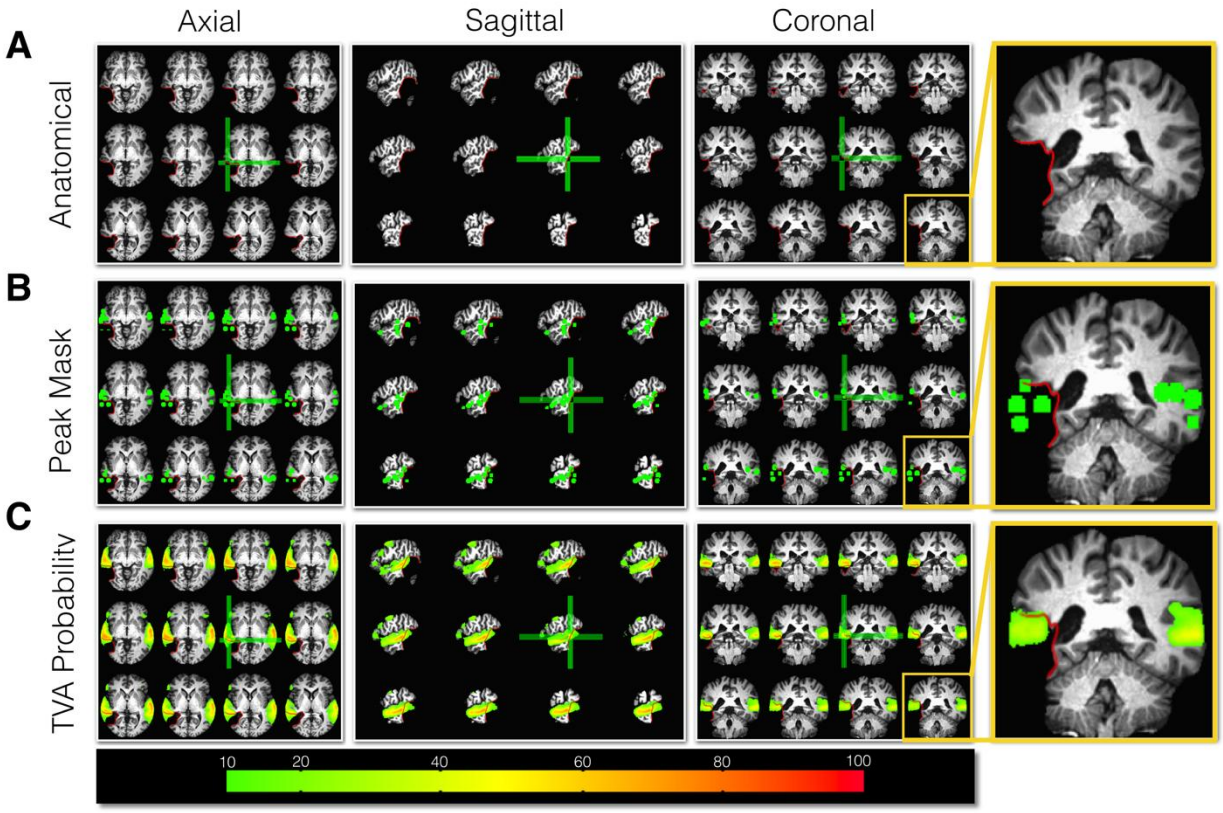
To demonstrate that Faith's lesion overlaps with voice-selective responses in the literature, peak voxels implicated in voice processing from five papers (Belin et al., 2000; Deen et al., 2015; von Kriegstein & Giraud, 2004; Watson, Latinus, Noguchi, et al., 2014; Watson, Latinus, Charest, et al., 2014) are displayed on Faith's brain. Coordinates for 29 peak voxels in the temporal lobe that were listed in the tables in the five papers were

extracted manually and plotted on the standard brain in Talairach space with a 5mm radius sphere centered at the peak voxel (Supplementary Table 1). The coordinates map was then converted to a brain mask and overlaid on Faith's brain. Figure 1B shows the overlap of the peak voxels with Faith's lesion.

2.2.3 Overlap with TVA probabilistic map

As another approach to determining which voice-selective regions may have been affected by Faith's resection, we compared her lesion to a probability map of the temporal voice areas (TVA) downloaded from http://vnl.psy.gla.ac.uk/resources.php. Belin and colleagues created the map based on individually-thresholded data from 152 participants. Each individual's T-map showing voice-selective voxels (voices > non-voice sounds) was corrected for multiple comparisons based on the spatial extent at $q < 0.05$ (Chumbley & Friston, 2009). They then applied a Gamma-Gaussian mixture model to separate the null voxels from the active voxels (Gorgolewski, Storkey, Bastin, & Pernet, 2012). Data from each participant was transformed into MNI space, binarized, summed, and normalized to 100 to create the group probability map. We converted this map from MNI space to Talairach space so it could be overlaid on Faith's anatomical scan. Voxels that were voice-selective in 10% or more of the participants are displayed in Figure 1C.

Figure 1. Faith's anatomical scan, the peak coordinates mask, and the TVA probabilistic map.



Panel A displays Faith's lesion on the axial (left), coronal (middle), and sagittal (right) planes with the spacing of one slice (around 1 mm). The right hemisphere lesion extended from the fusiform to the superior part of temporal lobe, encompassing a large part of her posterior superior temporal sulcus (pSTS). The estimated lesion size on the 3D axial, coronal, and sagittal axis is 43x37x37 (mm).

Panel B shows the peak coordinates mask of voices selective activations in the temporal lobe from the five studies. These peak coordinates mask is overlaid on Faith's brain. Each dot is a sphere centered at the peak voxel with a radius of 5 mm. These slices show that Faith's lesion has a large overlap with the voice selective activations.

Panel C shows a probabilistic map of temporal voice areas (TVA) overlaid on Faith's brain. The TVA probabilistic map was downloaded from http://vnl.psy.gla.ac.uk/resources.php and normalized to Talairach space. For a voxel to be shown on this map, it had to be voice-selective in at least 10% of the participants. The color bar shows the scale for the TVA-probability map. Like the peak coordinates mask in Panel B, Faith's lesion overlaps with the TVA probabilistic map.

2.3 Voice localizer to assess Faith's temporal voice areas

2.3.1 Stimuli and experiment procedures

To exam whether Faith shows voice-selective areas in her intact cortex we conducted a standard TVA localizer (Belin et al., 2000). Voice stimuli were designed by Belin and his colleagues and were downloaded from http://vnl.psy.gla.ac.uk/resources.php. This functional localizer lasts 10 minutes and contains one run in total. It contains 40 eight-seconds blocks of sounds (16 bit, mono, 22050 Hz sampling rate). Half of the blocks consists of vocal sounds (speech and nonspeech), and the other half consists of nonvocal sounds (industrial sounds, environmental sounds, and a few animal vocalizations). All sounds have been normalized and a 1kHz tone of similar energy was provided for calibration. The order of the sound blocks was provided on the website and optimized for the vocal vs. nonvocal contrast.

2.3.2 MRI acquisition

Faith was scanned on the same 3.0-T Phillips MR scanner as the anatomical scan (Philips Medical Systems, WA, USA) with a SENSE (SENSitivity Encoding) 32-channel head coil. Functional images were collected using echo-planar functional images (time to repeat = 2000ms, time echo = 35 ms, flip angle = 90°, voxel size = 3 × 3 × 3 mm). Each volume consisted of 36 interleaved 3 mm thick slices with 0 mm interslice gap. The slice volume was adjusted to cover most of the brain including the entire temporal lobe. Previous studies found that the location and extent of susceptibility effects are influenced by the slice orientation and phase-encoding direction (Ogawa, Lee, Kay, & Tank, 1990; Ojemann et al., 1997). In our study, we adopted oblique slice orientation aligned with each participant's anterior commissure–posterior commissure (AC–PC) line, because it produces fewer susceptibility artifacts than the commonly used transverse orientation (Ojemann et al., 1997) and at the same time provides better coverage of the brain. The phase-encoding direction (anterior–posterior) was chosen to move the signal loss away from the more anterior part of the brain.

2.3.3 Data analysis

Imaging data were analyzed using the AFNI software package (Cox, 1996). Before statistical analysis, the first volume was discarded to allow for magnetic saturation effects, and each volume was registered to the third volume. The EPI data were warped to align with the anatomical data and transformed to a standard space in the Talairach template (Talairach & Tournoux, 1988). Each volume was blurred with a 4-mm FWHM Gaussian kernel. Time series of each run were scaled by the mean of the baseline before passing into the

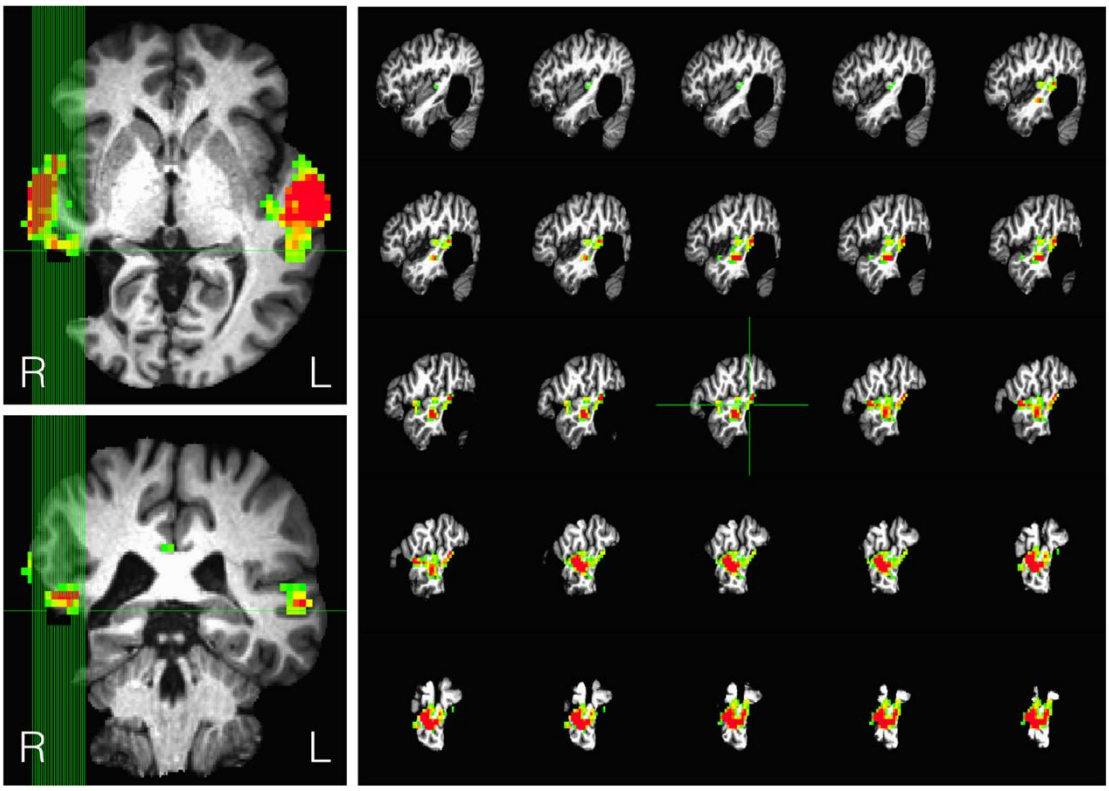
deconvolution analysis. Detrending and motion correction were carried out by including trends and head motion as regressors in the regression model. Repetition times with excessive motion (>0.3 mm) were removed. A general linear model procedure was used for ROI analysis. Voice-selective regions were identified with a vocal > non-vocal contrast.

2.3.4 Results

Two large clusters were localized at $p < 0.0001$ uncorrected on both temporal lobes (Figure 2). Although Faith's lesion impacted part of her right temporal lobe, a cluster of 484 voxels was still found on the right side (peak: 62, -10, 2). The other cluster was in the left temporal lobe and consisted of 461 voxels (peak: -64, -22, 5). Peak voxels in both hemispheres were on the superior temporal gyrus (STG) in the vicinity of the peak coordinates reported in previous experiments (Belin et al., 2000; Deen et al., 2015; von Kriegstein & Giraud, 2004; Watson, Latinus, Noguchi, et al., 2014; Watson, Latinus, Charest, et al., 2014). These voice localizer results indicate that, despite Faith's extensive right hemisphere lesion, her remaining right STS still shows typical voice-selectivity.

To summarize Faith's imaging results, Faith's lesion disrupted her right pSTS but left more anterior sections of STS intact. The lesioned part of Faith's right pSTS overlapped with voice-selective activations in normal participants but the unlesioned part of Faith's STS showed a typical voice-selective response.

Figure 2. Voice-selective areas in Faith.



This figure shows voxels in Faith that responded significantly more strongly to vocal sounds than to non-vocal sounds ($p < 0.0001$) in a standard functional TVA localizer (Belin et al., 2000).

2.4 Behavioral Testing

2.4.1 Control participants

Three groups of participants provided control data for the behavioral tasks. All were compensated with course credit or reimbursement.

The first group of controls (Vancouver group) were a subset of the 73 controls aged 19-70 years old reported in Liu et al. (2014). We selected 19 participants to create an age and sex-matched group for Faith (age range 36-70; mean = 51.7, S.D. = 11.9). All participants had no history of neurological or psychiatric diseases, and no visual or auditory complaints. Participants were required to be native English speakers that had lived in North America for five years or more.

The second group of controls (London group) were drawn from Garrido et al. (2009). This group consisted of eight women between 46 and 64 years old (mean = 56.6; S.D. = 5.5). They averaged 15.9 (S.D. = 2.0) years of education. All were native British English speakers

and reported normal hearing. One control (control 6) was taking fluoxetine, as well as ropinirole for treatment of restless leg syndrome. The other controls reported no neurological or psychiatric history.

The third control group (Dartmouth group) consisted of 18 undergraduate students tested at Dartmouth College. Their ages ranged from 18 to 22 (mean = 18.94; S.D. = 1.26), and eight were female. All reported normal hearing, and none reported neurological disorders.

2.4.2 General Material and Procedure

Faith wore headphones for all of the behavioral tasks. She did the voice identity discrimination task and the three-alternative vocal expressions task on a 13-inch MacBookPro and did the other tests were executed on a Dell laptop.

Faith's behavioral scores were compared to the controls' results using the modified t-test for single case studies developed by Crawford & Howell (1998). Differences between Faith and the controls were considered significant when the one-tailed probability was equal to or below 0.05 because our expectation was that Faith's performance on the voice processing tests would either be normal or impaired but not superior to the controls.

2.4.3 Voice Identity Discrimination

2.4.3.1 Material and Procedure

The auditory stimuli were created from 20 male and 20 female volunteers between the ages of 20–31 (Liu et al., 2014). Each stimulus was used only once as a target or as a distractor. Controls wore headphones and were tested with an IBM Lenovo laptop running SuperLab software.

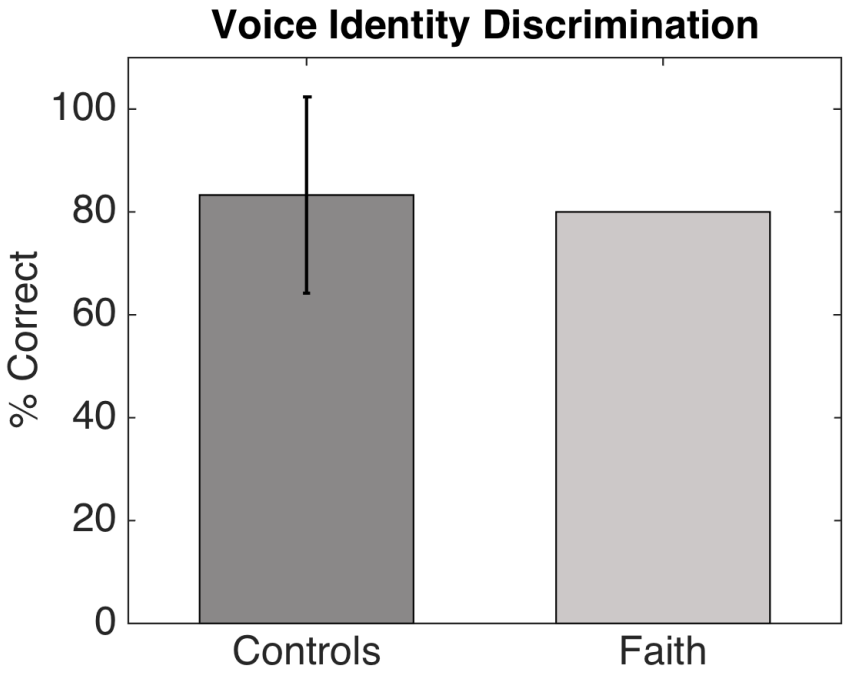
Each trial of the task consisted of a target voice and two choice voices. Audio stimuli consisted of two different texts that the volunteers read. For the initial target voice, the subjects read the phrase: "This is by far one of the most amazing books I have ever read, it tells the story of a Colombian family across generations." One choice voice was the target voice while the other was a distractor. Both choice voices read the phrase: "After a hearty breakfast, we decided to go for a walk on the beach. It was a lovely morning with the crisp smell of the ocean in the air." Volunteers were asked to read both texts at the same speed. All recordings were 10 s in duration.

Participants attempted to select the choice voice that matched the target voice. No feedback on performance was provided. In each trial, the participant heard the target voice first. After a 1.5-s pause, a ring tone sounded for 875 ms, which served as an auditory mask and separated the target from the test choices. Next, the participant heard the choice voices, sequentially. Choice voices were preceded by 1 s of silence and were followed by 1.5 s of silence. The participant was asked to press the number button "1" or "2" to indicate which choice voice matched the target voice. The 40 trials were divided into two equal blocks; male voices were presented in one block and females in the other.

2.4.3.2 Results

Faith's accuracy in the voice discrimination task was 80.0% (Figure 3). The average performance of the controls was 83.3% (S.D. = 9.5%). Faith's performance was comparable to the controls' mean ($p = 0.37 > 0.05$), and her z score was -0.35.

Figure 3. Faith's performance on voice identity discrimination task.



Faith's accuracy on the voice identity discrimination task compared with the Vancouver controls. The error bar shows ±2 standard deviations (S.D.).

2.4.4 Voice Identity Recognition

2.4.4.1 Learning six speakers

2.4.4.1.1 Material and procedure

At the start of the test, participants learned the voices of six unfamiliar young female speakers (Garrido et al., 2009). All voices were native British English speakers and had similar accents. Speakers read sentences with three key words taken from the BKB Sentence List (Bench, Kowal, & Bamford, 1979). All samples were recorded in an anechoic chamber using Cool Edit 96 (http://www.syntrillium.com) and were normalized for peak amplitude using the program PRAAT (Boersma & Weenink, 2005).

Participants attempted to learn name-voice pairings, and they were told that the names would also be necessary for a later task. They were first presented with the name of a speaker on the screen, each of which started with a different letter from A to F, and then they heard a sentence said by that speaker. After that, they heard a number of sentences and for each one, they decided if it was said by the same speaker or not. Half the sentences were said by the target speaker, while the other half were said by one of the other five speakers. This procedure was repeated for each of the six speakers.

In the practice block, participants were presented with two test sentences per speaker. In the first test block, six test sentences followed the sentence presenting each speaker's voice. In the second and third test blocks, there were ten test sentences. No sentences were repeated.

2.4.4.1.2 Results

Over the three learning blocks, the control participants' mean accuracy was 75.2% (S.D. = 2.9). Faith's overall performance was 75.6%, which corresponds to a z-score of 0.13. (Figure 4A). Taking a closer look at each learning block, controls responded correctly on 71.5% (S.D. = 4.1) of trials in the first block, 74.4% (S.D. = 5.3) in the second block, and 78.1% (S.D. = 5.5) in the third block. Faith scored 77.8%, 63.3% and 86.7% in these three blocks respectively. The z-score of the three blocks was 1.52, -2.09, and 1.57. Only the result of the second block was slightly lower than control performance ($t(7) = -1.98$, $p = 0.04$).

2.4.4.2 Naming the six speakers

2.4.4.2.1 Material and Procedure

Immediately after the learning task finished, participants were presented with 60 new sentences (6 speakers x 10 sentences per speaker) in a random order and were asked to select the name associated with that speaker in the previous task. The six names were shown on the computer screen. Feedback (one beep) was given for incorrect responses.

2.4.4.2.2 Results

With six choices, chance performance on this task was 16.7%. The mean percent correct for the London controls was 35.6% (S.D. = 10.9). Faith correctly identified 38.3% of the test items, which corresponds to a z-score of 0.25 (Figure 4B).

2.4.4.3 Old-new discrimination with the six speakers

2.4.4.3.1 Material and Procedure

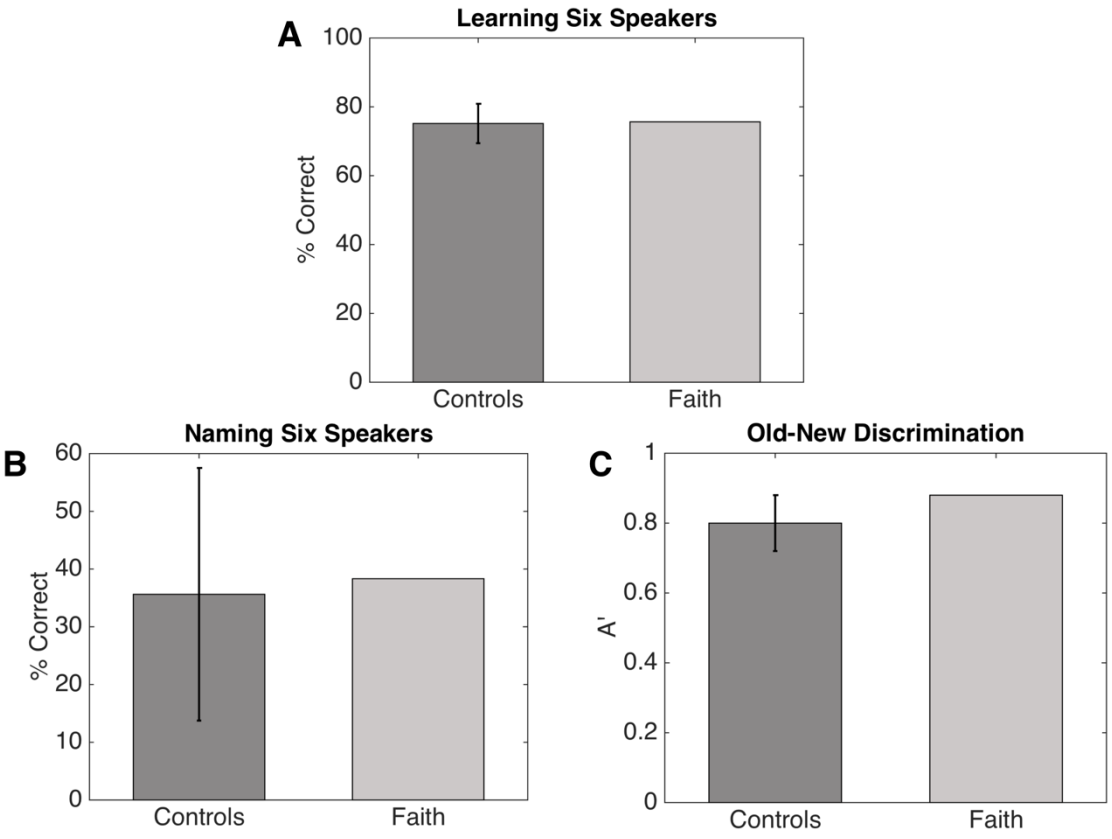
Immediately after the name identification task, participants' voice memory was tested with an old-new discrimination task. For each sentence, participants decided whether it was said by one of the six speakers used in the two previous tasks or a new speaker. New speakers were young females with accents similar to the six target speakers. New recordings were done for all speakers (targets and distracters) in a silent room. Peak amplitude for all stimuli was matched using PRAAT. The test trials included six sentences by each of the old speakers (36 'old' trials) and four sentences said by each of the nine new speakers (36 'new' trials).

2.4.4.3.2 Results

A' for the controls was 0.80 (S.D. = 0.04). Faith's A' of 0.88 was better than all the controls (z = 2.0) (Figure 4C).

In summary, the three voice identity recognition tasks demonstrated Faith had normal discrimination, recognition and familiarity for voice identity.

Figure 4. Faith's performance on voice identity recognition tasks



Performance of Faith and the London control group on the three voice tasks involving the six speakers. (A) Faith's accuracy compared with the control group on the three learning blocks. Scores from the three blocks were combined to calculate overall accuracy. (B) Faith's accuracy compared with the controls on the Naming the Six Speakers task. (C) Comparison of Faith and the controls on the Old-New Discrimination task. All error bars display the ±2 standard deviations (S.D.).

2.4.5 Vocal Sex Perception

2.4.5.1 Material and Procedure

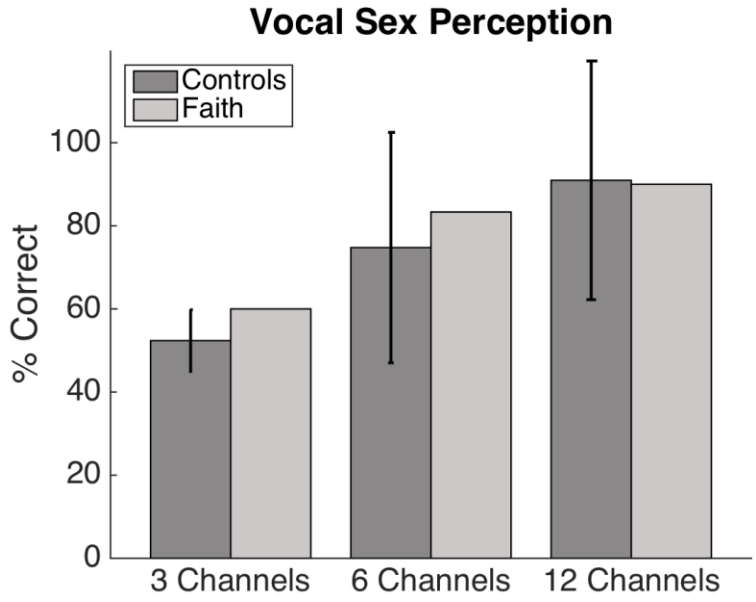
Twenty-six native English speakers read sentences aloud from the BKB Sentence List (stimuli were collected using a microphone and the program Cool Edit 96). Different speakers read different sets of sentences. Sentences were matched for peak amplitude and noise-vocoded using PRAAT to increase the difficulty of the task to avoid ceiling effects. Three, six, and twelve frequency channels were used.

On each trial, participants listened to one sentence and were asked to decide whether the speaker was a man or a woman. Half the sentences were spoken by males and half by females. Thirty sentences for each frequency channel were presented, making a total of 90 trials.

2.4.5.2 Results

Seven out of the eight London control participants performed this task. For the three noise-vocoded levels (3,6,12), participants correctly perceived the sex of the speaker on 52.4% (S.D. = 3.7), 74.8% (S.D. = 13.9) and 91.0% (S.D. = 14.4) of trials. Faith scored 60.0%, 83.3% and 90.0% (z = 2.05, 0.62, -0.07), indicating she categorizes vocal sex normally (Figure 5).

Figure 5. Faith's performance on vocal sex perception task



Faith's accuracy on the Vocal Sex Perception task compared with seven London controls. Error bars show ±2 standard deviations (S.D.).

2.4.6 Vocal Expression Perception

2.4.6.1 Recognition of vocal expression of emotion

2.4.6.1.1 Material and Procedure

Stimuli were 90 non-verbal sounds expressing one of the following emotions: achievement/triumph, amusement, anger, disgust, fear, pleasure, relief, sadness and surprise

(10 stimuli for each emotion) (Sauter, 2006; Sauter & Scott, 2007). After each stimulus, the list of nine possibilities was presented on the screen and participants selected the one that best described the expression of the voice.

2.4.6.1.2 Results

London controls selected the correct adjective on 82.2% (S.D. = 6.1) of trials. Faith's accuracy was 73.3% ($p = 0.16 > 0.05$), which placed her 1.47 standard deviations below the mean.

2.4.6.2 Recognition of vocal expression via paralinguistic cues in speech

2.4.6.2.1 Material and Procedure

Stimuli consisted of emotionally inflected spoken three-digit numbers. Like the previous task, there were ten stimuli for each of the nine emotion categories. There were another ten stimuli expressing contentment, to make a total of 100 trials in this task. All ten adjectives were presented on the computer screen in each trial, and participants were also asked to select the adjective that best described the emotion in the voice.

2.4.6.2.2 Results

The controls correctly identified the emotions in 72.1% (S.D. = 5.5) of trials. Faith's score of 66.0% ($z = -1.12$) did not differ significantly ($p = 0.10 > 0.05$) from the controls' mean score.

2.4.6.3 Three-alternative choices vocal expression of emotion

2.4.6.3.1 Material and Procedure

Stimuli consisted for speakers saying "Ah" so that it conveyed eight emotions (anger, disgust, fear, neutral, pain, pleasure, sadness, and surprise) from the Montreal Affective Voices (MAV) set (happiness was excluded due to ceiling effects). Each expression contained ten clips of audios recorded by five actors and five actresses (Belin, Fillion-Bilodeau, & Gosselin, 2008). Each trial consisted of three voices said by three different actors and actresses. Two voices expressed the same emotion while the other one expressed a different emotion. Participants were asked to pick the odd one out by pressing "1", "2", or "3". There were 72 trials, with a 2-s interstimulus interval and a 2.5-s intertrial interval. The

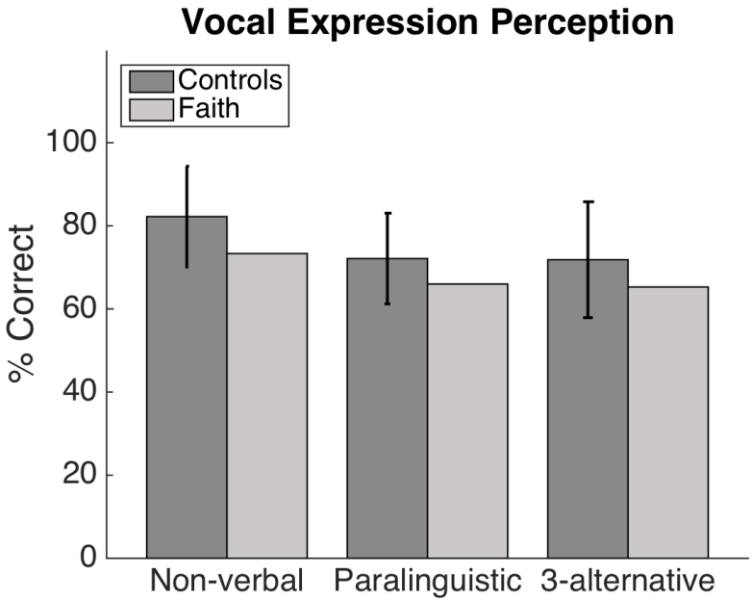
task was carried out using the online testing platform Testable (http://www.testable.org). Participants wore headphones.

2.4.6.3.2 Results

The controls' mean accuracy was 71.8% (S.D. = 7.0%). Faith responded correctly on 65.3% of trials (z = -0.94), which was comparable with the Dartmouth control groups' performance (p = 0.19 > 0.05) (Figure 6).

To exclude the possibility that Faith was impaired with particular vocal expressions, we analyzed each expression category separately in all three vocal expression tasks (supplementary material). None of these comparisons revealed a significant difference between Faith and the controls, so our results provide no evidence that Faith has expression-selective impairments (Supplementary Figure 1).

Figure 6. Faith's performance on vocal emotion perception task

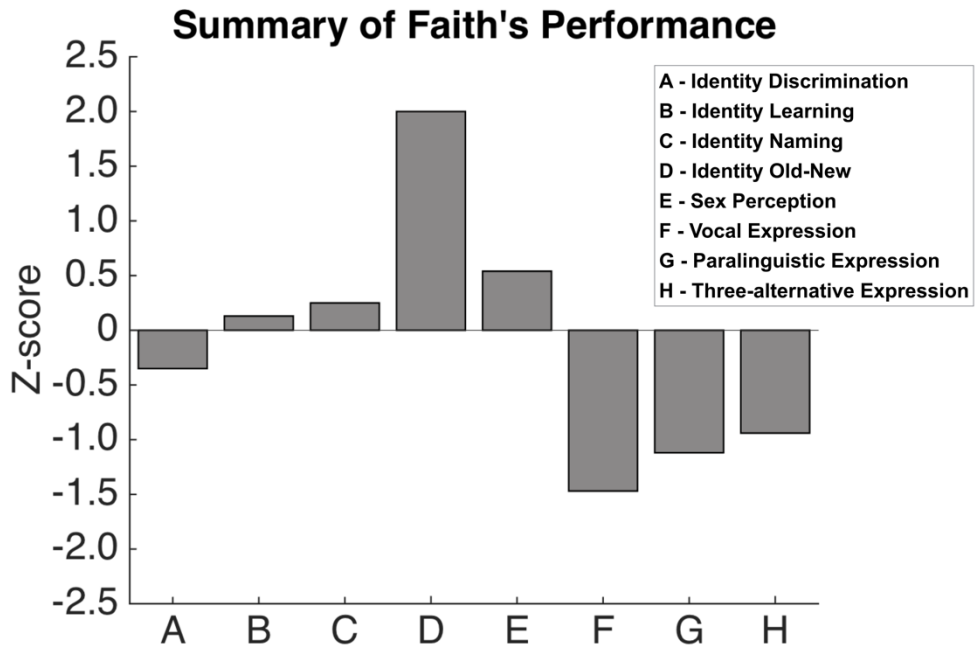


Faith's accuracy on the Vocal Expression Perception task. Her performance is compared with the London controls in the non-verbal and paralinguistic tasks and with 18 Dartmouth controls on the three-alternative task. Error bars display ±2 standard deviations (S.D.).

3. Discussion

We tested voice processing in Faith, a woman whose right pSTS was resected after surgery to remove a tumor. Eight voice tests measuring her abilities with identity discrimination, familiarity and recognition, sex perception, and expression perception in voices showed that her performance was comparable to that of control participants. Figure 7 presents a summary of Faith's behavioral performance relative to the controls. Her normal performance on all the vocal tasks demonstrates that voice processing ability can remain intact after a lesion to right pSTS.

Figure 7. Summary of Faith's Performance



Faith's z-scores on the behavioral voice tasks. Faith's performance was compared with the Vancouver controls on task A, with the London controls on task B to G, and with the Dartmouth controls on task H. The three blocks in the Learning the Six Speakers task were combined to calculate a single z-score, and the scores for each of the three channels in the sex perception test were also combined. The z-scores ranged from -1.5 to 2, and none of Faith's scores were significantly different from the controls.

Our investigation was motivated by fMRI results demonstrating increased signal in the pSTS in response to voice stimuli (Andics et al., 2010; Belin et al., 2000; Deen et al., 2015;

von Kriegstein & Giraud, 2004; Watson, Latinus, Noguchi, et al., 2014; Watson, Latinus, Charest, et al., 2014). We considered two explanations for these activations. According to the first account, pSTS plays a role in fundamental voice processing by representing vocal information and processing it to derive information about identity, sex, expression, and other characteristics. The second possibility is that pSTS does not contribute to basic voice processing but instead is involved in higher level voice processing (Belin et al., 2011; Campanella & Belin, 2007; Thurman et al., 2016; Yovel & O'Toole, 2016). Because Faith's voice perception is normal, her findings indicate that the right pSTS is not necessary for fundamental voice processing, and thus are not compatible with the first account.

Faith's results indicate pSTS activations to voices reflect higher-level processes, and several studies indicate that this region plays a role in integrating voice information with other kinds of social information (Belin et al., 2011; Campanella & Belin, 2007; Yovel & O'Toole, 2016). This evidence includes a meta-analysis that found that tasks involving integration led to responses that clustered in pSTS (Hein & Knight, 2008) as well as fMRI and MEG studies assessing the response to single (audio or visual) and multi-channel (audiovisual) social stimuli (Deen et al., 2015; Watson, Latinus, Charest, et al., 2014). The pSTS also shows a cross-modal fMRI adaptation effect when facial expressions were similar to the preceding vocal expression (Watson, Latinus, Noguchi, et al., 2014). Also consistent with an integrative role for pSTS are findings demonstrating that pSTS voxels selective for visual mouth movements also respond strongly and selectively to voices whereas voxels selective for moving eyes did not respond to voices (Zhu & Beauchamp, 2017).

What brain areas supported Faith's normal voice processing? Previous studies have demonstrated voice-selective regions in middle and anterior STS bilaterally, referred to as the temporal voice areas (TVAs) (Belin et al., 2011; Latinus, Crabbe, & Belin, 2011; Yovel & Belin, 2013). To see whether Faith's TVAs are preserved, we carried out a standard fMRI voice localizer in which blocks of voices and non-vocal sounds were presented (Belin et al., 2000) (Figure 2). Faith showed a typical voice-selective response in regions anterior to her lesion in both the right and left STS. Hence it is likely these more anterior STS regions as well as frontal regions (Andics et al., 2010; Watson, Latinus, Noguchi, et al., 2014; Watson, Latinus, Charest, et al., 2014) allowed her to perform normally in the lab and in daily life.

Furthermore, these results suggest that an intact right pSTS is not necessary to support voice-selective activity in these anterior regions.

Faith's range of behavioural impairments make it difficult to provide evidence that the pSTS integrates voice information with other social representations. Doing so would require demonstrating that disruption of right pSTS impairs the ability to integrate social information while not impairing the perceptual processes feeding an integrating mechanism. Because Faith's face perception (Susilo et al., 2015) and body perception (unpublished data) are severely disrupted by the lesions she has suffered, this makes it difficult to determine if there is integration of voice with face or body information. Testing a patient with a lesion selectively affecting pSTS or stimulating pSTS in neurologically intact participants may be effective ways to address this issue. However, this may be more useful in examining integration in identity perception rather than in expression perception, given Pitcher's (2014) demonstration that transcranial magnetic stimulation to right pSTS disrupted facial expression recognition.

In conclusion, Faith's consistently normal performance on a range of voice processing studies indicates that the right pSTS is not necessary for fundamental processing of voices. Her results are consistent with suggestions that right pSTS's responsiveness to voices reflects higher-level voice processing effects, such as the integration of voices with other social representations.

Acknowledgements

We are extremely grateful for Faith's participation in this project.

This project was supported by a Hitchcock Foundation grant to B.D. and CIHR grant MOP-102567, Canada Research Chair, and the Marianne Koerner Chair in Brain Diseases to J.B.

References

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage*, *52*(4), 1528–1540. https://doi.org/10.1016/j.neuroimage.2010.05.048

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531–539. https://doi.org/10.3758/BRM.40.2.531

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Bench, J., Kowal, Å., & Bamford, J. (1979). The Bkb (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, *13*(3), 108–112. https://doi.org/10.3109/03005367909078884

Bestelmeyer, P. E. G., Belin, P., & Grosbras, M.-H. (2011). Right temporal TMS impairs voice detection. *Current Biology*, *21*(20), R838–R839. https://doi.org/10.1016/j.cub.2011.08.046

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer. [Computer program]. <http://www.praat.org/>.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, *44*(1), 62–70. https://doi.org/10.1016/j.neuroimage.2008.05.021

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.

Crawford, J. R., & Howell, D. C. (1998). Comparing an Individual's Test Score Against Norms Derived from Small Samples. *The Clinical Neuropsychologist*, *12*(4), 482–486. https://doi.org/10.1076/clin.12.4.482.7241

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex*, *25*(11), 4596–4609. https://doi.org/10.1093/cercor/bhv111

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1431), 459–473. https://doi.org/10.1098/rstb.2002.1218

Gainotti, G. (2013). Laterality effects in normal subjects' recognition of familiar faces, voices and names. Perceptual and representational components. *Neuropsychologia*, *51*(7), 1151–1160. https://doi.org/10.1016/j.neuropsychologia.2013.03.009

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., … Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition.

Neuropsychologia, *47*(1), 123–131. https://doi.org/10.1016/j.neuropsychologia.2008.08.003

Gorgolewski, K., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, *6*, 245. https://doi.org/10.3389/fnhum.2012.00245

Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G. R., & Young, A. W. (2009). MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *106*(47), 20010–20015. https://doi.org/10.1073/pnas.0905792106

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hein, G., & Knight, R. T. (2008). Superior Temporal Sulcus—It's My Area: Or Is It? *Journal of Cognitive Neuroscience*, *20*(12), 2125–2136. https://doi.org/10.1162/jocn.2008.20148

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-Induced Changes in the Cerebral Processing of Voice Identity. *Cerebral Cortex*, *21*(12), 2820–2828. https://doi.org/10.1093/cercor/bhr077

Liu, R. R., Pancaroglu, R., Hills, C. S., Duchaine, B., & Barton, J. J. S. (2014). Voice Recognition in Face-Blind Patients. *Cerebral Cortex*, bhu240. https://doi.org/10.1093/cercor/bhu240

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(24), 9868–9872.

Ojemann, J. G., Akbudak, E., Snyder, A. Z., McKinstry, R. C., Raichle, M. E., & Conturo, T. E. (1997). Anatomic Localization and Quantitative Analysis of Gradient Refocused Echo-Planar fMRI Susceptibility Artifacts. *NeuroImage*, *6*(3), 156–167. https://doi.org/10.1006/nimg.1997.0289

Pitcher, D., Duchaine, B., & Walsh, V. (2014). Combined TMS and fMRI Reveal Dissociable Cortical Pathways for Dynamic and Static Face Perception. *Current Biology*, *24*(17), 2066–2070. https://doi.org/10.1016/j.cub.2014.07.060

Sauter, D. A. (2006). *An investigation into vocal expressions of emotions: the roles of valence, culture, and acoustic factors.* (Doctoral). University of London. Retrieved from http://discovery.ucl.ac.uk/1445045/

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion*, *31*(3), 192–199. https://doi.org/10.1007/s11031-007-9065-x

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice Identity Recognition: Functional Division of the Right STS and Its Behavioral Relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707

Susilo, T., Wright, V., Tree, J. J., & Duchaine, B. (2015). Acquired prosopagnosia without word recognition deficits. *Cognitive Neuropsychology*, *32*(6), 321–339. https://doi.org/10.1080/02643294.2015.1081882

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

Thurman, S. M., van Boxtel, J. J. A., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *NeuroImage*, *136*, 149–161. https://doi.org/10.1016/j.neuroimage.2016.05.015

von Kriegstein, K. V., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020

Watson, R., Latinus, M., Charest, I., Crabbe, F., & Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, *50*, 125–136. https://doi.org/10.1016/j.cortex.2013.07.011

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face–Voice Emotional Integration. *The Journal of Neuroscience*, *34*(20), 6813–6821. https://doi.org/10.1523/JNEUROSCI.4478-13.2014

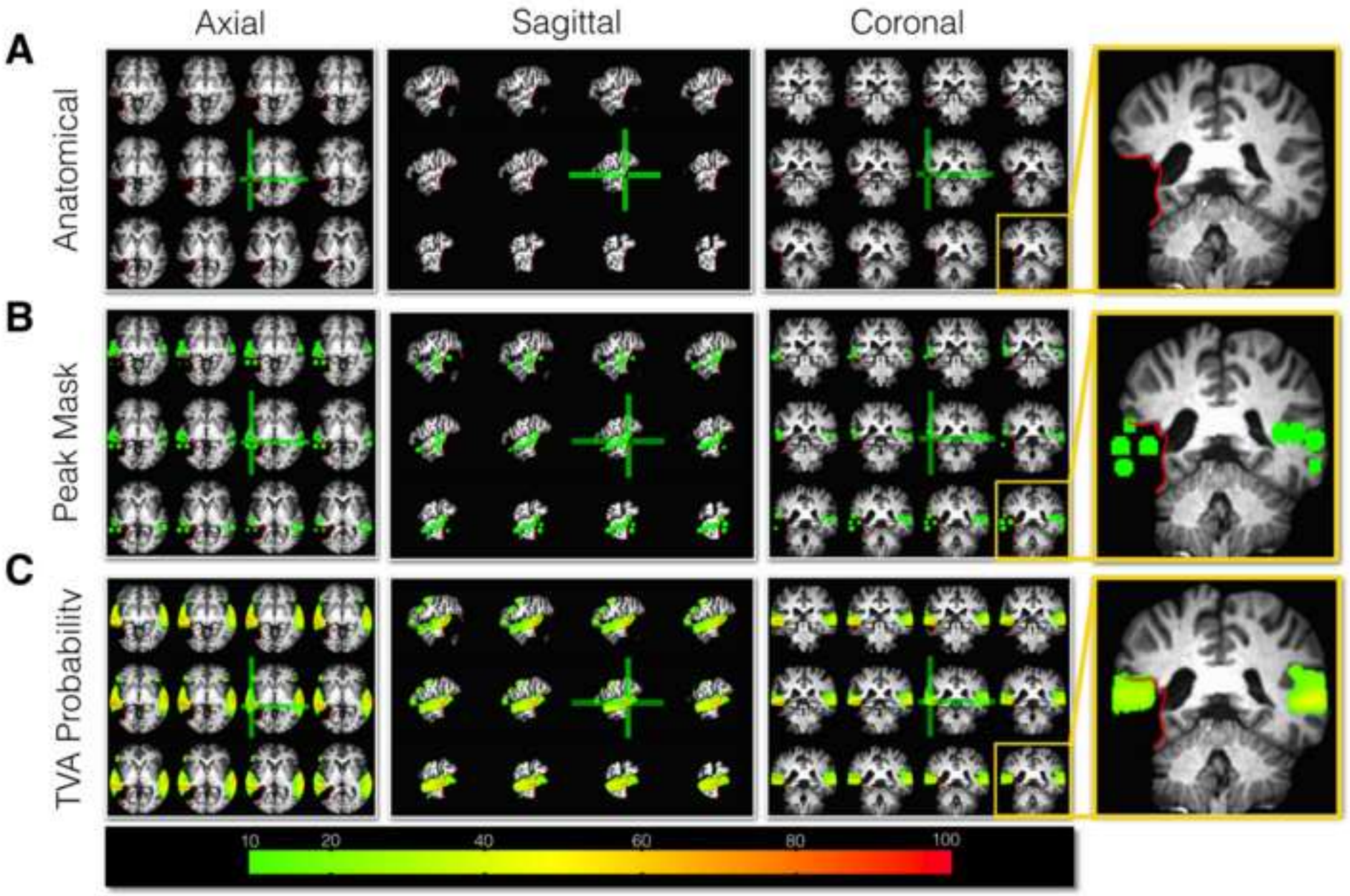
Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271. https://doi.org/10.1016/j.tics.2013.04.004

Yovel, G., & O'Toole, A. J. (2016). Recognizing People in Motion. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2016.02.005

Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, 2914–16. https://doi.org/10.1523/JNEUROSCI.2914-16.2017

Figure

Click here to download high resolution image



Supplementary Material

Click here to download Supplementary Material: Supplementary Jiahui, Garrido, Liu, Susilo, Barton, & Duchaine_revised2.docx