Cerebral Cortex April 2013;23:958–966
doi:10.1093/cercor/bhs090
Advance Access publication April 5, 2012

Cerebral Processing of Voice Gender Studied Using a Continuous Carryover fMRI Design

Ian Charest[1,2], Cyril Pernet[3], Marianne Latinus[2], Frances Crabbe[2] and Pascal Belin[2,4]

[1]Medical Research Council-Cognition and Brain Sciences Unit (MRC-CBU), Cambridge CB2 7EF, UK, [2]School of Psychology and Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, UK, [3]SFC Brain Imaging Research Centre, Division of Clinical Neuroscience, University of Edinburgh, Edinburgh EH4 2XU, UK and [4]BRAMS—International Laboratory for Brain, Music and Sound Research, University of Montreal and McGill University, Montreal, Quebec, Canada H3C 3J7

Address correspondence to Dr Ian Charest, Room 87, MRC-Cognition and Brain Science Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK. Email: ian.charest@mrc-cbu.cam.ac.uk.

Normal listeners effortlessly determine a person's gender by voice, but the cerebral mechanisms underlying this ability remain unclear. Here, we demonstrate 2 stages of cerebral processing during voice gender categorization. Using voice morphing along with an adaptation-optimized functional magnetic resonance imaging design, we found that secondary auditory cortex including the anterior part of the temporal voice areas in the right hemisphere responded primarily to acoustical distance with the previously heard stimulus. In contrast, a network of bilateral regions involving inferior prefrontal and anterior and posterior cingulate cortex reflected perceived stimulus ambiguity. These findings suggest that voice gender recognition involves neuronal populations along the auditory ventral stream responsible for auditory feature extraction, functioning in pair with the prefrontal cortex in voice gender perception.

Keywords: adaptation, auditory cortex, inferior prefrontal cortex, neuronal representation, superior temporal sulcus

Introduction

Voice gender is easily and accurately perceived by normal listeners (Childers and Wu 1991; Kreiman 1997), yet our brain's task is not as trivial as this ease of processing may suggest. The fundamental frequency of phonation (F0, perceived as the pitch of the voice) is highly variable within as well as between individuals. Despite being on average lower by nearly an octave in male compared to female voices, it shows considerable overlap between male and female speakers (Hillenbrand et al. 1995) suggesting that additional cues, such as formant frequencies (reflecting vocal tract length) as well as other sexually dimorphic acoustical cues (Wu and Childers 1991), are integrated. Yet the cerebral mechanisms underlying voice gender perception remain unclear.

Perceptual after effects caused by adaptation to voice gender have been observed using auditory adaptation techniques: brief exposure to voices of a given gender (adaptation) biases the perception of a subsequently presented gender-ambiguous voice toward the gender opposite to that of the adaptor (Mullennix et al. 1995; Schweinberger et al. 2008). Results from these 2 behavioral studies suggest the existence of neuronal populations involved in a plastic representation of voice gender. Two neuroimaging studies also directly compared activity elicited by male versus female voices, controlling for acoustical features by manipulating the fundamental frequency of the voices. Both studies suggested a right-hemispheric involvement in the cerebral processing of voice gender and report greater activity for female voices. Converging evidence for the involvement of specific cortical regions, including the temporal voice areas (TVAs), in voice gender recognition is, however, still missing (Lattner et al. 2005; Sokhi et al. 2005). This inability to find a persuasive link between localized cortical activity and gender perception could potentially be a consequence of the use of a subtraction approach, which constrains the search to brain regions more sensitive to voices of one gender over another. We suggest an alternative, more physiologically plausible model: voice gender representation could involve overlapping neuronal populations sensitive to male or female voices. Assuming equal proportions of male- and female-sensitive neurons in a given cortical area/voxel, the subtraction of male- versus female-related cerebral activity would fail to highlight them.

Here, we used an efficiency-optimized functional magnetic resonance imaging (fMRI) adaptation (Grill-Spector and Malach 2001) paradigm called a continuous carryover design (Aguirre 2007) to explore this alternative hypothesis. We took advantage of the recent development of audio morphing techniques (Kawahara 2003, 2006) to generate voice gender continua (Fig. 1*a*), providing 2 direct benefits over previous studies: 1) all stimuli sounded like natural voices and 2) changes in perceived gender can be examined at controlled physical differences. Subjects were scanned in a rapid event-related design while listening to voice stimuli drawn from male–female voice gender continua and performing a 2-alternative forced choice (2AFC) gender classification task. The continuous carryover design allows to examine in an optimally efficient way the repetition–suppression effect, that is, the effect of one stimulus on the cerebral response of the one presented immediately after. We used this adaptation paradigm as a means to test the hypothesis that the perception of male and female voices is carried out by overlapping neuronal populations: in that case, the repeated presentation of a male voice would be combined with a reduction of the response signal and a "recovery from adaptation" would be observed for a subsequently presented female voice. Furthermore, we examined the effects of stimulus differences based on perceived gender independently of their acoustical differences, providing a better understanding of the neural mechanisms involved in higher level voice gender perception.

Materials and Methods

Participants

Twenty young adult participants (10 females, mean age = 25.4 ± 6.3 years) with no history of neurological or psychiatric conditions participated in this study after giving written informed consent. The study was approved by the ethical committee from the faculty of information and mathematical sciences of the University of Glasgow. Subjects were paid £12 for participating in this study.

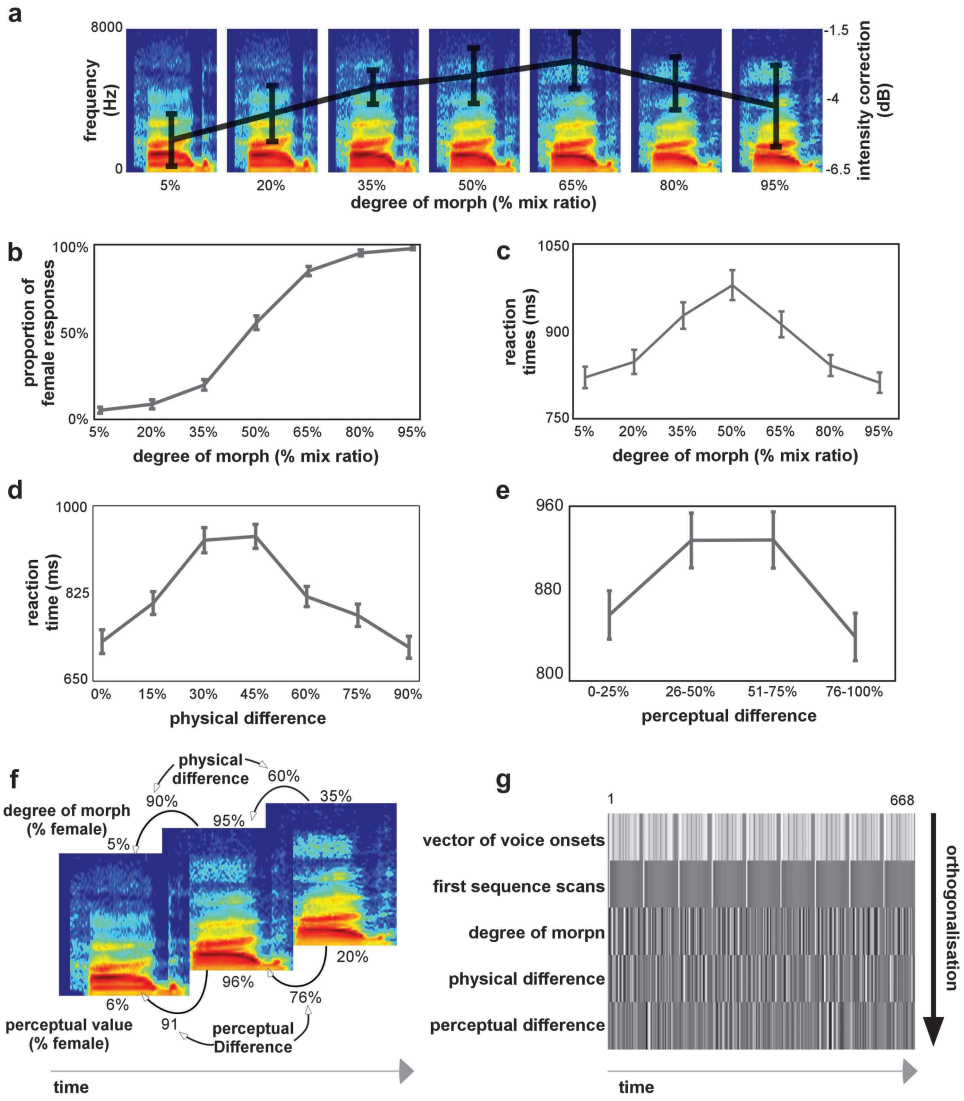

Figure 1. (*a*) Stimuli were voices derived from male-to-female voice gender continua. Example of a continuum where the physical interpolation was done with mix ratios increasing by 15%. Superimposed on the continua is the average intensity level correction for each degree of morph derived from the pilot study on perceived loudness. The error bars show the standard error computed from the average variation in intensity correction between the 9 different continua. (*b*) Voice gender psychophysical function: the group-average proportion of female responses is shown as a function of the degree of morph. For panels *b*–*e*, the error bars show the standard error computed from the individual subject's classifications. (*c*) Reaction times: the group-average reaction times as a function of the voices' degree of morph. (*d*) Reaction times: the group average reaction times as a function of the physical difference with the previously heard stimulus. (*e*) Reaction times: the group average reaction times as a function of the perceptual difference with the previously heard stimulus. (*f*) Illustration example for the definition of the parametric regressors: Spectrogram examples of 3 consecutive voices in the stimulation sequence. Shown above the spectrogram is the voice's respective degree of morph value and the physical difference with the previous stimulus. Shown below the spectrogram is the voice's perceptual value (for a given subject) and the perceptual difference with the previous stimulus. (*g*) Design matrix: the first row of the design matrix defines the stimulus onsets for the 9 continua (each continuum separated by baseline trials). The following rows represent the parametric regressors included in the general linear model. The first regressor models the first stimulation in a sequence. Since this stimulus is not preceded by another voice, we model it out of the regression. The second regressor models the voice's degree of morph. The third regressor models the physical difference between consecutive voices, and finally, the fourth regressor models the perceptual difference between consecutive voices.

Stimuli

Recordings of natural male and female voice stimuli were used to construct 9 voice gender continua via auditory morphing. These recordings consisted of male (n = 3) and female (n = 3) adult speakers uttering the syllables "had," "heed," or "hood," taken from the database of American English vowels described in Hillenbrand et al. (1995). Three female–male pairs were constituted by randomly assigning each female voice with a different male voice and were used to generate the continua (3 voices per gender * 3 vowels). The morphing procedure was performed using STRAIGHT (Kawahara 2003, 2006) in Matlab (The MathWorks, Inc., Natick, MA). STRAIGHT performs an instantaneous pitch-adaptive spectral smoothing in each stimulus to separate the contributions of the glottal source (including F0) versus supralaryngeal filtering (distribution of spectral peaks, including the first formant, F1; Ghazanfar and Rendall 2008) to the voice signal. Voice stimuli are decomposed by STRAIGHT into 5 parameters: fundamental frequency (F0), formant frequencies, duration, spectrotemporal density, and aperiodicity; each parameter can be independently manipulated. Anchor points, that is, time–frequency landmarks, were identified in each individual sound on the basis of landmarks easily recognizable on each spectrogram. Temporal anchors were onset and offset of phonation and burst of the "d." Spectrotemporal anchors were first, second, and third formants at onset of phonation, onset of formant transition, and end of phonation. Using the temporal landmarks, each continuum was equalized in duration (557 ms). Morphed stimuli were then generated by resynthesis based on a logarithmic interpolation of

female and male anchor templates and spectrogram in steps of 15%. We thus obtained, for each of the 9 male–female original voice pairings, a continuum of 7 voices ranging from 95% female (resynthesized female stimulus) to 95% male (resynthesized male stimulus) with 7 gender-interpolated voices in 15% steps (95% female–5% male; 80% female–20% male; . . .; 5% female–95% male; see Fig. 1*a*). Noteworthy, interpolated voices sounded natural, that is, as if produced by a real human being, as a result of the independent interpolation and resynthesis of the source and filter components of the voices. We further controlled for the potential contribution of differential frequency distributions in male and female voices (i.e., greater energy in higher frequencies for female voices) by matching all stimuli in perceived loudness (Fig. 1*a*). Intensity correction levels were obtained from a pilot experiment with 3 subjects, where each voice was compared in terms of perceived loudness with a random voice selected from the set of 63 voices. Examples of stimuli are provided as supplementary audio files.

Stimulus Presentation

Stimuli were presented using Media Control Functions (DigiVox, Montreal, Canada) via electrostatic headphones (NordicNeuroLab, Norway) at a sound pressure level of 80 dB as measured using a Lutron Sl-4010 sound level meter. Before they were scanned, subjects were presented with sound samples to verify that the sound pressure level was comfortable and loud enough considering the scanner noise.

Experimental Design and Task

We used a continuous carryover experimental design (Aguirre 2007). This design allows measuring both the direct effects (effect of voice gender) and the repetition suppression, which can be observed not only in pairs of voices (like the typical fMRI adaptation experiments) but also in the continuous modulation of response to voices presented in an unbroken stream (i.e., the modulation of activity to a stimulus by the preceding stimulus; Aguirre 2007). All voice gender continua (n = 9) were presented in one single echo-planar imaging (EPI) run of 24 min. The order of the continua was counterbalanced across subjects. The stimulus sequence within a continuum was determined using an n = 8 (7 morph steps plus 1 silent null event) type 1 index 1 sequence (ISI: 2s Nonyane and Theobald 2007), which shuffles stimuli within the continuum so that each stimulus is preceded by itself and every other within-continuum stimuli in a balanced manner. There were thus 8 repetitions of a stimulus per continuum. Each continuum sequence lasted around 2.25 min (71 volumes) and the sequences for the different continua were separated by a silent baseline of 18 s (9 volumes).

Task

Participants were instructed to perform a 2AFC voice gender classification task using 2 buttons of an MR compatible response pad (NNL technologies; button order counterbalanced across the subjects). Reaction times (relative to sound onset) were collected using MCF with a response window limited to the trial duration.

Magnetic Resonance Imaging

Localization of the TVAs (Functional Localizer Experiment)

A functional localizer of the TVAs was conducted for each subject. This consisted of a 10 min fMRI scan measuring the activity in response to either vocal or nonvocal sounds (Belin et al. 2000; Pernet et al. 2007) using an efficiency-optimized design. The comparison of responses to vocal and nonvocal sounds reliably highlights the TVAs: bilateral auditory cortical regions presenting greater activity in response to sounds of voice. Stimuli are available for download at http:\\vnl.psy.gla.ac.uk. The independent functional localizer was used in voxel selection/region of interest (ROI) definition. Furthermore, its aim was to identify whether statistical maps from the voice gender carryover experiment overlapped with the TVA.

Continuous Carryover Functional Measurements

Blood oxygen level–dependent (BOLD) measurements were performed using a 3.0-T Siemens TIM Trio scanner with a 12-channel head coil. We acquired 668 EPI image volumes for the carryover experiment (32 axial slices, time repetition [TR] = 2000 ms, time echo [TE] = 30 ms, flip angle [FA] = 77, 3 mm^3). The first 4 s of the functional run consisted of "dummy" gradient and radio frequency pulses to allow for steady state magnetization during which no stimuli were presented and no fMRI data collected. MRI was performed at the Centre for Cognitive Neuroimaging (CCNi) in Glasgow, United Kingdom.

Anatomical Measurements

High-resolution T_1-weighted structural images were collected in 192 axial slices and isotropic voxels (1 mm^3; field of view: 256 × 256 mm^2, TR =1900 ms, TE = 2.92 ms, time to inversion = 900 ms, FA = 9°).

Statistical Analysis

fMRI Data Preprocessing

Data analysis was performed using Statistical Parametric Mapping (SPM8; Welcome Department of Cognitive Neurology). All images were realigned to correct head motion with the first volume of the first session as reference. T_1-weighted structural images were coregistered to the mean image created by the realignment procedure and were used for normalization of functional images onto the Montreal Neurological Institute Atlas using normalization parameters derived from segmentation of the anatomical image. Finally, each image was smoothed with an isotropic 8 mm full-width at half-maximum Gaussian kernel.

General Linear Model

EPI time series were analyzed using the general linear model as implemented in SPM8. For each subject (first-level analysis), the localizer and the voice gender tasks were modeled separately.

For the voice localizer, voices and nonvoices were modeled as events using the canonical hemodynamic response function (HRF; SPM8), and one contrast per stimulus type was computed. A "voice greater than nonvoice" contrast was created for each subject, which was used at the group level (second-level analysis) in a one-sample t-test to identify the TVA (Fig. 2*a*).

For the voice gender task, the first analysis was to identify voxelwise signal changes that reflect direct and carryover effects of the voice continua. At the first level, each stimulus presentation (voice onsets) was modeled using the canonical HRF and parametric regressions were implemented using degree of morph (mix ratio, i.e., direct effect of voice gender), physical difference (absolute difference in mix ratio between 2 consecutive stimuli), and perceptual difference (absolute difference in perceived femaleness between 2 consecutive stimuli obtained from the behavioral task on an individual subject basis) as covariates (Figs 1*f*,*g* and 2*b*,*c*). A regressor modeling the first voice stimulation of each sequence was added in order to model the carryover effect of a stimulus following baseline EPI acquisition, which could potentially have added noise to the carryover effects. At the second level (i.e., across subjects), a one-sample t-test was performed on each regressor.

In order to visualize the parameter estimates related to the parametric regressions described above, regression coefficients were extracted using in-house Matlab (The MathWorks, Inc.) and SPM programs at the peak voxel of the significant ROI (single subject level) from the carryover analyses and the TVAs. The degree of morph, physical difference, and perceptual difference regressors (after convolution) were then multiplied by these coefficients. The between-subject average regression functions and standard error of the mean were then computed and are displayed in Figure 3 for each ROI and each parametric regressor (Fig. 3—side panels). Noteworthy, the regression coefficients for the parametric regressor modulating the perceived difference were "binned" in steps of 25% to account for interindividual differences in perceived gender and to allow averaging across subjects. The inflated cortical surfaces used for displaying results in Figure 2 were created using Caret (Van Essen C et al. 2001; Van Essen DC et al. 2001).

Behavioral Analysis

We computed a multiple regression to investigate the relative contribution of the degree of morph, the physical difference, and the

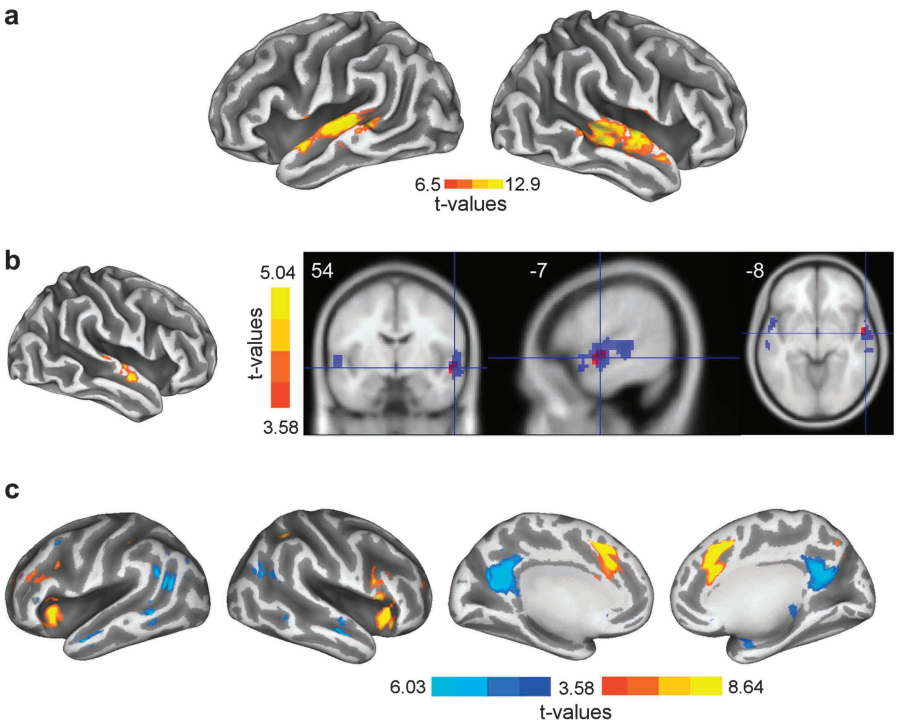

Figure 2. (*a*) Inflated cortical surface depicting the TVAs in both hemispheres. Throughout Figure 2, the color bar depicts the statistical lower boundary and global maxima. (*b*) Left: inflated cortical surface showing the activation map obtained with the physical difference regressor. Right: axial, sagittal, and coronal slices showing the TVAs (in blue) and the effect of physical difference (in red). Note how the effect of physical difference overlaps with the anterior parts of the TVAs. (*c*) Inflated cortical surfaces showing the activation maps for the perceptual difference regressor. On the 2 inflated cortical surfaces on the left, are shown the activity maps in the bilateral inferior frontal gyrus/insulae (positive; color map red-to-yellow) and middle temporal gyrus (negative; color map dark-to-light blue). On the 2 right most inflated cortical surfaces is depicted the bilateral activity maps for the anterior cingulate gyrus (positive; color map red-to-yellow) and precuneus (negative; color map dark blue-to-light blue).

perceptual difference between consecutive stimuli on the reaction times in individual subjects. The second order polynomial expansion of these regressors (degree of morph, physical difference, and perceptual difference) was included in our model. Regression coefficients were obtained for each subject independently, and a percentile bootstrap procedure was used on each parameter to test for between-subject significant contributions. The percentile bootstrap test was computed as follow: we sampled with replacement from the original distributions of between-subject regression coefficients and calculated the mean of each resampled distribution. This was performed 10 000 times and lower and upper confidence boundaries were obtained from this distribution of the bootstrapped means. The null hypothesis was rejected on the significance level alpha = 0.05 if 0 was not included in the two-tailed confidence interval.

Results

Behavioral Results

Behavioral results yielded the classical sigmoid-like psychometric function from the gender classification task, with a steeper slope at central portions of the continua (Fig. 1*b*). The percentages of female identification were of 6.1% (±1.7%) for the 5% female voice and 96.9% (±0.8%) for the 95% female voice of the continua. The 50% ambiguous male–female voice was identified as female 55.3 times of 100 (±3.9%). We observed faster reaction times on average at the extremities of the continua (801.8 ± 22.1 and 790.8 ± 20.7 ms) and the ambiguous 50% male–female voices needed more time to be classified on average (990.1 ± 30.6 ms; Fig. 1*c*). Because we were interested in carryover effects of a voice on the consecutive one, we computed the reaction times as a function of physical difference between 2 consecutive stimuli (Fig. 1*d*). For repeated consecutive voices (0% physical difference) or clear gender change (90% physical difference), the reaction times were 729.2 ± 23.5 and 717.9 ± 21.9 ms, respectively. On the other hand, for consecutive voices with an intermediate physical difference (45%), voice gender identification decisions were slower to achieve (938.8 ± 23.9 ms). This effect was also observed when computing the reaction times as a function of perceptual difference between 2 consecutive stimuli (Fig. 1*e*). For consecutive voices with low perceptual change (0–25% perceptual difference) or clear perceptual change (76–100% perceptual difference), the reaction times were 860.6 ± 21.9 and 840.8 ± 21.5 ms, respectively. On the other hand, for consecutive voices with intermediate perceptual changes (26–50% and 50–75%), voice gender identification decisions were slower to achieve (928.1 ± 24.8 and 928.4 ± 25.4 ms).

Effect of Degree of Morph on Reaction Times

The between-subject contribution of the degree of morph parameter on the reaction times was significant ($P < 0.05$; average coefficient value = –40.86 [–69.64 –12.98]). This indicates a significant longer response time for gender-ambiguous voices on the continua.

Effect of Physical and Perceptual Difference on Reaction Times

We also observed significant effects of the physical difference ($P < 0.05$; average coefficient value = –128.98 [–165.66 –89.69]) and the perceptual difference ($P < 0.05$; average coefficient value = 98.90 [69.11 128.89]) parameters on the reaction times.

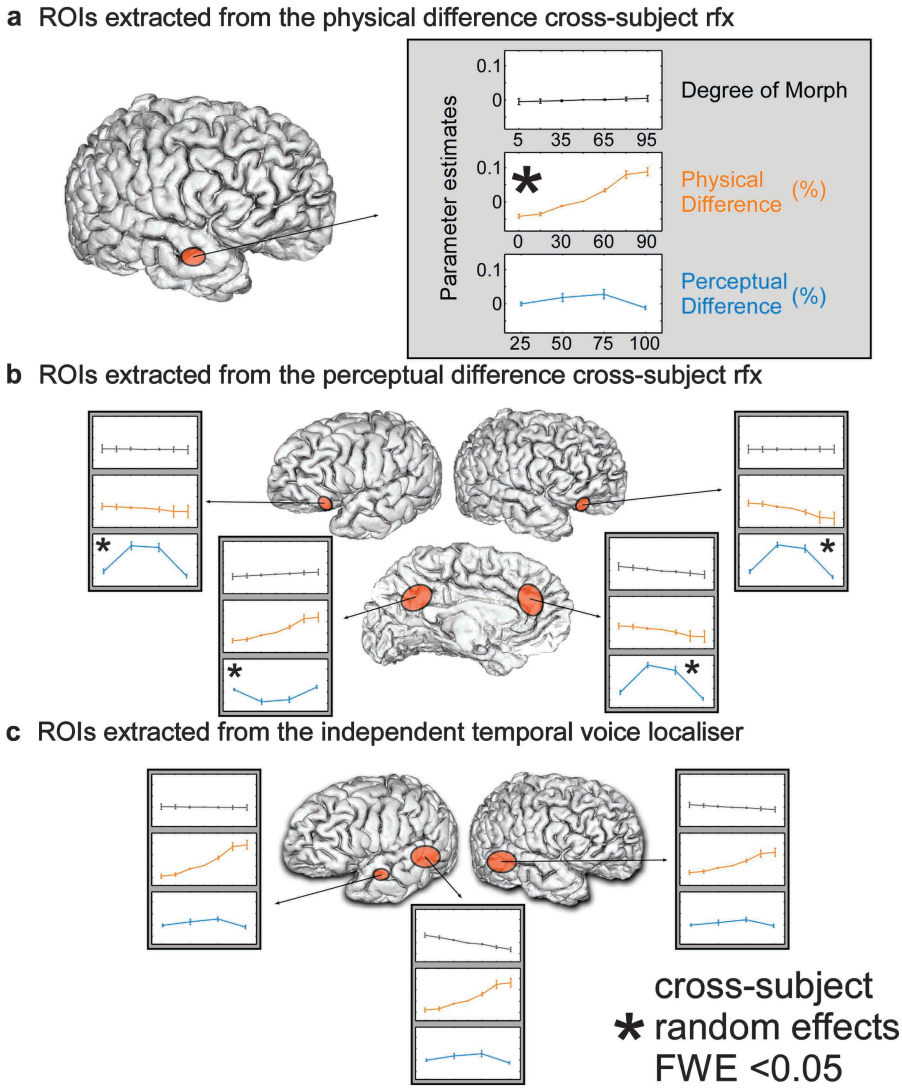

Figure 3. Parameter estimates and standard error for the contrasts investigated with the parametric regressors. (*a*) The parameter estimates for the anterior STS ROI in the right hemisphere that we obtained from the physical difference regressor. (*b*) The parameter estimates for the bilateral inferior frontal gyrus, precuneus, and ACC that we obtained from the perceptual difference regressor. (*c*) The parameter estimates for the left hemisphere posterior and anterior and right hemisphere posterior regions that we obtained from the functional voice localizer.

Altogether, this indicates an important influence of the previously heard voice on voice gender identification (Fig. 1*d*,*e*).

fMRI Results

Temporal Voice Areas

The TVAs identified by the independent functional localizer were located as expected along the upper bank of the superior temporal sulcus (STS); 3 clusters were identified surviving a threshold of 6.5 (threshold T value for a $P < 0.05$ familywise error [FWE] corrected, see Table 1 and Fig. 2*a*).

Effect of Voice Gender

As hypothesized by the overlapping neuronal population model, the regressor modeling the degree of morph did not reveal any regions showing greater activity to either one of the continuum end points ($P > 0.001$, uncorrected, i.e., no differences males vs. females). To further visualize this absence of effect, parameter estimates are displayed in Figure 3*a*–*c* (degree of morph panels).

Carryover Effect of Voice Gender Physical Difference

When analyzing the carryover effect, we observed significant repetition suppression effects in the anterior portions of the right STS, overlapping with the independently localized TVA (Fig. 2*b*): in this region, the smaller was the physical difference between stimuli, the lower/smaller was the BOLD signal $T_{1,19} = 4.55$, $P < 0.05$ (FWE-corrected cluster level; Fig. 3*a*, middle panel; Table 3).

Carryover Effect of Voice Gender Perceptual Difference

We then investigated the effects of perceptual difference between stimuli, included as an additional regressor in order to examine variance not explained by the physical difference (in the SPM design matrix, parametric regressors are orthogonalized, thus because the perceptual difference was entered after

Table 1
Acoustical properties of the continua

Utterance	Proportion of male voice (%)	F0 (Hz)	F1 (Hz)	F1 bandwidth (Hz)	F2 (Hz)	F2 bandwidth (Hz)	F3 (Hz)	F3 bandwidth (Hz)	F4 (Hz)	F4 bandwidth (Hz)	HNR (dB)	Jitter (μs)	Shimmer (dB)	Low-frequency energy (dB)	High-frequency energy (dB)
Had	5	208	935	131	1630	202	2773	228	4348	351	19	0.57	0.68	42	35
	20	192	902	129	1600	192	2749	226	4294	423	19	0.49	0.66	44	32
	35	177	878	120	1558	183	2703	194	4196	572	19	0.42	0.6	43	32
	50	163	851	103	1523	155	2661	183	4056	590	19	0.46	0.66	44	31
	65	151	820	101	1488	157	2612	185	3882	626	18	0.6	0.73	44	31
	80	139	801	111	1460	157	2568	158	3764	384	17	0.43	0.76	44	31
	95	129	786	117	1427	161	2527	170	3690	305	15	0.57	0.69	44	31
Heed	5	211	729	95	1995	205	2914	206	4523	396	21	0.71	0.59	44	24
	20	194	699	89	2003	177	2861	180	4545	342	21	0.65	0.54	44	25
	35	178	671	84	1890	177	2730	173	4257	404	21	0.51	0.58	44	26
	50	163	653	71	1916	190	2711	233	3933	2865	20	0.39	0.64	44	27
	65	149	629	59	1860	186	2696	191	3754	1382	19	0.84	0.74	44	27
	80	137	610	52	1817	154	2639	239	3799	902	18	0.56	0.77	44	28
	95	126	592	43	1782	139	2574	209	3736	917	17	0.67	0.8	44	28
Hood	5	228	506	65	1332	102	2773	143	4381	366	28	0.38	0.49	44	21
	20	207	508	80	1357	120	2724	138	4299	407	28	0.33	0.45	44	22
	35	188	522	70	1340	140	2662	136	4161	380	27	0.58	0.48	44	23
	50	171	511	54	1339	136	2590	143	3981	634	27	0.37	0.51	44	22
	65	155	495	58	1342	163	2534	152	3783	534	26	0.44	0.62	44	22
	80	141	499	80	1351	208	2488	165	3627	340	24	0.49	0.75	44	22
	95	128	483	71	1365	251	2445	167	3502	267	24	0.48	0.64	44	22

Note: F0: fundamental frequency in hertz. F1–F4: frequency of the first to the fourth formant in hertz. HNR: harmonic-to-noise ratio in decibel. Jitter and Shimmer reflect variation of pitch and loudness expressed in microseconds and decibel, respectively. The summed energy between 50 Hz–1 kHz (low-frequency energy in decibel) and 1–5 kHz (high-frequency energy in decibel) was computed from the long-term average spectrum between 0 and 6700 Hz.

Table 2
Temporal voice areas

Voice > nonvoice	Coordinates (mm)			*T* values	*P* values	Cluster size
	x	*y*	*z*			
Left						
Mid-STG	−60	−31	4	10.18	0.001	186
aSTS	−51	−5	11	11.25	0.001	52
Right						
Mid-STG	54	−28	4	12.93	0.001	289

Note: Whole-brain analysis. Clusters surviving a threshold of $T > 6.5$ (FWE, $P < 0.05$). STG, Superior Temporal Gyrus.

the physical difference, the effect observed corresponds to variations in the BOLD signal than cannot be explained by physical differences between stimuli—Fig. 1*g*). As we did for physical difference, we searched for regions showing repetition–suppression effects, that is, linear decrease of BOLD magnitude as the perceptual difference between consecutive stimuli decreased. This linear regression yielded bilateral effects in the inferior prefrontal cortices (IFGs), insulae, and the anterior cingulate cortex (ACC) (Fig. 2*c*; $T_{1,19} = 3.58$, $P < 0.001$ [FWE-corrected cluster level]). Interestingly, the BOLD signal magnitude as a function of perceptual difference between consecutive stimuli followed, in these regions, a quadratic polynomial expansion (Fig. 3*b*). When consecutive stimuli were perceived as very similar or very different (bins 0–25% or 76–100%), BOLD signal magnitude was lower than when consecutive stimuli were involving the 50% ambiguous voice (bins of 26–50% and 51–75%; Fig. 3*b*). This is in line with the subject's reaction times for which we observed an inverted U-shaped function where the 50% ambiguous stimuli led to slower voice gender decisions (Fig. 1*c*). We also observed a modulation of the BOLD signal in the precuneus, with greater (negative) magnitude for the 26–50% and 51–75% bins of perceptual difference (Fig. 3*b*; Table 4).

Note that for illustration purposes, we have plotted the parameter estimates for all of the regions that were observed in the fMRI analyses described above (carryover effects and functional localizer). For the degree of morph parameter, in most of the ROIs, the shape of the average regression function was flat, indicating the absence of effect of voice gender on the magnitude of the BOLD response (albeit a trend for stronger responses to male voices in the left posterior STS, which did not reach statistical significance).

For the physical difference parameter, we observed a trend for an increased magnitude of the BOLD signal as a function of physical difference in the ROIs defined from the TVA localizer and the precuneus and a decreased magnitude of BOLD signal in the right IFGs/insulae (Fig. 3—physical difference panels).

Finally, for the perceptual difference parameter, we observed a trend for a quadratic polynomial expansion of the average regression function in the ROIs defined from the temporal voice localizer and in the right anterior STS (aSTS) (Fig. 3—perceptual difference panels).

Discussion

We used auditory morphing technologies to generate voice gender continua in conjunction with a continuous carryover design to investigate the cerebral correlates of voice gender perception. Our aim was to disentangle between cerebral processes related to voice gender ("direct effect," i.e., spatially segregated neurons preferring male or female voices), voice gender repetition suppression effects (overlapping populations of male/female sensitive neurons), and higher order cognitive voice gender perception processes.

Voice Gender Behavioral Effect

We observed a good identification of the male and female portions of the continua, with slower reaction times on average for the voice gender ambiguous portions in line with recent behavioral data (Mullennix et al. 1995; Schweinberger et al. 2008). Furthermore, we observed a significant influence of context on the perception of voice gender indicated by

Table 3
Effects of physical difference

	Coordinates (mm)			*T* values	*P* values	Cluster size
	x	*y*	*z*			
aSTS	54	−7	−8	5.04	0.001	25

Note: Whole-brain analysis. Clusters surviving a threshold of $T > 3.58$ (FWE, $P < 0.05$).

Table 4
Effects of perceptual difference

	Coordinates (mm)			*T* values	*P* values	Cluster size
	x	*y*	*z*			
Left						
IFG	−33	20	4	5.73	0.001	186
Right						
IFG	51	20	1	6.18	0.001	240
ACC	−9	14	49	8.64	0.001	404
Precuneus	−6	−55	19	6.03	0.001	347

Note: Whole-brain analysis. Clusters surviving a threshold of $T > 3.58$ (FWE, $P < 0.05$).

changes in reaction times according to the physical difference and perceptual difference between consecutive stimuli.

Absence of Female Voice Effect in the Brain

A surprising result of this study is the absence of a larger brain response for female than male voices in the auditory cortex as reported by previous studies (Lattner et al. 2005; Sokhi et al. 2005). We did not observe a single brain region showing significant modulation of BOLD signal by the degree of morph of the voice in one direction (an increase of signal coupled with an increase of stimulus femaleness) or the other (an increase of signal coupled with an increase of stimulus maleness).

This difference could arise from differences in materials between the previous reports and this experiment. The vocal stimuli used in Lattner et al. (2005) and in Sokhi et al. (2005) were a combination of at least 2 words, whereas we used simple brief stimuli. Using a combination of words preserves information relative to the temporal dynamics that is largely absent from simple syllables, and the temporal dynamics of a voice, part of the prosody, is an important cue to categorize voice (Murry and Singh 1980; Andrews and Schmidt 1997). Thus, a simple explanation in term of processing temporal dynamics of the voice could partly justify the discrepancies between our study and previous reports. Another potential explanation of this result relies in the perceived loudness of the voice. Because the formant frequencies and F0 are both higher on average for female voices, female voices might be perceived as louder than the male voices, thus resulting in a larger brain activity (Langers et al. 2007). In the present study, the stimuli were controlled for perceived loudness via a pilot experiment in which subjects increased or decreased the intensity of the voices when comparing to a randomly selected reference voice from the stimulus set. Thus, using a well-controlled set of stimuli in terms of loudness, duration, and temporal variation, we did not replicate previous results showing larger activity for female than male voices, suggesting that these differences reflected more low-level differences than gender processing per se.

Repetition Suppression as a Function of Physical Difference

Lattner et al. (2005) and Sokhi et al. (2005) reported different brain regions processing male and female voices in the human brain. From a physiological point of view, it would make more sense if a single brain region would process voice gender. Here, we tested the hypothesis of overlapping neuronal populations encoding voice gender in the auditory cortex and the TVAs by including a regressor modeling the voice gender physical difference of 2 consecutive voices in the stimulation sequence. In an adaptation framework, 2 consecutive gender-similar voices (low physical difference) should lead to a reduction of BOLD signal. As the physical difference between 2 consecutive voices increases, the 2 voices become more distinctive on a gender basis (male or female) and recovery from adaptation should increase. Our data showed a significant linear modulation of BOLD signal in relation with increasing physical differences as observed in Figure 3*a* in the right anterior temporal lobe, along the upper bank of the STS.

Is the Right aSTS Voice Gender Specific?

Previous studies have shown the involvement of the anterior part of the STS in an acoustic-based representation of sounds in general (Zatorre et al. 2004; Leaver and Rauschecker 2010). Hence, our results should be interpreted with care in terms of voice gender selectivity. Indeed, fMRI adaptation results have often proven to be more complex than assumed, and only when combined with prior knowledge, perhaps some electrophysiological evidence and great care can unequivocal interpretations about domain specificity be put forward (for more detailed discussions on the interpretations of fMRI and fMRI adaptation results, see Grill-Spector et al. 2006; Krekelberg et al. 2006; Logothetis 2008; Mur et al. 2010).

Here, we would like to argue that the repetition suppression effects we observed in the anterior part of the right STS are related to acoustical feature extraction related to voice cognition, like previously shown for speaker identity (Imaizumi et al. 1997; Belin and Zatorre 2003; Andics et al. 2010; Latinus et al. 2011).

A recent study made use of cutting-edge multivariate pattern analysis (MVPA) and fMRI to investigate whether an abstract representation of a vowel or speaker emerges from the encoding of information in the human temporal lobes. Using spatially distributed activation patterns and a method based on support vector machine and recursive feature elimination, they were able to predict the nature (vowel or speaker) of the stimulus heard by the listener. Furthermore, they investigated the layout and consistency across subjects of the spatial patterns that made this decoding possible. They observed discriminative patterns distributed in early auditory regions and in specialized higher level regions that allow prediction of the nature of the stimuli. Noteworthy, they observed 3 clustered regions along the anterior–posterior axis of the right STS from which they could decode the speaker identity of the uttered vowels (Formisano et al. 2008). Interestingly, the most anterior right STS cluster in their discriminative maps resembles the region that we report here.

Sensitivity of Bilateral Inferior Frontal Gyrus and ACC to Task-Relevant Perceptual Changes

We observed a significant modulation of BOLD signal with perceptual differences between 2 consecutive items bilaterally in the inferior frontal gyrus covering part of the anterior insulae and in the ACC. This is consistent with recent voice perception studies conducted in macaques (Romanski et al. 2005; Cohen et al. 2006) and humans (Fecteau et al. 2005; Ethofer, Anders, Erb, Droll, et al. 2006) in which an involvement of prefrontal regions was reported.

More specifically, the inferior frontal gyrus was described to be involved in abstract self-representations (Nakamura et al. 2001; Kaplan et al. 2008), vocal affect evaluation (Imaizumi et al. 1997; Wildgruber et al. 2005; Ethofer, Anders, Erb, Herbert, et al. 2006; Johnstone et al. 2006), decision making, task difficulty, and attentional resources (Binder et al. 2004; Heekeren et al. 2004; Heekeren et al. 2008). The ACC has also been described to be involved in making decisions on highly ambiguous questions (Botvinick et al. 1999) and response competition/conflict (Carter et al. 1998; Kerns et al. 2004; Wendelken et al. 2009).

The voice gender perceptual difference effect that we observed involving bilateral IFG/insulae and the ACC is thus in line with most of the research describing their role as a higher cognitive function related to decision making, reasoning, sorting ambiguous stimuli in difficult decisions, etc. The longer reaction times and greater BOLD signal when presented with the 50% ambiguous male–female voices provide evidence for longer reasoning, increased attention, and more computation for selection procedure when hearing gender-ambiguous voices.

Finally, Andics et al. (2010) reported regions showing long-term neural sharpening effects induced by the explicit categorization feedback during training of voice identity recognition. They interpreted this reduction of BOLD signal as "trained category mean voice" representations, probably involved in a longer term categorical representation of voice identity (Andics et al. 2010). In a similar way, the prefrontal and anterior cingulate regions, which showed BOLD signal reductions when 2 consecutive voices had peripheral perceptual difference (either small or no change in voice gender or large or complete gender change), could therefore also be an indication of a long-term categorical representation of voice gender.

The inverse pattern of activity that we observed in the IFG/insulae and ACC was observed in the precuneus/posterior cingulate cortex (Fig. 3*b*). One possible interpretation is in terms of the "default network" (Shulman, Corbetta, et al. 1997; Shulman, Fiez, et al. 1997; Raichle et al. 2001; McKiernan et al. 2003; Buckner et al. 2008). In this framework, the greater is the stimulus complexity/ambiguity, reasoning necessity, task demands, the more negative the BOLD signal is (Kalbfleisch et al. 2007), consistent with our results.

Cerebral Organization of Voice Gender Perception

We observed an extraction of voice gender–related acoustical features in regions overlapping with the TVAs (repetition suppression as a function of physical difference—Figs 2 and 3—aSTS). This is in line with previous results where adaptation to voice identity along the anterior portions of the STS was reported (Belin and Zatorre 2003; Latinus et al. 2011). Recently, the anterior STS has been described as carrying an "acoustic signature" of sounds, in line with the processes of acoustic feature extraction related to voice gender that we describe in this experiment (Leaver and Rauschecker 2010). Second, we observed higher level cognitive processes related to voice gender perception in ACC/IFG/Insulae (repetition suppression as a function of perceptual difference—Figs 2*c* and 3*b*).

We suggest that the activity observed in the prefrontal cortex could be related to stimulus ambiguity and long-term voice gender representations because ambiguous voices were more difficult to rate as male or female, less categorically defined as one or the other gender, thus requiring more energy for decision making. Altogether, we suggest that the cerebral processing of voice and voice gender involves multiple stages, where acoustically relevant information is processed in the anterior part of the STS followed by an involvement of the IFG and ACC where higher level cognitive processes related to the perception of voice characteristics influence the subject's decision making.

Funding

United Kingdom Biotechnology and Biological Sciences Research Council (BBSRC) (BB/E003958/1).

Notes

We would like to thank Prof. Tim Griffiths, Dr Nikolaus Kriegeskorte, and Dr Geoffrey Aguirre for their insights and valuable comments on this work. We would also like to thank Eleftherios Garyfallidis. *Conflict of Interest*: None declared.

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.

Andrews ML, Schmidt CP. 1997. Gender presentation: perceptual and acoustical analyses of voice. J Voice. 11:307–313.

Belin P, Zatorre RJ. 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 14:2105–2109.

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

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD. 2004. Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci. 7:295–301.

Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD. 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature. 402:179–181.

Buckner RL, Andrews-Hanna JR, Schacter DL. 2008. The brain's default network: anatomy, function, and relevance to disease. Ann N Y Acad Sci. 1224:1–38.

Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD. 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science. 280:747–749.

Childers G, Wu K. 1991. Gender recognition from speech. Part II: fine analysis. J Acoust Soc Am. 90:1841–1856.

Cohen YE, Hauser MD, Russ BE. 2006. Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex. Biol Lett. 2:261–265.

Ethofer T, Anders S, Erb M, Droll C, Royen L, Saur R, Reiterer S, Grodd W, Wildgruber D. 2006. Impact of voice on emotional judgment of faces: an event-related fMRI study. Hum Brain Mapp. 27:707–714.

Ethofer T, Anders S, Erb M, Herbert C, Wiethoff S, Kissler J, Grodd W, Wildgruber D. 2006. Cerebral pathways in processing of affective prosody: a dynamic causal modeling study. Neuroimage. 30:580–587.
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.
Ghazanfar AA, Rendall D. 2008. Evolution of human vocal production. Curr Biol. 18:R457–R460.
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.
Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG. 2004. A general mechanism for perceptual decision-making in the human brain. Nature. 431:859–862.
Heekeren HR, Marrett S, Ungerleider LG. 2008. The neural systems that mediate human perceptual decision making. Nat Rev Neurosci. 9:467–479.
Hillenbrand J, Getty LA, Clark MJ, Wheeler K. 1995. Acoustic characteristics of American English vowels. J Acoust Soc Am. 97:3099–3111.
Imaizumi S, Mori K, Kawashima R, Sugiura M, Fukuda H, Itoh K, Kato T, Nakamura A, Hatano K, Kojima S, et al. 1997. Vocal identification of speaker and emotion activates different brain regions. Neuroreport. 18:2809–2819.
Johnstone T, van Reekum CM, Oakes TR, Davidson RJ. 2006. The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. Soc Cogn Affect Neurosci. 1:242–249.
Kalbfleisch ML, Van Meter JW, Zeffiro TA. 2007. The influences of task difficulty and response correctness on neural systems supporting fluid reasoning. Cogn Neurodyn. 1:71–84.
Kaplan JT, Aziz-Zadeh L, Uddin LQ, Iacoboni M. 2008. The self across the senses: an fMRI study of self-face and self-voice recognition. Soc Cogn Affect Neurosci. 3:218–223.
Kawahara H. 2003. Exemplar-based voice quality analysis and control using a high quality auditory morphing procedure based on straight. In: VoQual 03: Voice Quality: Functions, Analysis and Synthesis, 2003 August 27–29; Geneva (Switzerland): ISCA Tutorial and Research Workshop.
Kawahara H. 2006. STRAIGHT, exploitation of the other aspect of vocoder: perceptually isomorphic decomposition of speech sounds. Acoust Sci Technol. 27:349–353.
Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. Science. 303:1023–1026.
Kreiman J. 1997. Listening to voices: theory and practice in voice perception research. In: Johnson K, Mullennix JW, editors. Talker variability in speech processing. San Francisco, CA: Morgan Kaufmann Publishers. p. 85–108.
Krekelberg B, Boynton GM, van Wezel RJA. 2006. Adaptation: from single cells to BOLD signals. Trends Neurosci. 29:250–256.
Langers DRM, van Dijk P, Schoenmaker ES, Backes WH. 2007. fMRI activation in relation to sound intensity and loudness. Neuroimage. 35:709–718.
Latinus M, Crabbe F, Belin P. 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex. doi: 10.1093/cercor/bhr077.
Lattner S, Meyer ME, Friederici AD. 2005. Voice perception: sex, pitch, and the right hemisphere. Hum Brain Mapp. 24:11–20.
Leaver AM, Rauschecker JP. 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci. 30:7604–7612.
Logothetis NK. 2008. What we can do and what we cannot do with fMRI. Nature. 453:869–878.
McKiernan KA, Kaufman JN, Kucera-Thompson J, Binder JR. 2003. A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. J Cogn Neurosci. 15:394–408.
Mullennix JW, Johnson KA, Topcu-Durgun M, Farnsworth LM. 1995. The perceptual representation of voice gender. J Acoust Soc Am. 98:3080–3095.
Mur M, Ruff DA, Bodurka J, Bandettini PA, Kriegeskorte N. 2010. Face-identity change activation outside the face system: 'release from adaptation' may not always indicate neuronal selectivity. Cereb Cortex. 20:2027–2042.
Murry T, Singh S. 1980. Multidimensional analysis of male and female voices. J Acoust Soc Am. 68:1294–1300.
Nakamura K, Kawashima R, Sugiura M, Takashi K, Nakamura A, Hatano K, Nagumo S, Kubota K, Ito K, Kojima S. 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia. 39:1047–1054.
Nonyane BAS, Theobald CM. 2007. Design sequences for sensory studies: achieving balance for carry-over and position effects. Br J Math Stat Psychol. 60:339–349.
Pernet CR, Charest I, Belizaire G, Zatorre RJ, Belin P. 2007. The temporal voice areas: spatial characterization and variability. Neuroimage. 36:S1–S168.
Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. 2001. A default mode of brain function. Proc Natl Acad Sci U S A. 98:676–682.
Romanski LM, Averbeck BB, Diltz M. 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. J Neurophysiol. 93:734–747.
Schweinberger SR, Casper C, Hauthal N, Kaufmann JM, Kawahara H, Kloth N, Robertson DMC, Simpson AP, Zäske R. 2008. Auditory adaptation in voice perception. Curr Biol. 18:684–688.
Shulman GL, Corbetta M, Buckner RL, Fiez JA, Miezin FM, Raichle ME, Petersen SE. 1997. Common blood flow changes across visual tasks: I. Increases in subcortical structures and cerebellum but not in nonvisual cortex. J Cogn Neurosci. 9:624–647.
Shulman GL, Fiez JA, Corbetta M, Buckner RL, Miezin FM, Raichle ME, Petersen SE. 1997. Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. J Cogn Neurosci. 9:648–663.
Sokhi DS, Hunter MD, Wilkinson ID, Woodruff PW. 2005. Male and female voices activate distinct regions in the male brain. Neuroimage. 27:572–578.
Van Essen C, Lewis W, Drury A, Hadjikhani N, Tootell B, Bakircioglu M, Miller I. 2001. Mapping visual cortex in monkeys and humans using surface-based atlases. Vision Res. 41:1359–1378.
Van Essen DC, Dickson J, Hanlon D, Anderson CH, Drury HA. 2001. An Integrated software system for surface-based analyses of cerebral cortex. J Am Med Inform Assoc. 8:443–459.
Wendelken C, Ditterich J, Bunge SA, Carter CS. 2009. Stimulus and response conflict processing during perceptual decision making. Cogn Affect Behav Neurosci. 9:434–447.
Wildgruber D, Riecker A, Hertrich I, Erb M, Grodd W, Ethofer T, Ackermann H. 2005. Identification of emotional intonation evaluated by fMRI. Neuroimage. 24:1233–1241.
Wu K, Childers G. 1991. Gender recognition from speech. Part I: coarse analysis. J Acoust Soc Am. 90:1828–1840.
Zatorre RJ, Bouffard M, Belin P. 2004. Sensitivity to auditory object features in human temporal neocortex. J Neurosci. 24:3637–3642.