Cerebral response to 'voiceness': a functional magnetic resonance imaging study

Guylaine Bélizaire[a,b], Sarah Fillion-Bilodeau[a], Jean-Pierre Chartrand[a], Caroline Bertrand-Gauvin[b] and Pascal Belin[a,c]

[a]Department of Psychology, [b]Research Centre of the Montreal Geriatric Institute, University of Montreal, Montreal, Canada and [c]Center for Cognitive Neuroimaging, (CCNi), Department of Psychology, University of Glasgow, Glasgow, UK

Correspondence and requests for reprints to Guylaine Bélizaire, MSc, Laboratoire de Neurocognition vocale, Département de Psychologie, Centre de Recherche en Neuropsychologie et Cognition, C.P. 6128, succursale Centre-Ville, Montréal, Québec, Canada H3C 3J7
Tel: + 1 514 343 6111 ext. 1495; fax: + 1 514 343 5787; e-mail: guylaine.belizaire@umontreal.ca

Received 27 September 2006; accepted 13 October 2006

We evaluated the response of the voice-selective areas of the auditory cortex to sound 'voiceness', that is, the degree to which an auditory stimulus resembles human voice. Normal participants were scanned using event-related functional magnetic resonance imaging while passively listening to stimuli drawn from a 'voiceness' continuum generated via auditory morphing between sounds of voice and sounds of musical instruments. The voice-selective areas of the left and right superior temporal sulcus did not show the expected relation between 'voiceness' and size effect. Instead, superior temporal sulcus activity seemed mostly driven by sound naturalness, with largest activity differences observed for the intermediate, voice-instrument hybrid stimuli. *NeuroReport* 18:29–33 © 2007 Lippincott Williams & Wilkins.

Keywords: audition, auditory morphing, functional magnetic resonance imaging, naturalness, superior temporal sulcus, voice perception

Introduction

Among the entire spectrum of auditory stimuli we deal with daily, the human voice is probably the most prominent one. In fact, it can be suggested that voice is an 'auditory face' because like its visual alter ego, it conveys rich paralinguistic information about the identity, the sex and the emotional state of the speaker [1]. This establishes voice as an essential species-specific communication channel.

Considering the evidence pointing towards the presence of a specific system dedicated to the analysis of face stimuli [2,3] and the assumption that similar principles of functional organization could be shared between the different sensory modalities [4], it seems relevant to assume the existence of a cerebral system specialized in the analysis of vocal stimuli.

This assumption has been supported by several functional magnetic resonance imaging (fMRI) studies. Belin *et al.* [5] specifically showed for the first time the existence of 'voice-selective areas' (VSAs) in the human auditory cortex: they observed restricted areas along the anterior part of the superior temporal sulcus with greater response to sounds of voice compared with nonvocal sounds. In a follow-up study, Belin *et al.* [6] demonstrated that the regions along the anterior right superior temporal sulcus presented this enhanced response to voice even for nonspeech vocal sounds (e.g. laughs, cries, humming, etc.), suggesting that these areas would specifically be involved in the perception of paralinguistic aspects of voice.

Little is known, however, concerning the functional properties of the VSA. In particular, we ignore if these regions are mostly sensitive to the 'voiceness' of sounds, that is, if they respond more strongly to sounds that resemble vocal sounds. So far, investigating this issue has proved to be difficult owing to the lack of appropriate tools allowing a sufficiently precise control of the acoustic structure of the stimuli. Technological advances in voice analysis/synthesis methods now allow realistic audio morphing. STRAIGHT (Speech Transformation and Representation based on Adaptative Interpolation of weiGHTed spectrograms) [7] is one of those innovative instruments. This channel VOCODER (VOice CODER), based on the source-filter theory of voice production [8], allows flexible control of the relevant parameters present in voice while preserving a high level of quality and realism [7].

Here, we evaluated the functional properties of the voice-selective regions of the auditory cortex by creating a continuum of 'voiceness', that is, a continuum generated using STRAIGHT between natural vocal and nonvocal sounds (musical instruments). We used fMRI in normal volunteers to measure the response of the VSA to stimuli drawn randomly from the continuum. We hypothesized that the VSA would be more responsive to stimuli drawn from the voice-end of the continua.

Material and methods

Participants

Naturalness

To quantify the ecological nature of the morphed sounds, 10 French-speaking participants (four undergraduate and six graduate students; two men; 10 right-handed; mean ± SD age = 23.5 ± 0.7 years; education = 16.6 ± 0.6) were selected. All the participants were recruited at the Centre de Recherche en Neuropsychologie et Cognition (CERNEC, Université de Montréal) and gave written informed consent.

Functional magnetic resonance imaging

French-speaking participants were recruited ($n = 14$; seven men; 12 right-handed; age: 23.9 ± 0.9 years; education: 16.6 ± 0.8) through advertisements posted at the Université de Montréal and at the Université du Québec à Montréal. The only exclusion criterion retained was the self-reported presence of auditory problems. The handedness was determined by asking the participants what was their dominant hand. Participants were compensated 50CAD$/h for their participation in the study.

Stimuli

Natural stimuli

Vocal and nonvocal stimuli were used. Vocal stimuli were collected from 40 speakers and were composed of speech as well as nonspeech vocalizations. The speech stimuli comprised words and nonwords in English, French or foreign languages, whereas the nonspeech stimuli consisted of laughs, sighs and onomatopoeia. The nonvocal stimuli consisted of sounds from nature (e.g. wind stream, etc.), musical instruments and the modern environment (e.g. cars, telephones, aeroplanes, etc.). All the sounds were normalized for energy levels root mean square (RMS) and each vocal or nonvocal stimulus had a duration of 500 ms. Those sounds were used in a 'localizer' scan session to localize precisely the superior temporal sulcus VSA for each participant.

'Voiceness' continua

Emotionally neutral monosyllabic vocal sounds and musical instrument sounds were used to create 20 'voiceness' continua. The vocal sounds were selected from the recordings of Hillenbrand *et al.* [8]. The syllable used ([hæd]) does not have any meaning in French (the maternal language of the recruited participants). Twenty different speakers were selected to create the continua (10 men). The musical instrument sounds used belonged to three classes of musical instruments: strings, brass and wind, and were selected from the Soundfont Internet library (http://www.hammersound.net/). All instrument sounds were synthesized from natural stimuli and were selected on the basis of their naturalness; all the selected stimuli were normalized for energy levels (RMS). The musical notes morphed with the masculine voices had a lower pitch than the ones morphed with the feminine voices in an effort to minimize pitch differences within each continuum.

Twenty voice-to-musical instrument continua were created using the morphing technique described by Kawahara and Matsui [9]. A pitch-adaptive spectral envelope extractor as implemented in STRAIGHT (http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTtrial) was used to perform a decomposition of the vocal and nonvocal stimuli in source and frequency (f0) characteristics, yielding a smooth spectro-temporal representation of the signals [10]. A set of parametric representations derived from this decomposition was then linearly interpolated between the two extremes and resynthesized to obtain stimuli morphed between the original vocal and nonvocal stimuli. The duration of the 'voiceness' continua were, respectively, 638.44 ± 13.72 and 635.56 ± 9.47 ms for the male and the female speakers.

For the fMRI experiments, only five steps of the continua were presented: degrees 1 (90% vocal and 10% nonvocal), 3 (70% vocal and 30% nonvocal), 5 (50% vocal and 50% nonvocal), 7 (30% vocal and 70% nonvocal) and 9 (10% vocal and 90% nonvocal). This ensured that all stimuli played to the participants contained a part of both vocal and nonvocal information. A total of 220 stimuli were created (11 degrees of morphing × 20 continua); 100 of these morphs (five steps × 20 continua) were used in the fMRI experiment.

Procedures and data analyses

'Voiceness' and naturalness ratings

It was a concern that the intermediary stimuli along the 'voiceness' continuum might sound unnatural, as they would hardly correspond to any physical source in the real world. Their naturalness was thus assessed in order to quantify their ecological nature. Each of the 220 different morphs as well as the 40 original unmorphed stimuli (two stimuli × 20 continua) were presented to participants in a pseudo-random order while they were sitting in a sound-proof cabin (Laboratoire de neurocognition vocale, Université de Montréal). For each stimulus, they had to perform two judgments by selecting a point on a visual analogue scale: a naturalness judgment (from very natural to very unnatural) as well as a 'voiceness' judgment (from very vocal to very nonvocal). Ratings along the visual analogue scales were linearly converted to an integer number ranging from 0 to 370. Differences in average naturalness and 'voiceness' between stimuli were assessed using a repeated-measure analysis of variance (ANOVA) with degree of morphing as a factor.

Functional magnetic resonance imaging

In the fMRI experiment, sounds drawn from the 'voiceness' continua were presented in a pseudo-random order at a 4.1-s stimulus-onset-asynchrony according to an event-related experimental design. The participants were asked to simply lie still in the scanner and listen to the sounds. The auditory stimuli were presented at a mean of 85–90 dB pressure level using the MRI-compatible electrodynamic earphone (MR Confon HP SI01, Magdeburg, Germany; impedance = 25 Ω; sensitivity: 1 mV=158 V_{RMS} = 107 dB). The excellent electrodynamic quality of the headphones as well as the intensity at which the stimuli were presented ensured the accurate perception of the sounds by the participants in the scanner. Scanning was performed using a 1.5-T scanner (Siemens Vision imager; Siemens, Munich, Germany) at the Unité de Neuroimagerie of the Notre-Dame hospital (Montreal, Canada).

The experimental design comprised two 10-min functional blocks followed by an anatomical scan. During the first functional block (localizer), natural sounds (vocal and nonvocal) were presented with a 10% occurrence of null events serving as a baseline. The purpose of this localizer was to identify the VSA for each participant. During the second functional block, the sounds from the 'voiceness' continua were presented in a pseudo-random order (10% occurrence of null events). The scan parameters for the two functional imaging sessions were: TR=ISI=2.6 s, IOI=4.1 s; TE=44 ms, flip angle=90°, voxel size=$3.59 \times 3.59\,mm^3$;

matrix size=64 × 64, slice thickness=5 mm and number of slices=28. The use of an event-related protocol in which the auditory stimuli were presented unevenly to the image acquisition justifies the absence of jittering. At the end of the functional session, high-resolution, T1-weighted three-dimensional images were acquired. The scan parameters were the following: matrix size=256 × 256, voxel size=$1 \times 1 \times 1\,\mathrm{mm}^3$, slice thickness=1 mm, number of slices=160.

Functional magnetic resonance imaging data analysis

Preprocessing

Image processing and statistical analysis were performed using SPM99 according to the standard procedure (Wellcome Department of Cognitive Neurology; [11,12]). After localizing the anterior commissure for the first image of the first functional and anatomical volume in the scanning session, corrections for differences in slice acquisition were performed using sinc interpolation. Each functional time series was realigned to the first image of their corresponding first volume to correct for interscan movement. Images were then segmented and the grey–white matter information was used to coregister the anatomical and functional images. Images were then individually normalized in stereotaxic space [13] to the Montreal Neurological Institute T1 template [14] for group analysis. Finally, all the images were spatially smoothed using a 6-mm isotropic Gaussian kernel to compensate for residual interparticipant variability and to allow for the application of Gaussian random field theory in the statistical analysis [11].

Statistical analyses

Statistical analyses were performed on the realigned, time-corrected, normalized and smoothed data. A two-level random-effect analysis (RFX) was used to compare activation maps corresponding to auditory stimulation with the vocal and nonvocal condition of the localizer, in order to locate the group-average VSA. These areas permitted the generation of regions of interest (ROIs) in the left and right hemispheres using MARSBAR (http://www.marsbar.sourceforge.net/). The ROI defined were then applied to the second functional run ('voiceness') in order to measure the blood oxygen level-dependent (BOLD) response to the sounds of the 'voiceness' continua in each individual. An ANOVA was then performed to test for the effects of degree of morphing on activation of the VSA in each hemisphere. Post-hoc analyses (Tukey test) were performed when the ANOVA was significant. In order to see whether certain superior temporal sulcus VSA were linearly related to stimulus 'voiceness' or naturalness, we also performed voxel-wise linear regressions between the parameter estimates for each level of morphing and the averaged behavioural ratings of naturalness and 'voiceness'.

Results

'Voiceness' and naturalness

'Voiceness' ratings of the morphed stimuli are shown in Fig. 1. A significant main effect of morphing level on 'voiceness' ratings was observed [F(4,36)=124.442; $P<0.001$]. Tukey post-hoc analysis showed that the estimations made for all the degrees of morphing were significantly different from each other except degree 1 vs. 3 (Fig. 1).

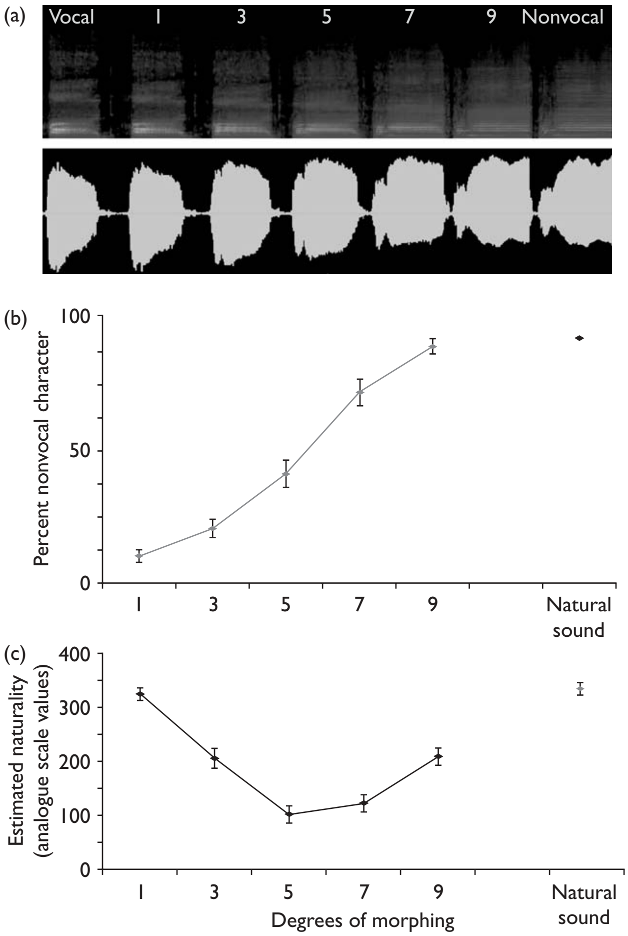

Fig. 1 'Voiceness' morphed continuum. (a) An example of spectrogram (upper panel) and waveform (lower panel) analysis of a 'voiceness' continuum resulting from the morphing of a masculine vocal ([hæd]) and a musical stimulus (bass guitar; 19.5 Hz). The natural stimuli are at the two extremes. (b) 'Voiceness' ratings: average proportion of nonvocal judgments for the five degrees of morphing used in the functional magnetic resonance imaging study. (c) Naturalness ratings for the same five morphed stimuli, and for the natural stimuli.

The perceived naturalness of the sounds was also modulated by the morphing procedure (Fig. 1). An ANOVA showed a significant main effect of morphing level on naturalness ratings [F(5,45)=59.753; $P<0.001$). Tukey post-hoc analysis showed that the estimated naturalness of morphs 3 (70% vocal), 5 (50% vocal), 7 (30% vocal) and 9 (10% vocal) were significantly different from the estimated naturalness of the natural sounds (vocal and nonvocal combined).

Functional magnetic resonance imaging

VSAs were identified through a random-effects group analysis on the basis of the contrast of vocal vs. nonvocal from the localizer scans. The significant ($P<0.005$ uncorrected) cortical areas were localized bilaterally along the superior temporal sulcus as in previous studies [5,6,14,15] (Talairach coordinates of centre of mass: right ROI: 50, −38, 1; volume=$864\,\mathrm{mm}^3$; left ROI: −56, −22, −4; volume=$1647\,\mathrm{mm}^3$) (Fig. 2).

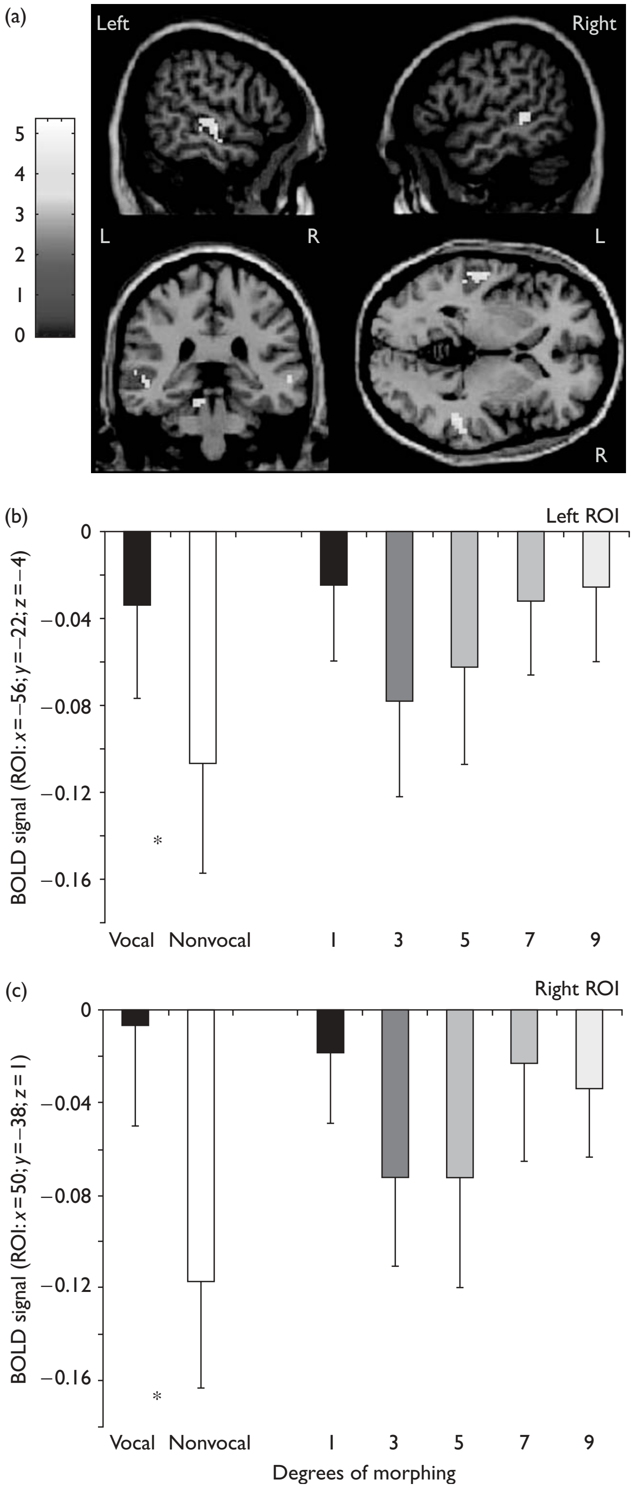

Fig. 2 Response of the voice-selective areas (VSAs) to the 'voiceness' continua. (a) VSAs resulting from the contrast vocal vs. nonvocal ($P < 0.005$ uncorrected) in the localizer scan. (b) Effect size (stimuli vs. baseline) for the VSAs of the left hemisphere. Left bars: localizer scan (vocal and nonvocal stimuli); right bars: voiceness scan, morphed stimuli. (c) Same as (b) for the VSAs of the right hemisphere. ROI, region of interest; BOLD, blood oxygen level-dependent. *$P < 0.01$.

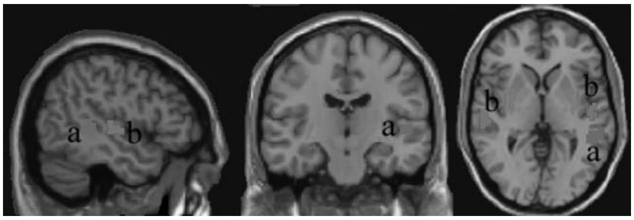

Fig. 3 Region of significant ($P < 0.05$ uncorrected) correlation between effect size and naturalness ratings (a) and superior temporal sulcus voice-selective areas (b).

The effect of 'voiceness' on the left and right voice-selective ROIs was tested using a repeated-measure ANOVA with degree of morphing as a factor. No main effect of 'voiceness' was observed in the left or right hemisphere [F(4,56)=1.85 and 2.86; P=0.132 and 0.062, respectively). For the right hemisphere only, a statistical tendency was, however, observed.

Correlation with naturalness

Voxel-wise linear regressions between individual effect sizes in the sounds vs. baseline comparison and average 'voiceness' and naturalness ratings were performed for the entire brain to investigate possible relationships between cerebral activation and subjective ratings. No significant correlation was found with the 'voiceness' ratings. One region of the right temporal lobe, posterior to the voice-selective ROI, however, was significantly correlated to the naturalness ratings of the stimuli (P=0.049, uncorrected; Talairach coordinates of centre of mass: 48, −16, 1; volume=459 mm^3) (Fig. 3).

Discussion

The purpose of this study was to investigate the response of the superior temporal sulcus VSA to 'voiceness', that is, the degree to which auditory stimuli resemble human voice. We measured the cortical responses of those areas to sounds drawn from 'voiceness' continua generated via auditory morphing between human voice and musical notes. We expected the left and right voice-selective superior temporal sulcus regions to preferentially respond to the vocal end of the continuum. Contrary to our expectations, no superior temporal sulcus regions were found to be modulated by 'voiceness'. A region of the right superior temporal sulcus close to the VSA, however, showed an activation that was correlated to sound naturalness, suggesting that sound naturalness is a better predictor of auditory cortical activity than sound 'voiceness'.

'Voiceness' and superior temporal sulcus voice-selective areas

In this study, an event-related protocol was used to localize the VSA. In agreement with previous studies that used a block design protocol [5], we found that the regions showing an enhanced response to sounds of voice compared with nonvocal sounds were localized bilaterally along the superior temporal sulcus, with maximum voice-selective response located close to the maxima observed in previous studies.

We had hypothesized a linear correlation between the behavioural ratings of 'voiceness' and cortical response to human voice as measured by the effect sizes in the sounds vs. baseline comparison. Yet, our results show a reversed bell-curve association between the parameters. As shown in Fig. 2, the smallest deactivations were found for the first (90% vocal), the seventh (30% vocal) and the ninth (10% vocal) degree of the continuum. This is clearly not what we expected: as our own results and past ones distinctly indicated that vocal sounds lead to greater activation than nonvocal sounds in the VSA, it was thus natural to assume that sounds of intermediate 'voiceness' would lead to intermediate values of activation. Instead of this pattern, we observe that intermediate sounds induce the largest deactivations.

Response to sound naturalness

A plausible explanation for this pattern of results is that another property of the intermediate 'voiceness' stimuli, sound naturalness, has influenced the BOLD signal. Whereas sounds drawn from both ends of the 'voiceness' continua (vocal and instrumental) did sound relatively natural, the stimuli in its middle tend to sound artificial because they did not correspond to existing auditory sources. Indeed, the naturalness ratings were lowest for these intermediate sounds (Fig. 1). Our fMRI results would thus indicate that the naturalness parameter of the stimuli explain more of the variance in the data set compared with 'voiceness'.

While evaluating the neural correlates underlying the processing of vocal information, Lattner *et al.* [16] observed results similar to ours. One of their experimental conditions consisted of assessing the role of the vocal 'prototypicality' by comparing the brain activations resulting from the presentation of natural voices and odd voices, that is, voices from which the pitch has been either artificially increased or decreased. The authors identified, only in the right hemisphere, a part of the anterior superior temporal gyrus that was significantly more activated by the processing of the odd voices. The cortical localizations of the area they identified and the one we did are very similar (our coordinates: x=48, y=−16, z=1; Lattner *et al.*'s [16] coordinates (2004): x=55, y=−1, z=0: 48; −16;1). Both Lattner *et al.*'s [16] study and ours show that among the several acoustic parameters that could be considered in the processing of human voice, naturalness seems to play an important role.

Naturalness and top-down cognitive processes

It has been suggested that the areas surrounding the superior temporal sulcus could also be involved in high-level cognitive operations like complex auditory mental imagery [17] as well as auditory vocal and musical hallucinations [18,19]. These experiments also suggest that our finding of superior temporal sulcus response to naturalness may reflect top-down processes such as sound identification.

Conclusion

In conclusion, we confirmed, using an event-related design, the localization of the VSA along the left and right superior temporal sulcus. Contrary to our initial hypothesis, our results also indicate that the naturalness of auditory stimuli seems to explain a greater part of the variance observed in the effects size of the BOLD signal of the superior temporal sulcus VSA than 'voiceness'.

References

1. Belin P, Fecteau S, Bedard C. Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci* 2004; **8**:129–135.
2. Bentin S, Allison T, Puce A, Perez E, McCarthy G. Electrophysiological studies of face perception in humans. *J Cogn Neurosci* 1996; **8**:551–565.
3. Allison T, Puce A, Spencer DD, McCarthy G. Electrophysiological studies of human face perception I: potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb Cortex* 1999; **9**:415–430.
4. Belin P, Zatorre RJ. 'What', 'where' and 'how' in auditory cortex. *Nat Neurosci* 2000; **3**:965–966.
5. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature* 2000; **403**:309–312.
6. Belin P, Zatorre RJ, Ahad P. Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 2000; **13**:17–26.
7. Kawahara H, Estill J, Fujimura O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. 2nd MAVEBA; 13–15 September 2001, Firenze, Italy.
8. Hillenbrand J, Getty LA, Clark MJ, Wheeler K. Acoustic characteristics of American English vowels. *J Acoustic Soc* 1995; **97**:3099–3111.
9. Kawahara H, Matsui H. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. Proceedings of the ICASSP2003; 6–10 April 2003, Hong Kong.
10. Fant G. *Acoustic theory of speech production*. The Hague: Mouton; 1960.
11. Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited – again. *Neuroimage* 1995; **2**:45–53.
12. Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SC, Frackowiak RS, Turner R. Analysis of fMRI time-series revisited. *Neuroimage* 1995; **2**: 173–181.
13. Talairach J, Tournoux P. *Co-planar stereotaxic atlas of the human brain*. New York: Thieme; 1988.
14. Evans AC, Kamber M, Collins DL, Macdonald D. *An MRIbased probabilistic atlas of neuroanatomy*. In: Shorvon S, Fish D, Andermann F, Bydder GM, Stefan H, editors. *Magnetic resonance scanning and epilepsy*. New York: Plenum; 1994. pp. 263–274.
15. Fecteau S, Armony JL, Joanette Y, Belin P. Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* 2000; **23**:840–848.
16. Lattner S, Meyer ME, Friederici AD. Voice perception: sex, pitch, and the right hemisphere. *Hum Brain Mapp* 2005; **24**:11–20.
17. Bunzeck N, Wuestenberg T, Lutz K, Heinze HJ, Jancke L. Scanning silence: mental imagery of complex sounds. *Neuroimage* 2005; **26**: 1119–1127.
18. Hunter MD, Griffiths TD, Farrow TF, Zheng Y, Wilkinson ID, Hegde N, *et al.* A neural basis for the perception of voices in external auditory space. *Brain* 2003; **126**:161–169.
19. Griffiths TD. Musical hallucinosis in acquired deafness. Phenomenology and brain substrate. *Brain* 2000; **123**:2065–2076.