ELSEVIER

Cognitive Brain Research 13 (2002) 17–26

COGNITIVE BRAIN RESEARCH

www.elsevier.com/locate/bres

Research report

Human temporal-lobe response to vocal sounds

Pascal Belin*, Robert J. Zatorre, Pierre Ahad

Neuropsychology/Cognitive Neuroscience Unit, Montreal Neurological Institute, McGill University, 3801 University street, Montreal, Québec H3A 2B4, Canada

Accepted 1 August 2001

Abstract

Voice is not only the vehicle of speech, it is also an 'auditory face' that conveys a wealth of information on a person's identity and affective state. In contrast to speech perception, little is known about the neural bases of our ability to perceive these various types of paralinguistic vocal information. Using functional magnetic resonance imaging (fMRI), we identified regions along the superior temporal sulcus (STS) that were not only sensitive, but also highly selective to vocal sounds. In the present study, we asked how neural activity in the voice areas was influenced by (i) the presence or not of linguistic information in the vocal input (speech vs. nonspeech) and (ii) frequency scrambling. Speech sounds were found to elicit greater responses than nonspeech vocalizations in most parts of auditory cortex, including primary auditory cortex (A1), on both sides of the brain. In contrast, response attenuation due to frequency scrambling was much more pronounced in anterior STS areas than at the level of A1. Importantly, only right anterior STS regions responded more strongly to nonspeech vocal sounds than to their scrambled version, suggesting that these regions could be specifically involved in paralinguistic aspects of voice perception. © 2002 Elsevier Science B.V. All rights reserved.

Theme: Sensory systems

Topic: Auditory systems: central physiology

Keywords: Auditory cortex; Speech; Voice; Cerebral lateralization; Superior temporal sulcus; Functional neuroimaging

1. Introduction

The human voice is a very common and ecologically important sound structure of our auditory environment. People probably spend more time listening to voices than to any other type of sound, first because it is the vehicle of speech and its virtually unlimited number of possible meanings. But the human voice carries much more than linguistic information. Voices can also be thought of as 'auditory faces': like a face, each voice contains in its physical structure a wealth of information on the speaker's identity and affective state, that we are all able to perceive with often good accuracy [22,32,38,41]. Besides the analyses specifically associated with speech perception, our auditory system thus performs a number of complex computations on vocal sounds that endow us with abilities ranging from merely detecting a human voice in a noisy background, to extracting information on unfamiliar speakers such as their gender, approximate age, emotional and motivational states and to recognizing familiar speakers. These different perceptual and cognitive abilities are hereafter grouped under the general term of 'voice perception'.

The different voice perception abilities pose different computational problems for the brain. Speech perception requires recognition of categories of phonemic gestures despite large inter-individual variability in production. In contrast, a voice perception ability such as speaker recognition emphasizes recognition of the invariants in the acoustic structure of a vocal production, characteristic of each speaker's unique vocal tract anatomy. These different processes are likely to be subserved by at least partially distinct neurophysiological substrates, as is most clearly illustrated by studies of 'phonagnosic' patients. Such patients show a severe deficit in speaker recognition after a cerebral lesion involving the temporal or parietal lobe, typically in the right cerebral hemisphere [2,3,39,40].

*Corresponding author. Tel.: +1-514-398-8519; fax: +1-514-398-1338.

E-mail addresses: pascal@bic.mni.mcgill.ca (P. Belin), http://www.zlab.mcgill.ca/.

PII: S0926-6410(01)00084-2

Interestingly, in most cases phonagnosia coexists with essentially spared speech perception; conversely, although voice perception abilities other than speech perception are typically not evaluated in aphasic patients, there exist reports of patients with receptive aphasia after left hemispheric lesion but essentially intact speaker recognition [2]. Thus clinical data suggest a functional dissociation between speech perception and speaker recognition. Yet, the anatomical information provided in these early studies was insufficient to fully characterize the probably complex and distributed neural network involved in the different voice perception abilities.

Functional neuroimaging techniques have allowed significant progress in our understanding of the neuronal bases of speech perception [7,12,42,46]. An increasing number of studies now use these techniques to investigate the structures involved in paralinguistic aspects of voice perception. Imaimuzi et al. [18], using positron emission tomography (PET), found that different brain areas were involved in the execution of speaker identification and of vocal emotion identification tasks, and that left and right temporal poles were more active during the speaker identification task. Morris et al. [25] used PET during a gender decision task on neutral or emotionally loaded nonspeech vocal stimuli, and showed that several brain areas, including the amygdala and ventral prefrontal cortex, responded differently to the neutral and emotional stimuli. Using a functional magnetic resonance imaging (fMRI) paradigm adapted to studies of the auditory function [4,16] we initiated a series of experiments, inspired from electrophysiological as well as neuroimaging studies of face perception [17,20,24,27], aimed at characterizing the functional architecture of the neural substrate involved in voice perception abilities. A first set of experiments identified several regions of secondary auditory cortex, mostly located along the superior temporal sulcus (STS), in which neuronal activity appeared not only sensitive but also highly selective to sounds of human voice [5]. We present here additional results from this series of experiments, focusing on (1) inter-individual variability in functional localization; (2) the effect of linguistic information on the voice-sensitive response; and (3) the effect of frequency-scrambling, a process that disrupts the typical spectral shape of voice while relatively preserving temporal information (Fig. 1).

2. Materials and methods

2.1. Subjects

Eight healthy adult subjects (age 22–47, four males and four females), all right-handed and with normal audition, gave written informed consent to participate in this study. The principle of these studies was approved by the

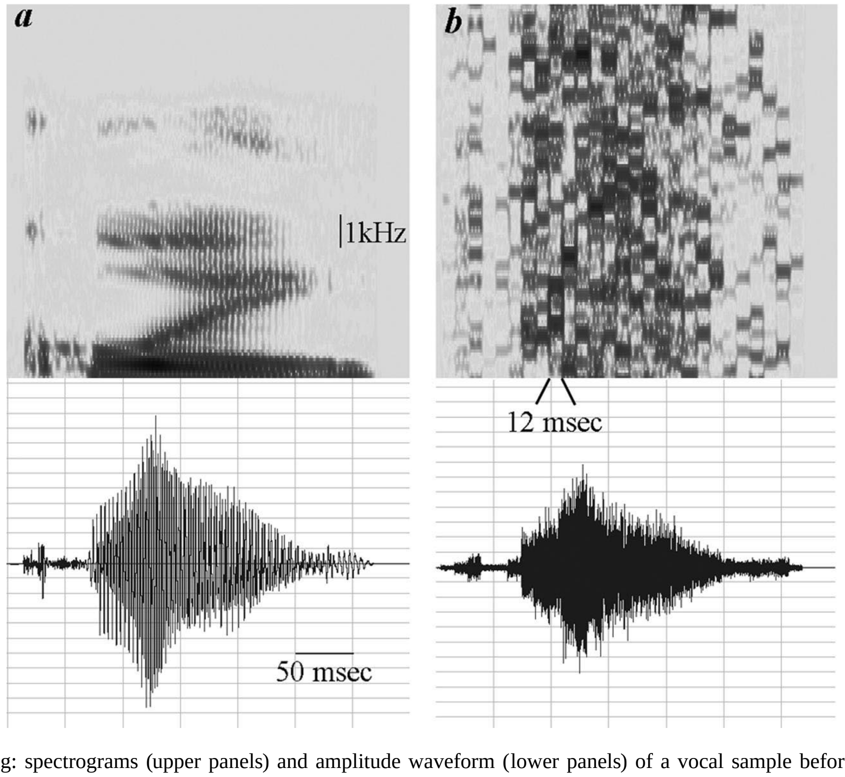

Fig. 1. Frequency scrambling: spectrograms (upper panels) and amplitude waveform (lower panels) of a vocal sample before (a) and after (b) frequency scrambling. The oblique lines in (b) indicate the size of the window used in the Fourier transformation (512 points).

Research Ethics Committee of the Montreal Neurological Institute.

2.2. Task and stimuli

In both Experiments 1 and 2, subjects were instructed to simply close their eyes and listen to the sounds that would be presented. Auditory stimuli were delivered binaurally at a mean 88–90 dB sound pressure level (SPL A), using foam insert earplugs (Etymotic Research, IL, USA) and an MR-compatible pneumatic sound transmission. Auditory stimuli consisted of digitized sound samples obtained from commercial and academic sources, or were recorded for the purpose of the study.They were arranged in 20-s blocks of similar number of sources and overall energy (RMS) using MITSYN (WLH, MA, USA) and COOLEDIT PRO (Syntrillium Software, AZ, USA). Vocal sounds were obtained from a large number of speakers ($n > 45$) of both sexes and of all ages (1 month to 80 years old). They consisted of speech sounds, such as isolated words, nonwords, connected speech in several languages (French, English, Finnish, Arabic, Chinese), and of nonspeech vocalizations such as laughs, cries, moans, sighs, etc. Sounds that did not involve vocal fold vibration were excluded (e.g. whistling, whispered voices). Non-vocal sounds consisted of a wide variety of environmental and musical sounds, including sounds from the elements such as wind, rain, streams; animal sounds such as vocalizations and footsteps; sounds of mechanical/electronic origin such as cars, telephones, planes; as well as musical sounds from a variety of instruments.

In Experiment 1 termed a 'localizer' scan a large number of sound samples was divided into two categories only, according to the vocal/non-vocal distinction, and randomly arranged in 21 blocks of each category. This small number of categories allowed us to maximize localization power by attributing a large number of images (42) to each sound category. Subcategories of vocal sounds (e.g. speech vs. nonspeech sounds) as well as non-vocal sounds (e.g. industrial vs. musical sounds) were mixed within each block in order to minimize effects specifically induced by a subcategory.

Experiment 2, conversely, used a greater number of sound categories at the expense of localization power, four of which are discussed here. There were four 20-s blocks for each category (eight brain volumes), each containing ten different sounds. Two categories consisted of vocal sounds, themselves separated in two categories along the speech/nonspeech distinction: one category consisted of only speech sounds, while the other category consisted of only nonspeech vocalizations. The other categories consisted of the scrambled versions of the two first categories: these sounds were obtained by performing short-term Fourier transforms on each vocal block (windows of 512 points, 11.6 ms at the sampling rate of 44.1 kHz), scrambling the values of the amplitude components (only with other amplitude components) and of the phase components (only with other phase components) in each window, and performing the inverse DFT on that signal (Fig. 1). Thus the amplitude values in the scrambled voices were the exact same set as those in the original signal (hence preserving the overall energy) but at different frequencies, which made these sounds totally unrecognizable as voices.

Stimuli are available for download at http://www.zlab.mcgill.ca.

2.3. MRI acquisition

Scanning was performed on a 1.5 T Siemens Vision imager. High-resolution T1 images were first acquired for anatomical localization and co-registration with functional series. One series of 128 functional images was then acquired for each experiment (gradient-echo, TE = 50 ms, TR = 10 s, head coil, matrix size: 64×64, voxel size: $4 \times 4 \times 5$ mm^3, ten slices parallel to the Sylvian fissure) for a scanning time of 21 min 40 s each. The long inter-acquisition interval (TR) ensures low signal contamination by image acquisition noise artefacts: the hemodynamic response induced by acquisition noise has come back to near-baseline level at the time of the next acquisition [4,16]. In both experiments, the 20-s auditory blocks (two brain volumes acquired for each auditory block) were presented in a randomized order with a 10-s silence inter-block interval (one brain volume), using Media Control Function (MCF, Digivox, Montreal, Canada); the beginning of each block was synchronized with acquisition of the first slice of each brain volume.

2.4. Data analysis

BOLD signal images were spatially smoothed (6-mm Gaussian kernel), corrected for motion artefacts and linearly transformed into standard stereotaxic space [35] using in-house software [9]. For Exp. 1, individual statistical maps were obtained by computing for each voxel the t value of the Spearman correlation of the voxel's value time-series with an ideal curve representing the desired contrast [36]. Group-average statistical images were obtained by computing an omnibus test on individual t maps using a pooled estimate of standard deviation and thresholded at $t = 5.7$ ($P < 0.001$), based upon the number of resolution elements in the acquisition volume (2880 resels) [43]. For Exp. 2, BOLD signal was sampled at the voxels that corresponded to local maxima of t value for the contrast of vocal vs. non-vocal sounds in Exp. 1 (thus selected independently). BOLD signal values for each condition were then converted to percent signal increases by reference to the mean value for the silence conditions, and subjected to analyses of variance with two factors: speech/nonspeech and scrambled/original.

3. Results

3.1. Experiment 1: sensitivity to sounds of human voice

In Experiment 1 we asked if regions of the human brain would show sensitivity to voice, i.e. greater activity during passive stimulation with vocal sounds relative to non-vocal sounds. During over 20 min, subjects passively heard a wide variety of sound samples from the environment arranged in 20-s blocks, simply separated into two categories along the vocal/non-vocal distinction. Each vocal block was composed of vocalizations from twelve different voices, consisting not only of speech samples but also of nonspeech vocalizations such as laughs, cries or throat clearings. These different types of vocalizations were all mixed together within the vocal blocks in order to avoid effects specifically related to a subcategory of vocal sounds (e.g. speech sounds). Similarly, the non-vocal blocks sounds were composed of sounds from a large variety of sources, including musical instruments, sounds from the natural environment, or mechanical/industrial sounds (see Methods).

In each subject, the contrast of images of blood oxygenation level dependent (BOLD) signal acquired during stimulation with the vocal sounds and with the non-vocal sounds revealed several regions of significantly ($t>5.7$, $P<0.001$) greater neuronal activation to vocal as compared to non-vocal sounds. In all subjects but one (no. 2), hemispheric maxima of vocal vs. non-vocal response were located in the upper bank of the STS, on both the left and right sides (Fig. 2). In subject no. 2, this maximum was located in planum temporale on both sides. Fig. 2 shows that location of hemispheric maxima varied widely across subjects in the y (antero–posterior) dimension: from $y=-6$ to $y=-41$ (in millimeters in standard stereotaxic space [35]) in the right hemisphere, and from $y=-14$ to $y=-46$ in the left hemisphere. The strongest difference between vocal and non-vocal sounds was observed in the left hemisphere for four subjects (nos. 1, 2, 5 and 7) and in the right for the four others (nos. 3, 4, 6 and 8). Interestingly, negative activation values, i.e. the extent to which some regions responded more to non-vocal than to vocal sounds, were always of much lesser significance than positive activations, except perhaps for subject no. 5 (Fig. 2).

The group-average map, obtained by pooling individual maps after transformation into stereotaxic space, yielded several regions of significantly greater ($P<0.001$) increase for the vocal sounds, but no regions of greater BOLD signal increase for the non-vocal sounds [5]. The voice-sensitive regions included several parts of the superior temporal gyrus, with several local maxima located along the superior temporal sulcus (STS), both anterior and posterior to Heschl's gyrus (Fig. 3). The maximum of voice sensitivity in the group was attained in both hemispheres in the anterior part of the STS (Talairach coordinates, left: $x=-62$; $y=-14$; $z=0$; right: $x=63$; $y=-13$, $z=-1$), with the strongest difference on the right side (Fig. 3).

An alternative way of combining information across subjects consists of sampling activity not from identical stereotaxic coordinates, but at each individual's maximum of voice-sensitivity, thus from different anatomical locations in each subject. This functional way of pooling data yielded BOLD signal percent changes relative to adjacent silence blocks of (group mean±S.E.): 3.0±0.8% for the vocal sounds and 1.6±0.5% for the non-vocal sounds in the right hemisphere; and 1.8±0.4% for vocal sounds and 0.6±0.2% for non-vocal sounds in the left hemisphere. Importantly, in all eight subjects the response to the vocal sounds was stronger at the right- than at the left-hemispheric maximum ($P<0.005$, Wilcoxon's signed-rank test).

3.2. Experiment 2: speech and nonspeech vocal sounds vs. scrambled voices

Experiment 2 was meant to evaluate the response selectivity of only a small number of cortical sites, independently identified by Exp. 1 (shown in colorscale in Fig. 3): we explored the locations that yielded a t value for the vocal vs. non-vocal contrast above an arbitrary value of $t=9$ (five regions in the right hemisphere, two in the left hemisphere). Two locations corresponding to primary auditory cortex (A1) were also sampled, as defined by the location of highest signal change within Heschl's gyrus [26] in the sound vs. silence contrast in Exp. 1 (left: $x=-38$; $y=-34$; $z=14$; right: $x=48$; $y=-22$; $z=6$). Several categories of control sounds were used, four of which are discussed here: speech sounds, nonspeech vocal sounds, and the frequency-scrambled versions of these two categories (see Methods).

Fig. 3 shows the group-average percent signal increases relative to the silence condition for the four categories of sounds, at each of nine locations identified in Exp. 1. To assess the effects of spectral and linguistic information on the voice-sensitive response, two-way repeated measures ANOVAs were conducted at each of the nine selected locations with speech/nonspeech and normal/scrambled as the two factors. Speech sounds were found to consistently induce greater responses than nonspeech vocal sounds at almost all voice-sensitive areas analyzed. This difference was strongest at the group-average maxima of voice-sensitivity yielded by Exp. 1 in both hemispheres (left: $F(1,7)=13.66$, $P<0.01$; right: $F(1,7)=34.72$, $P<0.001$). Importantly, the greater response to speech than nonspeech vocal sounds was already present at the level of A1, as shown in the two lower panels of Fig. 3, and indicated by the significant interaction between the speech/nonspeech and original/scrambled factors for both left and right A1

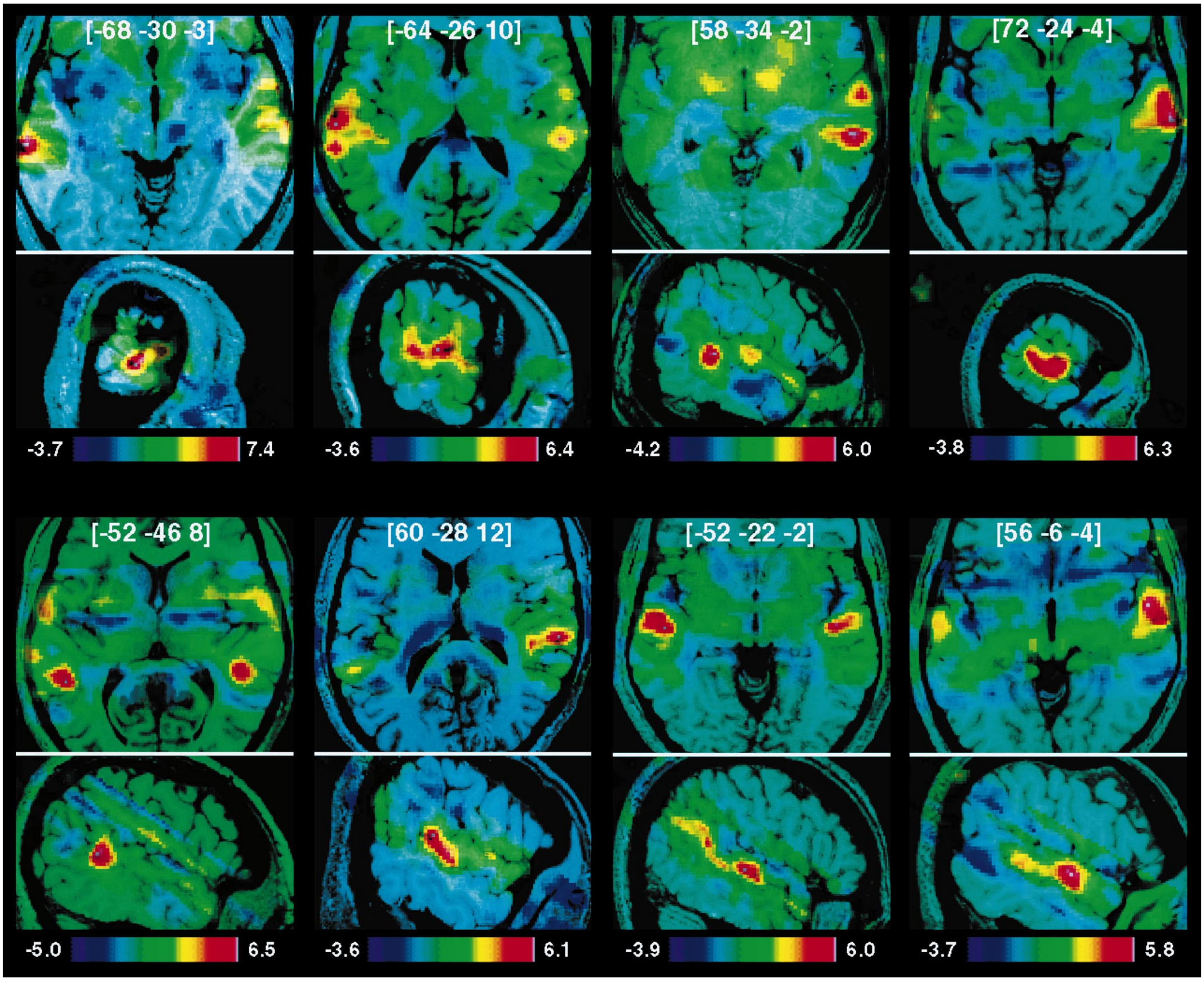

Fig. 2. Variability in individual voice-sensitivity maps. Unthresholded individual t statistics maps of the contrast between vocal and non-vocal listening conditions in Exp. 1 are shown for all eight subjects, overlaid in spectral colorscale on each individual's anatomical MR image transformed in standard stereotaxic space [35], in the axial (upper panels) and sagittal (lower panels) views passing through the subject's maximum of voice-sensitivity. Numbers beside colorscale indicate full range of t values. Numbers in brackets indicate coordinates of subject's voice-sensitivity maximum in stereotaxic space.

(right: $F(1,7)=7.77$, $P<0.05$; left: $F(1,7)=11.52$, $P<0.01$).

Frequency-scrambling was found to induce a significant decrease of activation relative to the original vocal sounds at most voice-sensitive locations analyzed. The three upper panels of Fig. 3 show that BOLD signal difference between the vocal sounds and their scrambled versions was greatest in the anterior STS regions that also yielded the highest difference between vocal and non-vocal sounds in Exp. 1 (right: $F(1,7)=17.48$, $P<0.005$; left: $F(1,7)=20.87$, $P<0.005$). Conversely, no significant effect of scrambling could be observed in either right or left primary auditory cortex when both speech and nonspeech sounds were grouped (right: $F(1,7)=4.73$, $P>0.05$; left: $F(1,7)=4.29$, $P>0.05$). Yet, the significant interaction of the two factors (see above) suggests that this was less true for the speech sounds, as evident in Fig. 3. Further insight was gained by comparing the effect of scrambling on the speech and nonspeech vocal sounds separately. Planned comparisons (paired t-tests) showed that speech sounds elicited greater response than their scrambled version in nearly all parts of auditory cortex including primary auditory cortex (right A1: $t=4.0$, right anterior STS: $t=8.47$; left A1: $t=2.41$, left anterior STS: $t=6.2$) as shown in Fig. 3. In contrast, nonspeech vocal sounds elicited greater response than their scrambled version only in the middle and anterior portions of the right STS ($t>2.7$).

4. Discussion

These experiments investigated the response of auditory

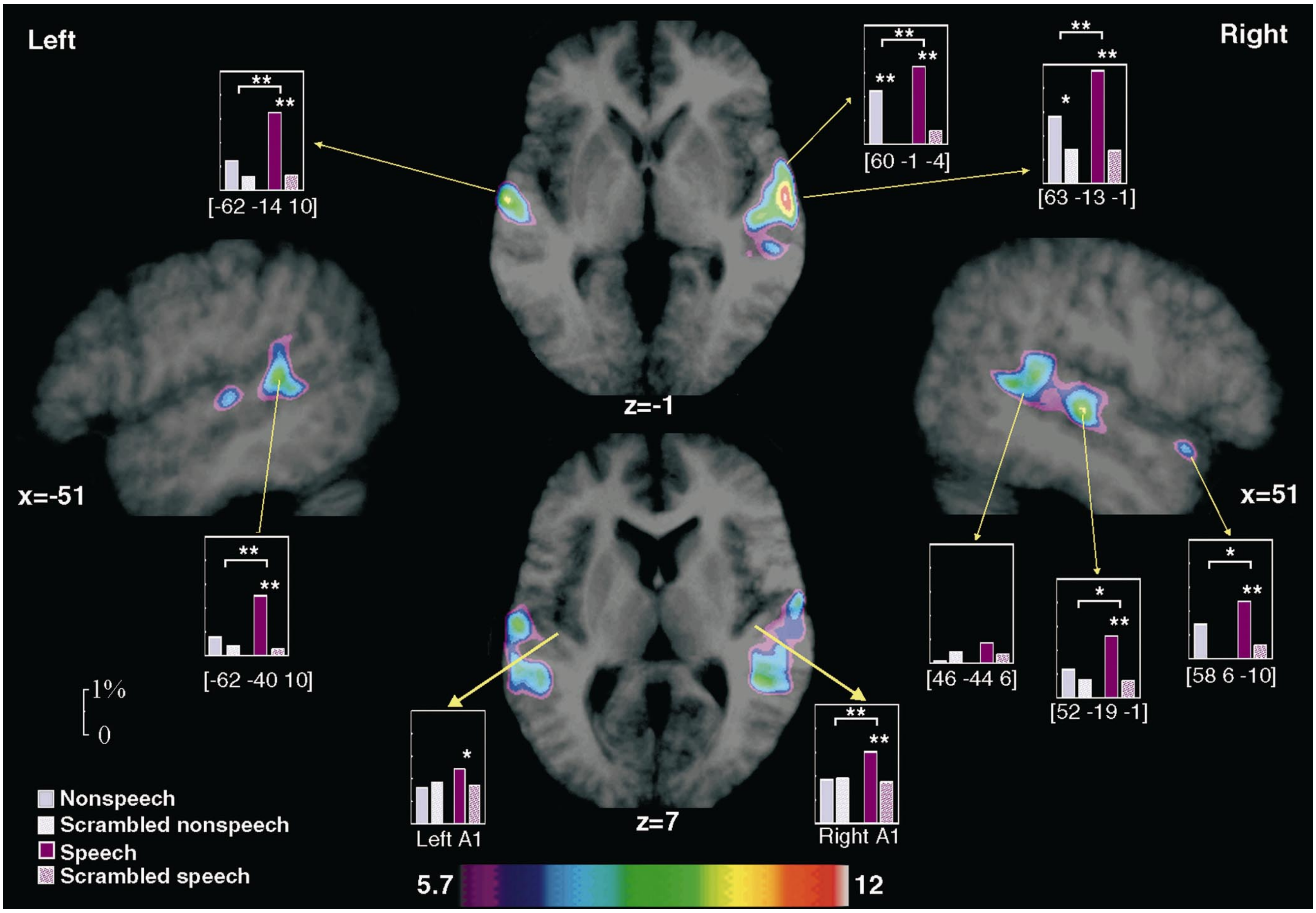

Fig. 3. Group-average response to original and scrambled vocal sounds. Regions of significantly greater ($P<0.001$) signal change to vocal than to non-vocal sounds in Experiment 1 are shown in colorscale (*t* statistics) overlaid on the subjects' average anatomical MR image in standard stereotaxic space, along slices in the axial (middle panels) and sagittal (lateral panels) orientations. Bar diagrams indicate mean signal change from silence for the four sound categories of Experiment 2: speech sounds, nonspeech vocal sounds, and their scrambled versions. *x*, *z* and number in brackets: coordinates in stereotaxic space. Stars above bracket: speech vs. nonspeech. Stars above bars: effect of scrambling, planned comparisons (paired *t*-tests). *, $P<0.05$; **, $P<0.01$.

cortex to sounds of human voices. They were characterized by two important methodological differences with conventional neuroimaging studies of speech perception. First, the vocal stimuli used in the stimulation blocks were not all speech sounds: they consisted of a large proportion (67% in Exp. 1, 50% in Exp. 2) of nonspeech vocalizations such as laughs, cries, coughs, etc. Second, instead of using vocal productions from a single person as is typically the case, we used a far greater range of speakers: vocal samples were collected from a large number of persons from both sexes and with a wide age range. These experiments were inspired by neurophysiological studies of object and face perception in the visual domain, where a large number of different exemplars are usually presented within each category [17,20,24,27]. In particular we used a two-step approach often used in those studies, where a first 'localizer' scan is used to highlight regions of interest that are then tested independently in subsequent scans in the same subject [20]. We first localized parts of auditory cortex that would show sensitivity to vocal sounds i.e. a stronger response to vocal than to non-vocal blocks. Both the large number of stimuli presented in each of these two categories (>500) and their diversity were meant to ensure that the vocal/non-vocal distinction was the most salient difference between these two sound categories, and thus would be reflected in the differential activation pattern. Then a second scan allowed to measure the effect of several parameters on the neuronal response at these selected voice-sensitive locations.

Three important results emerged. First, each subject showed a voice-sensitive response in regions of the belt/parabelt zone of auditory cortex, located in variable locations near the STS, with a trend to greater responses in the right hemisphere. Second, speech vocal sounds induced greater response than nonspeech vocalizations at most voice-sensitive location — as well as in primary auditory cortex. Finally, only anterior STS regions of the right hemisphere responded more strongly to nonspeech vocalizations than to their scrambled version.

4.1. Anatomical distribution of voice-sensitive regions

Experiment 1 was meant to reproduce the variety of sounds of our natural environment, simply divided into two categories according to whether sounds had been produced by a human vocal tract or not. In each subject, Exp. 1 revealed circumscribed regions of enhanced response to vocal sounds, as compared to non-vocal sounds matched in number of sources and overall energy. Interestingly, localization of these voice-sensitive regions was quite variable across subjects, as can be seen in Fig. 2: depending on the subject, hemispheric maxima of voice-sensitivity could be found anterior or posterior to the lateral extent of Heschl's gyrus, varying by as much as 35 mm in the antero–posterior position. These varying anatomical locations were at the same time highly consistent, since in most subjects (7/8) they consisted of gray matter of the upper bank of the STS in a restricted part of the anterior branch of this long and complex sulcal structure (also composed of two posterior branches extending into the parietal and occipital lobes [13]). The group-average data confirmed the above pattern, since the most significant peaks in the vocal vs. non-vocal statistical map were also located in the anterior STS region (Fig. 3). The cortical area of enhanced response to vocal sounds in the group also reached the upper surface of the superior temporal gyrus, especially in the left hemisphere where it extended posteriorly to planum temporale (Fig. 3).

The distribution of the voice-sensitive response observed in Exp. 1 was clearly biased towards the right hemisphere. At the individual level, the site of greatest activity difference between vocal and non-vocal sounds was equally distributed between the two hemispheres across subjects; yet when both left and right individual maxima were compared, responses to vocal sounds relative to silence were always greater on the right side. The group data also indicated a greater sensitivity in the right hemisphere: the hemispheric sites of greatest group-average difference between vocal and non-vocal sounds reached higher t values in the right hemispheres, as seen in colorscale at the level of anterior STS areas in Fig. 3. At these stereotaxic locations, selectivity to vocal sounds was also greater in the right hemisphere, as indicated by a more significant response to vocal sounds than to categories of control sounds such as bell sounds, human non-vocal sounds, amplitude-modulated noise and scrambled voices [5]. Thus the anatomical distribution of the voice-sensitive regions appears to be approximately symmetrical, but the magnitude of neuronal response to vocal sounds is higher in the right hemisphere.

Greater sensitivity to vocal sounds in the right hemisphere is not in conflict with the well-known superiority of the left hemisphere for processing speech sounds: a number of neuroimaging studies have shown quasisymmetrical activation of the temporal lobes when passive stimulation with speech sounds is compared to the silent baseline [7,8,11,23,31]; only comparison with tasks involving explicit linguistic tasks [12,45] or more specific control sounds [33,42] has been able to uncover the lateralized processing of speech sounds. The fact that speech sounds were mixed with nonspeech vocalizations in Exp. 1 probably contributed to further mask activations specifically related to speech processing. In addition, the vocal stimuli consisted of a large variety of vocalizations of different types and from different speakers; such variability in vocal patterns is likely to have maximized neuronal activity related to analysis of the speaker-related features of voice, which clinical studies suggest to be preferentially localized to the right hemisphere [2,3,39,40]. Thus the right-biased lateralization in the distribution of cortical areas sensitive to vocal sounds could be explained by the fact that both hemispheres process different parts of the wealth of information contained in the vocal signal.

One should note that the group-average image was not really representative of each individual's pattern of activation, as seen by a comparison of Figs. 2 and 3. The possibility afforded by fMRI to obtain individual functional images with a good signal-to-noise ratio reveals here an important inter-individual variability in functional anatomy, which can be related to the well-known structural variability across individuals after linear transformations in standard stereotaxic space [29]. Yet, anatomical variability can only account for a small part of the variability observed here, since the average location of the STS is clearly visible on the group-average anatomical image (Fig. 3). Such variability could reflect different cognitive states of the subjects during the passive listening tasks; the different subjects could possibly have attended preferentially to different acoustic features in the stimuli during the passive listening task, thus recruiting different cortical areas. An alternative explanation could be that at these higher levels of the cortical architecture, the exact localization of functionally defined neuronal populations (here, neurons 'tuned' to the acoustic structure of vocal sounds) is subject to larger variation than at more primary levels. It will be important to address these questions in future work, and in particular to assess intra-subject variability in the voice-sensitive response.

4.2. Effect of frequency-scrambling

Experiment 2 used 'scrambled voices' as control sounds. These sounds were obtained by performing on the signal a manipulation very similar to that performed on visual stimuli (e.g. 'scrambled faces') [1,15,21,28]. Scrambled sounds constitute good acoustic controls since their low-level acoustic structure is close to that of the original sounds (same amplitude and phase components in Fourier space, see Methods). Since phase and amplitude components were randomized separately, the scrambled sounds also have the same overall energy as the original sounds. Moreover, the 12-ms time-window used for the Fourier transformation ensured that the large-scale temporal structure of the signal was relatively preserved. Conversely, the spectral profile of the vocal sounds was much altered by randomization of phase and frequency components across the large number of frequency channels used. In particular, scrambled sounds had a much larger content of high frequencies than did the vocal sounds (Fig. 1). This manipulation had dramatic consequences on a listener's ability to recognize these sounds: they sounded like flowing water or modulated noise, but nothing like voices (see Methods).

As shown in Fig. 3, responses of primary auditory cortex and those of the voice-sensitive regions identified by Exp. 1 were markedly different. For both left and right A1, scrambled voices were found to induce signal changes comparable to those elicited by the original sounds, at least for the nonspeech sounds. In contrast, scrambling strongly attenuated neuronal response both for speech and nonspeech vocal sounds at more distant locations. This result is consistent with the notion that A1 is involved in extracting low-level acoustic components from the signal [10], some of which are similar in scrambled and original sounds. At higher stages of the auditory cortical architecture, however, where features extracted in A1 are probably combined into more complex representations [19,30], frequency scrambling was found to have a more severe effect on cortical response. These results are analogous to findings in the visual modality, of increasing disruption of the cortical response induced by frequency scrambling as one moves away from V1 in the functional architecture underlying visual object recognition [15]. They support the idea that auditory and visual cortices may be organized following similar principles of organization.

4.3. Functional role of anterior STS areas

Activations in the STS are often reported in studies of speech perception [7,33,42,45], but their functional significance remains unclear [6]. Binder et al., for example, found using fMRI that the STS regions consistently yielded higher signal changes in response to speech sounds than to frequency-modulated tones [7]. The interpretation of these activations is made complex by the fact that speech stimuli are also necessarily vocal sounds. Thus, neuronal activity in the STS regions could be driven by the phonemic–linguistic content of the speech stimuli, but also more simply by the acoustic structure common to all vocal sounds, speech or not. Recent results by Scott et al. are particularly relevant. These authors showed that left, but not right, anterior STS regions responded to filtered speech, transformed in a way that preserved temporal information and intelligibility [34], as much as to natural speech [33]. This result is in good line with results from Exp. 2 where response difference between speech and nonspeech vocalizations was greater in the left hemisphere, in particular in the anterior STS region (Fig. 3). Thus, linguistic information seems to be more important than vocal structure to yield activity in left anterior STS.

Conversely, right STS regions do not seem to require linguistic information to be responsive to vocal sounds. One important result from Exp. 2 is that whereas speech sounds induced a greater response than to their scrambled version at nearly all parts of auditory cortex, including A1, this was true for nonspeech vocalizations only in right anterior STS. Speech vocalizations were found to drive auditory cortex to higher levels of activity than nonspeech sounds, even at the level of A1. Such greater response was probably partly induced by the presence of linguistic information, but probably also partly by the fact that speech sounds generally contain a larger number of relevant acoustic features than other sounds, due to their high complexity and information transfer rate. The fact that the greater response to speech sounds was already observed at the level of A1 suggests that this second explanation might not be entirely false. Yet speech sounds do not explain all of the voice-sensitive response observed: in right anterior STS, even vocalizations devoid of linguistic content induced greater activation than to the scrambled controls. Thus, the voice-sensitive areas located in right anterior STS are probably not exclusively involved in linguistic analysis of the vocal signal. They constitute good candidates for cortical regions involved in other, paralinguistic aspects of voice perception.

As a whole, the above results suggest an interesting pattern of lateralization in the functional response to vocal sounds. Both anterior STS regions appear to be sensitive to vocal sounds, as suggested by Exp. 1, but left STS regions seem to show a preference for linguistic information, whereas right STS regions seem to show a preference for other information in the vocal structure, found in speech sounds as well as in nonspeech vocalizations. Such asymmetry has already been suggested by psychological and clinical studies: affective and identity information in faces and voices would be processed in the right hemisphere for both the visual and auditory modalities, while linguistic information would be processed in the left hemisphere, be it speech or lip motion during speaking [14]. These differences might emerge at high levels of auditory processing as a consequence of a lateralization of auditory processing at earlier cortical levels, perhaps reflecting the differential processing of spectral and temporal information [44]. Additional research is needed to further establish the validity of this model; in particular, it will be important to assess the exact function of right anterior STS regions in the analysis of vocal sounds.

5. Conclusions

One important implication of the present results is the parallel they draw with the face-sensitive areas of the visual system. Faces and voices are very similar in that both are characterized by a constrained physical structure, around which minute inter- and intra-individual variations convey rich information on the person's identity and affective state, as well as linguistic information. The present findings of voice-selective areas along the STS thus suggest that the similarities between face and voice perception might extend to their underlying neuronal architecture. Along with electrophysiological recordings in the primate [19,30], they support a model of human cortical architecture in which auditory object features, such as a speaker's vocal tract characteristics, would be processed in a dedicated ventral cortical pathway similar to the one existing in visual cortex [37]. They also suggest that such a system presents important lateral differences in the human brain.

Acknowledgements

We thank Bruce Pike and Alan Evans for access to neuroimaging facilities at the Brain Imaging Center, Bruce Rosen for useful discussion, and Marc Bouffard for his technical assistance. Supported by the Canadian Institute for Health Research, National Sciences and Engineering Research Council of Canada, McDonnell-Pew Cognitive Neuroscience Program, INSERM-FRSQ and Foundation France-Télécom.

References

[1] T. Allison, A. Puce, D.D. Spencer, G. McCarthy, Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli, Cereb. Cortex 9 (1999) 415–430.

[2] G. Assal, C. Aubert, J. Buttet, Asymétrie cérébrale et reconnaissance de la voix, Rev. Neurol. 137 (1981) 255–268.

[3] G. Assal, E. Zander, H. Kremin, J. Buttet, Discrimination des voix lors des lesions du cortex cerebral, Arch. Suisses Neurol. Neurochir. Psychiatr. 119 (1976) 307–315.

[4] P. Belin, R.J. Zatorre, R. Hoge, B. Pike, A.C. Evans, Event-related fMRI of the auditory cortex, Neuroimage 10 (1999) 417–429.

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

[6] J.R. Binder, J.A. Frost, P.S.F. Bellgowan, Superior temporal sulcus (STS) responses to speech and nonspeech auditory stimuli, J. Cog. Neurosci. 11 (1999) 99.

[7] J.R. Binder, J.A. Frost, T.A. Hammeke, P.S. Bellgowan, J.A. Springer, J.N. Kaufman, E.T. Possing, Human temporal lobe activation by speech and nonspeech sounds, Cereb. Cortex 10 (2000) 512–528.

[8] J.R. Binder, S.M. Rao, T.A. Hammeke, F.Z. Yetkin, A. Jesmanowicz, P.A. Bandettini, E.C. Wong, L.D. Estkowski, M.D. Goldstein, V.M. Haughton, J.S. Hyde, Functional magnetic resonance imaging of human auditory cortex, Ann. Neurol. 35 (1994) 662–672.

[9] D.L. Collins, P. Neelin, T.M. Peters, A.C. Evans, Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space, J. Comput. Assist. Tomogr. 18 (1994) 192–205.

[10] R.C. deCharms, D.T. Blake, M.M. Merzenich, Optimizing sound features for cortical neurons, Science 280 (1998) 1439–1443.

[11] S. Dehaene, E. Dupoux, J. Mehler, L. Cohen, E. Paulesu, D. Perani, P.-F. van de Moortele, S. Lehéricy, D. Le Bihan, Anatomical variability in the cortical representation of first and second language, NeuroReport 8 (1997) 3809–3815.

[12] J.F. Démonet, R. Chollet, S. Ramsay, D. Cardebat, J. Nespoulos, R. Wise, A. Rascol, R.S.J. Frackowiak, The anatomy of phonological and semantic processing in normal subjects, Brain 115 (1992) 1753–1768.

[13] H.M. Duvernoy, E.A. Cabanis, M.T. Iba-Zizen, J. Tamraz, J. Guyot, Le Cerveau Humain: Surface, Coupes Sériées Tridimensionnelles et Irm, Springer-Verlag, 1992.

[14] A.W. Ellis, Neuro-cognitive processing of faces and voices, in: A.W. Young, H.D. Ellis (Eds.), Handbook of Research on Face Processing, Elsevier, 1989, pp. 207–215.

[15] K. Grill-Spector, T. Kushnir, T. Hendler, S. Edelman, Y. Itzchak, R. Malach, A sequence of object-processing stages revealed by fMRI in the human occipital lobe, Hum. Brain Mapp. 6 (1998) 316–328.

[16] D. Hall, M.P. Haggard, M.A. Akeroyd, A.R. Palmer, A. Quentin Summerfield, M.R. Elliott, E.M. Gurney, R.W. Bowtell, 'Sparse' temporal sampling in auditory fMRI, Hum. Brain Mapp. 7 (1999) 213–223.

[17] J.V. Haxby, E.A. Hoffman, M. Ida Gobbini, The distributed human neural system for face perception, Trends Cog. Sci. 4 (2000) 223–233.

[18] S. Imaizumi, K. Mori, S. Kiritani, R. Kawashima, M. Sugiura, H. Fukuda, K. Itoh, T. Kato, A. Nakamura, K. Hatano, S. Kojima, K. Nakamura, Vocal identification of speaker and emotion activates different brain regions, NeuroReport 8 (1997) 2809–2812.

[19] J.H. Kaas, T.A. Hackett, M.J. Tramo, Auditory processing in primate cerebral cortex, Curr. Opin. Neurobiol. 9 (1999) 154–170.

[20] N. Kanwisher, J. McDermott, M.M. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception, J. Neurosci. 17 (11) (1997) 4302–4311.

[21] Z. Kourtzi, N. Kanwisher, Cortical regions involved in perceiving object shape, J. Neurosci. 20 (2000) 3310–3318.

[22] J. Kreiman, Listening to voices: theory and practice in voice perception research, in: K. Johnson, J. Mullenix (Eds.), Talker Variability in Speech Research, Academic Press, New York, 1997, pp. 85–108.

[23] B.M. Mazoyer, N. Tzourio, A. Syrota, M. Murayama, O. Levrier, G. Salamon, S. Dehaene, L. Cohen, J. Mehler, The cortical representation of speech, J. Cog. Neurosci. 5 (4) (1993) 467–479.

[24] G. McCarthy, A. Puce, J.C. Gore, T. Allison, Face-specific processing in the human fusiform gyrus, J. Cog. Neurosci. 9 (1997) 605–610.

[25] J.S. Morris, S.K. Scott, R.J. Dolan, Saying it with feeling: neural responses to emotional vocalizations, Neuropsychologia 37 (1999) 1155–1163.

[26] V.B. Penhune, R.J. Zatorre, J.D. MacDonald, A.C. Evans, Interhemispheric anatomical differences in human primary auditory cortex: probabilistic mapping and volume measurement from MR scans, Cereb. Cortex 6 (1996) 661–672.

[27] D.I. Perrett, P.A. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, M.A. Jeeves, Neurones responsive to faces in the temporal cortex: studies of functional organization, sensitivity to identity and relation to perception, Hum. Neurobiol. 3 (1984) 197–208.

[28] A. Puce, T. Allison, J.C. Gore, G. McCarthy, Face-sensitive regions in human extrastriate cortex studied by functional MRI, J. Neurophysiol. 74 (1995) 1192–1199.

[29] J. Rademacher, V.S. Caviness, H. Steinmetz, A.M. Galaburda, Topographical variations of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology, Cereb. Cortex 3 (1993) 313–329.

[30] J.P. Rauschecker, B. Tian, M. Hauser, Processing of complex sounds in the macaque nonprimary auditory cortex, Science 268 (1995) 111–114.

[31] Y. Samson, P. Belin, L. Thivard, N. Boddaert, S. Crozier, M. Zilbovicius, Perception auditive et langage: imagerie fonctionnelle du cortex auditif sensible au langage, Rev. Neurol. (in press).

[32] S.R. Schweinberger, A. Herholz, W. Sommer, Recognizing famous voices: influence of stimulus duration and different types of retrieval cues, J. Speech Lang. Hear. Res. 40 (1997) 453–463.

[33] S.K. Scott, C.C. Blank, S. Rosen, R.J. Wise, Identification of a pathway for intelligible speech in the left temporal lobe, Brain 123 (2000) 2400–2406.

[34] R.V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech recognition with primary temporal cues, Science 270 (1995) 303–304.

[35] J. Talairach, P. Tournoux, Co-Planar Stereotaxic Atlas of the Human Brain, Thieme, New York, 1988.

[36] R. Turner, Functional mapping of the human brain with magnetic resonance imaging, Semin. Neurosci. 7 (1995) 179–194.

[37] L.G. Ungerleider, J.V. Haxby, 'What' and 'where' in the human brain, Curr. Opin. Neurobiol. 4 (1994) 157–165.

[38] W.A. Van Dommelen, Acoustic parameters in human speaker recognition, Lang. Speech 33 (3) (1990) 259–272.

[39] D. Van Lancker, J. Kreiman, Voice discrimination and recognition are separate abilities, Neuropsychologia 25 (1987) 829–834.

[40] D.R. Van Lancker, G.J. Canter, Impairment of voice and face recognition in patients with hemispheric damage, Brain Cog. 1 (1982) 185–195.

[41] S.P. Whiteside, Identification of a speaker's sex: a study of vowels, Percept. Mot. Skills 86 (1998) 579–584.

[42] R.J. Wise, S.K. Scott, S.C. Blank, C.J. Mummery, K. Murphy, E.A. Warburton, Separate neural subsystems within 'Wernicke's area', Brain 124 (2001) 83–95.

[43] K.J. Worsley, A.C. Evans, S. Marrett, P. Neelin, A three-dimensional statistical analysis for CBF activation studies in human brain, J. Cereb. Blood Flow Metab. 12 (1992) 900–918.

[44] R.J. Zatorre, P. Belin, Spectral and temporal processing in human auditory cortex, Cereb. Cortex (in press).

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

[46] R.J. Zatorre, E. Meyer, A. Gjedde, A.C. Evans, PET studies of phonetic processing of speech: review, replication and reanalysis, Cereb. Cortex 6 (1996) 21–30.