Adaptation to speaker's voice in right anterior temporal lobe

Pascal Belin[CA] and Robert J. Zatorre[I]

Groupe de Recherche en Neuropsychologie et Cognition (GRENEC), Département de Psychologie, Université de Montréal CP 6128, Succ. Centre-Ville, Montréal (Québec), Canada H3C 3J7; [I]Neuropsychology/Cognitive Neuroscience Unit, Montreal Neurological Institute, McGill University, Montréal, Québec, Canada

[CA]Corresponding Author: pascal.belin@umontreal.ca

Little is known on how voices are represented in the brain. We used fMRI to investigate whether parts of auditory cortex would be sensitive to the repetition of a speaker's voice. Subjects were scanned while passively listening to spoken syllables, presented in blocs in which either syllable or speaker were repeated. Only one cortical region, located in the anterior part of the right superior temporal sulcus (STS), responded differently to the two conditions: activation relative to the silent baseline was significantly reduced when syllables were spoken by a single voice than when they were spoken by different voices. This result suggest that the right anterior STS plays an important role in the representation of individual voices. *NeuroReport* 14:2105–2109 © 2003 Lippincott Williams & Wilkins.

Keywords: Adaptation; Auditory cortex; fMRI; Neuroimaging; Repetition-suppression; Speaker recognition; Voice perception

INTRODUCTION

Speaker recognition is probably our most complex auditory cognitive ability, apart from speech perception. Our capacity to rapidly and effortlessly recognize an individual based on novel vocalizations suggests that our auditory system extracts the acoustic features of the vocal signal that present high inter-individual but little intra-individual variation, to combine them in long-term representations of vocal identities. However, the format and neural basis for these representations of vocal identity are still largely unknown.

Neuroimaging techniques are increasingly used to study the cerebral structures involved in para-linguisitic voice perception abilities such as speaker discrimination or perception of vocal emotion [1,2]. In a previous fMRI study, we showed the existence of voice-selective areas in the auditory cortex. Mostly located along the upper bank of the superior temporal sulcus (STS), these areas showed greater response to vocal sounds compared to non-vocal sounds. This voice-sensitive response was not observed with control sounds such as scrambled voices or frequency-filtered voices, suggesting a high degree of selectivity [3]. Right anterior STS regions, in particular, were found to respond strongly to non-speech vocalizations such as laughs and cries, suggesting a possible role of these regions in paralinguistic aspects of voice processing [4].

In this study, we used a variant of the adaptation paradigm recently introduced in fMRI studies [5], based on the property of neuronal populations to reduce their firing rate in response to repeated stimulation. This repetition-induced reduction of activity, termed adaptation or repetition-suppression in studies of visual object perception, is thought to be crucially involved in memory for visual objects [6,7], and might provide a neurophysiological basis for perceptual learning and priming [8]. Although the underlying neurophysiological mechanisms are still not entirely clear, neuronal adaptation is now increasingly used in fMRI experiments. Buckner *et al.* found that repeated visual presentation of objects induced a clear reduction of the amount of activation in areas in mid-levels of the processing hierarchy, including extrastriate cortex and inferior temporal-lobe regions, but not primary visual cortex [7]. Grill-Spector and Malach introduced a novel adaptation paradigm designed to induce selective repetition effects of different features of the visual presentation, to study level of object representation in visual cortex [5]. By selectively repeating features of the visual stimulation such as object size, or position, they showed the lateral-occipital cortex was more sensitive to changes in object illumination and viewpoint than to changes in size and position, suggesting that object representation at this level of the visual cortical architecture is already independent of size and position.

Here, two features of the vocal signal were symmetrically manipulated: linguistic content (syllable) and vocal identity (speaker's voice). Normal subjects were scanned while listening passively to 20 s auditory blocks composed of 12 vocal samples, for which one of these two features was kept

constant while the other feature was made different for each sample of the block. A sparse sampling method was used in order to reduce the scanning noise artefact [9,10]. In one condition (adapt-speaker), blocks were composed of 12 different syllables spoken by a single voice, thus maximally varying linguistic information while repeating the information related to speaker's identity. The symmetrical condition (adapt-syllable) used blocks composed of a same syllable spoken by 12 different speakers, thus minimizing variation in linguistic content while maximizing variability in speakers' vocal characteristics (Fig. 1). Overall, the exact same 144 stimuli were presented in the two conditions, simply with a different order of presentation. We predicted that although very similar, these two conditions would yield differences in brain activity related to the induced repetition effect.

MATERIALS AND METHODS

Subjects: Fourteen subjects (age 20–40 years, nine females and five males), all right-handed and with normal audition, gave written informed consent to participate in this study according to the declaration of Helsinki. The principle of these studies was approved by the Research Ethics Committee of the Montreal Neurological Institute.
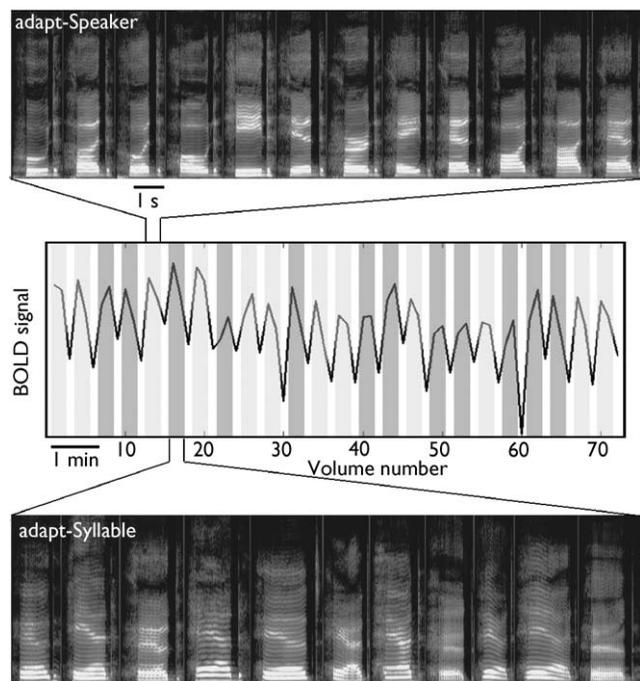
Fig. 1. Experimental design. Upper and lower panels: spectrograms (0–5 kHz, 20 s) of examples of auditory blocks for the two main adaptation conditions. Adapt-speaker: a same speaker says 12 different syllables. Adapt-syllable: the same syllable is spoken by 12 different speakers. Middle panel: BOLD signal time-course from one voxel in primary auditory cortex of a representative subject, across the whole scan duration (12 min). Dark and light gray vertical bars: adapt-speaker and adapt-syllable blocks, respectively (20 s); white bars: silence (10 s). Note features of the spectrogram that remain constant across different words for a same speaker (adapt-speaker).

Materials: Stimuli consisted of 144 spoken syllables: 12 American English vowels in /hVd/ syllables (had, hod, hawed, head, heard, haid, hid, heed, haod=/o/ as in boat, hood, hud, who'd) each spoken by 12 speakers: six adults (three men and three women) and six children (three boys and three girls). They were part of a database of American English vowels recorded in similar controlled conditions across a large number of speakers, described in detail elsewhere [11] and kindly made available to the authors. All 144 stimuli (16-bits, mono, 16 kHz sampling rate) were equated for RMS amplitude using Mitsyn (WLH, MA, USA).

Stimuli were arranged in 20 s auditory blocks of two kinds: adapt-syllable ($n = 12$) and adapt-speaker ($n = 12$), where phonological and speaker-related features were manipulated in a symmetrical way. Adapt-Syllable blocks consisted of a single one of the 12 syllables spoken by each of the 12 speakers; adapt-speaker blocks consisted of one single voice speaking all 12 words.

Scanning: 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 75 blood oxygenation-level dependent (BOLD) images was then acquired with the following characteristics: gradient-echo, TE = 50 ms, TR = 10 s, head coil, matrix size 64 × 64, voxel size 5 × 5 × 5 mm^3, 10 slices parallel to the Sylvian fissure (20–25 slices in the AC-PC orientation in three of the 14 subjects). The long, 10 s inter-acquisition interval (sparse sampling) ensures low signal contamination by image acquisition noise artifacts as the hemodynamic response induced by acquisition noise has returned to near-baseline level at the time of the next acquisition [9,10].

Subjects were instructed to keep their eyes closed and to listen to the sounds that would be presented. The 24 auditory blocks (12 adapt-syllable blocks and 12 adapt-speaker blocks) were presented at ~80 dB SPL in a pseudo-random order with a 10 s inter-block interval of silence, using Media Control Function (MCF, Digivox, Montreal). The beginning of each block was synchronized with acquisition of the first slice of every third brain volume. Thus two brain volumes were acquired for each block, one at 10 s and the other at 20 s post-onset. Each pair of brain volumes acquired during the blocks was separated by one brain volume acquired after 10 s of silence (Fig. 1).

Data analysis: BOLD signal images were spatially smoothed (6 mm Gaussian kernel), corrected for motion artefacts and linearly transformed into standard stereotaxic space [12] using in-house software [13]. Statistical analysis of the fMRI data was based on a linear model with correlated errors [14] (see URL: http://www.bic.mni. mcgill.ca/~keith/).

The *t*-statistics image comparing the two auditory stimulation conditions was thresholded using the minimum given by a Bonferroni correction and random field theory [15]. As differences between these two conditions were expected to be quite small, the search for significant differences was restricted to the cortical areas which showed

an activation value of $t > 2$ in the comparison of the pooled auditory stimulation conditions to the silent baseline. Within this search volume ($170\,cm^3$), with df = 896, any cluster of connected voxels with a volume $> 1324\,mm^3$ ($\sim$11 voxels) above a threshold of $t = 2$ was significant at $p < 0.05$.

RESULTS

Group-average comparison of images acquired during both adapt-speaker and adapt-syllable conditions to those acquired during silence yielded the typical pattern of auditory activation: strong bilateral activation of a large part of the superior temporal gyri centered around Heschl's gyrus. Primary auditory cortex (A1) was clearly identifiable in both left and right hemispheres as a marked local maximum of BOLD signal change (Talairach coordinates of peak: left −40,−32,16; right: 48,−18,10) located at the medial extremity of Heschl's gyrus [16,17].

The direct comparison of the two main adaptation conditions (adapt-speaker and adapt-syllable) yielded a significant BOLD signal difference in only one region of auditory cortex (Fig. 2). This area was located in the right

Fig. 2. Adapt-syllable *vs* adapt-speaker. The only region of significant BOLD signal difference between the two conditions is shown in colorscale (t-value) on axial (upper panel) and sagittal (lower panel) slices of the group-average anatomical MR image. x and z : coordinates in Talairach space. White arrow: Superior temporal sulcus (STS); Black arrow: Sylvian fissure. The right anterior STS region is less active when subjects are hearing one voice speaking several syllables, than several voices speaking a same syllable.

hemisphere, in the anterior part of the superior temporal gyrus near the STS (Talairach coordinates of peak: 58,2,−8). It showed significantly ($p < 0.05$) less activity during the adapt-speaker condition, when a single voice was speaking different words, than during the adapt-syllable condition, when several voices produced a same word. No brain region showed the reverse pattern.
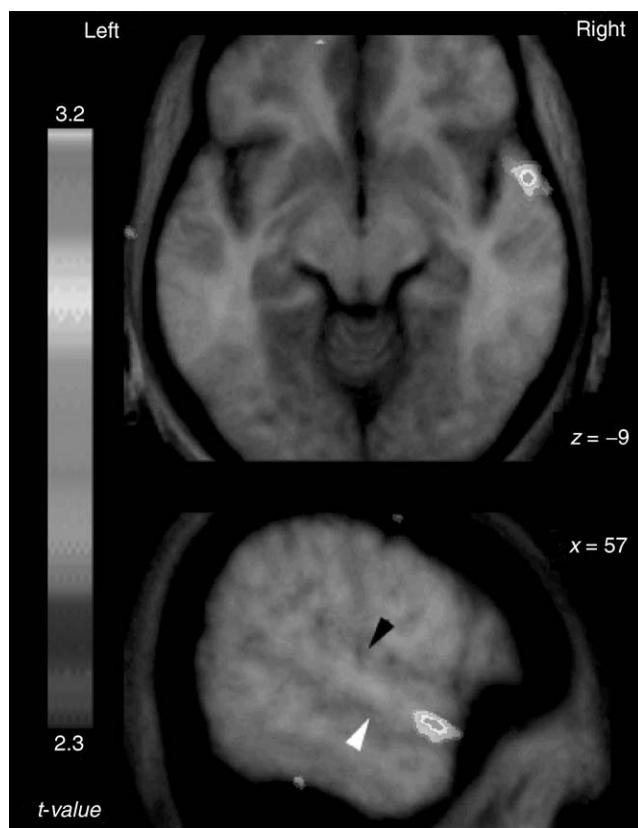
Figure 3 shows estimates of BOLD signal increase relative to silence in left and right A1 as well as in right anterior STS, split by condition (adapt-syllable *vs* adapt-speaker) and acquisition time (10 s and 20 s after block onset). A two-way ANOVA on values obtained from a 5 mm sphere centered on the STS location revealed a significant interaction between condition and time ($F_{(13,1)} = 5.54$, $p < 0.05$), reflecting a stronger difference between the two conditions at 10 s post-onset ($t = 2.15$, $p > 0.5$).
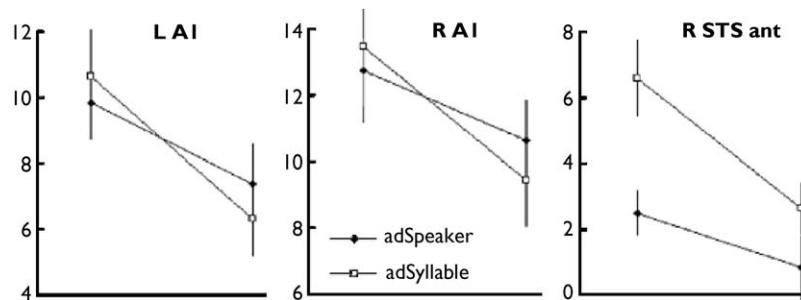
DISCUSSION

The two main auditory adaptation conditions, adapt-speaker and adapt-syllable, were almost identical to one another: subjects were in the same passive listening state, and they heard the exact same 144 stimuli (12 syllables spoken by 12 speakers), simply aranged in a different order to induce different repetition effects (Fig. 1). As expected, both conditions activated most of auditory cortex to similar extents when compared to silence. Only a single region of auditory cortex showed a significant difference in activity between the two conditions (Fig. 2). Auditory activation in this region was smaller when a single voice was heard in each block, than when several voices were heard.

The most straightforward explanation for this result is that it reflects neuronal adaptation in reaction to repetition of the speaker's voice. Some of the acoustic features in a vocal sound reflect the idiosyncratic properties of the unique vocal tract by which it is produced; these features were repeated for each syllable in an adapt-speaker block. This can be seen in the representative spectrogram shown in Fig. 1 (upper panel): although the clear bands corresponding to energy maxima (formants) have different trajectories for each different syllable, there are some other features that remain constant, such as the spacing between the horizontal striations corresponding to the harmonics of the fundamental frequency, or the dark band at mid-height corresponding to a hole at about 3 kHz in the spectral distribution of energy for this particular speaker. It is reasonable to assume that some neuronal populations in auditory cortex, at higher levels of the functional architecture, could be sensitive to the combination of acoustic features characteristic of each speaker's voice. These populations would be expected, then, to be sensitive to the repetition of these acoustic features, and to show adaptation or repetition-suppression, i.e. a reduction of their spiking rate relative to a condition where the voices would be all different. The smaller activity in right anterior STS for the adapt-speaker relative to the adapt-syllable condition probably reflects this phenomenon of adaptation to speaker's voice.

Sensitivity to speaker's voice: Neuronal populations sensitive to vocal identity in the anterior part of the right

Fig. 3. Auditory activation in AI and STS. Linear estimates (y axis, arbitrary units) of BOLD signal differences between auditory stimulation and silence are shown for the two volumes acquired in each block (x axis, 10 s and 20 s post-onset). Data averaged from 5 mm radius spheres centered around the right STS location (right panel), and left and right AI locations (left and center panels). Bars indicate sem.

superior temporal lobe would be in good line with the current understanding of the functional architecture of auditory cortex. Recent neuroimaging studies indeed suggest that anterior temporal regions might be involved in extracting paralinguistic information in vocal sounds. A PET study of emotion and speaker recognition showed that the anterior temporal lobes were more active bilaterally during speaker discrimination than during emotion discrimination [1]. In a follow-up PET study, the same group reported that the right anterior temporal pole was more active during discrimination of familiar voices than during control discriminations, and that activity in this region correlated with subject's identification performance [2]. Interestingly, the peak observed in the present study falls only 2 mm away from one of the voice-selective locations observed in earlier studies [3,4].

Neuronal representation of vocal identity: The notion of neuronal populations sensitive to vocal identity, although experimentally supported here for the first time, is quite plausible. Several authors have already suggested the existence of a representation of vocal identity in the cerebral cortex, mostly based on the analogy between face perception and voice perception [18,19]. A largely accepted model of face perception proposes the existence in visual cortex of face recognition units, which contain stored stuctural codes describing one of the faces known to the person [20]. Evidence for a neurophysiological counterpart of this representation has been obtained both by single-cell recording studies in the macaque brain [21] and by recent neuroimaging studies in the human brain [22–24]. In the light of the similarities between faces and voices, it is tempting to suggest that voices, as 'auditory faces,' could be analysed following similar schemes. Mesulam has thus suggested the existence of an area specialized for identifying individual voice patterns in a recent model of cerebral organization, by analogy with the area specialized for face encoding of visual cortex [18]. The present results suggest that such an area might be located in right anterior STS.

Although anterior STS were tested in the right as well as in the left hemispheres, no difference was found between the two listening conditions on the left side. This functional asymmetry is consistent with results suggesting that the right and left anterior STS regions could be sensitive to different types of vocal information: left STS regions seem to be more active by linguisitic information, and can even be activated by noise-vocoded speech, i.e. stimuli that keep some intelligibility but that have lost their vocal structure [25]. Conversely, we observed that right STS regions are the only parts of auditory cortex to show this voice-sensitive response even for non-linguisitic vocalizations such as laughs, cries or throat-clearings [4]. Thus, current evidence strongly suggests that anterior temporal-lobe regions, particularly in the right hemisphere, are involved in the extraction of paralinguistic information in the vocal signal, and in particular, vocal identity.

REFERENCES

1. Imaizumi S, Mori K, Kiritani S *et al. Neuroreport* **8**, 2809–2812 (1997).
2. Nakamura K, Kawashima R, Sugiura M *et al. Neuropsychologia* **39**, 1047–1054 (2001).
3. Belin P, Zatorre RJ, Lafaille P *et al. Nature* **403**, 309–312 (2000).
4. Belin P, Zatorre RJ and Ahad P. *Cogn Brain Res* **13**, 17–26 (2002).
5. Grill-Spector K and Malach R. *Acta Psychol (Amst)* **107**, 293–321 (2001).
6. Baylis GC and Rolls ET. *Exp Brain Res* **65**, 614–622 (1987).
7. Buckner RL, Goodman J, Burock M *et al. Neuron* **20**, 285–296 (1998).
8. Wiggs CL and Martin A. *Curr Opin Neurobiol* **8**, 227–233 (1998).
9. Belin P, Zatorre RJ, Hoge R *et al. Neuroimage* **10**, 417–429 (1999).
10. Hall D, Haggard MP, Akeroyd MA *et al. Hum Brain Mapp* **7**, 213–223 (1999).
11. Hillenbrand JM, Getty LA, Clark MJ and Wheeler K. *J Acoust Soc Am* **97**, 1300–1313 (1995).
12. Talairach J and Tournoux P. *Co-Planar Stereotaxic Atlas of the Human Brain.* New York: Thieme; 1988.
13. Collins DL, Neelin P, Peters TM and Evans AC. *J Comput Assist Tomogr* **18**, 192–205 (1994).
14. Worsley KJ, Liao C, Grabove M *et al.* A general statistical analysis for fMRI data. *Human Brain Mapping* 2000.
15. Worsley KJ, Marrett S, Neelin P *et al. Hum Brain Mapp* **4**, 58–73 (1996).
16. Penhune VB, Zatorre RJ, MacDonald JD and Evans AC. *Cerebr Cortex* **6**, 661–672 (1996).
17. Rademacher J, Morosan P, Schormann T *et al. Neuroimage* **13**, 669–683 (2001).
18. Mesulam MM. *Brain* **121**, 1013–1052 (1998).
19. Ellis AW. Neuro-cognitive processing of faces and voices. In: Young AW and Ellis HD (eds). *Handbook of Research on Face Processing.* Amsterdam: Elsevier; 1989, pp. 207–215.
20. Bruce V and Young A. *Br J Psychol* **77**, 305–327 (1986).
21. Perrett DI, Hietanen JK, Oram MW and Benson PJ. *Phil Trans R Soc Lond B Biol Sci* **335**, 23–30 (1992).

22. Puce A, Allison T, Gore JC and McCarthy G. *J Neurophysiol* **74**, 1192–1199 (1995).

23. Kanwisher N, McDermott J and Chun MM. *J Neurosci* **17**, 4302–4311 (1997).

24. Haxby JV, Hoffman EA and Ida Gobbini M. *Trends Cogn Sci* **4**, 223–233 (2000).

25. Scott SK, Blank CC, Rosen S and Wise RJ. *Brain* **123**, 2400–2406 (2000).