Brief Communications

Expertise with Artificial Nonspeech Sounds Recruits Speech-Sensitive Cortical Regions

Robert Leech,[1] Lori L. Holt,[2,3] Joseph T. Devlin,[4] and Frederic Dick[5,6]

[1]Division of Neuroscience and Mental Health, Imperial College London, Hammersmith Hospital, London W12 0NN, United Kingdom, [2]Department of Psychology, Carnegie Mellon University, and [3]Center for the Neural Basis of Cognition, 115 Mellon Institute, Pittsburgh, Pennsylvania 15213, [4]Cognitive, Perceptual and Brain Sciences, UCL Institute of Cognitive Neuroscience, UCL, London WC1H 0AP, United Kingdom, [5]School of Psychology, Birkbeck, London WC1E 7HX, United Kingdom, and [6]Center for Research in Language, University of California, San Diego, La Jolla, California 92093-0526

Regions of the human temporal lobe show greater activation for speech than for other sounds. These differences may reflect intrinsically specialized domain-specific adaptations for processing speech, or they may be driven by the significant expertise we have in listening to the speech signal. To test the expertise hypothesis, we used a video-game-based paradigm that tacitly trained listeners to categorize acoustically complex, artificial nonlinguistic sounds. Before and after training, we used functional MRI to measure how expertise with these sounds modulated temporal lobe activation. Participants' ability to explicitly categorize the nonspeech sounds predicted the change in pretraining to posttraining activation in speech-sensitive regions of the left posterior superior temporal sulcus, suggesting that emergent auditory expertise may help drive this functional regionalization. Thus, seemingly domain-specific patterns of neural activation in higher cortical regions may be driven in part by experience-based restructuring of high-dimensional perceptual space.

Introduction

Several brain regions are preferentially activated by specific categories of stimuli such as faces (Kanwisher et al., 1997; Tsao et al., 2003) or conspecific vocalizations (Petkov et al., 2008). For instance, in humans, parts of the left superior temporal sulcus (STS) show greater activation for speech than a range of other sounds (Belin et al., 2000; Binder et al., 2000; Scott et al., 2000) including spectrotemporally complex, meaningful nonlinguistic sounds like thunderclaps and dog barks (Dick et al., 2007). One possibility is that this preferential activation for speech reflects left STS specialization for specific acoustical and informational properties of speech (for review and discussion, see Price et al., 2005). An alternative hypothesis is that speech-sensitive activation in left STS reflects life-long expertise with decomposing, categorizing, and producing complex auditory stimuli (Diehl et al., 2004; Kuhl, 2004). If the latter hypothesis is true, then increased activation in left STS should not be specific to speech but should also emerge as a result of expertise with other complex auditory stimuli that the listener has learned to categorize.

Previous studies comparing musicians and nonmusicians have suggested greater left superior temporal activation for nonspeech stimuli (Ohnishi et al., 2001), implicating some role for auditory expertise in left STS. However, such retrospective studies are limited in addressing the causal role of experience, particularly as intrinsic differences between expert musician and control groups could drive patterns of results. Only a prospective study (e.g., a training study directly manipulating subjects' experience with a set of stimuli) can unambiguously establish that expertise can drive functional cortical reorganization in putatively speech-sensitive areas.

To test this expertise hypothesis, we investigated whether learning artificial, complex nonspeech auditory categories leads to more speech-like patterns of neural activation. The paradigm was intended to mimic the learning and development of phonetic categories, an ability associated with increases in activation in the left STS (Scott et al., 2000). Participants played a space-invaders-style video game involving visually presented aliens, each associated with a category of sounds. The auditory stimuli were designed to model some of the complexity of speech categories without sounding like human speech. To succeed in the game, participants had to learn the relationship between each alien and the accompanying category of sounds, although this was never made explicit to participants. Similar to the process of learning to treat acoustically distinct speech signals as members of the same phonetic category (Kuhl, 2004), listeners gradually learn that perceptually discriminable "alien" sounds are functionally equivalent in the game (Wade and Holt, 2005).

To investigate whether expertise with these artificial nonspeech sounds leads to a more speech-like neural signature, participants were scanned before and after five or more hours of

Received Dec. 3, 2008; revised Feb. 27, 2009; accepted March 21, 2009.

R.L. was supported by a Research Council UK academic fellowship; L.L.H. was supported by grants from the National Institutes of Health (R01DC004674), the National Science Foundation (BCS0746067), and the Bank of Sweden Tercentenary Foundation; J.D. was supported by The Wellcome Trust; and F.D was supported by the Medical Research Council (Grant G0400341). We thank Zarinah Agnew, Narly Golestani, and Marty Sereno for helpful comments.

Correspondence should be addressed to Dr. Robert Leech, Division of Neuroscience and Mental Health and Medical Research Council Clinical Sciences Centre, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK. E-mail: r.leech@imperial.ac.uk.

DOI:10.1523/JNEUROSCI.5758-08.2009

 0270-6474/09/295234-06$15.00/0

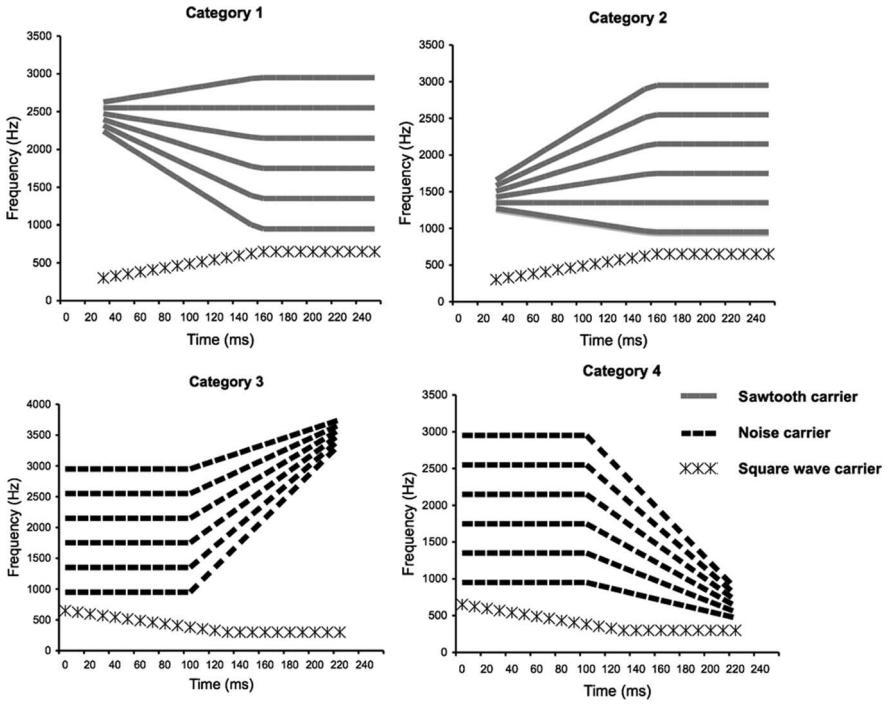

Figure 1. Schematic representations of the different sounds from the four auditory categories heard during the computer game training (taken from Wade and Holt, 2005). Each exemplar is comprised of the invariant lower spectral peak (a square wave) and one of six possible higher spectral peaks (either sawtooth or bandpass noise). Each of the four auditory categories was paired with a space-invader alien picture.

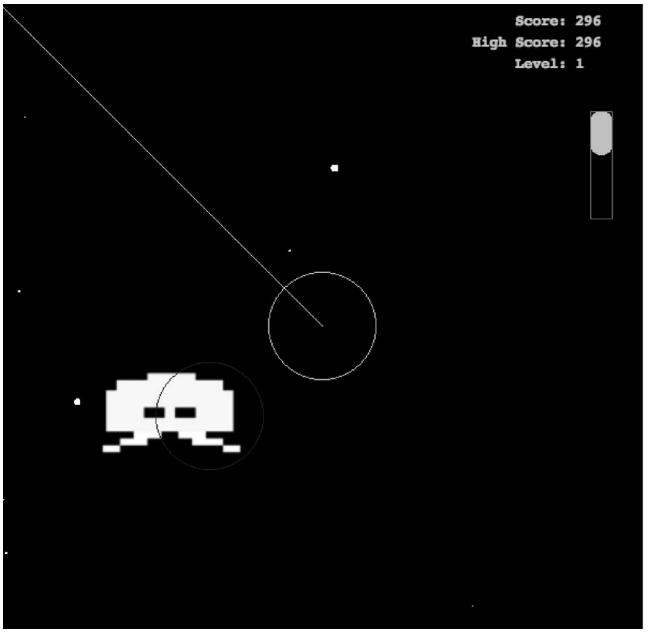

Figure 2. A typical screen shot (rendered in grayscale) of the space-invaders game; for a detailed description, see the study by Wade and Holt (2005). Subjects had to move their viewfinder toward the looming alien (here in white) and position it in the center of the screen so as to kill or capture it.

video-game-based training with the novel artificial, complex nonlinguistic sounds. We predicted that categorization expertise with these stimuli would lead to an increase in activation in left STS regions and that the degree of activation changes in this region would be modulated by how well participants had learned the novel auditory categories.

Materials and Methods

Subjects

Eight female and nine male participants (age range 19–36 years) took part in the study. Participants reported normal hearing and no history of neurological or psychiatric disease. All gave written informed consent and were paid for their participation. The study was approved by the UCL Research Ethics Committee.

Stimuli

The artificial complex sounds used during both scanning and training were grouped into four auditory categories. Each sound category was comprised of sounds with two spectral peaks with rapid onset or offset frequency transitions combined together additively (see Fig. 1). Across all four categories the acoustic source of the lower frequency spectral peak was a 143 Hz square-wave. For two of the categories the source of the higher spectral peak was a 150 Hz sawtooth wave whereas for the other two categories it was derived from uniform random (white) noise. These source signals were filtered to create spectral peaks approximately paralleling formant resonances of the human vocal tract in their change in frequency across time. Despite this similarity to speech, these sounds possessed a complex fine temporal structure completely unlike that of speech and participants do not report these sounds as being speech-like (Wade and Holt, 2005). Category exemplars varied along several spectrotemporal dimensions with some overlap across categories, as is observed in human speech categories (Kuhl, 2004). To successfully recognize a sound as an exemplar of a specific category, participants needed to simultaneously make use of multiple cues, i.e., to solve a nonlinear mapping involving both second spectral peak onset frequency and second spectral peak steady-state frequency (see Wade and Holt, 2005 for full details). Each category contained 11 sounds (see supplemental materials, available at www.jneurosci.org; http://www.psy.cmu.edu/~lholt/WadeHolt2005/gallery_irfbats.php). Six of the sounds (those in Fig. 1) were used both in training and scanning and an additional five novel sounds (spanning approximately the same spectral range) were heard only in scanning and during a behavioral posttest auditory categorization task.

Procedure

Training paradigm. Auditory training consisted of participants playing five or more hours of a space-invaders-type computer game (Wade and Holt, 2005) (see Fig. 2). Participants played the game unsupervised at home with records of both game performance and time playing logged. (For four participants, this information was lost because of data transfer errors). In the game there were four different visually presented aliens that participants had to either capture or shoot. Each alien was associated with a different auditory category, an exemplar of which was played repeatedly while the alien was on the screen. As the game progressed, the aliens moved faster and originated further from the center of the screen. Each alien consistently came from a particular direction, e.g., blue aliens always came from the top of the screen, with their exact position jittered within a quadrant of the screen. As the game became harder, aliens originated further from the center of the screen such that participants could hear the alien's characteristic sounds before seeing the alien. Thus, participants could use auditory category information to predict which screen quadrant the alien would come from, thereby improving overall game performance. Note that the video game was predominantly visual, and participants were given no instructions or hints to use or attend to auditory information (Wade and Holt, 2005). Success in learning the auditory categories was evaluated using two measures: (1) an indirect measure, which was the highest game level attained over the course of training, and (2) a more direct measure, which was a four alternative-

forced choice categorization task administered after the second scanning session, in which participants were continuously shown a "line-up" of the four alien characters and were played each auditory exemplar four times over the course of 176 randomly ordered trials. Participants were asked on each trial to identify which alien character corresponded to the sound they had just heard (for additional details, see Wade and Holt, 2005).

Imaging procedure. All volunteers participated in two scanning sessions, one before they had any experience with the novel, artificial sounds, and one after five or more hours of video-game based training. In both the pretraining and posttraining scanning sessions, participants underwent two 10 min runs in which they saw pictures of the four aliens and passively heard sounds from each auditory category. In half of the presentations, there was a mismatch between the picture of the alien and category of the accompanying sound. The order of the two runs was fully counterbalanced over participants. In the scanner, participants performed a visual oddball detection task, pressing a button when they saw an upside-down alien. This task focused attention on the visual alien figures to avoid explicit categorization of the sounds while helping to control and monitor cognitive and attentional state. To reduce auditory interference from the noise of the scanner, we used a semisparse sampling design in which stimuli were presented in silent periods between volume acquisitions. The visual and auditory stimuli were presented for 1.2 s, flanked by 100 ms of silence before and after the stimulus. Then, a functional volume was acquired for 2 s, for a total interstimulus interval [and repetition time (TR)] of 3.4 s. Each run consisted of 99 audio-visual trials and 79 silent trials in which subjects viewed a white fixation cross.

Pretraining and posttraining scans were separated in time by one to 4 weeks across participants. After the second posttraining run, an additional functional localizer scan was run to identify each participant's speech-sensitive brain regions. Participants passively listened to 50 trials of spoken words and 50 trials of environmental sounds that corresponded semantically to the words, while seeing color photographs of the object that matched the word or sound. As in the alien runs, participants performed a visual oddball detection task, pressing a button whenever they saw an upside-down picture. Auditory stimuli were taken from a previous study comparing speech and environmental sound processing (Cummings et al., 2006).

Scanning took place at the Birkbeck-UCL Centre for Neuroimaging (BUCNI) using a 1.5T Siemens Avanto scanner with a 12-element phased array head coil. Functional imaging consisted of 21 T2*-weighted prospective-motion-corrected echo-planar image slices [TR = 3400 ms, echo time (TE) = 41 ms, field of view = 224 × 224 mm], giving a notional 3.5 × 3.5 × 3.5 mm resolution. Oblique axial slices were automatically positioned using Siemens AutoAlign so as to consistently image peri-Sylvian cortex. (The AutoAlign protocol acquires several short anatomical images at the beginning of each scanning session to align the participant's brain to a standard template brain, where the slice planes are defined). The slice plane was approximately aligned with the Sylvian fissure, with the inferior-most slice passing through or under the inferior temporal gyrus, and the superior-most slice passing at least above the inferior frontal sulcus and the supramarginal gyrus. A total of 180 volumes were collected per run. An automated shimming algorithm was used to reduce magnetic field inhomogeneities. In addition, for anatomical localization purposes, a T1-weighted scan was acquired during the pretraining scan (MPRAGE, TR = 2730 ms, TE = 3.57 ms) with 1 mm^2 in-plane resolution and 1 mm slice thickness.

Analyses

Functional imaging data were analyzed using FMRIB Software Library (www.fmrib.ox.ac.uk/fsl). After removing the first four images of each session to allow for T1 equilibrium, functional images were realigned to correct for small head movements (Jenkinson and Smith, 2001) and then smoothed with a 6 mm full-width half-maximum Gaussian filter to increase the signal-to-noise ratio. The time series data were prewhitened to remove temporal auto-correlation (Woolrich et al., 2001). Images were then entered into a general linear model to compute participant-specific patterns of activations for both pretraining and posttraining sessions. The presence of audio-visual alien stimuli was modeled by convolving trial onsets with a double-gamma "canonical" hemodynamic response function (Glover, 1999). Oddball trials were also entered in the model but were not included in the subsequent analyses. Silent trials formed the implicit baseline condition. In addition, temporal derivatives and estimated motion parameters were included as covariates of no interest to increase statistical sensitivity.

First level results were transformed into standard space using a 12 degree-of-freedom affine registration, first registering each functional run to each subject's high-resolution anatomical scan (acquired during the pretraining scanning session) before registering this to the MNI152 template. At the second level, a paired *t* test was used to compare activation in the pretraining and posttraining session using a mixed-effect model (Beckmann et al., 2003; Woolrich et al., 2004). In addition, the posttest behavioral auditory categorization score was included as a covariate to assess the relation between the participant's skill at learning to categorize the auditory stimuli and her/his brain activation. Activations were thresholded at $Z > 2.3$, and were considered significant at $p < 0.05$ using a cluster-wise significance test. Analyses were conducted on the whole brain and also within a speech-sensitive mask to increase signal detection (Friston et al., 1994). This mask was created at the group level using the speech-localizer task and included only voxels for which speech showed significantly greater activation than semantically matched environmental sounds (at $p < 0.05$, whole brain, cluster-corrected). The resulting mask spanned much of the left superior temporal gyrus and sulcus (cluster size = 13,066 mm^3, $z = 5.13$, center of gravity = [−57, −31, −2]) (see Fig. 4*a*).

In addition to the voxelwise analyses of the group activation data, we also performed a region of interest (ROI)-based correlation analysis between participants' change in activation from pretraining to posttraining and their accuracy in explicitly categorizing the alien sounds. Here, we first created a speech-selective ROI for each individual participant using their own speech > environmental sounds contrast. These individually defined ROIs included only voxels that fell within the group-defined mask, and that were active at a significance level of $p < 0.01$, corrected for multiple comparisons within the volume of the group mask (Worsley et al., 1992). Four participants had no suprathreshold voxels and therefore were not included in the analyses. Within each participant's ROI, we then extracted the average parameter estimates (i.e., betas) for the posttraining minus pretraining contrast. Participants' average change in activation within their individually defined speech-selective ROI was then correlated with their scores on the postscan test of auditory categorization.

Results

Participants varied considerably in how well they mastered the computer game. After training, game performance was measured by the highest game level achieved and varied from 13 to 32 (mean = 20). This indirect measure of auditory learning correlated strongly with participants' skill in explicitly categorizing the novel sounds in the postscan behavioral categorization test, with accuracy scores ranging from 7% to 89% (mean = 45%, cross-measure correlation: $r(11) = 0.83$, $p < 0.001$, see supplemental Fig. 1, available at www.jneurosci.org as supplemental material). Importantly, the variation among participants was sizable, with 5 of 17 participants categorizing sounds at or below chance levels (25%) while 3 of 17 participants achieved >75% correct.

The aim of the initial imaging analysis was to identify group-level changes in activation between the posttraining and pretraining scans in an undirected, whole brain analysis. This analysis revealed two clusters where there was a decrease in deactivation from pretraining to posttraining scans: one was located in the precuneus bilaterally and spread into cingulate regions ($x = -3$, $y = -50$, $z = 31$, cluster size = 27,589 mm^3, $p < 0.001$ cluster-wise) and the other was found along the left angular gyrus ($x = -46$, $y = -65$, $z = 31$, cluster size = 6389 mm^3, $p < 0.05$ cluster-wise), see Figure 3. Interestingly, there were no changes found in canonical speech-sensitive regions, even when the analyses were constrained to speech-sensitive regions of interest. Finally, this

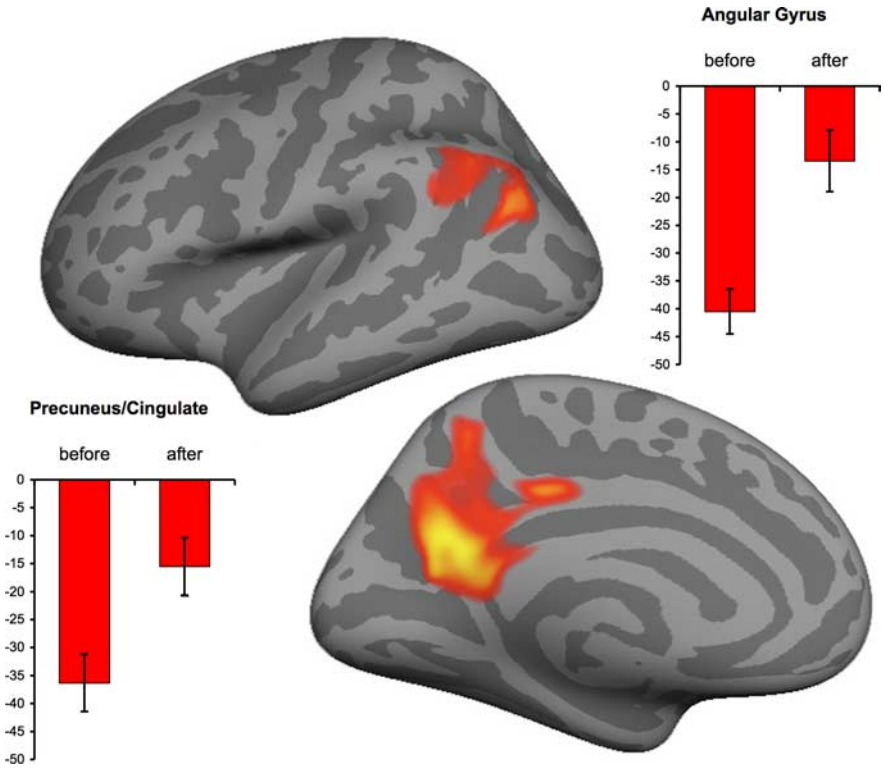

Figure 3. Overall differences in activation between pretraining and posttraining scans, painted in orange onto lateral (top) and medial (bottom) views of an average inflated cortical surface using FreeSurfer (Dale et al., 1999). Here, sulci are dark gray, and gyri are light gray. Bar graphs represent β weights at the peak voxel for each activation cluster; error bars represent $\pm$ 1 SE.

pattern did not change between trials in which the auditory and visual categories matched or mismatched. In other words, there was no evidence at a group level that training alone changed activation in speech-sensitive areas.

Given the considerable behavioral variability in how well participants learned the artificial auditory categories, we regressed activation induced by the artificial nonspeech sounds with each participant's performance on the behavioral categorization task. At the whole-brain level this regression revealed no regions with significant differences between pretraining and posttraining activation. However, within the left STS speech-sensitive region of interest (created via a separate scan as described in Materials and Methods), there was a cluster of voxels with a significant, positive correlation between participants' accuracy in categorizing the artificial sounds and their change in activation from pretraining to posttraining ($x = -54$, $y = -37$, $z = -1$, cluster size = 878 mm^3, $p < 0.05$ cluster-wise) (see Fig. 4*a*,*b*). The relationship between behavioral performance and change in pretraining to posttraining activation in left STS also held when speech-selective ROIs were defined on an individual-by-individual basis. As in the group-based analysis, behavioral categorization performance was positively correlated with an increase in activation in individual participant's speech-sensitive left temporal regions (Spearman's $r = 0.60$, $p = 0.014$, one-tailed) (Fig. 4*c*). In other words, participants who best learned the artificial auditory categories showed the greatest increase in activation in speech-sensitive cortex to these nonlinguistic sounds.

Discussion

These findings demonstrate that a region of left posterior STS (pSTS) commonly considered speech-selective is also recruited during passive perception of novel nonspeech sounds, but only in subjects who learned to perceive and use the category relationships among these sounds. This result is consistent with the hypothesis that speech-sensitive activation in left pSTS is in part driven by expertise in categorizing sounds.

Converging evidence that a left pSTS region is involved in categorizing complex sounds (whether they are speech or not) comes from reports of activation in similar regions in studies investigating sine-wave speech. For instance, Desai et al. (2008) and Dehaene-Lambertz et al. (2005) report activation of very similar pSTS regions during explicit categorization of sine-wave stimulus complexes, which under the right circumstances can be heard as either speech-like or nonspeech-like. These earlier studies focus on the change in activation that results from being instructed to categorize sine-wave stimuli as speech, and as such involve rapid (i.e., during the course of a single scanning session), possibly attentionally modulated shifts in cortical processing. The sine-wave speech studies suggest that this cortical region is not highly selective for full-spectrum speech acoustics, as sine-wave speech is an acoustic caricature of speech acoustics. One possibility is that left pSTS activation by sine-wave speech reflects a parasitic use of existing cortical speech processing regions.

The present stimuli are much different in this respect; they are not heard as speech-like and listeners cannot assign speech categories to them (Wade and Holt, 2005). What makes the current study markedly different from previous studies is the role it places on auditory category learning in modulating activation in the left posterior STS to nonspeech stimuli. Our study demonstrates that learning to categorize nonspeech sounds that have no obvious inherent category structure can induce a more speech-like neural profile. In contrast, the pretraining to posttraining analyses that did not account for individual variation in learning revealed activation changes in regions unlikely to be involved specifically with auditory expertise but rather more general effects, possibly relating to familiarity with the odd-ball task used during scanning. As noted previously, the activation changes in the precuneus, cingulate, and left angular gyrus actually involve a decrease in deactivation (rather than an increase in activation) from pretraining to posttraining. These regions are typically considered part of the default network (Buckner et al., 2008) with the activation changes in the present study possibly reflecting increasing familiarity with the task and stimuli used in the scanner. However, it is worth noting that other auditory and language training studies (Golestani and Zatorre, 2004) showed that posttraining activation in part of the left angular gyrus positively correlated with subjects' proficiency in learning a non-native speech contrast.

The current study does not address the issue of the spectrotemporal acoustic properties of the auditory stimuli and how they might interact with expertise-related changes in activation. Left-lateralized superior temporal regions implicated in speech processing have also been implicated in processing nonspeech

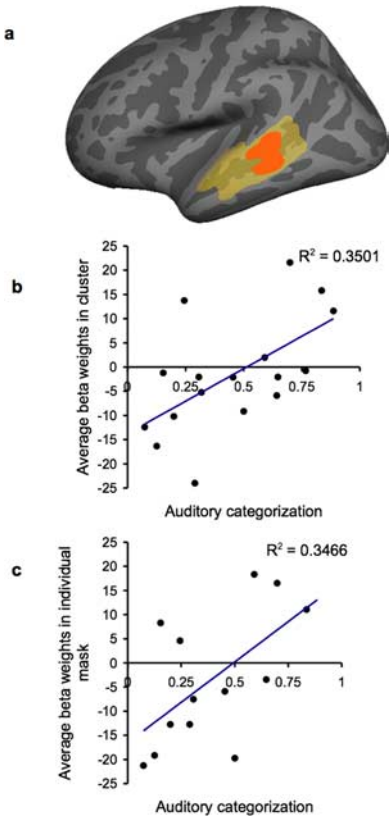

Figure 4. Increased activation in speech-sensitive superior temporal sulcus with auditory expertise. ***a***, The group-based speech > environmental sound localizer functional ROI, painted in transparent yellow onto the average inflated cortical surface. The location of increase in activation from pretraining to posttraining is painted in orange on the cortical surface. ***b***, The relationship between subjects' performance at categorizing the sounds outside the scanner and change in β weights from pretraining to posttraining (averaged across the cluster). ***c***, The relationship between auditory categorization ability and change in β weights from pretraining to posttraining, averaged across individually defined functional ROIs.

stimuli with rapid temporal changes (for a review, see Zatorre and Gandour, 2008). Therefore, there may be an interaction between training and the specific acoustic characteristics of nonspeech sound stimuli (Mirman et al., 2004). Expertise with complex stimuli involving spectrotemporal changes aligned with the time course and frequency range of speech may be reflected in modulation of activation in left pSTS region, but the present results cannot eliminate the possibility that expertise with acoustic signals radically different from speech would produce a different pattern of results. Nevertheless, the acoustic distinction between the alien sounds and speech is sufficiently great to suggest that any acoustic specialization within pSTS must be rather broadly tuned. It may be that left pSTS is particularly well suited for parcellating rapid temporally varying higher-dimensional acoustical space into sound categories, be they speech or nonspeech.

Although we find that activation in this left pSTS was modulated by listeners' expertise with auditory categories, we do not suggest this region is solely responsible for or dedicated to carrying out auditory categorization. The underlying auditory processing might be best understood in terms of more distributed accounts of neural processing possibly spanning a network of regions (Haxby et al., 2001). Indeed, consistent with this, a recent study taking advantage of multivoxel pattern classification techniques demonstrates that the processing underlying speech and voice categorization may be widely distributed across superior temporal regional (Formisano et al., 2008).

The results of the current study with expertise for auditory stimuli can be interpreted in the context of the "Greebles" studies from the face processing and visual expertise literature (Gauthier et al., 1999). In these studies, learning categories of complex nonface visual stimuli recruits activation in face-sensitive cortical regions, suggesting that seemingly domain-specific patterns of neural activation in higher cortical regions may be driven, in part, by experience-based restructuring of high-dimensional perceptual space. Similar results have also been observed retrospectively with experts in different domains, such as in discriminating cars and birds (Tarr and Gauthier, 2000; Xu, 2005). However, this has been a highly active research area and there is considerable debate about the replicability and the magnitude of expertise effects in face-selective regions of cortex involving nonface stimuli (for a review, see Kanwisher and Yovel, 2006). A similarly polarized debate is unlikely to be helpful in the context of speech processing, where there is growing recognition of the need to integrate evidence for general-purpose with domain-specific neural mechanisms (Zatorre and Gandour, 2008).

As with the Greebles studies, the current findings imply that at least some of what is taken to be part of a speech-specific cortical processing network is actually part of a more general network for processing and categorizing sound. The preferential left-lateralized STS activation typically observed for speech, in fact, may reflect ontogenetic changes to the neural mechanisms for processing auditory stimuli that result from expertise with speech sounds. The implication is that part of what makes the processing of speech-sounds special is how speech is used and the cortical fine-tuning this induces with the experience of speech, rather than a qualitative adaptation to speech sounds involving dedicated processing networks or representations.

References

Beckmann CF, Jenkinson M, Smith SM (2003) General multilevel linear modeling for group analysis in FMRI. Neuroimage 20:1052–1063.

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

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512–528.

Buckner R, Andrews-Hanna JR, Schacter DL (2008) The brain's default network: Anatomy, function, and relevance to disease. Ann N Y Acad Sci 1124:1–38.

Cummings A, Ceponiene R, Koyama A, Saygin AP, Townsend J, Dick F (2006) Auditory semantic networks for words and natural sounds. Brain Res 1115:92–107.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179–194.

Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A, Dehaene S (2005) Neural correlates of switching from auditory to speech perception. Neuroimage 24:21–33.

Desai R, Liebenthal E, Waldron E, Binder JR (2008) Left posterior temporal regions are sensitive to auditory categorization. J Cogn Neurosci 20:1174–1188.

Dick F, Saygin AP, Galati G, Pitzalis S, Bentrovato S, D'Amico S, Wilson S, Bates E, Pizzamiglio L (2007) What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: evidence from functional magnetic resonance imaging and lesion data. J Cogn Neurosci 19:799–816.

Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. Annu Rev Psychol 55:149–179.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–973.

Friston KJ, Worsley RSJ, Frackowiak JC, Mazziotta JC, Evans AC (1994)

Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1:214–220.
Gauthier I, Tarr MJ, Anderson AW, Skudlarski P, Gore JC (1999) Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. Nat Neurosci 2:568–573.
Glover GH (1999) Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 9:416–429.
Golestani N, Zatorre RJ (2004) Learning new sounds of speech: reallocation of neural substrates. Neuroimage 21:494–506.
Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.
Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156.
Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci 361:2109–2128.
Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.
Kuhl PK (2004) Early language acquisition: cracking the speech code. Nat Rev Neurosci 5:831–843.
Mirman D, Holt LL, McClelland JL (2004) Categorization and discrimination of non-speech sounds: differences between steady-state and rapidly-changing acoustic cues. J Acoust Soc Am 116:1198–1207.
Ohnishi T, Matsuda H, Asada T, Aruga M, Hirakata M, Nishikawa M, Katoh A, Imabayashi E (2001) Functional anatomy of musical perception in musicians. Cereb Cortex 11:754–760.
Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK (2008) A voice region in the monkey brain. Nat Neurosci 11:367–374.
Price C, Thierry G, Griffiths T (2005) Speech-specific auditory processing: where is it? Trends Cogn Sci 9:271–276.
Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123:2400–2406.
Tarr MJ, Gauthier I (2000) FFA: a flexible fusiform area for subordinate-level visual processing automated by expertise. Nat Neurosci 3:764–769.
Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RB (2003) Faces and objects in macaque cerebral cortex. Nat Neurosci 6:989–995.
Wade T, Holt LL (2005) Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. J Acoust Soc Am 118:2618–2633.
Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370–1386.
Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM (2004) Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21:1732–1747.
Worsley KJ, Evans AC, Marrett S, Neelin P (1992) A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 12:900–918.
Xu Y (2005) Revisiting the role of the fusiform face area in visual expertise. Cereb Cortex 15:1234–1242.
Zatorre RJ, Gandour JT (2008) Neural specializations for speech and pitch: moving beyond the dichotomies. Philos Trans R Soc Lond B Biol Sci 363:1087–1104.