



## Locating the initial stages of speech–sound processing in human temporal cortex

Stefan Uppenkamp,<sup>a,c,\*</sup> Ingrid S. Johnsrude,<sup>b,d,1</sup> Dennis Norris,<sup>b</sup>  
William Marslen-Wilson,<sup>b</sup> and Roy D. Patterson<sup>a</sup>

<sup>a</sup>Centre for the Neural Basis of Hearing, Department of Physiology, University of Cambridge, Downing Street, Cambridge, CB2 3EG, UK

<sup>b</sup>MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 2EF, UK

<sup>c</sup>Medizinische Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany

<sup>d</sup>Department of Psychology, Queen's University, Kingston, Canada K7L 4E8

Received 7 July 2005; revised 1 November 2005; accepted 16 January 2006

It is commonly assumed that, in the cochlea and the brainstem, the auditory system processes speech sounds without differentiating them from any other sounds. At some stage, however, it must treat speech sounds and nonspeech sounds differently, since we perceive them as different. The purpose of this study was to delimit the first location in the auditory pathway that makes this distinction using functional MRI, by identifying regions that are differentially sensitive to the internal structure of speech sounds as opposed to closely matched control sounds. We analyzed data from nine right-handed volunteers who were scanned while listening to natural and synthetic vowels, or to nonspeech stimuli matched to the vowel sounds in terms of their long-term energy and both their spectral and temporal profiles. The vowels produced more activation than nonspeech sounds in a bilateral region of the superior temporal sulcus, lateral and inferior to regions of auditory cortex that were activated by both vowels and nonspeech stimuli. The results suggest that the perception of vowel sounds is compatible with a hierarchical model of primate auditory processing in which early cortical stages of processing respond indiscriminately to speech and nonspeech sounds, and only higher regions, beyond anatomically defined auditory cortex, show selectivity for speech sounds.

© 2006 Elsevier Inc. All rights reserved.

**Keywords:** Auditory; Cortex; Speech; fMRI; Hearing

### Introduction

Processing in the ascending auditory pathway of the primate is largely sequential and hierarchical up to and including auditory

cortex with its core, belt and parabelt regions (Rauschecker, 1998; Kaas and Hackett, 2000). Speech is recoded by the auditory nuclei of the brainstem and thalamus before it reaches auditory cortex (Irvine, 1992; Eggermont, 2001; Frisina, 2001), and both the anatomical and physiological literature suggest that the recoding involves processing of acoustic features relevant to speech. However, these nuclei seem to apply general processes to all sounds independent of their source, and there is no indication that speech-specific processing begins before auditory cortex.

Neuroimaging evidence indicates that, in humans, information is processed in a hierarchical manner in auditory cortex, and that this hierarchical organization extends into the multimodal regions beyond auditory cortex (Hall et al., 2001, 2003; Wessinger et al., 2001; Patterson et al., 2002; Scott and Johnsrude, 2003). It is clear that speech sounds have to pass through primary auditory cortex (PAC), which is active in the presence of speech sounds (e.g., Belin et al., 2002; Ahissar et al., 2001). But the corresponding area in nonhuman mammals (A1) is also active in the presence of speech sounds (Steinschneider et al., 2003; Versnel and Shamma, 1998), so this activity appears to represent general auditory processing of acoustic features rather than speech-specific processes. Regions well beyond PAC, in the superior temporal sulcus and middle temporal gyrus, are sensitive to the intelligibility of sentences, particularly in the left hemisphere (Davis and Johnsrude, 2003; Narain et al., 2003; Scott et al., 2000; Giraud et al., 2004). However, intervening between PAC and these middle temporal regions, there are at least three anatomically differentiable regions that appear to be predominantly auditory, and the connectivity of these regions suggests at least three stages of processing (the human homologues of belt and parabelt cortex, and the upper STS region TAA; Kaas and Hackett, 2000; Seltzer and Pandya, 1991). We hypothesize that the earliest stages to show some specialization for the processing of speech will be within these three regions.

A large number of imaging studies have investigated temporal-lobe involvement in speech-specific processing (e.g., Binder et al.,

\* Corresponding author. Institut für Physik, Medizinische Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany.

E-mail address: stefan.uppenkamp@uni-oldenburg.de (S. Uppenkamp).

<sup>1</sup> The first two authors contributed equally to this work.

Available online on ScienceDirect (www.sciencedirect.com).

1997, 2000, 2004; Benson et al., 2001; Callan et al., 2004; Crinion et al., 2003; Davis and Johnsrude, 2003; Dehaene-Lambertz et al., 2005; Demonet et al., 1992; Gandour et al., 2003; Giraud and Price, 2001; Giraud et al., 2004; Hugdahl et al., 2003; Jancke et al., 2002; Joanisse and Gati, 2003; Liebenthal et al., 2003, 2005; Mummery et al., 1999; Narain et al., 2003; Poeppel et al., 2004; Rimol et al., 2005; Schlosser et al., 1998; Scott et al., 2000; Specht and Reul, 2003; Thierry et al., 2003; Vouloumanos et al., 2001; Zatorre et al., 1992; see also Belin et al., 2000; Price et al., 2005). However, many of these studies have investigated speech at the level of the word, phrase or sentence, and such stimuli would probably have engaged lexical, semantic and syntactic processes in addition to speech-sound processing (e.g., Crinion et al., 2003; Davis and Johnsrude, 2003; Giraud et al., 2004; Narain et al., 2003; Scott et al., 2000; Schlosser et al., 1998; Zatorre et al., 1992). In many studies, the acoustic characteristics of the speech stimuli differed substantially from the nonspeech stimuli (e.g., Benson et al., 2001; Binder et al., 2000, 2004; Demonet et al., 1992; Giraud and Price, 2001; Vouloumanos et al., 2001). Many studies include a task that required or encouraged the listener to make an explicit linguistic judgement about the sounds they were hearing (Callan et al., 2004; Jancke et al., 2002; Liebenthal et al., 2003; Binder et al., 2004; Dehaene-Lambertz et al., 2005). But none of these studies has compared the activity to elementary speech sounds (such as vowels) with that to acoustically matched nonspeech sounds, and done so while listeners are engaged on a task that encourages attention to the stimuli, but that does not require or encourage linguistic processing of the sounds.

In this paper, we identify the cortical regions where processing of speech and nonspeech sounds begins to diverge by comparing a new class of synthetic vowels with a set of nonspeech controls that are closely matched in terms of the distribution of energy across frequency and over time. The synthetic vowels, with distinctive properties that identify them as linguistically relevant sounds produced by a human vocal tract, are immediately heard as speech sounds, while the nonspeech controls cannot be heard as speech even with deliberate effort. We furthermore deliberately chose to use a task that would not preferentially engage speech processing—listeners performed a simple intensity monitoring task.

## Materials and methods

### *Stimulus development*

In creating synthetic vowels, we adhered to the following constraints: (1) Their spectra exhibit three to four relatively narrow, formants in the frequency region below 4000 Hz. (2) Stimuli are presented in sequences, where the individual elements within each sequence occur at a rate of about three per second. (3) The spectra of successive vowels within a sequence differ, but they all appear to come from the same vocal tract (e.g., they are different vowels spoken by the same person). All three constraints serve to enhance the speech-like quality of the vowel sequences. We created randomly ordered triplets of synthetic vowels with the three basic constraints described above, and they are always perceived as sequences of vowels.

For current purposes, vowels are streams of glottal pulses filtered by the vocal tract above the larynx. Each glottal pulse excites the four main resonances (or formants) of the vocal tract, and the response of each resonance to each glottal pulse is like a

‘damped sinusoid’, that is, a sinusoid with a decaying exponential envelope (Patterson, 1994). If the speaker is a man and the vowel is /a/, the four main formants will be around 730, 1090, 2400 and 3300 Hz, and the four streams of glottal pulses will look like those in the upper left-hand panel of Fig. 1. The amplitudes of the upper formants have been reduced relative to the lower formants as they are in normal speech. When summed, the four repeating damped sinusoids produce the wave shown in the lower part of this panel, and this wave is perceived as the vowel, /a/, as in the word ‘paw’. In this example, the period of the wave is 12 ms which corresponds to a pitch of 83 Hz.

To simulate sequences of these vowel sounds and matched controls, we developed a computer algorithm that constructs vowel sounds by calculating damped sinusoids with the appropriate carrier frequencies *on a cycle-by-cycle basis*, and which adds the component damped sinusoids to produce the vowel. We generated the five strong vowels /a/ (caw, dawn), /e/ (ate, bait), /i/ (beat, peel), /o/ (coat, wrote) and /u/ (boot, pool); the duration of the stimuli was 400 ms. These are referred to as ‘damped’ vowels (dv) and they elicit a clear speech percept (Patterson et al., 2000), especially when presented in sequences where the vowel type is varying.

The same algorithm was then used to create three classes of control stimuli, all of which had the same overall energy and the same long-term spectral and temporal characteristics as the vowel stimuli. These controls were produced by violating one spectral constraint of normal vowel production, one temporal constraint of normal vowel production or both. In natural vowels, formant frequencies are relatively stable, varying only slowly from cycle to cycle (spectral constraint), and the damped sinusoids all repeat regularly with the same period (temporal constraint). We violated the spectral constraint by randomizing the carrier frequency of each damped sinusoid at the start of each cycle over about an octave range around the formant frequency. We violated the temporal constraint by randomizing the start times of the individual damped sinusoids within each ‘cycle’, that is, within each 12-ms time segment.

When both constraints are violated simultaneously (as shown in the lower right-hand panel of Fig. 1), the resulting stimuli sound like a rapid spatter of overlapping tone pips, with an unusual but pleasant rain-like quality. We refer to the stimulus as “musical rain” (mr) to mark the fact that it does not produce a perception of speech (as documented in the next section). This is the crucial control stimulus in the design: as the results will show, it is sufficiently like the damped vowels physically to produce the same level of activation in all centers in the auditory pathway up to and including the primary receiving areas of auditory cortex in HG and PT—areas which it is argued deal primarily with the basic physical features of the sound. At the same time, musical rain is sufficiently different from damped vowels perceptually to produce differential activation in regions just beyond the primary receiving areas of cortex.

If the temporal constraint is preserved and the damped sinusoids all start at the same time (as in the lower left-hand panel of Fig. 1), the dominant perception is still one of musical rain but there is also a low buzzy pitch in the background. These stimuli are also not perceived as speech, and so, we refer to this stimulus as “musical rain with pitch” (mp). The imaging contrast between mp and mr should reveal any activation specific to the presence of a fixed voice-pitch in the absence of formants. Finally, if the spectral constraint is maintained but the temporal constraint is still violated

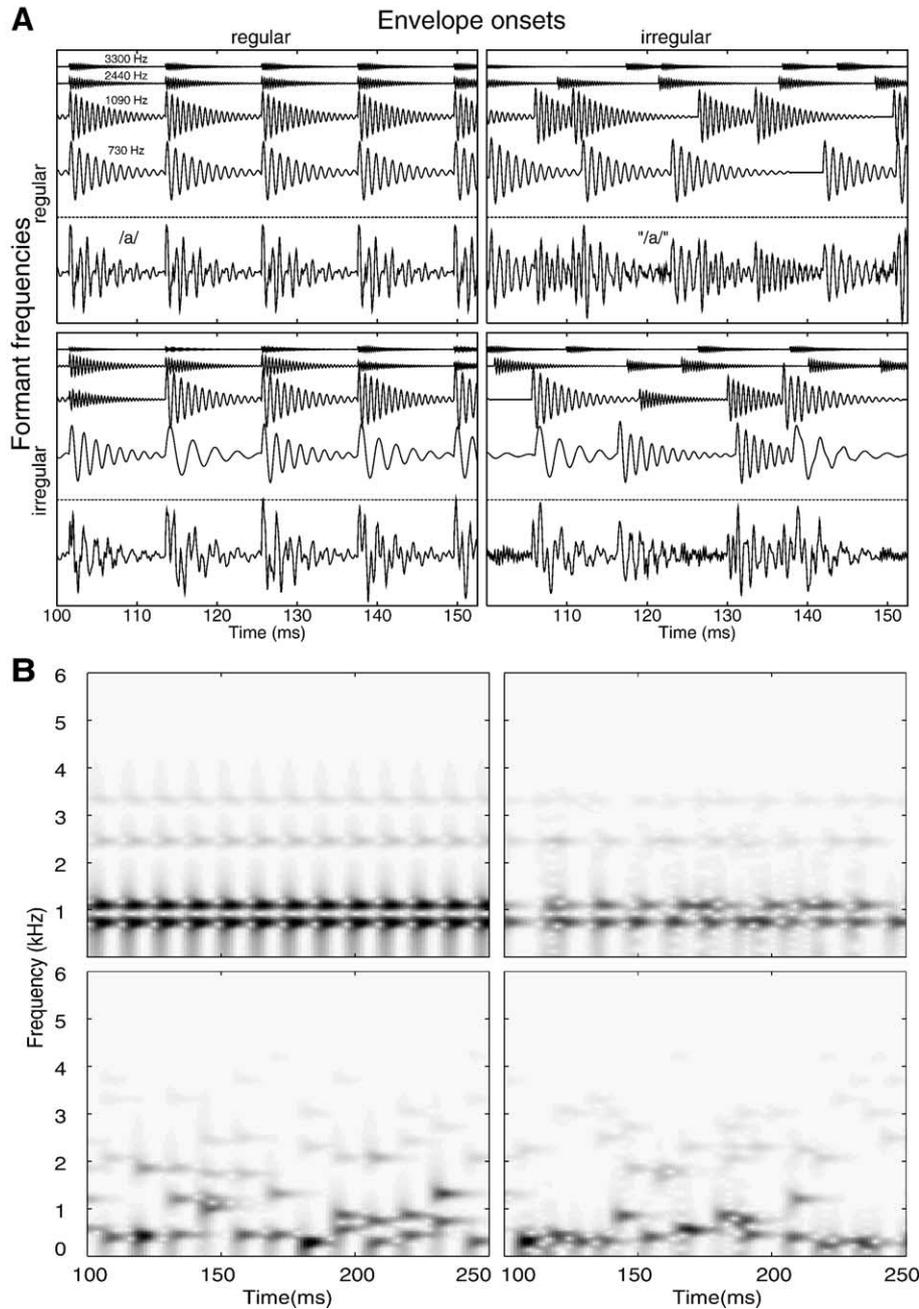


Fig. 1. Four classes of stimuli constructed from sets of isolated formants (damped sinusoids). Top left: “damped” vowel, dv, with regular envelope onsets and fixed formant frequencies. Top right: “raspy” vowel, rv, with fixed formants but randomized envelope onsets. Bottom left: damped sinusoids with synchronous onsets (which produces a pitch), but carrier frequencies which are randomized from cycle to cycle. It sounds like a gurgle with a pitch (“musical rain with pitch”, mp) Bottom right: both onsets and carrier frequencies are randomized from cycle to cycle, producing “musical rain”, mr. (B) Spectrograms of the same stimuli: top left: dv; top right: rv; bottom left: mp; bottom right: mr. Example stimuli can be downloaded from [http://medi.uni-oldenburg.de/members/stefan/phonology\\_1](http://medi.uni-oldenburg.de/members/stefan/phonology_1).

(as in the upper right-hand panel of Fig. 1), the perception of speech persists and the vowels are identifiable, but their voice quality is degraded. They have a raspy whispered quality with no pitch, as if the larynx was damaged. These are referred to as “raspy vowels” (rv). The imaging contrast between dv and rv should reveal any activation specific to a fixed pitch in the presence of formant structure. Basically, however, the purpose of these two extra conditions was to increase the variability of the stimuli within the speech and nonspeech classes, since this increases the power of

the experiment. Example stimuli can be downloaded from [http://medi.uni-oldenburg.de/members/stefan/phonology\\_1](http://medi.uni-oldenburg.de/members/stefan/phonology_1).

#### *Documenting the perceptual distinction between the speech and nonspeech stimuli*

A paired-comparison experiment was performed with nine normally hearing listeners to ensure that there was a clear-cut perceptual distinction between the ‘speech’ and ‘nonspeech’

sounds, with the former unambiguously more speech-like. We used the Bradley–Terry–Luce technique (David, 1988) which converts paired-comparison data into a relative perceptual scale—in this case a scale of ‘speechlikeness’. This technique requires that the stimulus set includes a reasonable diversity of perceptual conditions. We therefore constructed 14 different forms of synthetic vowel and other nonspeech sounds to compare with the four types of stimuli used in the imaging experiment. In some forms, the carrier sinusoids were replaced with narrow bands of noise, making them more like whispered speech. In some, the envelopes of the carrier sinusoids were not modulated, producing different versions of “sinusoidal speech” (Remez et al., 1981). In others, the number of formants was lowered to two or even one to reduce the speech-like quality of the sound. The forms are listed in Table 1 with a brief description of the manipulation used to produce them.

A two-interval, forced-choice task was used to gather the paired-comparison data. Each interval contained a sequence of three examples of one stimulus type (chosen at random). The sounds had durations of 400 ms and they were separated by 200-ms gaps. The onset of each interval was marked by a light and the listeners were asked to choose the interval that sounded most vowel-like. In the case of two completely nonvowel-like sounds, they were asked to choose the one that was more like speech. All of the stimuli were scaled to have the same intensity (RMS level). The sounds were played diotically via headphones (AKG 240D) at 50 dB HL by a TDT system II with sampling frequency of 20 kHz, through a low-pass filter with a cut-off frequency of 8 kHz. During the experiment, each condition was compared with each of the other conditions twice, once with the order A–B, and once with the order B–A, for a total of 306 trials. The order of the trials was randomized, and the experiment lasted approximately 1 h.

Table 1  
Summary of sound conditions used in the perceptual experiment

dmp_vow	* Damped vowels, four tracks of damped sinusoids at formant frequencies
dmp_two	As dmp_vow, but only first and second formant
dmp_fst	As dmp_vow, but first formant only
dmp_snd	As dmp_vow, but second formant only
flt_vow	As dmp_vow, but no lowpass slope in spectrum
jit_vow	As dmp_vow, 10% jitter in envelope timing
rsp_vow	* As dmp_vow, irregular envelopes (100% jitter in timing), i.e., no pitch
noi_vow	As dmp_vow, but narrow bands of noise as carriers of triangles
sin_vow	Four sinusoids at formant frequencies, no damped envelope
sin_two	As sin_vow, but only first and second formant
sin_fst	As sin_vow, but first formant only
sin_snd	As sin_vow, but second formant only
noi_pit	Four tracks of damped noise bands, one octave wide
noi_ran	As above, but irregular envelope (no pitch)
fxr_pit	Four tracks of damped sinusoids, random change of carrier frequencies within limited bandwidth, regular timing
fxr_ran	As above, but random timing (no pitch)
mus_pit	* Complete randomization of carrier frequencies for each track, regular timing
mus_ran	* Randomization of carrier frequencies and timing

The sounds shown in Fig. 1 are marked with an asterisk.

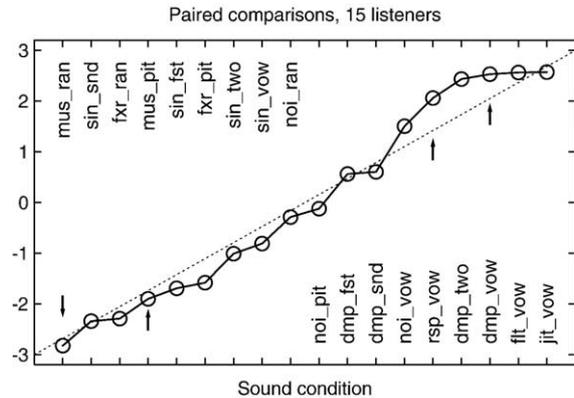


Fig. 2. Relative scale of preference for synthetic vowel and nonvowel sounds from the paired-comparison experiment. Damped vowels and musical rain have similar long-term distributions of energy in time and frequency, so they should activate primary auditory areas in a similar way. At the same time, they sound very different, so areas in the brain that are sensitive to the difference in their internal structure should show a strong contrast between these conditions.

A perceptual scale, reflecting the speech-like quality of the sounds, was constructed using a Bradley–Terry–Luce model (David, 1988) from the pooled paired comparisons of all nine listeners. The primary assumption is that the stimuli can be ordered linearly on the dimension of interest. The resulting scale of speechlikeness is presented in Fig. 2; it is a relative scale in which only differences between conditions have meaning. The scale shows clearly that the two synthetic-vowel conditions (damped and raspy) were among the most speech-like of the stimuli and the two musical-rain conditions were among the least speech-like. The conditions with flat envelopes and either sinusoidal or noise carriers come towards the middle of the order.

To be effective as control stimuli in the neuroimaging experiment, the critical property of the nonspeech sounds (the two forms of musical rain) is that they have the same energy as the damped and raspy vowels, and similar spectro-temporal profiles; that is, similar distributions of energy over frequency and time. The fact that this preliminary behavioral study showed such strong differences in speechlikeness indicates that it is the internal structure of the sound, and, in particular, the presence of formant frequencies that are fixed for the duration of the sound, that is crucial to the perception of a vowel sound. Neural centers that are not concerned with these specific aspects of the stimulus can be expected to produce the same degree of activation to the two classes of sounds, and thus they will not appear in the speech–nonspeech contrasts.

#### Functional magnetic resonance imaging experiment

##### Subjects

Ten right handed listeners between 20 and 50 years of age were scanned (three women and seven men). Volunteers were without any history of neurological illness, head injury or hearing impairment. The study was approved by the Addenbrooke’s Local Research Ethics Committee (Cambridge, UK), and written informed consent was obtained from all subjects.

##### Scanning procedure

We acquired imaging data using a Medspec (Bruker, Ettlingen, Germany) 3-T MRI system with a head gradient set. Echo-planar

imaging (EPI) volumes (192 in total) were acquired over four 8-min sessions. Each volume consisted of 21 slices (resolution  $1.95 \times 1.95 \times 5$  mm; echo time, 27 ms; acquisition time, 3.02 s). Acquisition was transverse oblique, angled away from the eyes and covered all of the brain except in a few cases; the very top of the superior parietal lobule, the anterior inferior edge of temporal cortex and the inferior aspect of the cerebellum were omitted in volunteers with larger brains. These areas were not included in the group analyses.

We used a sparse imaging technique (Edmister et al., 1999; Hall et al., 1999). On each trial, there was a 7-s stimulus interval followed by a 3-s scanning interval, making a total repetition time of 10 s. A sequence of 16 sounds from one of the experimental conditions was present on each trial; the tokens were selected randomly (with replacement) from the set that defined the condition and there were short gaps between the vowels. The complete sequence was approximately 9 s in duration, so it filled the stimulus interval and ran on into the start of the scanner noise to de-emphasize the abrupt onset of the scanner noise (see Fig. 3). The sparse imaging technique minimizes the interference of scanner noise on brain activity because the hemodynamic response to the experimental stimulus builds up while the scanner noise is off, and then the response is captured at its maximum strength by the scanner before the response to the scanner noise builds up. The technique has been shown to maximize statistical power in studies with auditory stimuli (Hall et al., 1999).

The volunteers heard *five* different kinds of stimuli, which consisted of the four conditions displayed in Fig. 1 and a condition with natural vowels (nv) produced by a native speaker of British English (author DN); the pitch of the natural vowels was the same as for the damped vowels (83 Hz). There was also a rest condition in which no sound was presented. Each of these six experimental conditions was repeated 32 times.

The volunteers were told at the outset that they would be hearing sequences of complex sounds, and they were asked to perform a level monitoring task to encourage them to remain alert throughout the course of the experiment. Every so often, two consecutive sounds would be presented at a lower level than the rest of the sounds in the sequence, and the volunteers were requested to press a button with their right hand when this occurred. These level targets (which were attenuated by 10 dB compared to other items in the sequence) were presented once every four trials on average. The execution of the task was monitored by the operator. The behavioral data were not recorded, however, since the purpose of the task was solely to encourage subjects to attend to the stimuli.

Four scanning sessions of 48 trials were performed. Sequences were presented in a pseudorandom order, which was the same for each listener. Subjects heard the sounds through high-quality electrostatic earphones mounted in sound-attenuating ear defenders (Palmer et al., 1998). Participants reported that the scanner noise was unobtrusive and that sounds were presented at a comfortable listening volume and at equal levels in both ears. Custom software (Palmer et al., 1998) was used to present the stimuli.

#### Analysis of fMRI data

The data processing and analysis were accomplished using Statistical Parametric Mapping (SPM99; Wellcome Department of Cognitive Neurology, London, U.K. <http://www.fil.ion.ucl.ac.uk/spm>). The BOLD time series was aligned to the first image of the series and then the structural image was co-registered with the BOLD series and resampled to  $2 \times 2 \times 2$  mm resolution. The realigned BOLD images were normalized to the standard SPM EPI template (ICBM 152, Brett et al., 2002) using affine and smoothly nonlinear spatial transformations, masking regions of susceptibility artifact to reduce tissue distortion (Brett et al., 2001). T1-weighted structural images were also spatially normalized. The resulting images are in standardized, “Talairach-like” space, the ICBM-152 template creates images that are a few millimeters displaced from the Talairach brain (Talairach and Tournoux, 1988), particularly in the superior–inferior ( $z$ ) dimension (Brett et al., 2002). Finally, the functional data were spatially smoothed using a Gaussian kernel whose full-width, half-maximum was 5 mm (Worsley et al., 1992).

All six conditions (five speech/nonspeech plus the rest condition) were entered separately into the design so that scan-to-scan movements identified in the realignment stage could be included as ‘covariates of no interest’ to further reduce the effect of movement on activation. Fixed-effects analyses were conducted on each listener’s data (192 scans), and across the whole group of subjects (with a total of 1728 scans) using the general linear model. Data from one subject were not used owing to excessive motion artifact in the single-subject fixed-effects analysis. The threshold for activation was  $P < 0.05$  corrected for multiple comparisons across the whole volume using the false discovery rate (FDR) correction procedure (Genovese et al., 2002).

The description of the stimuli summarized in Fig. 1 gives the impression that the experiment is a balanced  $2 \times 2$  factorial design: speech vs. nonspeech, and pitch vs. no pitch. However, this is not

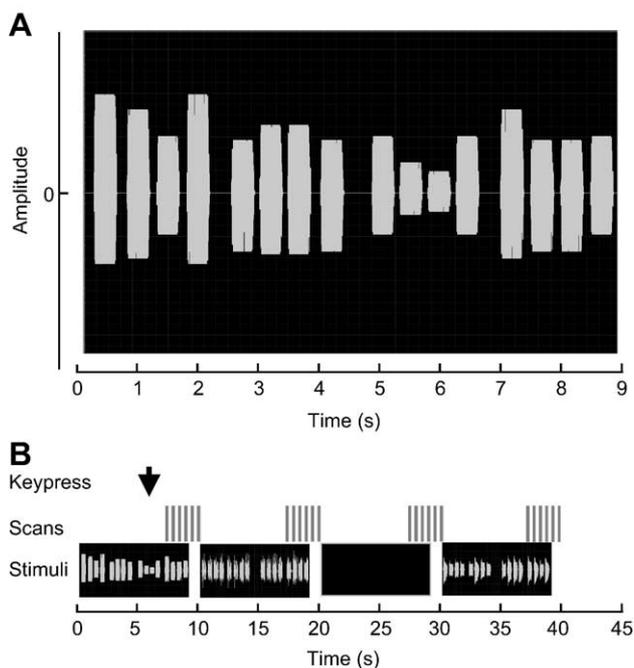


Fig. 3. The fMRI experimental procedure. (A) Stimulus amplitude as a function of time for one of the damped-sinusoid sequences used in the fMRI experiment. This particular sequence includes an attenuated pair (items 10 and 11), to which subjects were expected to make a keypress response. The attenuated pairs occurred every four sequences, on average. (B) The sparse-imaging design exploits the hemodynamic lag to acquire data at a time when stimulus-related activity is at its peak and activation resulting from the noise of the previous scan has decayed (Hall et al., 1999, 2000). Stimulus sequences were 9-s long and each scan volume took 3 s to acquire. The repetition time was 10 s.

the case for the experiment itself, because within the sequence that defines a trial, the vowel type (/a/, /e/, /i/, /o/, /u/) varied from sound to sound, whereas the pitch did not. This difference had the effect of making the vowel perception dominant and reducing the pitch perception to a component of the background. Nevertheless, in the analysis of the fMRI data, we included an analysis for the presence of a pitch effect because, in previous work, we had identified a bilateral region on lateral Heschl's gyrus (most probably an auditory belt region) that is sensitive to pitch salience in tonal stimuli (Griffiths et al., 2001; Patterson et al., 2002).

#### Evaluation of individual anatomy

The location of primary auditory cortex was identified in each listener, so that we could specify the relationship between functional activation and macroscopic anatomy precisely. Three of the authors (RP, SU, IJ) labeled Heschl's gyrus (HG) in both hemispheres of each listener, using MRIcro software (<http://www.psychology.nottingham.ac.uk/staff/crl/micro.html>) and the anatomical criteria suggested by Penhune et al. (1996). Both white and grey matter were included in the labeled volumes. In some listeners, HG is duplicated or partially duplicated (Penhune et al., 1996; Leonard et al., 1998). In this study, the labeling was restricted to the part of HG anterior to any dividing sulcus, since this is the most likely location of cytoarchitecturally defined primary auditory cortex (Rademacher et al., 1993; 2001). The labeling of the three judges was combined to produce a labeled volume of HG for each listener by including all voxels that two or

more judges had labeled as part of HG. The result is a three-dimensional HG map co-registered with the individual's functional data. Finally, a mean HG volume was created for the group by averaging across the individual labeled volumes for the nine subjects included in the functional analysis.

## Results

#### Activation in response to sound

When each of the five sound conditions was contrasted with silence, similar patterns of activation were observed (Fig. 4). The activation was largely confined to the temporal lobes, bilaterally, and there were always two main foci of activation in auditory cortex; one towards the medial end of Heschl's gyrus and one towards the lateral end of HG. These two foci are highly consistent in their locations across conditions.

#### Main effect of speechlikeness

No significant regions of activation were observed when any of the three vowel conditions was contrasted with each other. Activation for the two synthetic vowel classes was indistinguishable from the vowels produced by a human speaker. The fact that raspy vowels produce just as much activation as damped vowels and natural vowels suggests that the perception of these sounds as

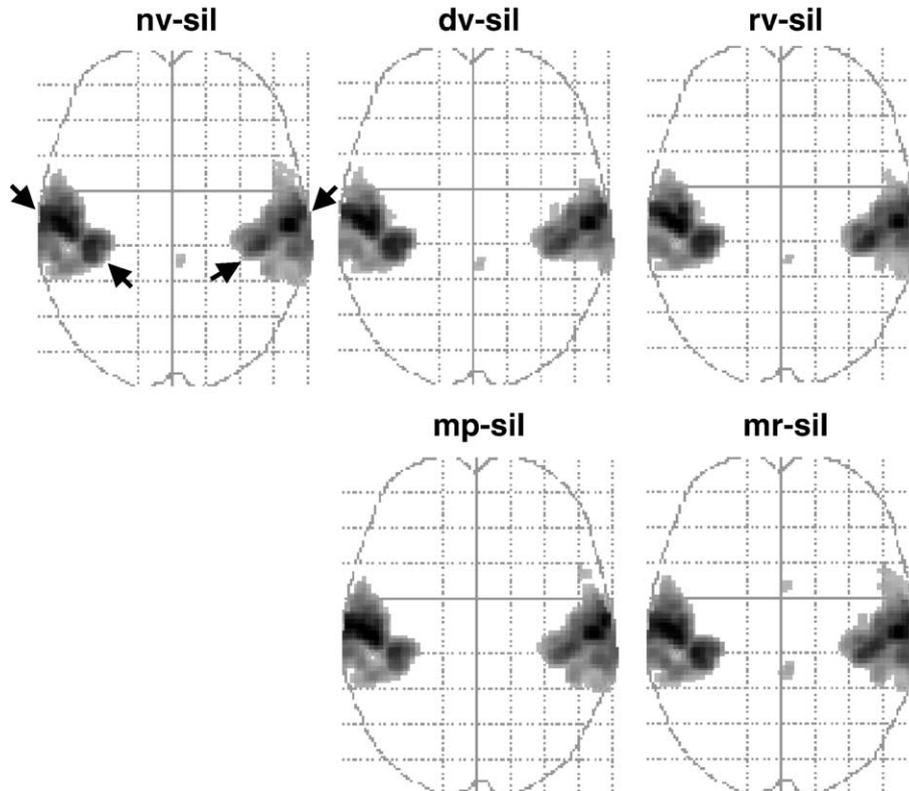


Fig. 4. Axial projection of group activation onto "glass brains", showing the response to individual stimulus conditions when compared to silence. Top row left to right: natural vowels (nv) minus silence (sil); damped vowels (dv) minus sil, and raspy vowels (rv) minus sil. Bottom row: musical rain with pitch (mp) minus sil (left) and musical rain (mr) minus sil (right). The arrows mark the approximate location of Heschl's gyrus in the two hemispheres. There are two activation foci on the long axis of Heschl's gyrus (posteromedial and anterolateral) in each contrast. Thus, these five types of sound produce activation in essentially the same cortical areas when contrasted with silence. In this and subsequent figures, the left hemisphere is on the left of the image in the coronal and horizontal slices.

vowels is the dominant feature for these regions of auditory cortex, not the particular realization of these vowel sounds.

Contrasting the three vowel conditions with the two nonspeech conditions revealed a highly significant bilateral region of activation in the superior temporal sulcus (STS) inferior to Heschl's gyrus (Fig. 5, left column, yellow; see Table 2). There were several other foci of activation observed bilaterally in anterior STS, on the superior temporal gyrus (STG), and in the mid-dorsolateral, precentral/premotor region, but their level of significance was much lower. Vowel stimuli did not produce differential activation in HG, either medially or laterally (Fig. 5). This was confirmed using the mean labeled volume of HG across subjects as a small volume within which we searched for significant activation in the contrast

of vowels (three conditions) vs. nonspeech (two conditions). This revealed only a single activated voxel, in the left hemisphere ( $-54, -8, 2, P = 0.031$  corrected for the small volume).

The reverse contrast, nonspeech (two conditions) vs. vowels (three conditions), yielded small regions of activation on left medial HG, the left superior temporal planum posterior to HG and the depths of the circular sulcus bilaterally (see Table 3).

We also compared the two synthetic-vowel conditions (damped vowels and raspy vowels) with the two nonspeech conditions to ensure that the activations observed for all three vowel conditions also held when the contrast was fully matched acoustically—that is, by removing the natural-vowel condition, which has no acoustically matched control condition, from the contrast. The results were essentially identical to the contrast with three vowel conditions (see Table 4). The reverse contrast, nonspeech vs. the two synthetic vowel conditions, revealed a single activation focus in left medial HG ( $-43, -18, -2, P = 0.002$  FDR whole-brain corrected level of significance), possibly as a result of rapid frequency transitions between cycles in the mp and mr stimuli (Zatorre and Belin, 2001).

#### Location of the activation specific to speech sounds

The left column of Fig. 5 shows three sections of the mean structural scan for the nine listeners in this study. The average location of HG in these subjects is shown in white. Activation maps of the main contrasts rendered on this brain are given in the figure in different colors. Activation specific to vowel sounds is clearly outside HG, inferior to the main area of activity in response to sound. Whereas this activation appears bilateral and symmetric in the group, it was not consistently bilateral in individual listeners, as shown in Fig. 6.

#### Main and simple effect of pitch

Griffiths et al. (2001) and Patterson et al. (2002) identified a bilateral region in lateral Heschl's gyrus that is sensitive to the presence of pitch in sound. This has been corroborated in MEG studies by Gutschalk et al. (2002) and Krumbholz et al. (2003), and also by Penagos et al. (2004) using fMRI. We examined the effect of pitch for the four conditions shown in Fig. 1, by comparing the activation for damped vowels plus musical rain with pitch with the activation for raspy vowels plus musical rain (without pitch). We employed a small-volume correction using a 10-mm sphere in the right and left hemisphere, centered on the peak pitch-sensitive voxels of Patterson et al. (2002); left:  $-55, -13, 2$ ; and right:  $57, -9, -2$ ) but no significantly activated voxels were observed.

The effect of pitch was, however, significant within these small-volume-corrected regions in a number of contrasts in which single conditions with pitch (including natural vowels, damped vowels, and musical rain with pitch) were compared to musical rain (without pitch); see Table 5. Natural vowels compared to musical rain yielded significant activation in the pitch region bilaterally; damped vowels compared to musical rain yielded significant pitch activation in the left hemisphere but not in the right. Comparing musical rain with pitch to musical rain without pitch yielded a significant activation in the pitch region in the right hemisphere, but only when a sphere of 5 mm around the activation focus reported by Patterson et al. (2002) (and not 10 as for the other contrasts) was used as the small volume. Interestingly, comparing either natural or damped vowels (with pitch) to pathological

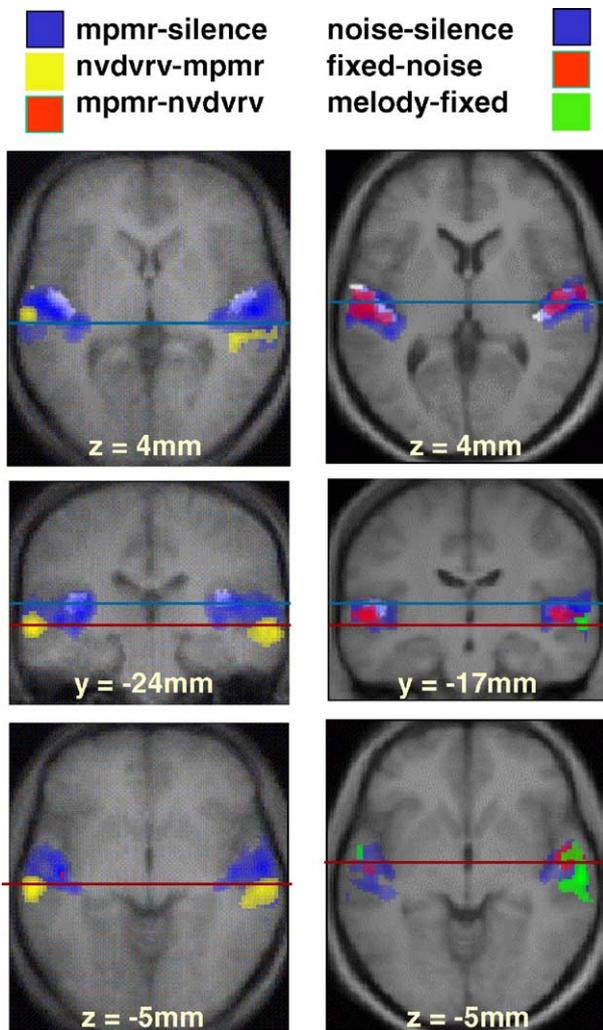


Fig. 5. (Left) Group activation in three contrasts using a fixed-effects model, rendered onto the average structural image of the group. Activation is shown at  $P < 0.05$  whole-brain corrected using a false-discovery-rate procedure. Blue: mp and mr (nonspeech) vs. sil. Yellow: vowel conditions (nv, dv and rv) compared to the two nonspeech conditions (mp, mr). Red: nonspeech compared to vowels. The white highlight shows the position of Heschl's gyrus (the average of the HG-labeled volumes for the nine individuals). (Right) For comparison, the location of activations observed in a previous study (Griffiths et al., 2001; Patterson et al., 2002) which were produced by all sound relative to silence (blue); sounds with pitch relative to noise (red); and melodies relative to fixed pitch (green).

Table 2

Vowel stimuli (natural vowels, nv; damped vowels, dv; and raspy vowels, rv) contrasted with nonspeech sounds (musical rain, mr; and musical rain with pitch, mp)

Location	<i>P</i> (FDR corrected)	Z score	<i>x</i>	<i>y</i>	<i>z</i>
Left STS	0	Inf	−66	−20	0
<i>Left posterior middle temporal gyrus</i>	<i>0.001</i>	<i>4.62</i>	<i>−54</i>	<i>−28</i>	<i>−2</i>
<i>Left STG</i>	<i>0.004</i>	<i>4.14</i>	<i>−66</i>	<i>−12</i>	<i>10</i>
Left posterior STG/supramarginal gyrus	0	5.09	−54	−44	24
Left anterior STG	0	5.03	−64	−4	4
Left precentral/premotor region	0.001	4.39	−50	−2	52
Left anterior STS	0.009	3.92	−60	−2	−8
Left postcentral gyrus	0.01	3.87	−48	−6	40
Left precentral/premotor region	0.018	3.72	−56	4	38
Left anterior STG	0.018	3.71	−64	4	2
Left STG	0.034	3.52	−54	−8	2
Right STS	0	Inf	66	−22	−2
<i>Right posterior STS</i>	<i>0</i>	<i>Inf</i>	<i>48</i>	<i>−34</i>	<i>4</i>
<i>Right posterior STS</i>	<i>0</i>	<i>Inf</i>	<i>60</i>	<i>−32</i>	<i>2</i>
Right precentral/premotor region	0	4.69	58	0	44
Right anterior STG	0.001	4.58	58	0	−4
<i>Right anterior STS</i>	<i>0.001</i>	<i>4.56</i>	<i>62</i>	<i>4</i>	<i>−10</i>
<i>Right anterior STS</i>	<i>0.027</i>	<i>3.58</i>	<i>62</i>	<i>8</i>	<i>−18</i>
Right anterior STS	0.01	3.89	58	14	−22
Right precentral/premotor region	0.012	3.83	58	10	40
Right globus pallidus	0.015	3.77	24	−4	0
Right posterior middle temporal gyrus	0.023	3.64	68	−50	18
Right posterior STS	0.023	3.63	44	−46	20

All significant peak activations more than 8 mm apart ( $P < 0.05$  corrected for multiple comparisons). Italics indicate subpeaks within a cluster. Abbreviations: STG—superior temporal gyrus; STS—superior temporal sulcus.

vowels (without pitch) did not yield significant activation in this pitch-sensitive region, again indicating that for these stimuli it was the perception of vowels per se, that was the dominant factor, and not the presence or absence of a pitch.

#### Interaction of speechlikeness and pitch

We also examined the brain for regions in which the pitch effect was greater for speech than for nonspeech stimuli, using the contrast (damped vowels – raspy vowels) – (musical rain with pitch – musical rain). This interaction analysis yielded no activation at a whole-brain corrected level of significance. As in the previous comparison, we used spheres of 10-mm radius around the peak voxels identified by Patterson et al. (2002) as a small-volume correction: this yielded significant activation in the pitch-sensitive region in the left hemisphere but not in the right (−54, −12, 12;  $P = 0.026$ , see Table 5).

#### Discussion

In this imaging study, volunteers heard natural vowels produced by a human speaker and four classes of synthetic sounds that were

closely matched to each other in terms of the distribution of energy over frequency and time. The two conditions with speech-like carrier frequencies (formants) were perceived as vowels. The two conditions perceived as nonspeech were produced from these synthesized vowels by simple manipulations in the time and frequency domains. When the speech conditions were compared to the nonspeech conditions, they were found to produce a pronounced region of differential activation in the superior temporal sulcus, bilaterally, midway along the temporal lobe in the anteroposterior direction, inferior to the location of Heschl's gyrus. It is difficult to say with certainty how similar the location of these activation foci in the two hemispheres is to those observed by others conducting speech–nonspeech contrasts. The STS is anatomically highly diverse, consisting of more than a dozen cytoarchitecturally and homologously distinguishable regions in the macaque monkey (e.g., Seltzer and Pandya, 1989a; Padberg et al., 2003). Although the applicability of this parcellation scheme to humans is highly uncertain and the microanatomical parcellation in humans is only now being established (Morosan et al., 2005), preliminary work by Morosan and colleagues demonstrates that, on a single coronal slice, at least 5 distinct cytoarchitectonic regions can be distinguished within the STS (Schleicher et al., 2005). This indicates that, as in nonhuman primates, the STS region in humans

Table 3

Nonspeech sounds (mp and mr) contrasted with vowel stimuli (nv, nvdv, rv)

Location	<i>P</i> (FDR corrected)	Z score	<i>x</i>	<i>y</i>	<i>z</i>
Left medial Heschl's gyrus (HG)	0.002	5.56	−46	−16	0
Left superior temporal plane posterior to HG	0.005	5.07	−36	−30	14
Left circular sulcus (depth)	0.035	4.37	−40	−22	−6
Right circular sulcus (depth)	0.006	4.95	38	−24	0

All significant peak activations more than 8 mm apart ( $P < 0.05$  corrected for multiple comparisons).

Table 4  
Synthetic vowel stimuli (dv, rv) contrasted with acoustically matched nonspeech sounds (mr, mp)

Location	<i>P</i> (FDR corrected)	Z score	<i>x</i>	<i>y</i>	<i>z</i>
Left STS	0	Inf	−64	−22	−2
<i>Left posterior middle temporal gyrus</i>	<i>0.009</i>	<i>4</i>	<i>−58</i>	<i>−32</i>	<i>0</i>
Left anterior STG	0	4.89	−64	−4	4
Left posterior STG/supramarginal gyrus	0.003	4.25	−52	−44	24
Left STG	0.011	3.93	−66	−12	12
Left precentral/premotor region	0.016	3.83	−48	−2	52
Left anterior STS	0.037	3.59	−58	−2	−6
Left precentral/premotor region	0.05	3.49	−56	4	38
Right STS	0	Inf	64	−24	−2
<i>Right posterior STS</i>	<i>0</i>	<i>7.72</i>	<i>60</i>	<i>−32</i>	<i>2</i>
<i>Right posterior STS</i>	<i>0</i>	<i>7.72</i>	<i>48</i>	<i>−34</i>	<i>4</i>
Right precentral/premotor region	0.001	4.57	58	0	44
Right anterior STS	0.01	3.97	64	4	−10
Right precentral/premotor region	0.021	3.75	58	10	40
Right anterior STG	0.038	3.58	56	−2	−2
Right precentral/premotor region	0.039	3.57	52	4	38
Right inferior temporal sulcus	0.046	3.51	54	−24	−16

All significant peak activations more than 8 mm apart ( $P < 0.05$  corrected for multiple comparisons). Italics indicate subpeaks within a cluster. Abbreviations: STG—superior temporal gyrus; STS—superior temporal sulcus.

is anatomically highly diverse, which makes comparison of activation foci across studies rather problematic given the spatial resolution of existing techniques (Price et al., 2005). Binder and colleagues (2000) calculated the center-of-mass location of peaks

obtained in their study and in three previous studies (Binder et al., 1997; Demonet et al., 1992; Zatorre et al., 1992). The average locus of activation across all four studies was within the STS bilaterally, but these peaks were about a centimeter medial to the

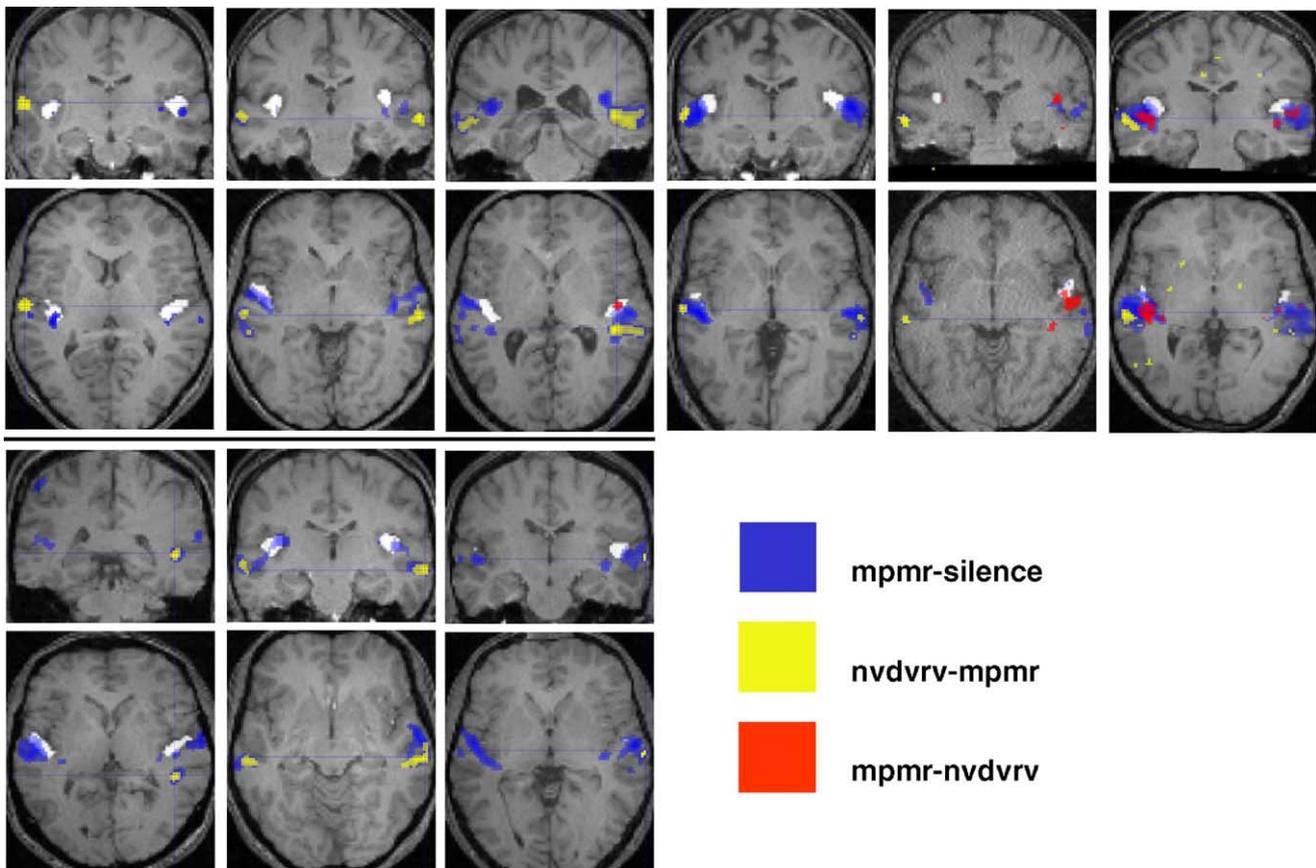


Fig. 6. Activation for three contrasts in individual listeners ( $n = 9$ ) rendered on sections of their individual structural images, shown in coronal (top) and horizontal (bottom) slices through the peak of the vowel-nonspeech activation in each case. The left hemisphere is on the left of each image. The labeled volumes of Heschl's gyri in each individual (see Materials and methods) are highlighted in white. Blue: mp and mr (nonspeech) vs. sil; yellow: vowel conditions (nv, dv, rv) compared to nonspeech conditions (mp, mr); red: nonspeech compared to vowels.

Table 5

Contrasts in which we searched for significant activation foci within a 10-mm radius of the peak voxels identified as sensitive to the presence of pitch by Patterson et al. (2002)

Contrast	Left					Right				
	<i>t</i>	<i>P</i>	<i>x</i>	<i>y</i>	<i>z</i>	<i>t</i>	<i>P</i>	<i>x</i>	<i>y</i>	<i>z</i>
<i>Main effect of pitch</i>										
(dv + mp) – (rv + mr)		ns					ns			
<i>Simple effects of pitch</i>										
nv – mr	3.56	0.008	–54	–10	0	4.24	0.004	58	0	–4
						3.97	0.011	64	–12	–2
dv – mr	4.73	0.001	–62	–18	2		ns			
nv – rv		ns					ns			
dv – rv		ns					ns			
mp – mr		ns				3.17 <sup>a</sup>	0.025	62	–10	2
<i>Interaction</i>										
(dv + mr) – (rv + mp)	3.71	0.026	–54	–12	12		ns			
(rv + mp) – (dv + mr)		ns					ns			

*t* values, associated *P* values and the *x*, *y*, *z* coordinates are listed when significant. dv: damped vowels; nv: natural vowels; rv: raspy vowels; mp: musical rain with pitch; mr: musical rain without pitch; ns not significant.

<sup>a</sup> 5-mm radius.

peaks reported here and thus plausibly in a separate region in the fundus of the sulcus.

In a previous study (Griffiths et al., 2001; Patterson et al., 2002), we observed a region on lateral Heschl's gyrus, most probably an auditory belt region, that is sensitive to pitch. However, in the current study, the comparison of the two synthetic-sound conditions with pitch and the two without pitch did not yield significant activation. This is probably because pitch did not vary throughout this experiment, either between successive vowels of one pitch condition or between pitch conditions. The left pitch region did show greater sensitivity to the difference between speech stimuli with and without pitch, when compared to nonspeech stimuli with and without pitch; this interaction probably merits further investigation, inasmuch as the left hemisphere is dominant for speech perception in most individuals (e.g., Knecht et al., 2000).

We also observed precentral/premotor regions in both left and right hemispheres that were slightly more active during perception of vowels than during perception of matched nonspeech sounds. This is consistent with several recent papers, indicating that passive listening to speech activates motor speech regions (Watkins et al., 2003; Wilson et al., 2004). Wilson and colleagues performed an fMRI study in which volunteers listened passively to monosyllables; activation in this condition overlapped with a region that was active when the volunteers produced those same speech sounds. The strongest premotor foci in the contrast of synthetic vowels vs. nonspeech in our study are within 6.7 mm (left) and 4.4 mm (right) of those observed by Wilson et al. (2004) for listening to monosyllables.

#### Location of the vowel-related activation relative to auditory cortex

The anatomical and functional organization of human auditory cortex is still poorly understood. Microelectrode recordings can be undertaken only in rare circumstances (e.g., during neurosurgery; Howard et al., 2000; Brugge et al., 2003), postmortem histological material is difficult to obtain (Hackett et al., 2001; Wallace et al., 2002), and in vivo tracer studies in humans are currently not possible. However, a 'working model' of auditory cortical

organization can be derived from work in the macaque (Kaas and Hackett, 2000; Kaas et al., 1999; Rauschecker, 1998), combined with the few studies in humans (Hackett et al., 2001; Liegeois-Chauvel et al., 1991; Howard et al., 2000; Morosan et al., 2001; Rademacher et al., 2001; Rivier and Clarke, 1997; Wallace et al., 2002; see Hall et al., 2003; Scott and Johnsrude, 2003, for reviews). This working model includes a 'core' of primary-like auditory cortices which are recipients of a dominant projection from the ventral medial geniculate complex of the thalamus (Rauschecker et al., 1997), and which are located on Heschl's gyrus. This is approximately encircled by a 'belt' of tissue receiving strong projections from the dorsal medial geniculate complex of the thalamus, from core areas and from other auditory belt regions. These belt areas project to adjacent parabelt zones which do not receive significant projections from auditory core. Belt and parabelt regions project to the dorsal bank of the superior temporal sulcus (STS), to other sites on the superior and middle temporal gyri and to prefrontal cortex (Kaas and Hackett, 2000; Kaas et al., 1999; Romanski et al., 1999). The strikingly hierarchical connections among these auditory regions suggest that these areas may support up to four discrete levels of processing between auditory core and the dorsal bank of the superior temporal sulcus (Kaas et al., 1999; Kaas and Hackett, 2000).

The activation that we observe in the superior temporal sulcus when vowels are compared to nonspeech sounds probably reflects the outputs of at least three previous stages of cortical processing in core, belt and parabelt regions. The STS in the macaque is anatomically highly heterogeneous, but the upper bank, continuing the length of the STS, is largely comprised of a region (area TAA) that receives its input mainly from the auditory areas of the superior temporal gyrus (Seltzer and Pandya, 1978, 1989a). This region projects into adjacent polysensory cortex in the fundus of the STS as well as to prefrontal cortex (Seltzer and Pandya, 1989a,b). Thus, the upper bank of the STS contains cortex that subserves the penultimate or final stage of unimodal auditory processing. The receptive field properties of neurons become more integrative and complex in higher-order regions (e.g., Rauschecker and Tian, 2004; Rauschecker et al., 1995; Tian et al., 2001; Wessinger et al., 2001). Within the working model, these sulcal

regions are good candidates for the centers that identify the characteristics of speech sounds. That is, this may well be where the system first identifies that the individual sounds have four formants below 4000 Hz, characteristic of the human vocal tract, and that the positions of the lower three vary from sound to sound in the manner peculiar to human speech. Differential activity for vowels over nonspeech sounds was not observed at an earlier cortical stage in this processing pathway, suggesting that sensitivity to the specific internal organization of vowel sounds, and the systematic vowel changes over the sequence, arises only after extensive processing. These results lead us to speculate that speechlikeness may not be a unitary attribute of a sound that is identified at an early stage of auditory processing in order for the sound to be accorded specialized processing in cortex; rather, speechlikeness may build up over stages of processing and be abstracted after the acoustic analysis is largely complete.

#### *Interpretation of the vowel-related activation*

It is possible to gain some insight into what the stages of processing might involve by considering the relationship of our speech stimuli to other common categories of sounds, and the grouping principles thought to be involved in source categorization and/or identification.

By far, the strongest grouping principle is common onset; channels that come on together are bound together in perception (Darwin and Carlyon, 1995). All of the stimuli in our study have common onset in the traditional sense. The millisecond delays of some of the formants in the first cycle, or burst, of musical rain are much, much smaller than the 100–200 ms delays that produce onset asynchrony effects (Darwin and Carlyon, 1995). Within each stimulus sequence (Fig. 3), the elements all come from one stimulus condition, and they all sound like they come from a single, distinctive source. This is partly because of the cadence of the sequence; the sounds come in sets of four regularly spaced bursts, with longer gaps between successive sets. All of the stimulus sequences would be expected to activate neural centers involved in trying to analyze and identify sound sources, and the common onset principle probably does not play a role in this experiment. We hypothesize that much of the processing in auditory cortex on the surface of the temporal lobe (Heschl's gyrus and planum temporale) is probably involved in basic feature extraction and source analysis, and that this is why the centers in this region are strongly activated, but not differentiated, by the stimuli in this experiment.

The stimuli do produce differential activation in the superior temporal sulcus, bilaterally, and we would argue that the contrasts suggest that the differential activation has to do with identifying the speech sequences. It is the case, that the natural vowels and damped vowels of the current experiment share two characteristics with animal calls and musical notes; they have a single stable pitch and a relatively fixed spectral envelope over durations of 100–400 ms. Thus, the contrast between these speech-sound conditions (nv and dv) and the nonspeech sounds (mp and mr) could represent processing that is common to animal calls and musical instruments as well as speech. However, the contrast between raspy vowels and the nonspeech sounds also reveals differential activation in the same region of STS, and this is crucial. Animals and musical instruments do not use 'whisper mode', although such sounds would appear to be acoustically possible. Only humans produce sequences of pitch-less stimuli with well defined spectral enve-

lopes similar to those of voiced vowels. It is this interaction that suggests that the activation in STS is related to processing of speech sounds.

The 'musical rain with pitch' stimuli (mp) not only have common onset, but they also have a form of harmonicity, and they produce a clearly identifiable temporal pitch. Harmonicity is an important grouping cue and a property of voiced speech, but it does not on its own result in differential activation in STS; its pattern of activation is much more similar to that of musical rain (mr). What the two nonspeech conditions share in acoustic terms is a lack of fixed formant tracks. The presence of fixed formant tracks would almost undoubtedly prove to be a grouping cue if experiments were performed on the effect of consistency of formant tracks, but it has not been investigated systematically.

Fixed formant tracks are characteristic of many animal calls and musical instruments; however, the formants are much broader and less well defined for animals and musical instruments than in vowels. This is what unites the voiced vowel stimuli and the raspy vowel stimuli, and the processing of this information is arguably what causes the differential activation in this region of the STS. These characteristics are largely restricted to human speech sounds and so this region is a candidate for the first speech-specific center in the brain.

#### **Summary**

This paper has identified regions of the brain that appear to be differentially involved in the processing of those stimulus characteristics that make sequences of vowels unique in the natural environment. With the aid of carefully controlled sets of stimuli, we have identified candidate areas that appear to be specifically involved in processing those properties of the acoustic signal that indicate whether or not the sound will be perceived as speech. The most active of these centers is in the superior temporal sulcus just lateral to and below Heschl's gyrus.

#### **Acknowledgments**

This work was supported by grants from the Medical Research Council (G9900369, G9901257, G0500221; authors RP and SU). The functional MRI was carried out at the Wolfson Brain Imaging Centre in Cambridge; we thank the radiographers for their assistance with data acquisition and Chris Wood for his help with data preprocessing.

#### **References**

- Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13367–13372.
- Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. *Nature* 403, 309–312.
- Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. *Cogn. Brain Res.* 13, 17–26.
- Benson, R.R., Whalen, D.H., Richardson, M., Swainson, B., Clark, V.P., Lai, S., Liberman, A.M., 2001. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain Lang.* 78, 364–396.

- Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., Prieto, T., 1997. Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* 17, 353–362.
- Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufmann, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. *Cereb. Cortex* 10, 512–528.
- Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., Ward, B.D., 2004. Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci.* 7, 295–301.
- Brett, M., Leff, A.P., Rorden, C., Ashburner, J., 2001. Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage* 14, 486–500.
- Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. *Nat. Rev., Neurosci.* 3, 243–249.
- Brugge, J.F., Volkov, I.O., Garell, P.C., Reale, R.A., Howard, M.A., 2003. Functional connections between auditory cortex on Heschl's gyrus and on the lateral superior temporal gyrus in humans. *J. Neurophysiol.* 90, 3750–3763.
- Callan, D.E., Jones, J.A., Callan, A.M., Akahane-Yamada, R., 2004. Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage* 22, 1182–1194.
- Crinion, J.T., Lambon-Ralph, M.A., Warburton, E.A., Howard, D., Wise, R.J.S., 2003. Temporal lobe regions engaged during normal speech comprehension. *Brain* 126, 1193–1201.
- Darwin, C.J., Carlyon, R.P., 1995. Auditory grouping. In: Moore, B.C.J. (Ed.), *Hearing*. Academic Press, London, pp. 387–424.
- David, H.A., 1988. *The Method of Paired Comparisons*, 2nd ed. Oxford Univ. Press, New York.
- Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. *J. Neurosci.* 23, 3423–3431.
- 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.
- Demonet, J.F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.L., Wise, R., Rascol, A., Frackowiak, R., 1992. The anatomy of phonological and semantic processing in normal subjects. *Brain* 115, 1753–1768.
- Edmister, W.B., Talavage, T.M., Ledden, P.J., Weisskoff, R.M., 1999. Improved auditory cortex imaging using clustered volume acquisitions. *Hum. Brain Mapp.* 7, 89–97.
- Eggermont, J.J., 2001. Between sound and perception: reviewing the search for a neural code. *Hear. Res.* 157, 1–42.
- Frisina, R.D., 2001. Subcortical neural coding mechanisms for auditory temporal processing. *Hear. Res.* 158, 1–27.
- Gandour, J., Xu, Y., Wong, D., Dzemidzic, M., Lowe, M., Li, X.J., Tong, Y.X., 2003. Neural correlates of segmental and tonal information in speech perception. *Hum. Brain Mapp.* 20, 185–200.
- Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* 15, 870–878.
- Giraud, A.L., Price, C.J., 2001. The constraints functional neuroimaging places on classical models of auditory word processing. *J. Cogn. Neurosci.* 13, 754–765.
- Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cereb. Cortex* 14, 247–255.
- Griffiths, T.D., Uppenkamp, S., Johnsrude, I., Josephs, O., Patterson, R.D., 2001. Encoding of the temporal regularity of sound in the human brainstem. *Nat. Neurosci.* 4, 633–637.
- Gutschalk, A., Patterson, R.D., Rupp, A., Uppenkamp, S., Scherg, M., 2002. Sustained magnetic fields reveal separate sites for sound level and temporal regularity in human auditory cortex. *NeuroImage* 15, 207–216.
- Hackett, T.A., Preuss, T.M., Kaas, J.H., 2001. Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. *J. Comp. Neurol.* 441, 197–222.
- Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. “Sparse” temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223.
- Hall, D.A., Summerfield, A.Q., Gonçalves, M.S., Foster, J.R., Palmer, A.R., Bowtell, R.W., 2000. Time-course of the auditory BOLD response to scanner noise. *Magn. Reson. Med.* 43, 601–606.
- Hall, D.A., Haggard, M.P., Summerfield, A.Q., Akeroyd, M.A., Palmer, A.R., Bowtell, R.W., 2001. Functional magnetic resonance imaging measurements of sound-level encoding in the absence of background scanner noise. *J. Acoust. Soc. Am.* 109, 1559–1570.
- Hall, D.A., Hart, H.C., Johnsrude, I.S., 2003. Relationships between human auditory cortical structure and function. *Audiol. Neuro-otol.* 8, 1–18.
- Howard, M.A., Volkov, I.O., Mirsky, R., Garell, P.C., Noh, M.D., Granner, M., Damasio, H., Steinschneider, M., Reale, R.A., Hind, J.E., Brugge, J.F., 2000. Auditory cortex on the human posterior superior temporal gyrus. *J. Comp. Neurol.* 416, 79–92.
- Hugdahl, K., Thomsen, T., Erslund, L., Rimol, L.M., Niemi, J., 2003. The effects of attention on speech perception: an fMRI study. *Brain Lang.* 85, 37–48.
- Irvine, D.R.F., 1992. Physiology of the auditory brainstem. In: Popper, A.N., Fay, R.R. (Eds.), *The Mammalian Auditory Pathway: Neurophysiology*. Springer, New York, pp. 153–231.
- Jancke, L., Wustenberg, T., Scheich, H., Heinze, H.J., 2002. Phonetic perception and the temporal cortex. *NeuroImage* 15, 733–746.
- Joanisse, M.F., Gati, J.S., 2003. Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. *NeuroImage* 19, 64–79.
- Kaas, J.H., Hackett, T.A., 2000. Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11793–11799.
- Kaas, J.H., Hackett, T.A., Tramo, M.J., 1999. Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* 9, 164–170.
- Knecht, S., Drager, B., Deppe, M., Bobe, L., Lohmann, H., Floel, A., Ringelstein, E.B., Henningsen, H., 2000. Handedness and hemispheric language dominance in healthy humans. *Brain* 123, 2512–2518.
- Krumbholz, K., Patterson, R.D., Seither-Preisler, A., Lammertmann, C., Lütkenhöner, B., 2003. Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. *Cereb. Cortex* 13, 765–772.
- Leonard, C.M., Puranik, C., Kuldau, J.M., Lombardino, L.J., 1998. Normal variation in the frequency and location of human auditory cortex landmarks. Heschl's gyrus: where is it? *Cereb. Cortex* 8, 397–406.
- Liebenthal, E., Binder, J.R., Piorkowski, R.L., Remez, R.E., 2003. Short-term reorganization of auditory analysis induced by phonetic experience. *J. Cogn. Neurosci.* 15, 549–558.
- Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631.
- Liegeois-Chauvel, C., Musolino, A., Chauvel, P., 1991. Localization of the primary auditory area in man. *Brain* 114, 139–153.
- Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage* 13, 684–701.
- Morosan, P., Schleicher, A., et al., 2005. Multimodal architectonic mapping of human superior temporal gyrus. *Anat. Embryol. (Berl.)*, 1–6.
- Mummery, C.J., Ashburner, J., Scott, S.K., Wise, R.J.S., 1999. Functional neuroimaging of speech perception in six normal and two aphasic subjects. *J. Acoust. Soc. Am.* 106, 449–457.
- Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. *Cereb. Cortex* 13, 1362–1368.
- Padberg, J., Seltzer, B., Cusick, C.G., 2003. Architectonics and cortical connections of the upper bank of the superior temporal sulcus in the rhesus monkey: an analysis in the tangential plane. *J. Comp. Neurol.* 467, 418–434.
- Palmer, A.R., Bullock, D.C., Chambers, J.D., 1998. A high-output, high-quality sound system for use in auditory fMRI. *NeuroImage* 7, S359.

- Patterson, R.D., 1994. The sound of a sinusoid: spectral models. *J. Acoust. Soc. Am.* 96, 1409–1418.
- Patterson, R.D., Uppenkamp, S., Norris, D., Marslen-Wilson, W., Johnsrude, I., Williams, E., 2000. Phonological processing in the auditory system: a new class of stimuli and advances in fMRI techniques. *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP) Beijing, China, Vol. II.*, pp. 1–4.
- Patterson, R.D., Uppenkamp, S., Johnsrude, I., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36, 767–776.
- Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *J. Neurosci.* 24, 6810–6815.
- Penhune, V.B., Zatorre, R.J., MacDonald, J.D., Evans, A.C., 1996. Interhemispheric anatomical differences in human primary auditory cortex: probabilistic mapping and volume measurement from magnetic resonance scans. *Cereb. Cortex* 6, 661–672.
- Poeppl, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., Braun, A.R., 2004. Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. *Neuropsychologia* 42, 183–200.
- Price, C., Thierry, G., Griffiths, T.D., 2005. Speech-specific auditory processing: where is it? *Trends Cogn. Sci.* 9, 271–276.
- Rademacher, J., Caviness, V.S., Steinmetz, H., Galaburda, A.M., 1993. Topographical variation of the human primary cortices and its relevance to brain mapping and neuroimaging studies. *Cereb. Cortex* 3, 313–329.
- Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H.J., Zilles, K., 2001. Probabilistic mapping and volume measurement of human primary auditory cortex. *NeuroImage* 13, 669–683.
- Rauschecker, J.P., 1998. Cortical processing of complex sounds. *Curr. Opin. Neurobiol.* 8, 516–521.
- Rauschecker, J.P., Tian, B., 2004. Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 91, 2578–2589.
- Rauschecker, J.P., Tian, B., Hauser, M., 1995. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114.
- Rauschecker, J.P., Tian, B., Pons, T., Mishkin, M., 1997. Serial and parallel processing in rhesus monkey auditory cortex. *J. Comp. Neurol.* 382, 89–103.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. *Science* 212, 947–950.
- Rivier, F., Clarke, S., 1997. Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *NeuroImage* 6, 288–304.
- Rimol, L.M., Specht, K., Weis, S., Savoy, R., Hugdahl, K., 2005. Processing of sub-syllabic speech units in the posterior temporal lobe: an fMRI study. *NeuroImage* 26, 1059–1067.
- Romanski, L.M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P.S., Rauschecker, J.P., 1999. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136.
- Schleicher, A., Palomero-Gallagher, N., et al., 2005. Quantitative architectural analysis: a new approach to cortical mapping. *Anat. Embryol. (Berl.)*, 1–14.
- Schlosser, M.J., Aoyagi, N., Fullbright, R.K., Gore, J.C., McCarthy, G., 1998. Functional MRI studies of auditory comprehension. *Hum. Brain Mapp.* 6, 1–13.
- Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* 26, 100–107.
- Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123, 2400–2406.
- Seltzer, B., Pandya, D.N., 1978. Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Res.* 149, 1–24.
- Seltzer, B., Pandya, D.N., 1989a. Intrinsic connections and architectonics of the superior temporal sulcus in the rhesus monkey. *J. Comp. Neurol.* 290, 451–471.
- Seltzer, B., Pandya, D.N., 1989b. Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. *J. Comp. Neurol.* 281, 97–113.
- Seltzer, B., Pandya, D.N., 1991. Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. *J. Comp. Neurol.* 312, 625–640.
- Specht, K., Reul, J., 2003. Functional segregation of the temporal lobes into highly differentiated subsystems for auditory perception: an auditory rapid event-related fMRI-task. *NeuroImage* 20, 1944–1954.
- Steinschneider, M., Fishman, Y.I., Arezzo, J.C., 2003. Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. *J. Acoust. Soc. Am.* 114, 307–321.
- Talairach, P., Tournoux, J., 1988. *Co-Planar Stereotaxic Atlas of the Human Brain*. Thieme, Stuttgart.
- Thierry, G., Giraud, A.L., Price, C., 2003. Hemispheric dissociation in access to the human semantic system. *Neuron* 38, 499–506.
- Tian, B., Reser, D., Durham, A., Kustov, A., Rauschecker, J.P., 2001. Functional specialization in rhesus monkey auditory cortex. *Science* 292, 290–293.
- Versnel, H., Shamma, S.A., 1998. Spectral-ripple representation of steady-state vowels in primary auditory cortex. *J. Acoust. Soc. Am.* 103, 2502–2514.
- Vouloumanos, A., Kiehl, K.A., Werker, J.F., Liddle, P.F., 2001. Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. *J. Cogn. Neurosci.* 13, 994–1005.
- Wallace, M.N., Johnston, P.N., Palmer, A.R., 2002. Histochemical identification of cortical areas in the auditory region of the human brain. *Exp. Brain Res.* 143, 499–508.
- Watkins, K.E., Strafella, A.P., Paus, T., 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994.
- Wessinger, C.M., van Meter, J., Tian, J., van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *J. Cogn. Neurosci.* 13, 1–7.
- Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702.
- Worsley, K.J., Evans, A.C., 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.
- Zatorre, R.J., Belin, P., 2001. Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 11, 946–953.
- Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A., 1992. Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256, 846–849.