Hierarchical Processing in Spoken Language Comprehension

Matthew H. Davis and Ingrid S. Johnsrude
Medical Research Council Cognition and Brain Sciences Unit, Cambridge, United Kingdom CB2 2EF

Understanding spoken language requires a complex series of processing stages to translate speech sounds into meaning. In this study, we use functional magnetic resonance imaging to explore the brain regions that are involved in spoken language comprehension, fractionating this system into sound-based and more abstract higher-level processes. We distorted English sentences in three acoustically different ways, applying each distortion to varying degrees to produce a range of intelligibility (quantified as the number of words that could be reported) and collected whole-brain echo-planar imaging data from 12 listeners using sparse imaging. The blood oxygenation level-dependent signal correlated with intelligibility along the superior and middle temporal gyri in the left hemisphere and in a less-extensive homologous area on the right, the left inferior frontal gyrus (LIFG), and the left hippocampus. Regions surrounding auditory cortex, bilaterally, were sensitive to intelligibility but also showed a differential response to the three forms of distortion, consistent with sound-form-based processes. More distant intelligibility-sensitive regions within the superior and middle temporal gyri, hippocampus, and LIFG were insensitive to the acoustic form of sentences, suggesting more abstract nonacoustic processes. The hierarchical organization suggested by these results is consistent with cognitive models and auditory processing in nonhuman primates. Areas that were particularly active for distorted speech conditions and, thus, might be involved in compensating for distortion, were found exclusively in the left hemisphere and partially overlapped with areas sensitive to intelligibility, perhaps reflecting attentional modulation of auditory and linguistic processes.

Key words: **speech; language; auditory cortex; hierarchical processing; primate; human; inferior frontal gyrus; temporal lobe; hippocampus; sentence processing; fMRI**

Introduction

Understanding spoken language is a rapid and seemingly automatic process. The translation of speech sounds (in our native language) into meaning is generally achieved without awareness of intervening processes, despite the background noise and inter-speaker variability that is characteristic of everyday speech. This robustness reflects the multiple acoustic means by which stable elements (such as phonetic features or syllables) are coded in clear speech; this redundancy permits comprehension when some acoustic information is lost. Robustness in speech comprehension may derive from the operation of compensatory mechanisms that are recruited when speech becomes difficult to understand, such as listening to loudspeaker announcements at a busy train station or a radio with poor reception. In this study, we use functional magnetic resonance imaging (fMRI) to explore the functional organization of brain regions involved in spoken language comprehension, with a view to understanding the neural basis for normal comprehension and processes that are recruited when speech becomes more difficult to understand.

Several different levels of representation (e.g., phonetic features, phonemes, morphemes, and words) have been proposed to mediate between an incoming speech signal and the computation of its meaning (McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997). Models of spoken language comprehension assume that processing is hierarchically organized, with greater abstraction from the surface (acoustic) properties of speech at higher processing levels. However, the degree to which higher-level linguistic processes can be distinguished from less-specialized auditory and sound-form-based processes remains unclear (Whalen and Liberman, 1987; Remez et al., 1994; Scott et al., 2000).

This hierarchy of processing stages may map onto auditory anatomy. The auditory cortex in primates comprises several cortical fields, organized into core (primary), belt (secondary), and parabelt regions. Anatomical and electrophysiological studies indicate that adjacent regions are interconnected and information proceeds from core, to belt, to parabelt, and to more distal areas as processing demands become more complex (for review, see Rauschecker, 1998; Kaas and Hackett, 2000). Neuroimaging studies have suggested a similar processing hierarchy in humans, but, to date, such studies have been limited to nonlinguistic stimuli (e.g., frequency-modulated tones or bandpassed noise) (Wessinger et al., 2001; Hall et al., 2002).

In this study, we alter (distort) the specific surface properties of speech in three different ways (see Fig. 1) and use a correlation design to relate brain activity to intelligibility. We operationalize "intelligibility" as the amount of a sentence that is understood, an aggregate measure of the multiple hierarchically organized processes involved in comprehension. Within areas that are sensitive to intelligibility, we can differentiate regions that are also sensitive to the type of distortion used (form-dependent) and, thus, probably involved in acoustic analysis, and those that are insensitive to

Received Sept. 5, 2002; revised Jan. 7, 2003; accepted Jan. 7, 2003.

This work was supported by the Medical Research Council (UK). We thank Paul Boersma and Chris Darwin for assistance with Praat scripts, Philip Dilks and Iain Turnbull for their help with the pilot study, the staff of the Wolfson Brain Imaging Centre, University of Cambridge, for their help with data acquisition, Matthew Brett and Ian Nimmo-Smith for advice on image processing and statistical analysis, and Brian Cox for his assistance with figures. We are also grateful to Daniel Bor, John Duncan, Stefan Kohler, William Marslen-Wilson, Dennis Norris, Sophie Scott, and anonymous reviewers for comments and suggestions. Example sounds can be found on the Internet at: www.mrc-cbu.cam.ac.uk/~matt.davis/jneurosci/.

Correspondence should be addressed to Matt Davis, Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, UK CB2 2EF. E-mail: matt.davis@mrc-cbu.cam.ac.uk.

 0270-6474/03/233423-09$15.00/0

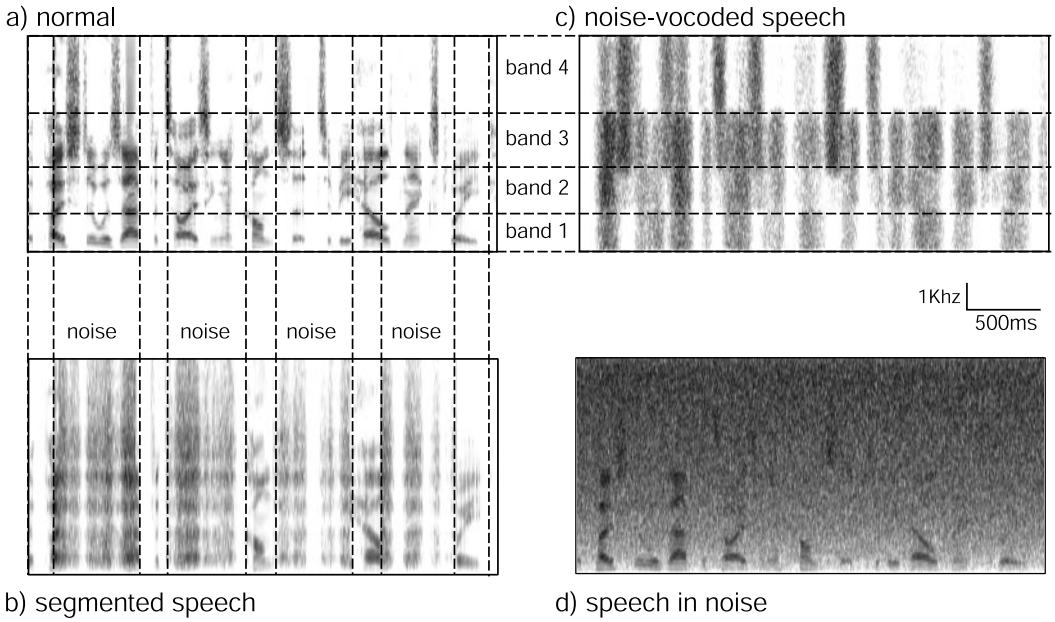

Figure 1. Spectrograms of sample experimental stimuli. *a*, The sentence "The poster was advertising a concert to be held next week" in undistorted normal form. *b*, Segmented with 500 msec noise bursts alternating with 200 msec of undistorted speech. *c*, Vocoded speech shown with four frequency bands. *d*, Speech in noise with a background of speech-spectrum noise at a signal-to-noise ratio of −1 dB.

distortion type (form-independent); these areas may be involved in higher-level linguistic processes.

We can also identify the neural correlates of effortful understanding by contrasting the neural response to distorted (yet still intelligible) sentences for which comprehension is difficult with conditions that involve less effort (cf. Poldrack et al., 2001). Activation in this contrast may reflect the action of compensatory mechanisms that modulate activity at an acoustic level or at higher levels of processing.

Materials and Methods

Stimulus preparation

There were 190 declarative English sentences ranging in topics, comprising 5–17 words (1.7–4.3 sec in duration), and digitized at a sampling rate of 22.1 kHz taken from the test and filler sentences used in a previous behavioral study (Fig. 1*a*) (Davis et al., 2002). Three forms of distortion were applied to these sentences using Praat software (Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, The Netherlands) (available at www.praat.org). All three forms of distortion preserved the duration, amplitude, and average spectral composition of the original sentences, although the acoustic form of sentences processed with the three types of distortion were markedly different.

Segmented. Segmented speech was created by dividing the speech waveform into short chunks at fixed intervals and replacing even-numbered chunks of speech with a signal-correlated noise (SCN) version of the original speech (Bashford et al., 1996). Signal-correlated noise is a waveform with the same spectral profile and amplitude envelope as the original speech but consisting entirely of noise. Although it retains some physical properties of the speech that it replaces (e.g., a speech-like rhythmical structure), these periods of signal-correlated noise do not contain any intelligible speech sounds (Schroeder, 1968). The duration of clear speech was fixed at 200 msec and 500, 200, or 100 msec sections of speech were replaced by signal-correlated noise (Fig. 1*b*).

Vocoded. Noise-vocoded speech (Shannon et al., 1995) was created by dividing the speech signal between 50 and 8000 Hz into 4, 7, or 15 bandpass-filtered frequency bands. Sentences were resynthesized by replacing information in each frequency band with amplitude-modulated bandpass noise (Fig. 1*c*). Frequency bands were approximately lineally spaced (i.e., the width of each band was proportional to the center frequency of that band). Noise vocoded speech sounds like a harsh robotic whisper.

Noise. Speech in noise was generated by adding a continuous speech-spectrum noise background to sentences at three signal-to-noise ratios (−1, −4, or −6 dB) (Fig. 1*d*). The overall amplitude of each speech-in-noise stimulus was reduced to match the amplitude of the original sentence.

In addition to these three forms of distortion, a signal-correlated noise baseline was generated using the same algorithm as that for segmented speech but without periods of clear speech. Sentences processed in this way sound like a rhythmic sequence of noise bursts, carry no linguistic information, and are entirely unintelligible (Schroeder, 1968).

Pilot study

A pilot behavioral study was conducted to ensure that a continuum of intelligibility was obtained for each form of distortion. Eighteen native English speakers heard single-stimulus sentences over closed-ear headphones (model DT770; Beyerdynamic, West Sussex, UK) played from the soundcard of a Dell laptop PC (Dell Computer Company, Round Rock, TX). Participants were required to either type as many words as they could understand or to rate intelligibility (on a nine-point scale) immediately after each item. Sentences were pseudorandomly assigned to a type and level of distortion (three versions of the test were created with the same sentences assigned to different conditions). Each subject was tested on one version of this behavioral study and therefore heard each sentence only once. Word-report performance (calculated as the proportion of words per sentence that were reported correctly) and rated intelligibility were averaged over five items per condition per subject. A total of six levels of intelligibility were tested for each form of distortion. For the 19 conditions tested (six levels of three types of distortion and clear speech), word-report scores and rated intelligibility were reliably correlated ($r = 0.99$; $p < 0.001$).

We selected three levels of each form of distortion described above: a low-intelligibility condition (~20% of words reported correctly), a medium-intelligibility condition (~65% of words reported correctly), and a high-intelligibility condition (~90% of words reported correctly) (Fig. 2). ANOVA comparing intelligibility ratings showed no significant difference between the three types of distortion at each level of intelligibility (low intelligibility, $F_{(2,34)} = 1.58$, $p > 0.1$; medium and high intelligibility, both F values < 1). However, for word-report scores, some differences between types of distortions were reliable at each level of intelligibility. For low intelligibililty ($F_{(2,34)} = 8.75$; $p < 0.001$) pairwise comparisons indicated significantly reduced intelligibility for vocoded speech, medium intelligibility($F_{(2,34)} = 4.35$; $p < 0.05$), with increased

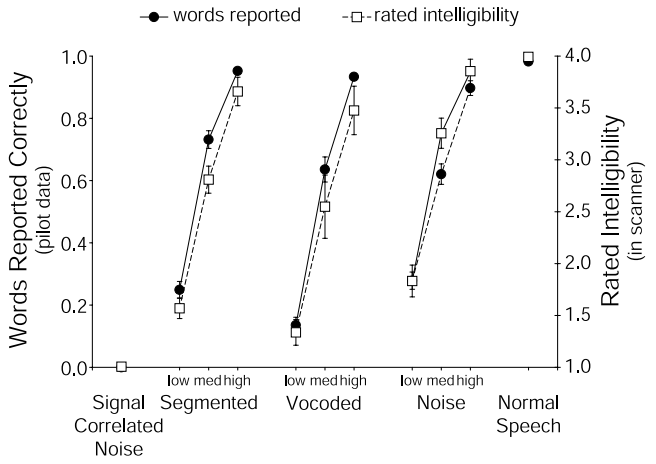

Figure 2. Word-report scores from the pilot study and subject ratings during scanning for the 11 different kinds of stimuli. Each *point* represents the average (and SE) for 18 subjects in the pilot study (percentage of words reported correctly) or eight subjects in the fMRI study (fMRI ratings). Correlation between word report and ratings was $r = 0.98$; $p < 0.001$.

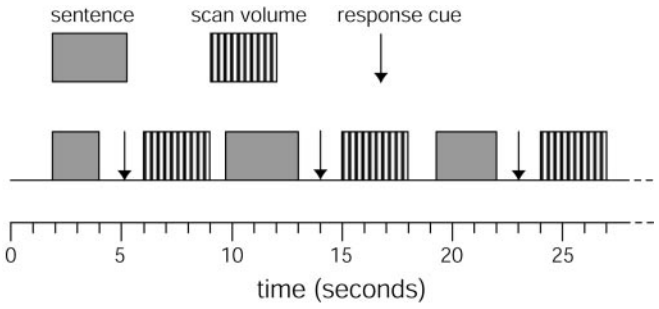

Figure 3. Details of the scanning procedure. A sparse imaging technique was used (see Materials and Methods) in which a single stimulus item was presented in the silent periods between scans. A tone pip after each sentence cued the subject's intelligibility judgment (button press). Timings of sentence onset and offset and tone cue relative to scan onset were jittered across trials.

intelligibility for segmented speech, and for high intelligibility stimuli ($F_{(2,34)} = 3.76$; $p < 0.05$) marginally reduced intelligibility for speech in noise.

Scanning procedure

Twelve right-handed volunteers (five females) between 18 and 42 years of age were scanned. All subjects were native speakers of English, without any history of neurological illness, head injury, or hearing impairment. This study was approved by the Addenbrooke's Local Research Ethics Committee (Cambridge, UK), and written informed consent was obtained from all subjects. Volunteers were told that they would be listening to sentences that were distorted with different amounts and types of noise and were asked to rate the intelligibility of each item using a four-alternative button press with their right hand. The alternatives ranged from understanding most or all of the sentence (index finger; button 4) to none or not very much of the sentence (little finger; button 1). Volunteers were given a short period of practice in the scanner with a different set of sentences that were processed in the same way as the experimental items.

We acquired imaging data using a Medspec (Bruker, Ettlingen, Germany) 3 tesla MRI system with a head gradient set. Echo-planar imaging (EPI) volumes (228 in total) were acquired over two 17 min sessions. Each volume consisted of 21 × 4 mm thick slices with an interslice gap of 1 mm; field of view, 25 × 25 cm; matrix size, 128 × 128; echo time, 27 msec; acquisition time, 3.02 sec; and actual repetition time, 9 sec. We used a sparse imaging technique in which stimuli are presented in the silent period between successive scans, minimizing acoustic interference (Edmister et al., 1999; Hall et al., 1999). 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 temporal cortex, and the inferior aspect of the cerebellum).

Two scanning sessions of 114 trials were performed. Each trial comprised a stimulus item followed by a tone pip and a single EPI volume (Fig. 3). Stimulus items in other trials were pseudorandomly drawn from the 11 experimental conditions (low-, medium-, and high- intelligibility conditions for each of three forms of distortion, plus signal-correlated noise and clear speech). There were 19 trials of each stimulus type and an additional 19 silent trials. Stimulus onset and offset were jittered relative to scan onset by temporally aligning the midpoint of the stimulus item (0.8–2.1 sec after sentence onset) with the midpoint of the gap between scans (6 sec), thus ensuring that scans were obtained 3–6 sec after stimulation. This coincided with the peak of the hemodynamic response evoked by the stimulus (Edmister et al., 1999; Hall et al., 1999). The tone pip occurred 1 sec after stimulus offset (or at a matched position in silent trials) and cued the subject to rate the intelligibility of the item just presented [or a self-determined (random) button press for silent scans]. Items from the 190-sentence corpus were pseudorandomly assigned to different distortion conditions using three different forms of randomization. Sentences presented as SCN were chosen from the other 12 conditions; no other items were presented more than once in the experiment. Stimuli were presented diotically, using a high-fidelity auditory stimulus-delivery system incorporating flat-response electrostatic headphones inserted into sound-attenuating ear defenders (Palmer et al., 1998). To further attenuate scanner noise, participants wore insert earplugs (E.A.R. Supersoft; Aearo Company, Indianapolis, IN) rated to attenuate by ~30 dB. When wearing earplugs and ear defenders, participants reported that the scanner noise was unobtrusive and that sentences 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 stimulus items, and DMDX (Forster and Forster, 2003) was used to record button-press responses.

Analysis of fMRI data

Data processing and analysis was accomplished using Statistical Parametric Mapping (SPM99; Wellcome Department of Cognitive Neurology, London, UK). Preprocessing steps included within-subject realignment, spatial normalization of the functional images to a standard EPI template (masking regions of susceptibility artifact to reduce tissue distortion) (Brett et al., 2001), and spatial smoothing using a Gaussian kernel of 12 mm, suitable for random-effects analysis (Xiong et al., 2000).

We were interested, first of all, in identifying areas in which activation correlated with intelligibility (see Fig. 4*a*). Within these intelligibility-sensitive areas, we then wanted to differentiate between areas of form dependence (activation that was sensitive to the acoustic form of the stimulus, as shown in Fig. 4*b*) and areas of form independence (areas that responded equivalently to the different forms of distortion). This distinction might plausibly separate areas involved in lower-level acoustic processes from higher-order linguistic levels of processing. In addition, we thought it would be informative to establish the overlap, if any, between intelligibility-sensitive form-dependent areas and primary-like cortical auditory areas. Such cortical auditory areas were identified as those exhibiting elevated response to signal-correlated noise over silence. Some spatial segregation of the two response types might indicate a hierarchy of processing within auditory cortices as stimulus characteristics become more complex, such as has been observed in the macaque (Rauschecker et al., 1995; Rauschecker, 1998).

We also wanted to identify brain areas exhibiting increased response to degraded speech stimuli, over and above any correlation with intelligibility. We hypothesized that, when speech is difficult to understand (i.e., when speech is distorted yet still potentially intelligible), listeners will make additional effort to extract as much meaning as possible. This might be reflected in an increased brain response to distorted speech compared with clear speech (which, with sparse imaging, a high-fidelity sound delivery system, and comfortable listening volume, was presented under near-ideal conditions for effortless comprehension). Because this elevated response for more distorted speech could also arise from pro-

cesses that are recruited as the auditory stimulus becomes less comprehensible and therefore less engaging (cf. Stark and Squire, 2001), we also compared activation for distorted speech with signal-correlated noise, which, as is immediately evident to the subjects, is not speech (see Fig. 4*c*). An elevated response to the distorted conditions, over normal speech and SCN, would be consistent with mechanisms acting to enhance comprehension for potentially intelligible input. Such mechanisms may be domain general (e.g., attentional modulation of auditory processing) or more specific to speech processing. Overlap with other kinds of response would be informative in this regard; a distortion-elevated response that overlapped with sensitivity to distortion type might indicate compensation acting on an acoustic level (consistent with attentional facilitation). In contrast, areas exhibiting an elevated response to distorted input and sensitivity to intelligibility, independent of distortion type, might indicate that distortion places additional "load" on higher-level linguistic processes (these alternatives are presented in more detail in Discussion).

Two separate design matrices were constructed for each listener to optimize sensitivity to our effects of interest. The primary analysis included two columns indicating both the presentation of a sentence before each scan (as opposed to a silent period) and the mean proportion of words reported correctly for that type and level of distortion in the behavioral pilot. [We used word-report scores as a covariate for two reasons: (1) to compensate for the small but significant differences in the intelligibility of the three types of distortion identified in the pilot study and (2) report scores provide a more objective measure than the ratings obtained during scanning, and the two measures are highly correlated (see Fig. 2).] Three additional columns were included in the design matrix that coded which of the three types of distortion was applied to each sentence (scans following signal-correlated noise and clear-speech stimuli were modeled only in the first two columns). Realignment parameters and a dummy variable coding the two scanning sessions were included as covariates of no interest, and a correction for global signal magnitude was made. The second design matrix was used to evaluate activation for signal-correlated noise sentences compared with silence and to obtain signal change estimates for each condition (see Figs. 5*f*, 7*d*); it included a single column for each of the twelve conditions in the experiment and covariates of no interest as before.

The parameter estimates for each subject, derived from the least-mean-squares fit of these models, were entered into second-level group analyses in which *t* values were calculated for each voxel, treating inter-subject variation as a random effect. For main effects of intelligibility and compensation for distortion, we report activation foci that survive a whole-brain false discovery rate (FDR) (Genovese et al., 2002) correction at $p < 0.05$. This procedure controls the expected proportion of false positives among suprathreshold voxels to the specified rate (0.05). Where the null hypothesis is true (i.e., there are no activated voxels), the FDR procedure produces identical results to a Bonferroni correction, providing stringent control of familywise-error rate (Benjamini and Hochberg, 1995). Contrasts that were used to evaluate sensitivity to acoustic form (i.e., detecting differences between the three forms of distortion) were applied over the whole brain (with appropriate correction) and within regions of interest defined by the areas revealed as active by the intelligibility and compensation-for-distortion contrasts.

Results

Behavioral data were not available from four subjects in the fMRI study. Mean four-point ratings from the remaining eight subjects correlated highly ($r = 0.98$; $p < 0.001$) with report scores from the 18 subjects in the pilot study (Fig. 2). (Report scores of zero were assumed for signal-correlated noise in the pilot study.) Given the greater accuracy and consistency of the report scores compared with the ratings obtained during scanning, we used these as regressors in the analysis of the fMRI data. Intelligibility-sensitive areas were identified as those voxels in which blood oxygenation level-dependent (BOLD) signal was correlated with word-report scores (Fig. 4*a*).

The comparison of SCN versus silence across subjects yielded activation bilaterally, in Heschl's gyrus and surrounding areas, consistent with recruitment of core and belt auditory cortex, as predicted (Fig. 5*a*,*b*, *pink-blue intensity scale*).

Correlation with intelligibility

The BOLD signal was positively correlated with word-report score in voxels along the length of the superior and middle temporal gyri in the left hemisphere, extending outward from auditory cortex toward the temporal pole and the temporoparietal junction (Fig. 5*a*,*b*, *red* and *yellow scale*). Similar less-extensive activation was observed in the right superior and middle temporal gyri. A portion of left inferior frontal gyrus also showed a positive correlation with intelligibility, as did the body of the left hippocampal complex (Fig. 5*d*). Within the superior temporal gyri, a correlation with intelligibility was observed in areas adjacent to those activated in the SCN–silence contrast (Fig. 5*a*,*b*).

To test for brain areas sensitive to differences between the three forms of distortion, the intelligibility-responsive region was masked by all six possible contrasts between pairs of the three distortion types (Fig. 4*b*). Setting the threshold for each of these six contrasts to $p = 0.00851$ results in a combined α level of $p < 0.05$ [because, by binomial expansion, $0.95 = (1 - 0.00851)^6$]. Intelligibility-responsive areas in which none of these contrasts reach significance at $p < 0.00851$ can therefore be considered to be form independent (i.e., insensitive to differences between

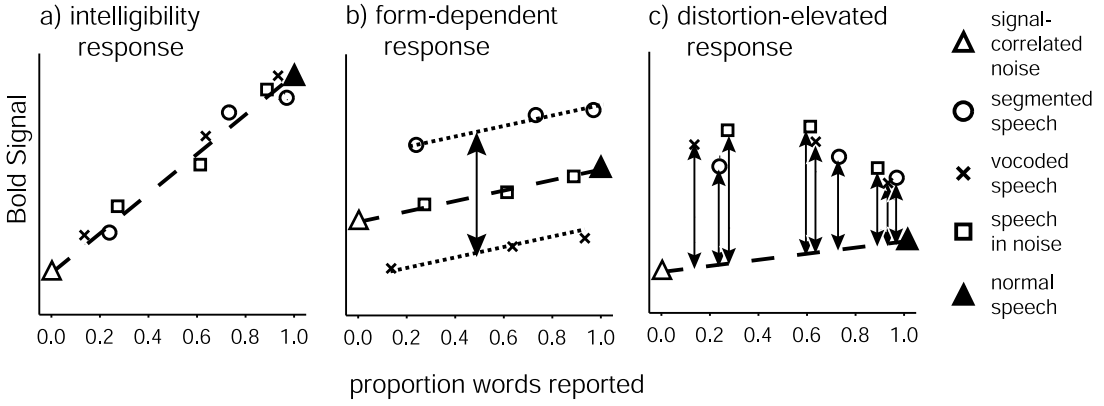

Figure 4. *a–c*, Predicted BOLD signal for three contrasts: linear correlation between BOLD signal and intelligibility (*a*), a differential response to the three forms of distortion (*b*), and an elevated response to all forms of distorted speech (*c*), compared with clear speech and signal-correlated noise. *Open triangles*, Signal-correlated noise; *circles*, segmented speech; *x symbols*, vocoded speech; *squares*, speech in noise; *filled triangles*, normal speech.

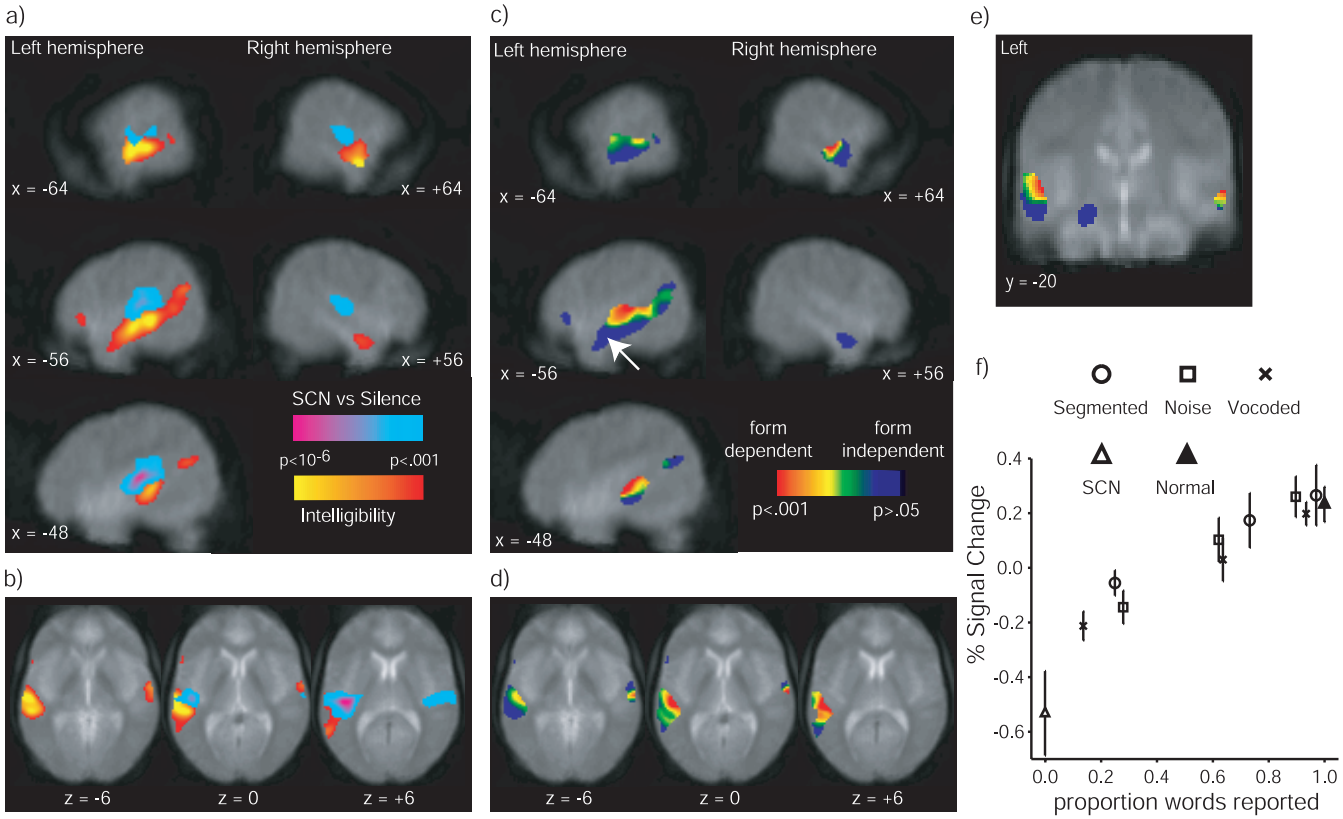

Figure 5. Areas showing a significant linear response to increasing intelligibility. Activations are shown superimposed on the mean EPI image across subjects and thresholded at $p < 0.001$, uncorrected for multiple comparisons. *a*, Sagittal sections depicting areas in which activation was observed for signal-correlated noise relative to rest (*pink-blue scale*) and areas in which activation correlates with report score (intelligibility response; *yellow-red scale*). *b*, Axial sections through Heschl's gyrus. Activation to sound is predominant on Heschl's gyrus, whereas correlation with intelligibility is observed posteriorly, inferiorly, and laterally to Heschl's gyrus. *c*, The intelligibility response pattern shown in *a* but with a mask to show voxels that are sensitive to the acoustic properties of the stimulus (dependent on distortion type). Because the identification of form-independent regions depends on the absence of any reliable difference between distortions, we cannot precisely localize the transition between form-dependent and form-independent regions and so present the mask in *graded color scale*. The 95% contour of the mask (form dependence; $p < 0.05$) corresponds to the boundary between *blue* and *green*; the 99.9% contour (form dependence at $p < 0.001$) corresponds to the boundary between *orange* and *red*. Activation foci related to intelligibility and common to all forms of distortion (form independent) are shown in *blue* and listed in Table 1 (top). Arrow indicates the approximate location of the form-independent left anterior temporal lobe voxel from which data are plotted in *f*. *d*, Axial sections through Heschl's gyrus. Extending posteriorly and laterally from primary auditory cortex into belt and parabelt cortices, form-dependent (*graded color*) intelligibility responses give way to form-independent (*blue*) responses. *e*, Coronal section ($y = -20$) depicting a form-independent intelligibility response in left hippocampus. *f*, The *graph* shows the size of the response (percentage of signal change from the mean) for a form-independent voxel (→) in the left anterior temporal lobe (−58, −2, −24) against word-report score for the different listening conditions. *Open triangle*, Signal-correlated noise; *circles*, segmented speech; *x symbols*, vocoded speech; *squares*, speech in noise; *filled triangle*, normal speech. Error bars indicate SEM across subjects.

types of distortion) at an uncorrected $p > 0.05$ (Fig. 5*c*,*d*,*e*, *blue*; Table 1, top). A form-independent correlation with intelligibility was observed in the anterior middle temporal gyrus bilaterally and in posterior superior temporal sulcus, inferior frontal gyrus, hippocampus, and precuneus in the left hemisphere (for a typical response profile, see Fig. 5*f*).

By increasing the statistical threshold for the form-dependent contrasts to a more stringent significance level ($p < 0.001$ uncorrected, corresponding to $p < 1.67 \times 10^{-4}$ in each of the six contrasts, or to an FDR corrected $p < 0.00851$ within the region of interest), we identified two areas in which activation not only reflects a response to intelligibility but also shows form dependence at a corrected level of significance. This contrast identifies areas that are engaged in spoken language comprehension (as shown by the significant correlation between activation and intelligibility) but also shows differential activation depending on the acoustic form(s) of distorted speech presented. Such a response was observed bilaterally in the superior temporal gyrus (Fig. 5*c*,*d*,*e*, *red*).

To explore the nature of the sensitivity to acoustic form observed in the form-dependent regions, an estimate of the effect of different distortion types was calculated for peak voxels in the left and right superior temporal gyrus (Fig. 6*a*,*b*; see figure legend for coordinates). In both hemispheres, this difference arises through an elevated response to segmented speech compared with the other two forms of distortion, particularly to speech in noise.

Compensation for distortion

We reasoned that mechanisms involved in compensating for distortion are not required in either the clear-speech condition or the signal-correlated-noise condition, and we identified relevant areas by comparing activation for potentially intelligible conditions with both fully intelligible and completely unintelligible conditions. In addition, because differences in intelligibility between conditions were included in the model, elevated activity for distorted speech is statistically independent of a response that is correlated with intelligibility. A distortion-elevated response was observed in left-hemisphere areas, including the middle and superior temporal gyri, the opercular part of the inferior frontal gyrus, the lateral inferior frontal gyrus, the posterior middle frontal gyrus (premotor cortex), and the ventral anterior nucleus of the thalamus (Fig. 7*a*; Table 1, bottom). Both a distortion-

Table 1. fMRI activations

Region	Coordinates			Z-score
	x	*y*	*z*	
Form-independent correlates of activity with intelligibility				
R MTG	64	−8	−16	5.50*
L MTG	−60	−34	−2	5.20*
L Hippocampus	−22	−20	−12	4.02*
L post MTG	−62	−56	6	3.79*
L angular gyrus	−54	−60	22	3.71*
L frontal operculum	−58	16	−2	3.35*
L SFS	−10	56	30	3.26*
L Precuneus	−8	−50	30	3.26*
Form-independent activity increases for distorted speech relative to normal speech and signal-correlated noise				
Vathal	−12	−6	10	4.25*
L frontal operculum	−48	14	−6	4.22*
L MTG	−66	−20	−4	3.72*
L posterior STP	−52	−46	16	3.54*
L orbitofrontal	−32	52	−4	3.40
R intraparietal sulcus	40	−46	38	3.39
L MFG (premotor)	−40	2	44	3.35
Cingulate gyrus	−2	16	46	3.29
L orbitofrontal	−36	42	−14	3.23
L anterior STG	−58	−8	−6	3.20
L posterior STG	−56	−54	8	3.15
R MFG (premotor)	46	8	32	3.14
L substantia nigra	−6	−20	−16	3.12

We present coordinates of activation foci together with Z-scores and an estimate of location relative to gross anatomy for each contrast of interest. MFG, Middle frontal gyrus; MTG, middle temporal gyrus; SFS, superior frontal sulcus; STP, superior temporal planum; STG, superior temporal gyrus; Vathal, ventral anterior thalamic nucleus; R, right; L, left. All peak voxels exceeding $p < 0.001$ are reported. Voxels marked with an asterisk reach whole-brain FDR correction at $p = 0.05$. Note that, because the FDR correction is an adaptive procedure, statistical thresholds are at different values for the two contrasts.

elevated response and a correlation with intelligibility was observed in the left temporal cortex and left frontal operculum.

As for the correlation with intelligibility, the distortion-elevated response was masked with a combined map showing sensitivity to different forms of distortion. Regions surviving this exclusive masking procedure (at $p > 0.00851$, uncorrected for each of six contrasts, equivalent to a combined $p > 0.05$) are considered to be insensitive to acoustic form (Fig. 7*b*,*c*, *blue*; for the response profile in a typical voxel, see *d*). This analysis also revealed that a subset of the areas showing a distortion-elevated response was also sensitive to acoustic form (Fig. 7*b*,*c*, *red*). Interestingly, the distortion-elevated response in the temporal lobe was primarily form dependent, except for its most posterior, anterior, and inferior aspects. As discussed before (and shown in Fig. 6*a*), this form-dependent response arises from an elevated response to segmented speech. In contrast, most of the frontal-lobe distortion-elevated activation was form independent, except for a region in the frontal operculum, just lateral to the insula. This region exhibited a significantly elevated response for vocoded speech compared with segmented speech (Fig. 6*c*).

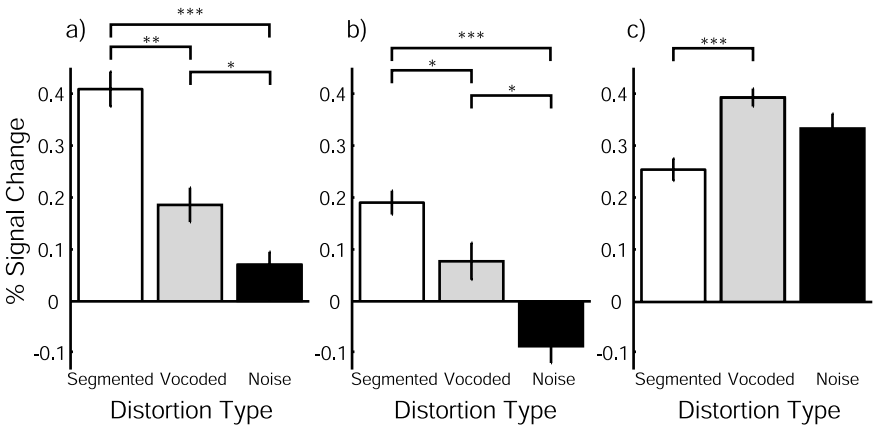

Figure 6. BOLD signal in peak voxels showing a form-dependent response. *Graphs* show mean percentage signal change for each distortion type compared with the mean response to signal-correlated noise and normal speech. Error bars indicate SEM after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (cf. Loftus and Masson, 1994). *Braces* show significance of paired *t* tests comparing the three distortion types (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). *a*, Form-dependent response in left superior temporal gyrus (−52, −28, 6). *b*, Right superior temporal gyrus (66, −16, 0). *c*, Left inferior frontal gyrus (−42, 20, −6).

Discussion

Our observation of intelligibility-sensitive regions in the lateral temporal lobe replicates and extends the findings of previous functional imaging studies (Binder et al., 2000; Scott et al., 2000; Vouloumanos et al., 2001). These studies used subtractive designs; regions that were active for intelligible speech were identified by comparison with nonspeech baseline conditions that may not be of equivalent acoustic structure or complexity. Although Scott et al. (2000) used more than one form of intelligible speech with a well controlled baseline, their design did not permit them to fractionate areas that respond to speech intelligibility into regions that are involved in low-level acoustic and higher-level nonacoustic processing. Our correlational design reduces the importance of subtractions from baseline conditions. Comparing among different forms of distortion allows us to test for acoustic processes within intelligibility-responsive areas.

Sound (compared with silence) produced activation in the probable location of primary auditory cortex (Rivier and Clarke, 1997; Morosan et al., 2001; Rademacher et al., 2001). Importantly, activation in primary auditory cortex did not correlate reliably with intelligibility; instead, the bilateral temporal-lobe region in which activation correlated with intelligibility is adjacent to primary auditory cortex. The form-dependent portion of this intelligibility-sensitive region includes both auditory belt and parabelt areas (and beyond) and, therefore, probably more than one processing stage (Rauschecker et al., 1995; Rauschecker, 1998; Kaas and Hackett, 2000), although these data cannot speak to further functional segregation.

The form-dependent profile observed in these periauditory areas arises from an increased response to segmented speech compared with other forms of distortion that are matched on intelligibility. This may reflect differential sensitivity of neurons to particular acoustic features of speech. For instance, neurons sensitive to rapid spectral changes or formant transitions that are present in clear speech might respond more strongly to segmented speech. Alternatively, neurons that are sensitive to transitions between periodic and aperiodic sounds might respond more strongly to segmented speech, because these transitions are absent from the other forms of distortion.

Surrounding this periauditory form-dependent region anteriorly, posteriorly, and inferolaterally, we observed areas in which activation correlated significantly with intelligibility but was insensitive to acoustic differences among types of distortion. We conclude that these form-independent areas are involved in process-

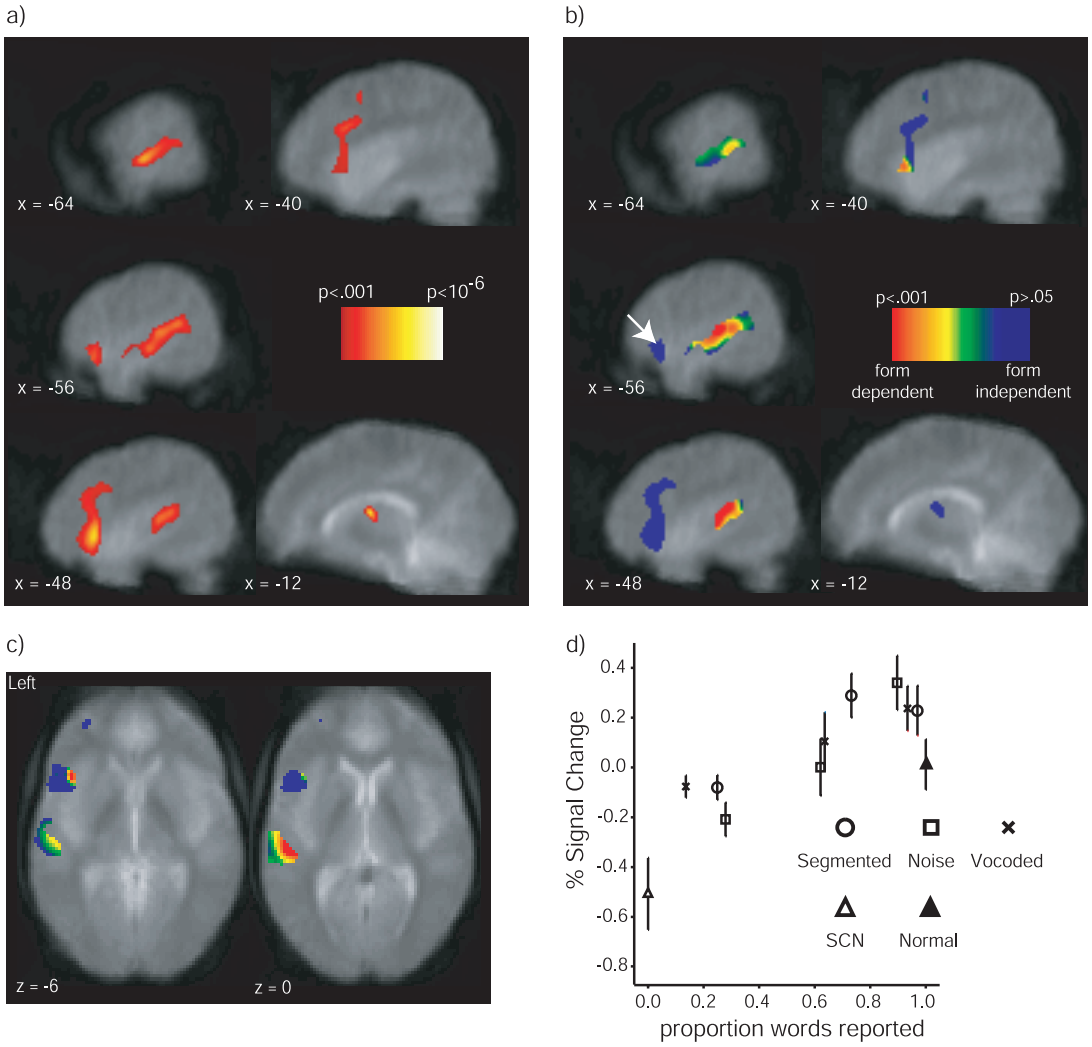

Figure 7. Areas showing an elevated response to distortion. Activations are shown superimposed on the mean EPI image across subjects and thresholded at $p < 0.001$, uncorrected for multiple comparisons. *a*, Sagittal sections depicting areas in which activation is significantly elevated for distorted speech relative to clear speech and signal-correlated noise and covaried for intelligibility (distortion-elevated response). *b*, The activation pattern shown in *a* but with a mask to show voxels that are sensitive to the acoustic properties of the stimulus (showing a distortion-elevated response that is form dependent). The mask is shown in *graded color scale* as in Figure 5. Areas exhibiting distortion-elevated activation that is form independent are shown in *blue* and listed in Table 1 (bottom). Arrow indicates the approximate location of the form-independent left inferior frontal voxel from which data are plotted in *d*. *c*, Axial sections through Heschl's gyrus showing distortion-elevated responses in temporal and frontal lobe regions. Form-dependent (*graded color*) and form-independent (*blue*) responses are observed in temporal and frontal lobe regions. *d*, The *graph* shows the response (percentage of signal change) of a voxel (⟶) in left inferior gyrus (−56, 16, −6) against word-report score. Error bars indicate SEM across subjects.

ing speech at more abstract nonacoustic levels of representation. The hierarchical structure that we infer from these results is consistent with cognitive accounts of spoken language comprehension (McClelland and Elman, 1986; Gaskell and Marslen-Wilson, 1997) in which lexical and semantic processes are driven by the output of lower-level acoustic and phonetic processes. This finding also mirrors what is known of the anatomical and functional organization of the auditory system in nonhuman primates. Whereas form-dependent responses were observed in both core and belt areas of auditory cortex, it is only in the parabelt and more distant polymodal cortex that we see a form-independent response to the intelligibility of speech signals.

A stream of processing, specialized for sound–object identification, has been documented previously in nonhuman primates. This extends anteriorly within lateral temporal neocortex (Rauschecker and Tian, 2000; Tian et al., 2001), similar to the anterior temporal portion of the form-independent intelligibility response. Future work to determine the functional specialization of these anterior temporal regions might therefore focus on whether responses in these regions are affected by the lexical and semantic content of sentences. An additional inferior frontal area exhibited a similar form-independent intelligibility profile. This area in humans, as in other primates, may receive projections from anterior auditory areas and anterior temporal lobe, extending the anteroventral-processing stream into ventrolateral frontal cortex (Hackett et al., 1999; Romanski et al., 1999a,b; Mamata et al., 2002).

We also observed a form-independent, intelligibility-related response in left posterior superior temporal gyrus and left angular gyrus. These activations may be indicative of other parallel streams of processing, extending posteriorly from auditory and form-dependent regions (Hickok and Poeppel, 2000; Scott and Johnsrude, 2003). Although the functional significance of such posterior streams has yet to be firmly established, one proposal common to these accounts is that a stream running dorsally to the sylvian fissure may play a role in linking the perception and production of speech. In support of this account, a number of previous cognitive models have proposed separate processing path-

ways involved in phonological versus lexical processing of speech (Gaskell and Marslen-Wilson, 1997; Norris et al., 2000).

An additional region in which the BOLD signal was correlated with intelligibility in a form-independent manner was the left anterior hippocampus. The medial temporal-lobe structures of the left (usually language-dominant) hemisphere are known to be required for the encoding and retention of verbal material (Milner, 1958; Johnsrude, 2001; Strange et al., 2002). Results from neuroimaging studies suggest that activation in the left anterior hippocampus is sensitive to the presence of meaning in verbal stimuli (Martin et al., 1997; Otten et al., 2001). The correlation that we observe may thus reflect the increasing memorability of sound sequences as they become more meaningful.

In addition to establishing anatomical specialization for speech comprehension, we wanted to identify brain areas that exhibited an increased response to degraded speech stimuli compared with both clear speech and an unintelligible baseline. We observed a left-lateralized frontal and temporal lobe system that showed this profile. This network of areas is consistent with anatomical connectivity. Auditory belt and parabelt are known (from work in nonhuman primates) to be reciprocally connected with prefrontal areas, including premotor cortex and areas in the inferior frontal gyrus, such as Brodmann area 45 (Hackett et al., 1999; Romanski et al., 1999a,b). These connections may provide a means by which frontal areas can modulate the operation of lower-level auditory areas in the temporal lobe during effortful comprehension of spoken language, thereby assisting in the recovery of meaning from distorted speech input.

Mechanisms to compensate for degraded input need not be speech specific. Low-level auditory or attentional processes that segregate speech from noise or restore continuity to speech briefly masked by noise (Cherry, 1953; Warren, 1970; Bregman, 1990) may assist in perceiving speech heard in noisy environments. Because the distortion-elevated activation in the temporal lobe was primarily form dependent (i.e., sensitive to distortion type) and particularly pronounced in areas adjacent to the probable location of primary auditory cortex, this response may reflect increased allocation of attention to spoken input (Grady et al., 1997). Because the response of this temporal lobe region was particularly pronounced for segmented speech, we speculate that perceiving speech in the "gaps" between noise bursts places a particular demand on this attentional system (cf. Warren, 1970; Bashford et al., 1996).

In contrast to the response profile observed in temporal lobe regions, the elevated response to distorted speech in the left frontal cortex was primarily insensitive to the form of distortion applied, as might be expected for compensatory processes that apply at a nonacoustic level. Although some of these frontal regions may not be involved in processes specific to language comprehension (such as decision processes involved in assessing intelligibility), we propose that a restricted portion of this activated area, the inferior frontal region in which responses were also correlated with intelligibility, contributes to the linguistic processes involved in accessing and combining word meanings (Thompson-Schill et al., 1997; Wagner et al., 2001). All three forms of distortion might be expected to draw more heavily on processes in which semantic or syntactic context is used to recover words and meanings that cannot be identified from bottom-up information alone (Miller et al., 1951; Gordon-Salant and Fitzgibbons, 1997). The results of previous work, in which sentence comprehension is challenged by the inclusion of more complex grammatical structures or lexical ambiguity, are consistent with this hypothesized role for left inferior frontal regions (Kaan and Swaab, 2002; Rodd et al., 2002).

Finally, we observed a focal region in the frontal operculum that showed an elevated response to distortion that was form dependent (particularly sensitive to noise-vocoded speech). This was the only form-dependent region that we observed outside of the temporal lobe and may correspond to an area previously identified electrophysiologically in macaques, which responds specifically to auditory stimuli, including both vocal and nonvocal sounds (Romanski and Goldman-Rakic, 2002). Additional behavioral work investigating comprehension of noise-vocoded speech may be informative in assessing the role of this region of elevated activation.

References

Bashford Jr JA, Warren RM, Brown CA (1996) Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. Percept Psychophys 58:342–350.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc [B] 57:289–300.

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

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound, pp 345–393. Cambridge, MA: MIT.

Brett M, Leff AP, Rorden C, Ashburner J (2001) Spatial normalization of brain images with focal lesions using cost function masking. NeuroImage 14:486–500.

Cherry EC (1953) Some experiments on the recognition of speech with one and two ears. J Acoust Soc Am 25:975–979.

Davis MH, Marslen-Wilson WD, Gaskell MG (2002) Leading up the lexical garden-path: segmentation and ambiguity in spoken word recognition. J Exp Psychol Hum Percept Perform 28:218–244.

Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM (1999) Improved auditory cortex imaging using clustered volume acquisitions. Hum Brain Mapp 7:89–97.

Forster KI, Forster JC (2003) DMDX: a windows display program with millisecond accuracy. Behav Res Methods Instrum Comput, in press.

Gaskell MG, Marslen-Wilson WD (1997) Integrating form and meaning: a distributed model of speech perception. Lang Cognit Process 12:613–656.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15:870–878.

Gordon-Salant S, Fitzgibbons PJ (1997) Selected cognitive factors and speech recognition performance among young and elder listeners. J Speech Lang Hear Res 40:423–431.

Grady C, Van Meter J, Maisog J, Pietrini P, Krasuski J, Rauschecker J (1997) Attention-related modulation of activity in primary and secondary auditory cortex. NeuroReport 8:2511–2516.

Hackett T, Stepniewska I, Kaas J (1999) Prefrontal connections of the parabelt auditory cortex in macaque monkeys. Brain Res 817:45–58.

Hall D, Johnsrude I, Haggard M, Palmer A, Akeroyd M, Summerfield A (2002) Spectral and temporal processing in human auditory cortex. Cereb Cortex 12:140–149.

Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999) "Sparse" temporal sampling in auditory fMRI. Hum Brain Mapp 7:213–223.

Hickok G, Poeppel D (2000) Towards a functional neuroanatomy of speech perception. Trends Cogn Sci 4:131–138.

Johnsrude IS (2001) The neuropsychological consequences of temporal lobe lesions. In: Cognitive deficits in brain disorders (Harrison JE, Owen AM, eds), pp 37–53. London: Dunitz.

Kaan E, Swaab TY (2002) The brain circuitry of syntactic comprehension. Trends Cogn Sci 6:350–356.

Kaas J, Hackett T (2000) Subdivisions of auditory cortex and processing streams in primates. Proc Natl Acad Sci USA 97:11793–11799.

Loftus GR, Masson MEJ (1994) Using confidence-intervals in within-subject designs. Psychon Bull Rev 1:476–490.

Mamata H, Mamata Y, Westin CF, Shenton ME, Kikinis R, Jolesz FA, Maier SE (2002) High-resolution line scan diffusion tensor MR imaging of white matter fiber tract anatomy. AJNR Am J Neuroradiol 23:67–75.

Martin A, Wiggs CL, Weisberg J (1997) Modulation of human medial temporal lobe activity by form, meaning, and experience. Hippocampus 7:587–593.

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognit Psychol 18:1–86.

Miller GA, Heise GA, Lichten W (1951) The intelligibility of speech as a function of the context of the test materials. J Exp Psychol 41:329–335.

Milner BA (1958) Psychological defects produced by temporal lobe excision. The brain and human behaviour. Proc Assoc Res Nerv Ment Dis 36:244–257.

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.

Norris D, McQueen J, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. Behav Brain Sci 23:299–370.

Otten LJ, Henson RNA, Rugg MD (2001) Depth of processing effects on the neural correlates of memory encoding: relationship between findings from accross- and within-task comparisons. Brain 124:399–412.

Palmer AR, Bullock DC, Chambers JD (1998) A high-output, high-quality sound system for use in auditory fMRI. NeuroImage 7:S359.

Poldrack RA, Temple E, Protopapas A, Nagarajan S, Tallal P, Merzenich M, Gabrieli JDE (2001) Relations between the neural bases of dynamic auditory processing and phonological processing: evidence from fMRI. J Cognit Neurosci 13:687–697.

Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, Zilles K (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. NeuroImage 13:669–683.

Rauschecker JP (1998) Cortical processing of complex sounds. Curr Opin Neurobiol 8:516–521.

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci USA 97:11800–11806.

Rauschecker JP, Tian B, Hauser M (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268:111–114.

Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM (1994) On the perceptual organization of speech. Psychol Rev 101:129–156.

Rivier F, Clarke S (1997) Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex. NeuroImage 6:288–304.

Rodd JM, Davis MH, Johnsrude IS (2002) An fMRI study of semantic ambiguity. Soc Neurosci Abstr 32:17–11.

Romanski L, Goldman-Rakic P (2002) An auditory domain in primate prefrontal cortex. Nat Neurosci 5:15–16.

Romanski L, Bates J, Goldman-Rakic P (1999a) Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. J Comp Neurol 403:141–157.

Romanski L, Tian B, Fritz J, Mishkin M, Goldman-Rakic P, Rauschecker J (1999b) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2:1131–1136.

Schroeder MR (1968) Reference signal for signal quality studies. J Acoust Soc Am 44:1735–1736.

Scott S, Johnsrude I (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107.

Scott SK, Blank CC, Rosen S, Wise RJS (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123:2400–2406.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304.

Stark CEL, Squire LR (2001) When zero is not zero: the problem of ambiguous baseline conditions in fMRI. Proc Natl Acad Sci USA 98:12760–12765.

Strange BA, Otten LJ, Josephs O, Rugg MD, Dolan RJ (2002) Dissociable human perirhinal, hippocampal, and parahippocampal roles during verbal encoding. J Neurosci 22:523–528.

Thompson-Schill S, D'Esposito M, Aguirre G, Farah M (1997) Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. Proc Natl Acad Sci USA 94:14792–14979.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290–293.

Vouloumanos A, Kiehl KA, Werker JF, Liddle PF (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.

Wagner A, Pare-Blagoev E, Clark J, Poldrack R (2001) Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. Neuron 31:329–338.

Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392–393.

Wessinger C, VanMeter J, Tian B, Van Lare J, Pekar J, Rauschecker J (2001) Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J Cogn Neurosci 13:1–7.

Whalen DH, Liberman AM (1987) Speech perception takes precedence over nonspeech perception. Science 237:169–171.

Xiong J, Rao S, Jerabek P, Zamarripa F, Woldorff M, Lancaster J, Fox PT (2000) Intersubject variability in cortical activations during a complex language task. NeuroImage 12:326–339.