A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception

Sophie K. Scott[a)]
Departments of Psychology and Phonetics & Linguistics, University College London, London WC1E 6BT, United Kingdom

Stuart Rosen
Department of Phonetics & Linguistics, University College London, London WC1E 6BT, United Kingdom

Lindsay Wickham
Department of Psychology, University College London, London WC1E 6BT, United Kingdom

Richard J. S. Wise
MRC Clinical Sciences Centre, London W12 0NN, United Kingdom

(Received 27 September 2002; revised 14 November 2003; accepted 17 November 2003)

Positron emission tomography (PET) was used to investigate the neural basis of the comprehension of speech in unmodulated noise (''energetic'' masking, dominated by effects at the auditory periphery), and when presented with another speaker (''informational'' masking, dominated by more central effects). Each type of signal was presented at four different signal-to-noise ratios (SNRs) (+3, 0, −3, −6 dB for the speech-in-speech, +6, +3, 0, −3 dB for the speech-in-noise), with listeners instructed to listen for meaning to the target speaker. Consistent with behavioral studies, there was SNR-dependent activation associated with the comprehension of speech in noise, with no SNR-dependent activity for the comprehension of speech-in-speech (at low or negative SNRs). There was, in addition, activation in bilateral superior temporal gyri which was associated with the informational masking condition. The extent to which this activation of classical ''speech'' areas of the temporal lobes might delineate the neural basis of the informational masking is considered, as is the relationship of these findings to the interfering effects of unattended speech and sound on more explicit working memory tasks. This study is a novel demonstration of candidate neural systems involved in the perception of speech in noisy environments, and of the processing of multiple speakers in the dorso-lateral temporal lobes. © *2004 Acoustical Society of America.*
[DOI: 10.1121/1.1639336]

PACS numbers: 43.64.Sj [PFA] Pages: 813–821

I. INTRODUCTION

Conversing at a cocktail party has been used as a classic demonstration of how we can listen to one person while ignoring the surrounding babble, yet also detect relevant spoken information such as one's name (Cherry, 1953; Conway *et al.*, 2001; Moray, 1959). This apparent processing of ignored auditory information is consistent with evidence that irrelevant speech signals can disrupt performance in auditory-verbal tasks, such as those involving verbal working memory (Tremblay *et al.* 2000). In trying to understand at least some aspects of the ''cocktail party effect,'' a distinction has been made between ''energetic'' and ''informational'' masking [see Brungart (2001) for a review].

''Energetic'' masking is demonstrated most clearly when a speech signal is presented together with a steady-state wideband noise. Here, the interfering effects of the masker arise primarily in the cochlea, reflecting the frequency analysis exacted on the basilar membrane. Just as for the energetic masking of tones by noise, the degree of energetic masking of speech by noise is primarily determined by the frequency spectrum of the noise, and its relative intensity (assumptions incorporated in articulation index theory) which computes SNR values within frequency channels (French and Steinberg, 1947). In particular, the lower the signal-to-noise ratio (SNR), the greater the masking of the signal.

''Informational'' masking, on the other hand, is supposed to reflect the interference of a masker on a target signal due to the masker having similar (and perhaps competing) informational content (Dirks and Bower, 1968; Festen and Plomp, 1990). The prime example of this is when speech is used to mask speech. Of course, energetic masking must also occur in such situations, though the amplitude modulations of natural speech alone impair its effectiveness as an energetic masker. The degree to which informational masking differs from energetic masking is demonstrated by the finding that the intelligibility of attended speech masked by other speech is relatively constant over a range of SNRs between −12 and 0 dB, and increases with increasingly positive SNRs (Brungardt, 2001). This level independency may arise from other sources, including the possibility that the amplitude modulations of a masking voice may afford ''glimpses'' of the target signal (Festen and Plomp, 1990). Such glimpses would allow the perception of the target signal during modulations of the masker, and reduce the effects of overall masking speech level.

a)Electronic mail: sophie.scott@ucl.ac.uk

The aim of this study was to investigate the neural basis of these masking effects using a functional neuroimaging technique [positron emission tomography (PET)], which tracks neural activity in terms of regional cerebral blood flow changes. In particular, we wanted to determine the neural correlates of the perception of speech masked by different signals, and also to determine the extent of SNR-dependent neural responses within this. The maskers chosen were speech and steady-state noise (with the same long-term spectral shape as speech). These two maskers were chosen as clear exemplars of informational and energetic masking, respectively. Within each condition, four SNR levels were selected, such that performance differences in the profile of behavioral responses (comprehension of sentences spoken by the target speaker) were clear between the speech and noise maskers without overall intelligibility falling below 60%. [See Davis and Johnsrude (2003) for an example of using noise masking as a method to manipulate intelligibility over a wider range of intelligibility levels.] The aim was also to use enough SNR levels, over a sufficient range, to maximize the possibility that neural correlates of level-dependent effects could be identified. To satisfy these constraints different SNR levels were chosen for each condition: -6, -3, 0, and $+3$ dB for the speech masker condition, and -3, 0, $+3$ and $+6$ dB for the noise masker condition. Pilot testing suggested that a SNR of -6 for the noise masker reduced intelligibility too much, while using a $+6$ dB SNR for the speech masker led to ceiling effects on the sentence comprehension task [consistent with positive SNRs being associated with an increase in intelligibility (Brungart 2001)]. The masking speech need not be intelligible to lead to interference effects in masking: studies have shown that reversed speech can also be an effective masker (Brungart and Simpson, 2002).

PET has several advantages for such a study, since it is *relatively* quiet compared to fMRI, and it does not suffer from signal loss due to susceptibility artefacts and geometric distortion (Devlin *et al.*, 2000). Susceptibility artefacts can be particularly problematic in the anterior temporal lobes, which we have previously demonstrated to be important in speech processing (Scott *et al.*, 2000). However, a disadvantage of PET is that the number of possible scans is limited by the total amount of radioactivity that can be administered. This study is thus preliminary in the sense that extensive testing of different masking conditions and different levels was not possible. Most importantly, this design did not permit the presentation of a baseline condition of listening to speech with no masking signal. In partial control for this, the results were contrasted with a previous study (Mummery *et al.*, 1999), in which passive speech perception [relative to signal correlated noise (Schroeder, 1969)] was studied.

There are several candidate neural systems that might be recruited during the attentional control of speech perception in energetic and informational masking. There could be a modulation of activation in primary auditory cortex (PAC) by different listening contexts, as reported by Ulanovsky *et al.* (2003). There could also be an alteration of the profile of activity in auditory association cortex, potentially linked to functional subsystems within this. For example, Griffiths and Warren (2002) have proposed that the planum temporale (posterior to PAC) operates as an informational hub for incoming auditory information, and that this might be associated with a distinct role of informational masking. Zatorre and colleagues (2002) have also emphasized the role of posterior auditory fields in the spatial representation of auditory objects. In contrast (though not necessarily contradiction) a meta-analysis of functional imaging studies has identified a role for the superior temporal gyrus (STG), lateral to primary auditory cortex (PAC), which is important in the processing of different aspects of auditory structure (e.g., AM, FM, harmonic structure), and which forms part of the acoustic processing of the speech signal (Scott and Johnsrude, 2003). Important in normal speech perception, this might be a candidate for the parallel processing of acoustic cues that are important for tracking a target voice and a masking signal in an informational masking context. There is thus the possibility that informational masking might be associated with a modulation of activity in the lateral STG and regions anterior to this, if informational masking results from competition between the processing of the target and the unattended voice within this system. Considering regions outwith the auditory system, it is possible that more generic, amodal attentional mechanisms could be recruited when speech is presented in a masking context. This would be associated with activation in prefrontal and parietal regions, commonly seen in cognitive tasks requiring the control of attention across modalities.

II. METHOD: SIGNAL PROCESSING AND STIMULI

All stimulus materials were drawn from digital representations (sampled originally at 22.05 kHz) of simple sentences recorded in an anechoic chamber by one male and one female speaker of standard Southern British English. The target sentences were always BKB sentences spoken by the female speaker whereas maskers were based on the IHR ASL sentences spoken by the male speaker (more details are given in the next section). All sentences were low-pass filtered at 3.8 kHz (sixth-order elliptical filter, both forward and backwards, so as to ensure zero-phase filtering equivalent to a 12th-order filter), and then downsampled to 11.025 kHz to save space. The masker and target signals were played together diotically (target and masker summed together and presented to both ears).

In the speech-in-noise condition, the target speech was played together with unmodulated noise, with the same long-term average spectrum as the masking male speaker. These calculations began with a spectral analysis of all 270 masker sentences, sampled at 22.05 kHz. Analyses used a FFT of length 512 sample points (23.22 ms), with windows overlapping by 256 points, giving a value for the spectrum at multiples of 43.1 Hz. The spectrum was then smoothed (in the frequency domain) with a 27-point Hamming window that was two octaves wide, over the frequency range 50 Hz to 7 kHz. The smoothed spectrum was then used to construct an amplitude spectrum for an inverse FFT (component phases randomized with a uniform distribution over the range $0-2\pi$) in order to create the speech-shaped noise.

Different SNRs were determined by a simple rms calculation across the entire waveform (e.g., target sentence), and

all combined waves were normalized to the same rms value.

The target and speech-masker stimuli were the BKB and IHR sentences respectively (Foster *et al.*, 1993; MacLeod and Summerfield, 1987). These are sets of syntactically simple sentences used to test intelligibility; they are scored according to the number of key words (two to three words per sentence) that are repeated correctly (e.g., ''she's brushing her hair,'' ''the clown had a funny face,'' ''the bag was very heavy''—the key words are underlined). The use of a female target speaker and a male masking speaker was likely to lead to less extensive informational masking than two same sex speakers (Brungart, 2001), but this was chosen to enable the instruction ''listen to the female speaker'' to be used throughout, rather than to train the subjects on the identity of the target speaker.

There were eight scans for each stimulus condition (noise masker and speech masker), with four different SNRs for each of these (+3, 0, −3, −6 dB for the speech-in-speech, +6, +3, 0, −3 dB for the speech-in-noise), presented twice in a random order. In the pretesting (see below) the stimuli were presented one target sentence (plus masker) at a time; during PET scanning the target sentences (plus maskers) ran continuously, for approximately 60 s. No overt response was required. During PET scanning, no sentences were repeated, either as targets or as maskers.

III. METHOD: BEHAVIORAL TESTING

Pilot testing was used to determine the intelligibility of the different masker and SNR conditions: ten normal-hearing adults (age range 26–50, five men, none of whom participated in the PET study) were presented with sentences over headphones and asked to repeat back the words that they could hear. This was done for a range of SNRs for both masker types (−12 to +9 dB SNR for the speech masker, −3-to +6 dB SNR for the noise masker). Sixteen sentences were used for each condition. These data were used to select the SNR conditions in which intelligibility was above a threshold (60%) to be used for the PET scanning.

The subjects for the PET study were presented with the stimuli prior to scanning. They were played individual BKB sentences and the masking stimuli over headphones and repeated back what they could hear. Eight sentences were presented per condition. Intelligibility was scored by an experimenter, who recorded the number of correct key words per condition. Since there was some variation in the number of key words per sentence, this score varied between a maximum of 18–20 key words per condition. This gave a score for each subject, masking condition, and SNR. The order of conditions was randomized.

IV. METHOD: PET SCANNING

Seven right-handed native English-speaking male volunteers were recruited and scanned. The mean age was 42, with a range of 35–52. Each participant gave informed consent prior to participation in the study, which was approved by the Research Ethics Committee of Imperial College School of Medicine/Hammersmith, Queen Charlotte's & Chelsea & Acton Hospitals. Permission to administer radioisotopes was given by the Department of Health (UK).

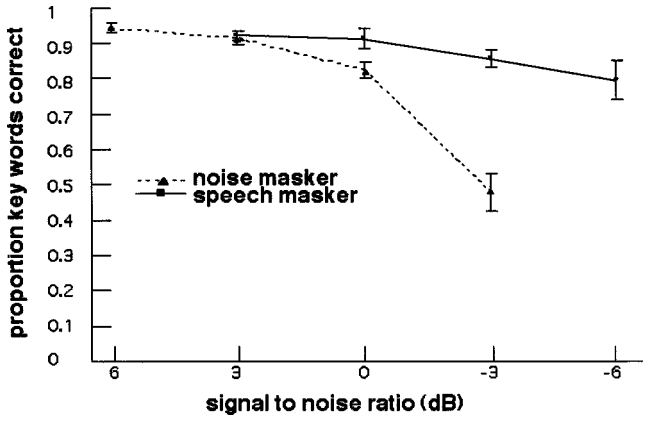

FIG. 1. The mean intelligibility of speech for the two maskers (speech and noise) as a function of SNR for the seven scanned subjects. Standard errors of the mean are shown by the error bars.

None of the subjects reported hearing problems and all were able to perceive speech in the different conditions during prescan training, though performance was poorer than the pilot subjects, such that performance in the lowest speech-in-noise SNR was lower than that in the pilot, with an average of 48% (see Fig. 1).

PET scanning was performed with a Siemens HR++ (966) PET scanner operated in high-sensitivity 3D mode. Sixteen scans were performed on each subject, using the oxygen-15-labeled water bolus technique. All subjects were scanned while lying supine in a darkened room with their eyes closed.

The stimuli were presented at a comfortable level determined for each subject, and this level was kept constant over the scanning sessions. The sentence presentations began 15 s before the scanning commenced, and each sentence presented was novel (i.e., there were no repeats). The subjects were instructed to listen passively to the female speaker ''for meaning'' in the scanning sessions. Passive listening reduces the likelihood that activation seen is due to controlled processing aspects of the task, which would be involved if the subjects were required to make explicit responses or try and remember the sentences they heard (Scott and Wise, 2003). Since this study is novel in focusing on normal speech perception in complex sound situations, this ensured that the activation seen was related to this, and not to some other aspect of the task requirements.

V. ANALYSIS

The images were analyzed using statistical parametric mapping (SPM99, Wellcome Department of Cognitive Neurology, http://www.fil.ion.ucl.ac.uk/spm), which allowed manipulation and statistical analysis of the grouped data. All scans from each subject were realigned to eliminate head movements between scans and normalized into a standard stereotactic space (the Montreal Neurological Institute template was used, which is constructed from anatomical MRI scans obtained on 305 normal subjects). Images were then

smoothed using an isotropic 10-mm, full width at half-maximum, Gaussian kernel, to allow for variation in gyral anatomy and to improve the signal-to-noise ratio.

In the statistical analysis of functional imaging data, the unit of analysis is the voxel, a cube of 2 mm^3. As voxel size is smaller than the resolution of the PET scanner, activity within adjacent voxels is not independent. Specific changes in regional cerebral blood flow (rCBF) were investigated by comparing the activity at each voxel in standardized space across scans (and, therefore, behavioral conditions). Specific, voxel-by-voxel analyses were performed using appropriate contrasts (e.g., which voxels are showing a greater response in speech masking contexts, relative to noise making contexts) to create statistical parametric maps of the t statistic, which were subsequently transformed into *Z* scores. The precise contrasts used are outlined in the next section. The threshold for significance was set at $P<0.05$, corrected for analyses across the whole volume of the brain ($P<0.000001$, uncorrected; *Z*-score >4.7). For contrasts where there is a prior hypothesis about the anatomical involvement, uncorrected p values of less than 0.0001 can be accepted.

The analysis included a blocked analysis of covariance (ANCOVA) with global counts as confound to remove the effect of global changes in perfusion across scans. This is necessary because there is considerable scan-to-scan variance in the number of counts (i.e., the precise amount of radio-labeled water infused) and in the blood flow of the subject (e.g., blood pressure tends to drop the longer one lies prone, and in a PET scan the subjects lie down for at least 1.5 h).

VI. RESULTS: BEHAVIORAL DATA

The intelligibility data from the pretesting of the scanned subjects are shown in Fig. 1. This demonstrates that, as predicted by previous studies, the speech-in-noise shows evidence of a level-dependent effect below a SNR of 3 dB; in contrast the speech-in-speech conditions show much less variation at 0 dB and below. A logistic regression was used to compare the trend in performance with SNR for the speech and noise masker. Analyses proceeded from the total number of key words correctly identified by each listener in each condition of SNR and masker type (six sentences containing a total of 18–20 key words), that is, 64 data points (8 listeners$\times$2 maskers ypes$\times$4 SNRs). SNR was the only continuous variate. The first analysis used listener, masker type, and SNR as predictors of performance (but with +3 dB added to the SNRs for the conditions with the speech masker, so that the 4 SNRs for each masker type were aligned). The triple interaction was not significant ($p\approx0.88$), but all the second-order interactions were ($p<0.005$). Significant terms involving listeners indicate that listeners varied in terms of their sensitivity to changes in SNR and masker type. More importantly, there was a significant interaction between masker type and SNR ($p<0.001$), indicating that performance depends upon SNR to a greater degree for the noise than the speech masker. A separate analysis of the role of SNR for the speech masker still found performance to depend significantly on SNR ($p<0.001$).

VII. RESULTS: PET SCANNING

Four different types of analyses were performed to investigate the neural correlates of the behavioural effects. The first were ''subtraction'' contrasts that reveal the activations which were greater when the participants listened to speech-in-speech, relative to speech-in-noise, and vice versa. These contrasts are insensitive to the different SNR levels. Since the design required the subjects to listen to the female speaker throughout the experiment, activations associated with the female speaker were ''subtracted'' out of the contrast. The second set of contrasts was parametric and investigated SNR-dependent responses by using the SNR as a covariate in each condition (speech-in-noise and speech-in-speech) separately. The third analysis was also parametric and used the intelligibility scores across all scans as a variable to identify neural activations that correlate positively with this. Finally, a statistical comparison was made between the speech-in-speech$>$speech-in-noise conditions and a previous study (Mummery *et al.*, 1999) which contrasted the perception of speech (single words) with a nonspeech baseline (signal correlated noise). This was an attempt to reveal the extent to which regions activated by ''unattended'' speech activated brain regions seen in passive speech percep-

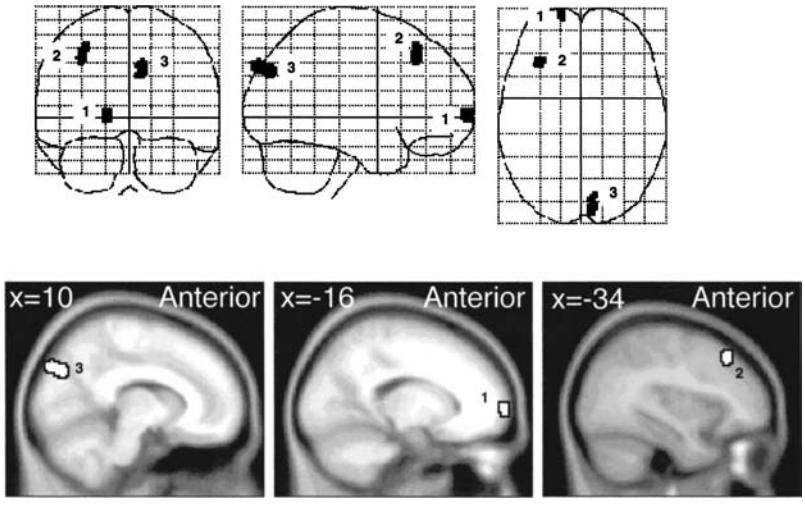

FIG. 2. Glass brain projections (upper panels) and sagittal slices on an average T1 weighted magnetic resonance imaging (MRI) image (lower panels) showing the activity seen for the speech-in-noise over speech-in-speech conditions, thresholded at $p<0.0001$, number of coactivated voxels >40. Labels: 1—left frontal pole, 2—left dorsolateral prefrontal cortex, 3—right posterior parietal cortex.

TABLE I. The locations, *Z* scores, corrected *p* values, and coordinates of the rCBF changes seen in the different contrasts (*indicates uncorrected *p* values).

Contrast	Region	*Z* score	*P* (corr)	*x*	*y*	*z*
Noise> speech	R posterior parietal	**5.22**	**0.009**	**10**	**−86**	**38**
	L rostral prefrontal cortex	**5.20**	**0.009**	**−16**	**68**	**2**
	L dorsolateral prefrontal cortex	**5.42**	**0.021**	**−34**	**28**	**44**
Speech> noise	L superior temporal gyrus	**7.30**	**0.000**	**−64**	**−20**	**2**
		6.73	0.000	−58	−8	2
		6.19	0.000	−68	−30	10
	R superior temporal gyrus	**7.18**	**0.000**	**64**	**−18**	**2**
		5.79	0.001	70	−26	4
		5.28	0.007	66	−8	0
Increasing intelligibility	L anterior superior temporal gyrus	**5.10**	**0.0015**	**−58**	**0**	**0**
(speech>SCN) +(sp-in-speech>sp-in-noise)	L lateral STG	**5.38**	**0.022**	**−52**	**−14**	**2**
	L posterior STS	5.23	0.039	−56	−32	6
	L lateral STG	5.16	0.053	−66	−16	0
	R lateral STG	4.99	0.096	66	−24	2
	L anterior STG	4.90	0.000*	−60	8	−2

tion. An important proviso is that these two studies were performed on different PET scanners, which may make the comparison less sensitive.

A comparison of the activity seen in the speech-in-noise condition, relative to the speech-in-speech condition, revealed SNR-independent responses in the left frontal pole, left dorsolateral prefrontal cortex, and the right posterior parietal cortex. (Fig. 2, Table I) In contrast, strong responses were seen for speech-in-speech over speech-in-noise in the left and right lateral superior temporal gyri (STG) and sulci (STS), running posterior, lateral and anterior to primary auditory cortex (Fig. 3, Table I). The activation also extends into Heschl's gyrus on the right, although this is not a separate peak of activation. Since the behavioral data indicates

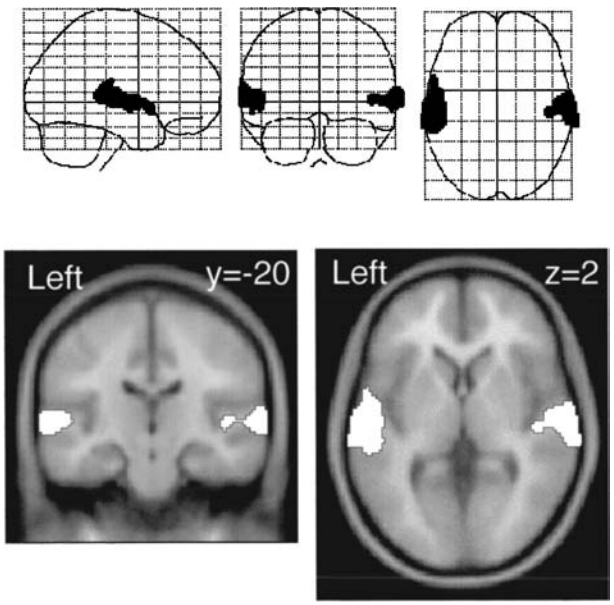

FIG. 3. Neural activity in the speech-in-speech condition, relative to the speech-in-noise condition; the rCBF changes are shown on glass brain projections and coronal and transaxial slices of an average T1 weighted MRI image, thresholded at $p<0.0001$, number of co-activated voxels >40.

that the lowest SNR for speech-in-noise (−3 dB) results in the poorest comprehension (see Fig. 1), the analysis was repeated excluding this condition and thus avoiding any activations due to gross comprehension differences between the speech-in-noise and in the speech-in-speech conditions. With this condition excluded, the peak changes little, moving just 2 mm up in the *z* plane, with the other coordinates remaining unaffected ($Z=7.04$).

The analysis of rCBF changes with SNR revealed no significant changes for speech-in-speech, either for SNR increases or decreases. SNR-dependent analysis of speech-in-noise revealed activations in left inferior prefrontal cortex [−42, 20, −12, $Z=4.15$, $P(uncorrected)<0.001$] and left dorso-medial premotor area [−14, 2, 70, $Z=4.15$, $P(uncorrected)<0.001$] (Fig. 4). The rCBF levels appear to vary linearly with the decreasing SNR conditions in both regions [Figs. 4(a) and (b)], and reflect neural responses that increase as the listening task becomes harder. These results are included, however, with the proviso that the *p* values fall below the level of significance for whole brain comparisons. There was no significant rCBF change associated with increasing SNR levels (i.e., as speech perception becomes easier).

The use of the subjects' average intelligibility scores across all the masker and SNR conditions as a covariate revealed a left lateralized peak, in anterior STG (Fig. 5, Table I), where activity increased positively with intelligibility. To demonstrate that this effect was seen across all the subjects the analysis was repeated with the subjects' individual intelligibility scores entered as subject specific covariates, which gave one peak at the same location in the anterior STG (−58, 0, 0, $Z=4.59$). Individual subjects' rCBF values for this peak voxel were plotted against their intelligibility scores and the regression lines for each plot. This showed that for all but one of the subjects the relationship is a positive one, which is why a positive correlation between rCBF and intelligibility comes out in the mean intelligibility analysis. The R^2 values (and corresponding *p* values) for the linear

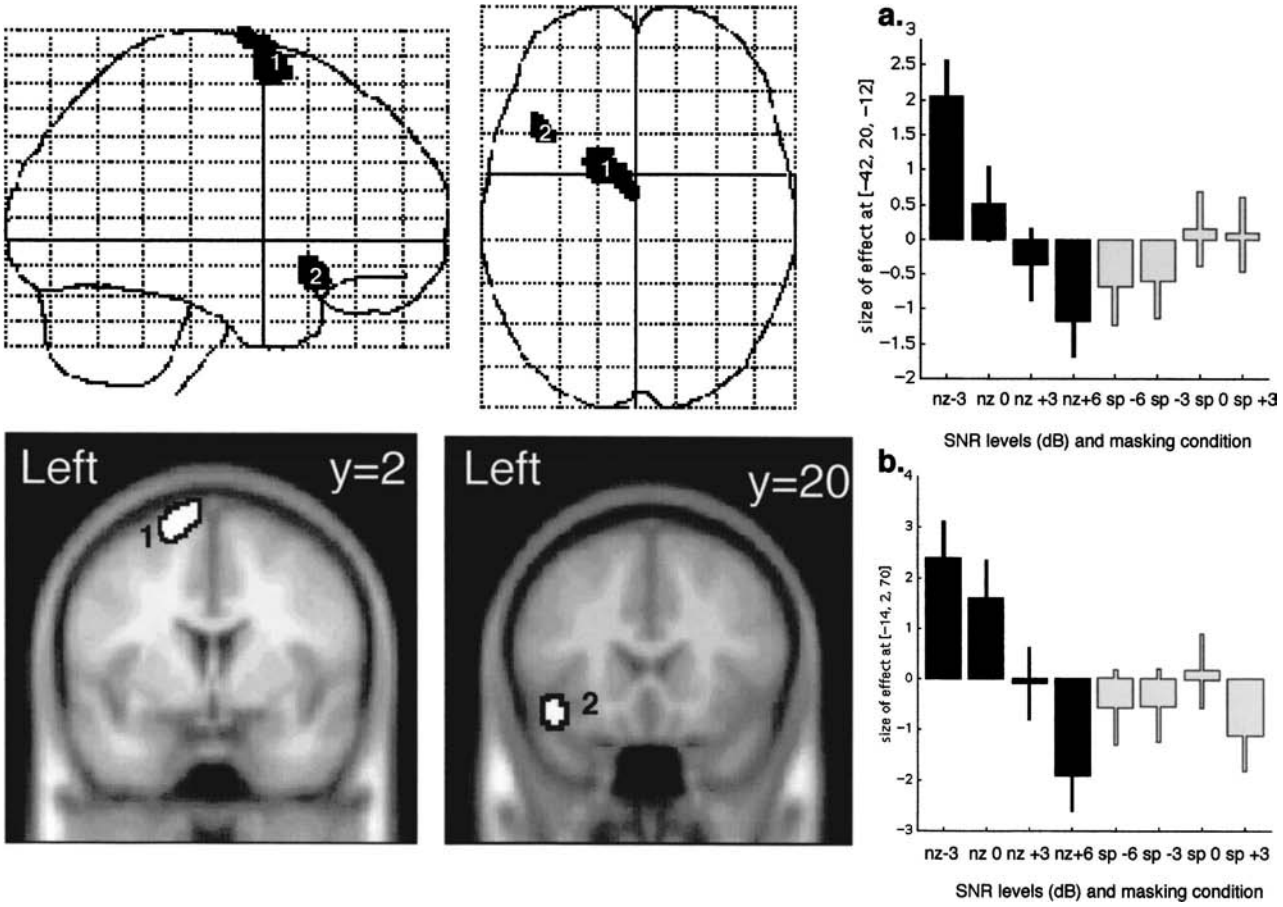

FIG. 4. Glass brain projections (upper left panels) and coronal slices of an average T1 weighted MRI image of rCBF changes that correlate with the difficulty of the speech-in-noise condition—i.e., negatively correlated with the SNR values, thresholded at $p<0.001$, number of coactivated voxels >40. Key: 1—left dorso-medial premotor cortex, 2—left inferior lateral prefrontal cortex. The graphs on the right show the rCBF changes associated with the different SNR values for the speech-in-noise condition at the peak voxel in the inferior prefrontal region (a) and the dorso-medial premotor activation (b).

regression for each subject (1–7) respectively are 0.123 ($p=0.18$), 0.447 ($p=0.0042$), 0.252 ($p=0.05$), 0.436 ($p=0.005$), 0.074 ($p=0.30$), 0.466 ($p=0.003$), 0.0014 ($p=0.87$). These R^2 values indicate there is a considerable amount of rCBF variation accounted for by the intelligibility of the sentences, for five out of the seven subjects, and that this was significant at $p<0.05$ for four of the seven subjects.

A second level random effects analysis was used to compare the activation for speech in speech>speech-in-noise and the speech>signal correlated noise contrast from Mummery *et al.* (1999) (also smoothed to 10 mm). This tested for regions coactivated by both contrasts, and for those significantly more activated by the speech in speech >speech-in-noise contrast. This revealed extensive coactivations in bilateral STG/STS and the supratemporal planes, with peaks in left posterior STS, bilateral STG, and left anterior STS. (Fig. 6, Table I)

VIII. DISCUSSION

The results of this study reveal a clear difference between the neural processing of speech when it is masked by speech versus noise. The former is associated with extensive activation in bilateral superior temporal gyri, the latter with the recruitment of brain regions remote from those classically associated with speech perception. Previous behavioral studies have suggested that speech and noise act as maskers in distinctly different ways (e.g., Brungart, 2001), and this neuroimaging evidence is consistent with such observations.

The use of unmodulated speech-spectrum noise as a masking stimulus, although at SNR levels above those that eliminate comprehension, shows a distributed network of neural regions, consisting of right parietal cortex and left prefrontal cortex (when contrasted with listening to speech-in-speech). The responses in these regions are independent of SNR level. Such patterns of activation are not seen in normal speech perception when a passive listening task is used (e.g., Binder *et al.*, 2000; Mummery *et al.*, 1999; Scott *et al.*, 2000, Wise *et al.*, 1991; 2001); this suggests that these areas are recruited to facilitate the perception of speech specifically in the context of unmodulated noise, which makes the speech difficult to hear due to masking at the auditory periphery. In this context the activation might be associated with some degree of attention; indeed prefrontal and parietal activations are associated with online, controlled cognitive processing and the involvement of explicit attentional mechanisms, albeit those that are not specific to auditory processing [e.g., prospective memory (Burgess *et al.*, 2001); auditory vigilance (Paus *et al.*, 1997); rapid visual information processing (Coull *et al.*, 1996)].

When the subjects listen to the female speaker in the

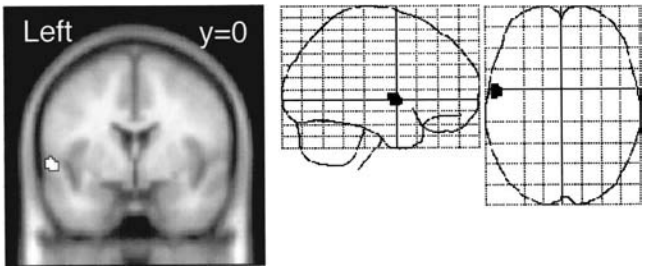

FIG. 5. Coronal slice of an average T1 weighted MRI image and glass brain projections of the rCBF regions that correlate with the increases in intelligibility, across all conditions, thresholded at $p<0.0001$ with number of coactivated voxels >40.

context of another (male) speaker a different pattern of activation is seen. Despite the fact that this contrast, by comparing the perception of speech-in-speech with speech in noise, ''subtracts'' the mental process of perceiving speech (as it is present in both conditions), the activation in left and right STG/STS is extensive, and independent of the SNR level (Fig. 3). There is considerable correspondence between the glass brain images for this contrast and that of a PET study contrasting speech with signal correlated noise (SCN); the activations extend along the lateral STG/STS, and their extents are similar in the anterior and posterior dimensions. This similarity was confirmed by a statistical comparison with this previous PET study (Fig. 6). Regions which are activated by both speech>SCN and speech-in-speech

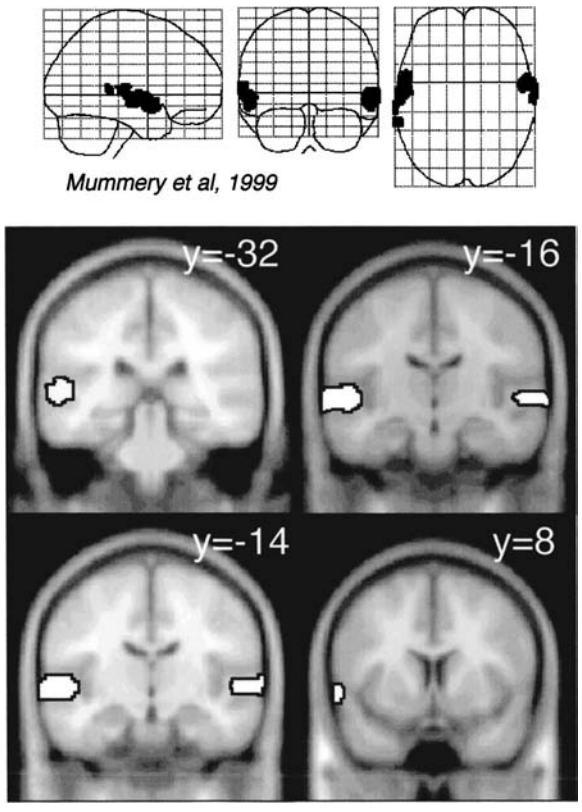

FIG. 6. Glass brain projections of the speech specific responses seen in an earlier study [reanalysis of Mummery *et al.* (1999)] are shown on the top panels: The activations seen for speech-in-speech over speech-in-noise and for speech>signal correlated noise are shown in the lower figures, on coronal slices of an average T1 weighted MRI image. Both sets of activations are thresholded at $p<0.0001$ with number of co-activated voxels >40.

>speech-in-noise run bilaterally along the supra temporal plane and lateral STG/STS, with greater anterior and posterior extent on the left.

There are several potential explanations of this result. The peaks in posterior STG are consistent with claims that posterior auditory fields are involved in the analysis of auditory objects and their location (Griffiths and Warren, 2002; Zatorre *et al.*, 2002), even though there were no cues to spatial location in this study. This finding may thus reflect the neural basis of grouping auditory objects, here different speakers, a processing demand which is increased in the speech-in-speech condition. The peaks in bilateral lateral STG/STS and left anterior STG are consistent with the ''unattended'' speech being processed along the same stream of processing as attended speech. This argument suggests regions antero-lateral to PAC may be involved in processing more than one concurrent speech source, and that informational masking occurs as a result of these competing speech related cues. The use of an informational masking signal which is not intelligible, but which is acoustically as complex in its structure as speech [e.g., spectrally rotated speech (Scott *et al.*, 2000)] will be able to reveal the extent to which such activation is a result of the acoustic overlap between the two signals, or the semantic content of the masking speech.

A different interpretation is that ''glimpsing'' of the target signal, due to amplitude modulations in the masking speech, may lead to greater activation in these regions than is seen when perceiving speech in unmodulated noise. This would result in increased activations of speech processing regions as more of the speech signal is available for cortical processing (due to the ''gaps'' in the masker). This can be explicitly addressed in further studies using amplitude-modulated noise maskers (to enable some ''glimpsing''). It is also the case that the speech-in-speech condition is considerably more complex than speech alone, in terms of its modulation spectrum and spectral profile: future studies with unintelligible masking stimuli as complex as speech will be able to address the role of the acoustic profile of the masking stimulus in the pattern of activation seen.

There is some evidence from this study for an involvement of primary auditory cortex in informational or energetic masking: the speech-in-speech>speech-in noise contrast shows right STG activation that extends medially into Heschl's gyrus, which is the location of primary auditory cortex in man (Fig. 3). However, this was not a separate peak of activation, making it harder to be certain about the validity of this finding. This equivocal result could, however, be due to the power of the analysis, as a consequence of resolution of the technique and the design of the current study. Blood flow measures may not be able to resolve small changes due to such modulation in a study where auditory stimulation is present in each condition.

One tentative conclusion, therefore, is that in the informational masking conditions there is some evidence that both auditory object segregation regions and speech processing regions are involved, and that these may contribute to the central auditory processes associated with informational masking. This may also suggest that the bilateral STG/STS regions commonly seen in functional imaging studies of

speech perception are also capable of processing in parallel other, unattended, speech information; the converse of this is that when the female speaker is ''streamed'' out of the auditory scene, the selection is not occurring ''early'' in the processing of the perceptual stream. The activations common to speech perception and speech-in-speech>speech-in-noise run along the STG/STS into regions which, on the left, are associated with the processing of intelligible speech (Scott *et al.*, 2000), consistent with the suggestion that the unattended voice is processed, to some degree, for meaning. Further studies will be able to address how these posterior and lateral auditory regions interact in informational masking.

Several cognitive processing tasks have revealed the consequences of obligatory perceptual processing of unattended speech. Early studies indicated that selective attention can be directed to the meaning of speech, with the potential for interference if concurrent ''unattended'' speech overlaps in semantic content (e.g., Treisman, 1960). There is also evidence that that concurrent irrelevant speech can disrupt performance on verbal working memory tasks (Tremblay *et al.*, 2000). This suggests that the perceptual competition underlying informational masking can also affect the cognitive processes dependent upon speech perception. There is also evidence that intelligible irrelevant speech is processed for meaning: recent work, for example, has shown that the meaning of unattended speech can interfere with semantic processing (e.g., Neely and LeCompte, 1999). The current findings may demonstrate a candidate neural basis for these behavioral effects. Importantly, as with informational masking, aspects of irrelevant speech disruptions in working memory tasks appear not to be speech specific, and effects can be seen with tone sweeps and reversed speech, although there is a relationship between the acoustic features of the irrelevant signal and the amount of distraction it causes (e.g., Tremblay *et al.*, 2000). As noted, further functional imaging studies can determine the extent to which the to-be-ignored voice is processed because it is meaningful speech, or because it is acoustically similar to speech.

The investigation of a SNR-dependent response for speech-in-speech revealed no activation that correlated with the SNR, although this is arguing to a null result and future work with more sensitive techniques [e.g., same sex speakers, which will increase the amount of informational masking (Brungart, 2001)] may find a difference. This finding is, however, consistent with the hypothesis that there would be no such response, since the behavioral data indicates that informational masking is SNR independent at SNRs of 0 dB and below (Brungart, 2001). In contrast, there were activations in the speech-in-noise that correlated with SNR, consistent with the original hypothesis and the behavioral data. Since there was no prediction about likely regions associated with SNR-dependent responses, and the activations did not survive correction for whole brain comparisons, the activations seen in left dorso-medial premotor cortex and left inferior prefrontal cortex (shown in Fig. 4) must be treated with some caution. However, there is evidence for an involvement of both these regions in speech processing. In a recent paper on speech production (Blank *et al.*, 2002) extensive medial prefrontal activation was associated with the production of ''propositional'' speech, (e.g., a verbal description of a relative that you see often but don't live with), in contrast to ''automatic'' speech (e.g., speaking a very familiar nursery rhyme repeatedly). This activation extended dorsally to the premotor region which shows the SNR-dependent response; as the speech is harder to hear in the noise, there is greater activation in this region. This premotor response potentially reflects the use of articulatory strategies (i.e., ''sounding out'' the heard words), which may or may not be explicit (i.e., associated with a deliberate strategy), and which are recruited to facilitate speech perception in the context of energetic masking. Previous studies have implicated ''mirror'' neurones (Rizzolatti and Arbib, 1998) in lateral premotor regions (posterior Broca's area) in aspects of speech perception when the task makes more motoric demands, e.g., segmentation (Burton *et al.*, 2000). The activation seen in this study is dorsal and medial to such activations, and this possibly reflects the fact that subjects simply had to comprehend the speech in the current study without doing such complex phoneme monitoring. In other words, no meta-linguistic processing was required.

With respect to the inferior prefrontal cortical response, another recent study (Crinion *et al.*, 2003) showed a response in ventral prefrontal cortex to hearing speech (children's stories) compared to reversed speech. The peak was very close (−44 26 −16) to that seen here. Crinion *et al.* (2003) associated this response with prefrontal regions that receive projections from anterior temporal lobe regions implicated in the perceptual processing of intelligible speech (Scott *et al.*, 2000) and they identified this region as a more executive component of story comprehension. Certainly the neuroanatomy is consistent with this claim, and it is intriguing to speculate that the ventral prefrontal activation seen in the current study demonstrates ''top down'' efforts to support speech comprehension with semantic information.

The correlation of intelligibility scores with rCBF across both conditions revealed a left lateralized response, lying lateral and anterior to primary auditory cortex in the STG. A previous study from our group, which investigated the neural correlates of speech processing while controlling for auditory complexity, showed responses associated with intelligibility of speech in left anterior STS (Scott *et al.*, 2000). The peak is superior and posterior to the most anterior peak shown in the Scott *et al.* (2000) study, which may be due to the task; subjects in the previous study were not forced to stream the speech out from a noisy background. Another difference is one of sensitivity, since more of the variation behaviorally comes from reduced comprehension in the most difficult speech-in-noise condition, and the other conditions do not differ greatly. Nonetheless, the activation which is associated with increasing comprehension of the attended speech lies in the anterior temporal lobe and not in posterior temporal lobe regions often claimed to be central for the processing of speech for meaning (Hickok and Poeppel, 2000). In the context of the current study and our work on intelligibility, this confirms that anterior temporal lobe regions are associated with intelligibility in speech.

This study thus presents evidence that different masking contexts for speech perception recruit different neural sys-

tems. Regions in rostral and dorsolateral prefrontal cortex and posterior parietal cortex are recruited, in a SNR-independent fashion, when subjects listen to speech-in-noise. In contrast, speech-in-speech activates bilateral STG/STS, in addition to the activation associated with the perception of the attended speech, indicating parallel processing of the speakers. Further studies are needed to determine whether this occurs because the unattended speech is meaningful, or whether it is due to more basic acoustic properties of the masking signal. Future studies will also be able to determine whether the SNR-dependent effects seen for speech-in-noise reflect the automatic recruitment of articulatory and semantic mechanisms in difficult speech perceptual conditions, or whether these reflect more explicit and deliberate cognitive strategies. The overall correlation of intelligibility of the signal with the activity in left anterior STG is further evidence for the role of anterior regions in the comprehension of intelligible speech.

ACKNOWLEDGMENTS

SKS and RJSW are both funded by grants from the Wellcome Trust. We would like to thank Phil Beaman for helpful comments on the manuscript.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., and Possing, E. T. (**2000**). ''Human temporal lobe activation by speech and nonspeech sounds,'' Cereb. Cortex **10**, 512–528.

Blank, S. C., Scott, S. K., Murphy, K., Warburton, E., and Wise, R. J. S. (**2002**). ''Propositional speech production: Broca, Wernicke and beyond,'' Brain **125**, 1829–1838.

Brungart, D. S. (**2001**). ''Informational and energetic masking effect in the perception of two simultaneous talkers,'' J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D. S., and Simpson, B. D. (**2002**). ''Within-ear and across-ear interference in a cocktail-party listening task,'' J. Acoust. Soc. Am. **112**, 2985–2995.

Burgess, P. W., Quayle, A., and Frith, C. D. (**2001**). ''Brain regions involved in prospective memory as determiend by positron emission tomography,'' Neuropsychologia **39**, 545–555.

Burton, M. W., Small, S. L., and Blumstein, S. E. (**2000**). ''The role of segmentation in phonological processing: An fMRI investigation,'' J. Cogn Neurosci. **12**, 679–690.

Cherry, E. C. (**1953**). ''Some experiments on the recognition of speech, with one or two ears,'' J. Acoust. Soc. Am. **25**, 975–979.

Conway, A. R., Cowan, N., and Bunting, M. F. (**2001**). ''The cocktail party phenomenon revisited: the importance of working memory capacity,'' Psychonomic Bull. Rev. **8**, 331–335.

Coull, J. T., Frith, C. D., Frackowiak, R. S. J., and Grasby, P. M. (**1996**). ''A fronto-parietal network for rapid visual information processing: A PET study of sustained attention and working memory,'' Neuropsychologia **34**, 1085–1095.

Crinion, J. T., Lambon-Ralph, M. A., Warburton, E. A., Howard, D., Wise, R. J. (**2003**). ''Temporal lobe regions engaged during normal speech comprehension.'' Brain, **126**, 1193–1201.

Davis, M. H., and Johnsrude, I. S. (**2003**). ''Hierarchical processing in spoken language comprehension,'' J. Neurosci. **23**, 3423–3431.

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Wilson, J., Moss, H. E., Matthews, P. M., and Tyler, L. K. (**2000**). ''Susceptibility-induced loss of signal: comparing PET and fMRI on a semantic task,'' Neuroimage **11**, 589–600.

Dirks, D., and Bower, D. (**1968**). ''Masking effects of speech competing messages,'' J. Speech Hear. Res. **12**, 229–245.

Festen, J., and Plomp, R. (**1990**). ''Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing,'' J. Acoust. Soc. Am. **88**, 1725–1736.

Foster, J. R., Summerfield, A. Q., Marshall, D. H., Palmer, L., Ball, V., and Rosen, S. (**1993**). ''Lip-Reading the BKB sentence lists—corrections for list and practice effects,'' Br. J. Audiol. **27**, 233–246.

French, N., and Steinberg, J. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Griffiths, T. D., and Warren, J. D. (**2002**). ''The planum temporale as a computational hub,'' Trends Neurosci. **25**, 348–353.

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

MacLeod, A., and Summerfield, Q. (**1987**). ''Quantifying the contribution of vision to speech perception in noise,'' Br. J. Audiol. **21**, 131–141.

Moray, N. (**1959**). ''Attention in dichotic listening: Affective cues and the influence of instructions,'' Q. J. Exp. Psychol. **11**, 56–60.

Mummery, C. J., Ashburner, J., Scott, S. K., and Wise, R. J. S. (**1999**). ''Functional neuroimaging of speech perception in six normal and two aphasic patients,'' J. Acoust. Soc. Am. **106**, 449–457.

Neely, C. B., and LeCompte, D. C. (**1999**). ''The importance of semantic similarity to the irrelevant speech effect,'' Mem. Cognit. **27**, 37–44.

Paus, T., Zatorre, R. J., Hofle, N., Caramanos, Z., Gotman, J., Petrides, M., and Evans, A. C. (**1997**). ''Time-related changes in neural systems underlying attention and arousal during the performance of an auditory vigilance task,'' J. Cogn Neurosci. **9**, 392–408.

Rizzolatti, G., and Arbib, M. A. (**1998**). ''Language within our grasp,'' Trends Neurosci. **21**, 188–194.

Schroeder, M. R. (**1969**). ''Reference signal for signal quality studies,'' J. Acoust. Soc. Am. **44**, 1735–1736.

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

Scott, S. K., and Wise, R. J. S. (**in press**). ''Functional Imaging and Language: A Critical Guide to Methodology and Analysis,'' to appear in Speech Commun.

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

Treisman, A. M. (**1960**). ''Contextual cues in selective listening,'' Q. J. Exp. Psychol. **12**, 242–248.

Tremblay, S., Nicholls, A. P., Alford, D., and Jones, D. M. (**2000**). ''The irrelevant sound effect: does speech play a special role?'' J. Exp. Psychol. Learn. Mem. Cogn. **26**, 1750–1754.

Ulanovsky, N., Las, L., and Nelken, I. (**2003**). ''Processing of low-probability sounds by cortical neurons,'' Nat. Neurosci. **6**, 391–398.

Wise, R. J. S., Scott, S. K., Blank, S. C., Mummery, C. J., and Warburton, E. (**2001**). ''Identifying separate neural sub-systems within, ''Wernicke's area,'' Brain **124**, 83–95.

Wise, R., Chollet, F., Hadar, U., Friston, K., Hoffner, E., and Frackowiak, R. (**1991**). ''Distribution of cortical neural networks involved in word comprehension and word retrieval,'' Brain **114**, 1803–1817.

Zatorre, R. J., Bouffard, M., Ahad, P., and Belin, P. (**2002**). ''Where is 'where' in the human auditory cortex?'' Nat. Neurosci. **5**, 905–909.