Neural correlates of intelligibility in speech investigated with noise vocoded speech—A positron emission tomography study

Sophie K. Scott
Department of Psychology and Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, United Kingdom

Stuart Rosen and Harriet Lang
Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, United Kingdom

Richard J. S. Wise
MRC CSC, Hammersmith Hospital, Ducane Road, London W12 0NN, United Kingdom

(Received 13 August 2003; accepted 25 April 2006)

Functional imaging studies of speech perception in the human brain have identified a key role for auditory association areas in the temporal lobes (bilateral superior temporal gyri and sulci) in the perceptual processing of the speech signal. This is extended to suggest some functional specialization within this bilateral system, with a particular role for the left anterior superior temporal sulcus (STS) in processing intelligible speech. In the current study, noise-vocoded speech was used to vary the intelligibility of speech parametrically. This replicated the finding of a selective response to intelligibility in speech in the left anterior superior temporal sulcus, in contrast to the posterior superior temporal sulcus, which showed a response profile insensitive to the degree of intelligibility. These results are related to theories of functional organization in the human auditory system, which have indicated that there are separate processing streams, with different functional roles, running anterior and posterior to primary auditory cortex. Specifically, it is suggested that an anterior stream processing intelligibility can be distinguished from a posterior stream associated with transient representations, important in spoken repetition and working memory. © 2006 Acoustical Society of America. [DOI: 10.1121/1.2216725]

PACS number: 43.71.Rt, 43.64.Sj [PFA]

I. INTRODUCTION

The use of functional neuroimaging techniques that allow relatively precise anatomical localization [positron emission tomography (PET) and functional magnetic resonance imaging (fMRI)] has developed our understanding of the cortical basis of human speech perception, and the relationship with the functional neuroanatomy of audition. Work on the anatomy of the auditory cortex of the nonhuman primate (mostly in the rhesus macaque) has shown both parallel and serial (hierarchical) anatomical connections and neural processing in core, belt, and parabelt auditory regions (Rauschecker, 1998; Kaas and Hackett, 1999). This has been usefully applied to the findings from studies that have shown a similar architecture for sound processing in the human brain [e.g., Talavage et al., 2000, Hall et al., 2002; see Scott and Johnsrude (2003) for a review]. Importantly, these functional-anatomical studies have highlighted the rostral-caudal organization of the primate auditory cortex. In particular, the response to conspecific vocalizations more strongly activates temporal lobe regions lateral and anterior to the primary auditory cortex (Rauschecker, 1998). By contrast, Tian and colleagues demonstrated neurons posterior to primary auditory cortex that responded to the spatial localization of a vocalization but were largely insensitive to specific calls. These results, along with others, have led to the hypothesis that there are at least two “streams” of auditory processing: one directed anteriorly to identify the sound and another directed posteriorly that represents the spatial localization of the sound. Although this framework is not universally accepted (e.g., Middlebrooks, 2002), this functional-anatomical interpretation gains support from detailed anatomical connectivity studies in the macaque. These have shown segregation of superior temporal-prefrontal projections for anterior and posterior white matter tracts, which could form the basis for separate streams of auditory processing (Romanski et al., 1999).

In the current study, we aimed to elaborate the roles of the rostral and caudal auditory cortex in speech processing in the human primate. Functional imaging studies of speech perception have shown extensive involvement of bilateral superior temporal lobe regions, including the primary auditory cortex (PAC), and over the lateral superior temporal gyrus (STG), a region corresponding to the parabelt cortex in nonhuman primates (Binder et al., 2000). Activation typically extends down to the superior temporal sulcus (STS).
However, the neural activation measured in these studies (as indexed by regional cerebral blood flow) is always determined relative to a baseline condition, and the precise pattern of activation seen depends on the baseline condition selected. The selection of the baseline condition(s) can influence the degree to which the neural response to speech is left lateralized. Speech relative to silence (Wise et al., 1991, 1999) typically shows a response in the bilateral primary auditory cortex, as there is no control for general auditory stimulation. Studies that employ an auditory baseline condition, which attempts to control for some of the acoustic structure in the speech signal, have found that speech-specific responses are seen bilaterally along much of the length of the lateral STG and STS. This has been shown for human vocalizations contrasted with signal-correlated noise (Mummery et al., 1999) and its variants (Belin et al., 2000), where the amplitude envelope variation (but not the spectral detail) of the speech signal is matched in the baseline stimuli.

This apparent bilateral representation of speech has led to the claim that the processing of speech is mediated bilaterally, and that a left hemisphere specialization for language occurs later in the neurolinguistic system (Hickok and Poeppel, 2000). However, symmetry of activation in the left and right dorsolateral temporal lobes need not imply that both cerebral hemispheres are performing identical computations on the input. A recent study of phonological structure indicated a role for the left lateral STG in the processing of phonological structure in speech (Jacquemot et al., 2003) and a paper from our group indicated that neural responses as a function of intelligibility in speech (both normal and noise-vocoded) are also lateralized to the left temporal lobe (Scott et al., 2000).

Scott et al. (2000) contrasted two forms of intelligible speech (normal and six-channel noise-vocoded) (Shannon et al., 1995) against spectrally rotated equivalents (Blesser, 1972), in an attempt to use a baseline condition that more adequately controlled for the acoustic structure in speech. Spectral rotation preserves the overall pattern of spectro-temporal variation in speech (i.e., formants are clearly visible) as well as the original pitch contours of the speech, but the inversion in the spectral domain renders the speech unintelligible. Scott et al. (2000) demonstrated a left lateralized stream of speech processing, running lateral and anterior to the PAC, in which the processing of speech, noise vocoded speech, and rotated speech was observed in the bilateral STG and posterior STS and the processing of intelligible speech alone was seen more rostrally, in the anterior left STS. In contrast, responses in the right anterior STG/STS were greater to both speech-based stimuli (normal and rotated speech) than the two noise excited conditions, suggesting a preferential processing in right STG/STS for stimuli with dynamic, melodic pitch variation. This interpretation is supported by studies of the neural processing of pitch variation (Zatorre and Belin, 2001) and musical patterns (Patterson et al., 2002), both of which have demonstrated a predominantly right lateralized response to pitch variation. The responses of the left and right anterior STG/STS can thus be distinguished. In addition, these anterior responses were compatible with hypotheses, based on nonhuman primate research, that there is an anterior “what” stream of auditory processing (Rauschecker, 1998). The left lateralized, anterior response to intelligible speech has been replicated and extended using fMRI (Narain et al., 2003; Specht and Ruel, 2003) and has also been demonstrated in primate neural responses to consonant-specific vocalizations. (Poremba et al., 2004)

In contrast to this emphasis on anterior temporal lobe regions, many (but not all) clinical studies on speech perception, based on lesion-deficit analyses in aphasic patients, emphasize the left posterior STS as a region that maps sound on to meaning. In our previous PET study (Scott et al., 2000) this region did not show a speech-specific response, as it had the same stimulus response profile as the lateral STG, responding to speech, noise vocoded speech and rotated speech irrespective of intelligibility. We have suggested (Scott et al., 2000; Wise et al., 2001) that the posterior STS is involved in the short-term representation of sound input that might be mapped onto the production of words, rather than an acoustic-phonetic mapping onto lexical structures. Consistent with this potential distinction between anterior and posterior STS functions, small lesions of the left posterior STS result in conduction aphasia, where patients can understand spoken language, but cannot repeat words (Anderson et al., 1999; Quigg and Fountain, 1999). These findings suggest that it might be possible to identify different response profiles to acoustic stimuli between the lateral STG and posterior STS, which was not possible in our previous study (Scott et al., 2000), as both regions were equally activated by speech, noise vocoded speech, and rotated speech.

The current study was a replication and elaboration of our previous study. A parametric design was used, rather than the paired contrasts used previously. In parametric designs, a stimulus property is varied along a dimension in a number of steps, and regional cerebral blood flow (rCBF) changes that vary linearly with this variable are identified using a covariate analysis. This increases the sensitivity of the design, relative to functional imaging studies that rely on subtraction techniques. The number of channels in noise-vocoded speech (Shannon et al., 1995) was varied from 1 to 16, to give a dimension of intelligibility and permit a correlational analysis of the rCBF changes that covary with this manipulation. Spectral rotation was again used as a control for complexity, since increasing the number of channels in noise-vocoded speech necessarily increases the spectro-temporal detail. In addition, to compare responses in the anterior and posterior STG and STS, a contrast of conditions in which some (if not all) of the lexical items were intelligible, against those where little if any of the stimuli can be understood was planned, to test the hypothesis that the posterior STS might show sensitivity to any potentially relevant spectro-temporal detail in the signal, rather than to overall degree of intelligibility.

Positron emission tomography (PET) was used to index neural activity. A disadvantage of PET is that the total number of scans performed is limited by the dose of radioactivity administered, which means that single-subject analyses are not normally possible and that grouped data are typically presented. PET has, however, several advantages for such investigations. It is relatively quiet (relative to the high levels of acoustic noise generated in fMRI). Most critically, the
signal recorded in PET is equipotent across the cortex and subcortical regions, in contrast with fMRI, where the echo planar imaging (EPI) sequence normally used to measure the blood oxygen level dependent (BOLD) response is subject to signal loss due to artefact and geometric distortion (Devlin et al., 2000). This effect is most dramatic in the anterior temporal lobes, regions which are of critical importance in speech perception (Scott et al., 2000). Finally, the rCBF values which are measured to index neural activity are linear with respect to activity, unlike the BOLD response typically measured in fMRI (Dhankhar et al., 1997).

II. METHOD

A. Stimuli

Speech was noise vocoded to 1, 2, 3, 4, 8, and 16 channels, over the frequency range 70 Hz to 4 kHz using essentially the technique described by Shannon et al. (1995). The input waveform was passed through a bank of 1, 2, 4, 8, or 16 analysis filters (each a sixth-order Butterworth IIR, with three orders per upper and lower side) with frequency responses that crossed 3 dB down from the pass-band peak. Envelope detection occurred at the output of each analysis filter by full-wave rectification and second-order Butterworth low-pass filtering at 20 Hz. These envelopes were then multiplied by a white noise, and each filtered by a sixth-order Butterworth IIR output filter identical to the analysis filter. The rms level from each output filter was then set to be equal to the rms level of the original analysis outputs, before being summed together. Cross-over frequencies for both the analysis and output filters were calculated using an equation (and its inverse) relating position on the basilar membrane to its best frequency (Greenwood, 1990).

Spectrograms of the stimuli are shown in Fig. 1.

The control for complexity was noise-vocoded spectrally rotated speech. The original (low-pass filtered) speech was spectrally rotated around 2 kHz (Blesser, 1972), filtered so as to have the same long-term average spectrum as ordinary speech, then noise vocoded with 3 and 16 channels. These two conditions provide a control for neural activity that has an increase in spectro-temporal complexity without an increase in intelligibility. The stimuli were the BKB (Foster et al., 1993) and IHR (MacLeod and Summerfield, 1987) sentences, spoken by an adult British English speaking male. These are short concrete sentences with three or four key words.

B. Participants

There were seven right-handed participants (six men and one woman), all with normal hearing. The average age was 38 (range 30–62). All the participants were able to understand the noise-vocoded speech after pretraining and in later testing (see Sec. IV). All gave written informed consent, 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).

Participants were pretrained on five channel noise-vocoded speech (NVC speech), by presenting them with a sentence and asking them to repeat back what they heard (these sentences were not included in the later scanning and intelligibility testing). If participants were incorrect, then they were given feedback and played the stimulus again until all sentences could be repeated correctly on the first presentation. This took fewer than 20 sentences for each participant. A different number of channels was used for pretraining to ensure that all the noise-vocoding conditions during scanning were novel (i.e., that no one condition was more familiar than the others), while also providing training that fell within the mid range of the overall number of channels presented.

C. PET scanning

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. There were two scans for each stimulus condition, 1-, 2-, 3-, 4-, 8-, and 16-channel noise-vocoded speech and 3- and 16-channel rotated noise-vocoded speech. Stimulus presentation was for 1 min, throughout which the sentences were played continuously. The stimulus conditions were presented in a random order. The loudness was set 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 “for meaning” in the scanning sessions. Passive listening reduces the likelihood that the observed activations are due to task-specific controlled processing, 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). Thus, this study focused on the implicit, automatic mechanisms of speech perception. After the completion of scanning, each subject was tested on each condition, using a measure of the number of key words repeated correctly (16 sentences were presented per condition, scored as number...
correct out of 50 key words). The sentences used were different for each testing condition and different from those used in training.

III. ANALYSIS

The images were analyzed using statistical parametric mapping (SPM99b, Wellcome Department of Cognitive Neurology, http://www.fil.ion.ucl.ac.uk/spm). All scans from each subject were realigned to eliminate head movements between scans and normalized into a standard stereotactic space. Images were then smoothed using an isotropic 10-mm, full width half-maximum, Gaussian kernel, to allow for variation in gyral anatomy and to improve the signal-to-noise ratio. Specific effects were investigated, voxel-by-voxel, using appropriate contrasts to create statistical parametric maps of the t statistic, which were subsequently transformed into Z scores. The analysis included a blocked analysis of covariance (AnCova) with global counts as a covariate to remove the effect of global changes in perfusion across scans. 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) [further details of this analysis technique are given in Scott et al. (2004)].

The main statistical analysis of the rCBF data aimed to identify the rCBF changes that correlate with the number of channels in NVC, but not with the rotated NVC stimuli. Intelligibility is most simply related to the logarithm of the number of channels (Faulkner et al., 2001). The use of the logarithmic transform was supported by a logistic regression on the data presented in Fig. 2. With a single continuous predictor variable, using the logarithm of the number of channels led to a much better fit than the use of the number of channels directly. Therefore, we modeled for rCBF changes that correlate positively with the log number of channels and masked inclusively for regions where the two rotated conditions (16R and 3R) were significantly less responsive than their unrotated equivalents $(3+16 > 3R +16R)$. The use of inclusive masking was an attempt to identify voxels that are both more activated by unrotated NVC speech than by rotated speech, and which also show a sensitivity to the log number of channels in the noise vocoded speech. This contrast does not, however, explicitly exclude voxels in which there is a difference between the rotated conditions, i.e., the possibility remains that the analysis does not rule out voxels driven solely by acoustic complexity. A parallel analysis was therefore used in which the voxels that increased with the log number of channels were identified, but an exclusive mask was used to remove voxels which were sensitive to the differences between the 1, 3R, and 16R stimuli $[16R > 3R > 1]$, i.e., which were driven by increases in acoustic complexity. This analysis was thus a second attempt to identify voxels which correlate with the number of channels, but which are not solely responding to acoustic properties of the stimuli. The SPM package does not allow the simultaneous application of multiple masking conditions, so these two analyses were performed separately. The results of these parallel analyses are shown in Table I.

To test the hypothesis that the posterior STS might be responding to the conditions where there was the potential for intelligible speech, regardless of the actual degree of intelligibility, a contrast was planned comparing all the conditions where there was some intelligibility (3, 4, 8, and 16 channels) over those where there was little or none (1, 2, 3R, and 16R). Exclusive masking was used to exclude neural responses which were significantly activated overall by increasing numbers of channels. This subtraction contrast was designed therefore to reveal responses which are sensitive to a degree of spectro-temporal structure, sufficient for some speech perception, but not sensitive to specific increases in the number of channels.

IV. RESULTS

The behavioral data are shown in Fig. 2. This clearly demonstrates a relationship between the number of channels and the intelligibility of the speech, consistent with previous studies (Shannon et al., 1995). Likewise, the intelligibility of the rotated sequences was negligible, as in previous studies (Scott et al., 2000; Narain et al., 2003). Figures 3 and 4 and Table I show the neural correlates of increasing numbers of channels of NVC speech in the input, controlling for overall spectro-temporal complexity (by inclusive masking for $(3+16 > 3R+16R)$). rCBF increases can be seen in both temporal lobes, over the lateral STG, running lateral and anterior to the primary auditory cortex (PAC). In the left temporal lobe, however, the response to intelligibility passes rostrally down the temporal lobe, extending as far as the temporal pole. There was a peak in the left lateral superior temporal gyrus (STG), the left STS, and the left anterolateral temporal pole. The peak in the right temporal lobe was in the antero-lateral STG. The plots for the rCBF changes for each peak are shown in Figs. 4(a)–4(d). There is some sensitivity to acoustic complexity in the left lateral STG/STS peaks [Fig. 4(b) and 4(c)]; the response to 16R is greater than that to 3R, but the response to...
the unrotated stimuli is greater still (indeed, no contrast in this study identified peak responses where the response to the unrotated stimuli was the same or smaller than to the rotated stimuli). When exclusive masking for acoustic structure (16R > 3R > 1) is used with the intelligibility contrast, the more posterior STG peak is not seen, confirming that it is sensitive to acoustic structure (Table I). The left temporal pole response shows no difference between 3R and 16R, suggesting insensitivity to acoustic complexity [Fig. 4(d)]. The left anterior STS peak shows a “midway” response: the pattern of activation follows the behavioral relationship between the number of channels and intelligibility, and the peak is not excluded by the 16R > 3R > 1 mask for acoustic sensitivity. This peak does, however, show some acoustic sensitivity to the difference between the 16R and 3R conditions [Fig. 4(c)].

<table>
<thead>
<tr>
<th>Region</th>
<th>Z score</th>
<th>MMN coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right antero-lateral STG</td>
<td>5.96</td>
<td>64 −4 −2</td>
</tr>
<tr>
<td>Left lateral STG</td>
<td>4.52</td>
<td>−64 −28 8</td>
</tr>
<tr>
<td>Left anterior STS</td>
<td>5.6</td>
<td>−62 −10 0</td>
</tr>
<tr>
<td>Left temporal pole</td>
<td>4.73</td>
<td>−48 16 −16</td>
</tr>
</tbody>
</table>

Contrast 2: peaks increasing with log number of channels, masking exclusively for regions where (16R > 3R > 1)

<table>
<thead>
<tr>
<th>Region</th>
<th>Z score</th>
<th>MMN coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right lateral STG</td>
<td>5.36</td>
<td>64 −6 −4</td>
</tr>
<tr>
<td>Left anterior STG/STS</td>
<td>5.01</td>
<td>−64 −8 4</td>
</tr>
<tr>
<td>Left temporal pole</td>
<td>4.73</td>
<td>−48 16 −16</td>
</tr>
</tbody>
</table>

FIG. 3. Regions where activity increased with the number of channels (expressed logarithmically), masked inclusively for regions in which the unrotated stimuli were significantly higher than their rotated equivalents (voxel-level significance $P < 0.0001$, unconnected for whole brain analysis). In the panels on the left, the activation is superimposed on coronal slices of an average T1 weighted MRI image (the Montreal Neurological Institute template, used in SPM99). The panel on the top left shows the most posterior activation, and the subsequent panels show the activation at slices, 0.5 cm apart, moving in an anterior direction. These panels show the activation running along the left STS, extending into STG in a posterior direction and into the medial and ventral temporal lobe in the anterior direction. The numbers below each panel refer to the distance along the y plane of MNI coordinates. The panels also show right STS activation in the slices between −15 and 0 mm. The panels on the right show the same activation, this time projected onto a “glass brain;” this lacks anatomical detail but shows all the activations on the same figure. Panel (a) shows a sagittal view, with the brain viewed from the side; panel (b) shows a coronal view, with the brain viewed from the back, and panel (c) shows an axial view, with the brain viewed from above.
The peak in the right STG is principally driven by differences between the single channel condition and the higher channel conditions, i.e., by some degree of spectral structure and variation [Fig. 4(a)]. There is also an elevated response to the 16-channel condition. If the contrast is repeated, with an exclusive mask for regions showing some sensitivity to spectral structure (e.g., \([2, 3, 4, 8, 16, 3R, \text{and } 16R]\) > 1), then the right STG activation is not seen. In contrast, the peaks in the left temporal lobe, which vary with the number of channels in a more linear fashion (though not with the rotated conditions), remain unchanged by this mask.

The planned contrast of conditions with sufficient spectro-temporal structure for some comprehension (3, 4, 8, and 16) over unintelligible conditions (1, 2, 3R, and 16R), with exclusive masking for regions that respond to increasing number of channels, shows a purely left lateralized response (Fig. 5). The activation runs posterior and lateral to the PAC, into the posterior STS. There are peaks at \(x = -60\), \(y = -38\), and \(z = 4.20\).
and the plot for the latter is shown in Fig. 5. These peaks reflect activations which are not solely driven by increasing numbers of channels.

V. DISCUSSION

This study has thus presented strong evidence that, as in our previous study, intelligible speech is (predominantly) processed in the left anterior temporal lobe, and that there is a rostrally directed stream of processing of the speech signal. Importantly, there is a sensitivity to increases in the number of channels that are associated with intelligibility in regions lateral to PAC, in very similar regions to those which respond to amplitude modulations (contrasted with unmodulated noise) in fMRI studies (Giraud et al., 2000; Hart et al., 2003). Hart and colleagues specifically identified regions in the lateral STG, adjacent to the PAC, that respond selectively to amplitude modulations in the speech AM range. This lateral STG region shows sensitivity to phonological structure in speech (Jacquemot et al., 2003), and in the current study the response profile indicates some specialization for speech specific information, although it is also sensitive to aspects of acoustic structure. This suggests that the neural regions involved in processing amplitude modulation are also important initial stages in the processing of the speech signal. Descriptively, there is also the appearance of hierarchical processing of the signal (Scott and Johnsrude, 2003): the most posterior peak in the left STG shows a linear relationship between the number of channels from 1 to 8 (but not to the rotated stimuli), whereas the more anterior peaks (STS, temporal pole) have profiles more similar to the intelligibility data, where 1 and 2 channels, and 3 and 4 channels activate this area to a similar degree.

This study also revealed a response to intelligible speech in the right lateral STG, not seen in our previous study. This may be a consequence of the use of a more sensitive parametric design. There are three other potential factors which may affect this result. There is an enhanced response in the right lateral STG to all the NVC conditions relative to the single channel condition—that is, there is increased activation for each condition with some spectral structure, relative to the one channel condition. This suggests that the right STG is sensitive to changes in spectral structure, or perhaps to the pitch sensation consequent upon this. There is also an enhanced response in the right lateral STG to the 16-channel condition. Since the intelligible speech with 16 channels sounds recognizably like the original speaker [and informal listening suggests those up to 8 do not (Warren et al., under review)], right temporal lobe mechanisms involved in speaker recognition might be activated along with increased intelligibility (Van Lancker et al., 1989). It is also possible that, since the increased number of channels leads to a stronger sense of pitch (Xu et al., 2002), the activation may reflect processing of the variation in pitch (due to the original speech intonation), which is known to lead to greater right STG/STS involvement (Patterson et al., 2002). An effect of pitch variation should be attenuated by use of rotated stimuli as a baseline, since these preserve dynamic pitch variation. However the rotated stimuli do evoke a weaker pitch sensation than the unrotated stimuli. Further studies will allow us to clarify the reasons for this finding.

The use of a parametric design allows us to draw a distinction within the left temporal lobe between the lateral STG and the posterior STS. In our previous study (Scott et al., 2000), these both responded to speech, rotated speech, and NVC speech equally. Lateral STG, like the anterior temporal lobe, shows a sensitivity to the number of channels in the NVC speech and a reduced sensitivity to the rotated stimuli, consistent with these regions forming an auditory “what” pathway [Figs. 4(b) and 4(c)] (Rauschecker, 1998; Scott et al., 2000; Scott and Johnsrude, 2003). The posterior STS, however, is activated equally by the 3-, 4-, 8-, and 16-channel NVC speech conditions, despite considerable intelligibility differences across these conditions.

The left posterior STS therefore does not appear to be part of the same functional system as the antero-laterally directed pattern of hierarchical acoustic processing (Scott and Johnsrude, 2003; Wise et al., 2001). Unlike the anterior STG, the posterior STS does not receive direct monosynaptic input from primary auditory areas, being at least two synapses away (Galuske et al., 1999). Left posterior STS has been identified as having an important role in the rehearsal or production of spoken repetition (e.g., Wise et al., 2001), and in the representation of sequences in auditory short-term memory rehearsal (Buchsbaum et al., 2001). The posterior STS is activated by some acoustic stimuli (Mummery et al., 1999; Hall et al., 2002; Giraud and Price, 2001) and is associated with plasticity in the recovery of speech perception after aphasic stroke (Left et al., 2002). We have speculated previously that the posterior STS will show sensitivity to stimulus properties in potentially pronounceable sequences (Scott et al., 2000), where the stimuli with some potential were speech, NVC speech, and rotated speech. Rotated speech is not necessarily pronounceable per se, but it preserves the temporal structure of the speech and thus the overall suprasegmental structure of the speech, and can indeed be understood after lengthy training (Blesser, 1972). The left posterior STS may therefore be responding to the saliency of this speechlike signal rather than to its overall intelligibility. The current study supports this position: the peak in the posterior STS is revealed by contrasting the four conditions in which there were some intelligible items with those in which there were none. The peak does not, however, show sensitivity to the number of channels and consequent variation in intelligibility. Thus the posterior STS shows a sensitivity to aspects of the acoustic structure of the sound sequences and the potential for meaningful structure to emerge from this. Such a response would be predicted by models which represent the acoustic structure of incoming sequences in some form of “buffer,” for example the syllable processing model of Hartley and Houghton (1996). In this model, based on the errors made in word and nonword repetition tasks, the incoming speech signal is represented in a manner consistent with the underlying phonological structure of that language at the syllable (i.e., suprasegmental) level, with an emphasis on the sonority profile and the onset/rhyme distinction. If the posterior STS does play an important part in the processing.
of such structure, then it would not be expected to simply detect intelligibility *per se*, but more structural, suprasegmental elements that relate the incoming sound to how it might be spoken, and which are preserved in NVC speech with only a few channels.

Speech perception can also lead to activation in more posterior STS/inferior parietal cortex. We have recently replicated the original Scott *et al.* (2000) study with fMRI (Narain *et al.*, 2003) and found, in addition to a left anterior STS response, intelligibility-related activation which lies at the junction of the STS and in the left inferior parietal lobe, posterior (\(y=-.64\)) to the caudal STS activation described in this paper (\(y=-.44, -.38\)) and that in our earlier PET study (\(y=-.38\)) (Scott *et al.*, 2000). The role of this more elusive posterior activation, and its relation to semantic processing, will be elaborated in further studies.

A recent fMRI study (Davis and Johnsrude, 2003) also used varying numbers of channels in NVC speech resulting in low, medium, and high intelligibility. In their study three different manipulations of intelligibility were constructed: NVC speech of 4, 8, and 15 channels, varying the signal-to-noise ratio of speech in noise and interrupting sentences with random sections of signal correlated noise (Schroeder, 1969). In addition to extensive pilot testing, performance during scanning was measured on line by using an explicit intelligibility judgment (not subsequently used for analysis). The contrast of increasing intelligibility (from the pilot testing) revealed a distributed system running anterior and posterior to PAC, into the ventral temporal lobes and prefrontal cortex. Evidence was found for a distinction between form-dependent processes (those that showed a response dependent on a particular manipulation condition) and form-independent processing (responses which were intelligibility specific but not sensitive to the type of stimulus manipulation). The design, however, included no formal control for the acoustic structure of the stimuli, which varied greatly both across manipulations and with intelligibility. Thus this study was not able to distinguish different patterns of activation within the hierarchical pattern of responses seen. For example, some regions in bilateral STG were reported to show a mixture of form-dependent and -independent responses: a control for acoustic aspects of the stimuli would have enabled these responses to be more clearly described. Without this, the “form-independent” regions are only independent of the form variations included in the study itself. This design also assumes that there are no important differences between the neural resources recruited by the three types of speech manipulation. There is some evidence that this assumption may be incorrect. We have previously demonstrated that the perception of speech in noise (energetic masking) is associated with activity in the prefrontal and parietal cortex, and that SNR level-independent responses are seen in the left ventral prefrontal and supplementary motor regions (Scott *et al.*, 2004). These regions are not activated by the manipulation of channel number in the current study, suggesting that there is not a complete overlap of the neural regions associated with these two manipulations of intelligibility. Furthermore, as discussed in the Introduction, the BOLD response typically measured in fMRI is associated with severe signal loss and distortion in the anterior temporal lobes. The most anterior peak reported for the intelligibility contrast in Davis and Johnsrude’s study had a \(y\) coordinate of \(-34\), and the most anterior peak associated with intelligibility in the current study is \(+16\). This 5-cm difference may well be the result of signal loss in these more anterior regions arising from the use of fMRI.

Our current study of noise vocoded speech differs in several ways. Our use of a passive paradigm ensures that the activations seen will be associated with auditory processing of the input, rather than with explicit decision-making aspects of the task. It would be interesting to determine the extent to which the prefrontal activations seen in the Davis and Johnsrude (2003) paper would remain if a passive task were used. Our use of a wider range of channels also enables us to delineate brain regions which respond to increases across a wider range of NVC speech conditions, with special reference to the lower number of channel conditions, where intelligibility varies greatly. The smallest number of channels used by Davis and Johnsrude (2003) was four. From Fig. 2, and the plots in Fig. 4, we can identify both behavioral and neural responses that are sensitive to variation in the numbers of channels from one to four. Our use of two rotated NVC speech conditions enables us to explicitly distinguish between neural responses to auditory structural complexity changes and those which are associated with intelligibility. The use of PET enables us to identify peaks of activity that lie in very anterior temporal lobe regions (Table I, Figs. 3 and 4).

These factors explain why we are able to identify neural responses, associated specifically with intelligibility, in anterior temporal lobe regions not reported by Davis and Johnsrude (2003). Our design also permits the delineation of a difference between anterior and posterior auditory association cortical responses, consistent with the differing functions ascribed to these (Scott and Johnsrude, 2003; Wise *et al.*, 2001). This is also a consequence of the use of a passive design: if the posterior STS is important in transient representations of auditory inputs, then its role might be expected to change in an explicit, decision making task, as outlined by Buchsbaum and colleagues.

Thus we have argued for a distributed neural system involved in the processing of intelligible speech, which divides into two broad streams of processing: an anterior lateral stream associated with the acoustic-phonetic, semantic, and syntactic information which contributes to meaning in the speech signal (but not to all acoustic structure) and a posterior STS region sensitive to any potentially relevant structure in the incoming signal, regardless of the specific degree of acoustic detail. This is broadly within the parameters of the known neuroanatomy of the primate auditory cortex. Of course both anterior and posterior routes converge in prefrontal and premotor cortical regions (reviewed in Scott and Johnsrude, 2003) to form an interacting network.

**ACKNOWLEDGMENTS**

SKS and RJSW are both supported by grants from the Wellcome Trust. Many thanks to Sam Eaton-Rosen for help in creating Fig. 1.