The Anterior Pathway for Intelligible Speech: Insights from Univariate and Multivariate Methods

Samuel Evans

Institute of Cognitive Neuroscience, UCL

Thesis submitted for the degree of Doctor of Philosophy

University College London, September 2011

Primary supervisor: Professor Sophie Scott

Secondary supervisor: Professor Stuart Rosen

Declaration:

I, Samuel Evans, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis

Abstract:

Whilst there is broad agreement concerning the existence of an anterior processing stream in the human brain concerned with extracting meaning from speech, there is an ongoing controversy as to whether intelligible speech is first resolved in left anterior or bilateral posterior temporal fields (Hickok and Poeppel, 2007;Rauschecker and Scott, 2009). Proponents of the bilateral processing model argue that bilateral responses are driven by the acoustic properties of the speech signal, whilst proponents of the left lateralised model suggest that left lateralisation is driven by access to linguistic representations. This thesis directly addresses these controversies using Functional Magnetic Resonance Imaging (fMRI) and univariate and multivariate analysis methods. Two main questions are addressed: (1) where are responses to intelligible, and intelligible but degraded speech, separated from responses to acoustic complexity and (2) does the resulting pattern of lateralisation, or otherwise, derive from the acoustic properties or the linguistic status of speech. The results of this thesis reconcile, to some degree, the two theoretical positions. I show that the most consistent and largest amplitude responses to intelligible, and degraded but intelligible speech, are found in the left anterior Superior Temporal Sulcus (STS). Additional responses were also found in right anterior and left posterior STS, however, these were less consistently identified. Regions of the left posterior STS showed sensitivity to resolved intelligible speech and also showed a response likely to reflect acoustic-phonetic processing supporting the resolving of intelligibility. Right posterior STS responses to intelligible speech were noticeably absent across all studies. No evidence was found for a relative acoustic basis for hemispheric lateralisation in the case of speech derived manipulations of spectrum and amplitude, but evidence was found in support of a left hemisphere specialism for resolving intelligible speech, supporting a relative left lateralisation to speech driven by linguistic rather than acoustic factors.

Acknowledgments:

I'd like to say a big thank you to Sophie Scott and Stuart Rosen. I couldn't have asked for better supervisors and I feel very privileged to have worked closely with them both during the course of my PhD. They have been a wonderful source of information, advice and enthusiasm, and most importantly they have been great fun to work with too.

Also a big thanks to Alex Leff and Janaina Mourão-Miranda for answering a multitude of technical questions and to Narly Golestani for help in defining ROIs. I'd also like to thank Carolyn McGettigan and Zarinah Agnew for their help and advice, and the fun times shared in our lab.

Finally the biggest thank you goes to my wife, Abi, and my daughter, Amelia, who have put up with me during the last three years. I promise that you'll see much more of me now and I am really looking forward to spending more time with you both. This is my last PhD I promise!

Contents

Figures

Tables

Abbreviations

AG	Angular Gyrus
BMA	Bayesian Modelling Averaging
BOLD	Blood Oxygenation Level Dependent (response)
BMS	Bayesian Model Selection
DCM	Dynamic Causal Modelling
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
GLM	General Linear Modelling
HG	Heschl's Gyrus
IOG	Inferior Occipital Gyrus
ITG	Inferior Temporal Gyrus
LDA	Linear Discriminant Analysis
LTASS	Long Term Average Speech Spectrum
MFG	Middle Frontal Gyrus
MTG	Middle Temporal Gyrus
PT	Planum Temporale
PAC	Primary Auditory Cortex
RFE	Recursive Feature Elimination
RMS	Root Mean Squre
SMA	Supplementary Motor Area
SNR	Signal to Noise Ratio
SPM	Statistical Parametric Map
STG	Superior Temporal Gyrus
SVM	Support Vector Machine
VOI	Volume of Interest

Chapter 1 : INTRODUCTION

1.1 The complexity of the speech signal

Spoken languages consist of an inventory of phonemic units which function contrastively to construct linguistic meaning. The speech signal, which carries this phonemic code, is best understood by a source filter-model. The sound source has a regular harmonic structure in the case of voiced sounds or is aperiodic when generated by turbulent airflow in their voiceless equivalents. This source signal is shaped by resonances in the vocal tract which are generated by articulatory gestures, these resonances amplify and attenuate particular frequency components that compose the source signal, giving rise to a complex spectro-temporally varying signal. This speech signal is incredibly acoustically complex; composed of modulations in amplitude and frequency, durations of periodicity and aperiodicity, silence and excitation, and spectral structure such as harmonics and formants, relative spectral prominences in collections of harmonics.

It is possible to demonstrate by synthesizing speech tokens that differ along an acoustic dimension, that a single acoustic cue, such as the trajectory of the formants, are sufficient to distinguish between two or more speech sounds (Liberman et al., 1957). Ordinarily however multiple cues are shown to support the perception of phonetic features (Lisker, 1977) and these cues are shown to interact (Theintun, 1987;Kluender and Alexander, 2010). The relationship between phonemes and their surface acoustic form is complex and inherently context dependent. For example, the duration of the vowel in the words “bead” and “beat” differ as a function of whether the consonant in coda position is voiced or voiceless. Further the same utterance spoken by different individuals will vary due to differences in vocal tract size and accent, and even the same utterance spoken by the same individual will vary from instance to instance due to differences in speaking style. However despite the lack of an exact one to one relationship between acoustic form and phonemic representation, often termed the “acoustic invariance” problem, individuals effortlessly understand speech.

Manipulating speech intelligibility by removing or distorting acoustic information provides a window into the processes underlying speech perception. Such research demonstrates that no single acoustic feature underlies intelligibility; indeed speech can be degraded in a number of different ways whilst still maintaining intelligibility. For example, subjects can learn to understand speech when spectral detail has been greatly reduced (Shannon et al., 1995;Remez et al., 1981) and when frequency information is shifted to mismatch expectations (Eisner et al., 2010), in addition to when temporal information is severely distorted (Dupoux and Green, 1997) and when speech is masked by other sounds (Brungart, 2001). The fact that there is a multiplicity of acoustic cues available in the speech stream that are sufficient, but not necessary, for intelligibility is referred to as cue redundancy. It is this redundancy that makes speech perception so robust, likely allowing us to exploit co-varying information from multiple acoustic sources (Scott and Evans, 2010;Kluender and Alexander, 2010).

The inherent complexity and lack of invariance in the speech signal led researchers in the 1960s to suggest that speech is in some way privileged or special compared to other sounds. One key finding that contributed to this conclusion was the discovery that speech sounds were perceived categorically. Categorical perception is a phenomena described when subjects hear a continuum of sounds which differ in equal acoustic steps and abruptly report a change in the identity of the speech sound at a specific point on the continuum, rather than exhibiting a smooth identification function that mirrors the acoustic distance between the speech tokens (Liberman et al., 1957). The strength of the conclusions drawn from this phenomenon have since been qualified by the finding that categorical perception is not specific to speech and is also true of non-speech sounds (Pisoni, 1977), occurs in other animals (Kluender et al., 1987) and is strongly influenced by the way a task is presented (Schouten et al., 2003). The strong view of this argument has been transferred to auditory neuroscience research with the suggestion that speech is processed differently to other sounds at the earliest stages of processing (Benson et al., 2001). A more moderate, and widely held view, is that the neural substrates involved in processing speech and other sounds are shared up to a certain stage in processing at which point they diverge (Rauschecker and Scott, 2009;Hickok, 2009), this is the perspective presented in this thesis. The earliest point at which processing diverges is debatable; it is clear however that processing must diverge at some stage as speech uniquely interfaces with linguistic

representations. A central theme within this thesis is the identification of neural regions which respond specifically to speech as contrasted with acoustic complexity.

1.2 Neural Basis of Speech Perception – Hierarchical Structure & Multiple Streams

Acoustic information is encoded by the pattern of firing in the auditory nerve caused by the shearing of inner hair cells against the tectorial membrane within the organ of corti residing in the cochlea. The auditory nerve projects to the cochlea nucleus, and subsets of ascending fibres cross to the contralateral superior olive and inferior colliculus, and other fibres synapse on the ipsilateral superior olive. Projections from the superior olive are directed through the lateral lemniscus, reach the inferior colliculus, and continue through the medial geniculate nucleus of the thalamus to primary auditory cortex (PAC). By the time acoustic information reaches the cortex it has been highly processed and re-coded by subcortical structures (Patterson and Johnsrude, 2008). It is commonly assumed that the processing occurring in subcortical regions is general to all sounds and that speech specific processing does not emerge until the level of the cortex (Scott and Johnsrude, 2003).

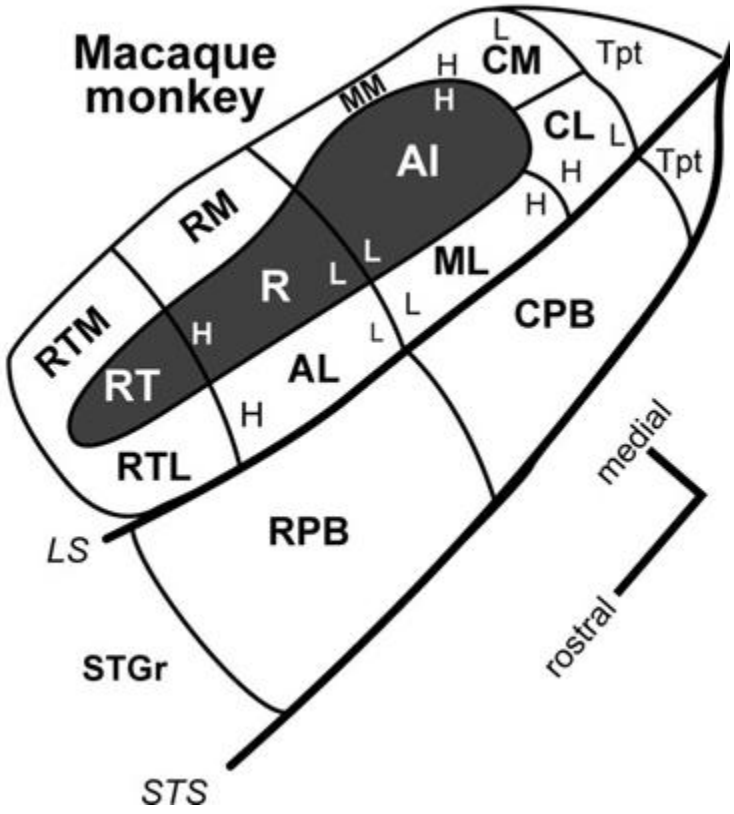
Much has been learnt concerning human auditory processing from the study of non-human primates. The advantage of conducting research with non-human primates is that invasive methods can be conducted much more readily, although there is obviously a limit in how far analogies between humans and monkeys can be taken with respect to language processing. Many of the insights gained using these methods have since been confirmed with functional imaging in humans. One such fundamental insight is the observation that auditory information is processed hierarchically in the auditory system. In the monkey three core primary auditory areas have been identified: RT, R and A1. These regions receive dense parallel thalamic input and can be delineated from each other by their tonotopic response and cytoarchitectonic structure. They are surrounded by eight belt regions; the four lateral belt regions can be differentiated by their physiological response to frequency

reversals (Romanski and Averbeck, 2009). These lateral belt regions are bordered by a parabelt region which has rostral and caudal divisions (see Figure 1.1).

The delineation of three separate levels across core-belt-parabelt can be identified architectonically by a stepwise reduction in staining for parvalbumin, acetylcholinesterase, and cytochrome oxidase along the core-belt-parabelt axis, with staining heaviest in the core, moderate in the belt, and lightest in the parabelt (Hackett, 2011). A hierarchy is inferred from the observation that belt regions are densely connected to both core and parabelt, whereas only sparse connections link core to the parabelt suggestive that the majority of information is transmitted serially from core to belt to parabelt (Kaas and Hackett, 2000). The functional response of the different regions can also be differentiated. For example, whilst core regions respond more vigorously than belt regions to pure tones, belt responds best to more complex stimuli like narrow band noise (Kaas and Hackett, 2000). Neural responses also become less well tuned to surface acoustic structure with distance to core regions. For example frequency tuning bandwidths and response latencies increase, and temporal precision decreases with poorer entrainment shown to amplitude modulation, as one moves along this axis (Hackett, 2011). This is suggestive that information is transformed along the pathway perhaps reflecting a greater abstraction from surface acoustic structure in later areas.

Figure 1.1 Schematic of the auditory cortex of the Macaque.

Reproduced with permission: Hackett TA (2011) Information flow in the auditory cortical network. Hearing Research In Press, Correct Proof.

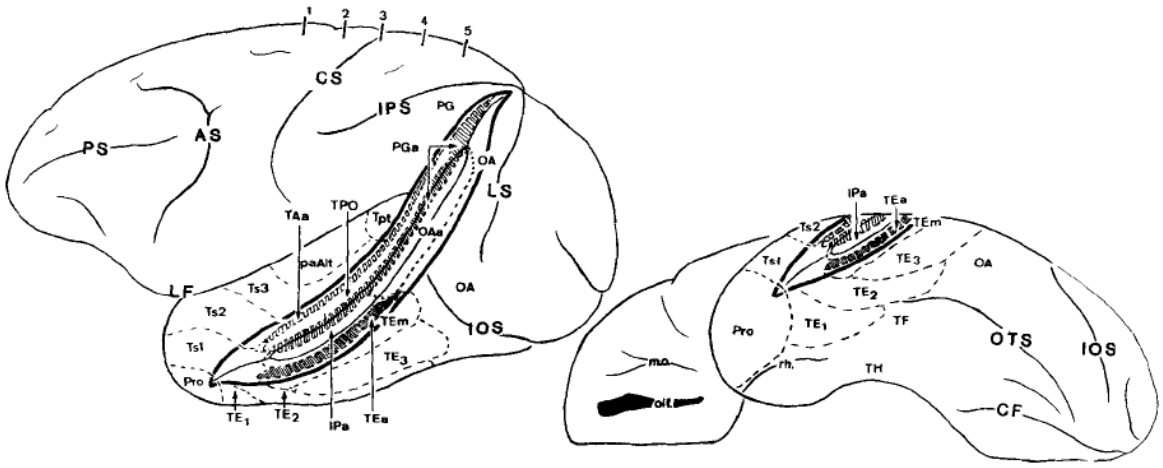


A second fundamental insight has been the identification of discrete processing streams which begin in auditory cortex and extend to frontal cortex. Each core region is most densely connected with its adjacent belt. Lesion studies show that information flows from core area AI to medial belt area CM, and that these connections are separate to those connecting core area R to belt regions RM and AL (Romanski and Averbeck, 2009). This is suggestive that information flows in parallel separate streams from core to belt, with each core region receiving largely independent afferents from the thalamus. Belt regions maintain this topographic connectivity, with for example, caudal lateral belt areas ML and CL most densely connected to caudal parabelt, and rostral lateral belt AL and rostral ML likewise connected to rostral parabelt. Rostral parabelt has been shown to connect to rostral STG areas Ts1 and Ts2, whilst caudal parabelt connects to area TpT located on the caudal STG

(Figure 1.2, note in this scheme, paAlt, TS3 and TpT be thought to correspond to belt and parabelt areas, Seltzer & Pandya (1989)).

Figure 1.2 Schematic of the Macaque temporal lobe

Reproduced with permission, Seltzer B, Pandya DN (1989) Intrinsic connections and Architectonics of the Superior Temporal Sulcus in the Rhesus Monkey. Journal of Comparative Neurology 290:451-471.

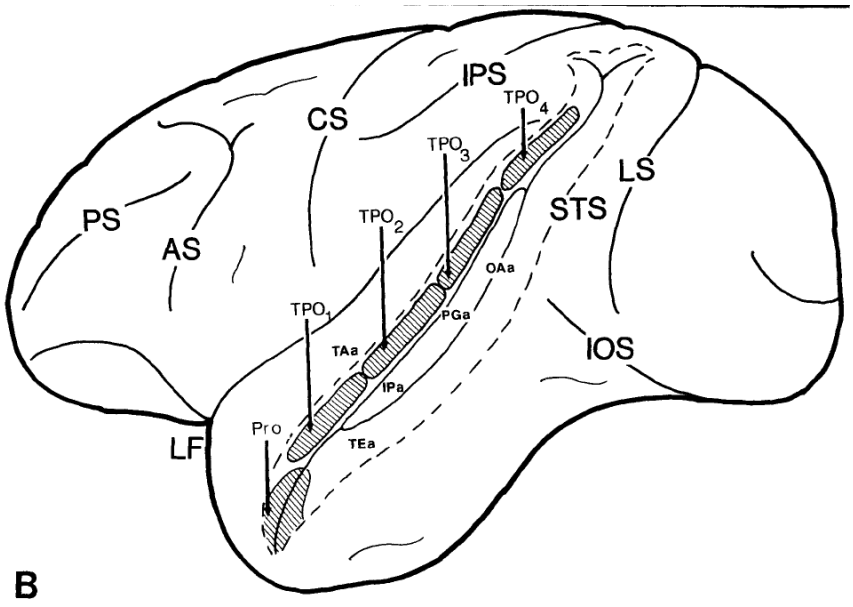


The upper bank of the STS consists of three zones: Area TAa, TPO and PGa(Seltzer and Pandya, 1989). Area TAa lies within the upper bank of the sulcus and only receives input from the superior temporal gyrus (STG), indicating that is likely to have a unimodal auditory function. Area TPO is medial to TAa, to which it is reciprocally connected, and is a multimodal region which receives input from auditory, visual and somatosensory areas. TPO consists of as many as four rostral to caudal architectonic subdivisions (TPO1-4) running the length of the sulcus (see Figure 1.3). These are serially interconnected via feedforward and back connections, with links also between non adjacent subdivisions; TPO1 and TPO2 are linked to the temporal pole (area Pro), and each TPO is in turn reciprocally connected to medially adjoining PGa. PGa is situated medially at the junction with the depth of the sulcus and caudally expands to occupy almost the entire upper bank of the sulcus.

The rostral-caudal distinction is maintained beyond temporal cortex in the connections to frontal cortex, with rostral frontal cortex densely connected to rostral belt and parabelt, and caudal belt and parabelt reciprocally connected with caudal and dorsolateral prefrontal frontal cortex (Romanski and Averbeck, 2009). The rostral-caudal patterns of anatomical connectivity have been shown to support largely separate functional streams, with neurons in the belt area AL shown to be more selective for call type and neurons in CL more selective for spatial localizations (Tian et al., 2001). This has led researchers to conclude in favour of the existence of an anterior "what" stream, concerned with extracting meaning from auditory input, and a posterior stream "where" stream, specialised for spatial localization. Within in this framework it seems likely that AL and CL mark the initial stages of these streams, with the "what" stream likely extending to the left temporal pole (Poremba et al., 2004).

Figure 1.3 Schematic of the Macaque STS.

Reproduced with permission: Seltzer B, Pandya DN (1989) Intrinsic Connections and Architectonics of the Superior Temporal Sulcus in the Rhesus-Monkey. Journal of Comparative Neurology 290:451-471.



Functional imaging studies conducted with humans suggest a similar hierarchical structure may exist in humans. Primary auditory cortex in humans, located within Heschl's Gyri (HG), corresponds to the core regions in the macaque. It contains three separate cytoarchitectonic regions TE1.0, TE1.1, TE1.2 (Morosan et al., 2001) and is located on the dorsal surface of the Superior Temporal Gyrus, largely hidden within the sylvian fissure. HG is highly variable and can contain up to three gyri, in the case of multiple gyri the anterior most is defined as primary (Penhune et al., 1996). The planum polare is situated anterior to HG. Whilst posterior to HG is the Planum Temporale (PT), with its anterior border defined by Heschl's Sulcus, its lateral border defined as the superolateral margin of the superior temporal Gyrus and the posterior border defined as the posterior termination of the horizontal stem of the sylvian fissure (Vadlamudi et al., 2006).

It is difficult to directly map belt and parabelt regions in the macaque to the superior temporal plane in humans due to the inherent differences in anatomy. Rivier & Clarke (1997) carried out cytochrome oxidase, acetylcholinesterase and NADPH-Diaohorase staining on the human supra-temporal plane and found five separate cortical areas on the supratemporal plane (A1, AA, PA, LA, MA) and one on the posterior part of the superior temporal gyrus (STA). They found that whilst the A1 region, equivalent to macaque core, had a chrome oxidase profile compatible with a primary sensory area, STA had the profile of a high order association area, and LA, PA, MA, AA and AIA had intermediate profiles consistent with an anatomical hierarchical organization. Primary auditory cortex in humans has been shown to respond strongly to pure tones and is tonotopically organized (Formisano et al., 2003). Wessinger et al. (2001) showed that whilst primary auditory cortex responds to both tones and band pass noise, regions anterior-lateral and medial to PAC responded to band pass noise alone. Davis & Johnsrude (2003) demonstrated that this hierarchy extends to higher level language processes, with regions surrounding primary auditory cortex shown to be both sensitive to intelligibility and the acoustic form of the intelligibility distortion, and more distant intelligibility regions invariant to the acoustic form of distortion.

Anterior and posterior streams have also been demonstrated in humans. Upadhyay et al. (2008) demonstrated using structural connectivity, separate pathways from anterior HG to anterior

STG, and from posterior HG to posterior STG. The anterior superior temporal sulcus (STS) has been consistently associated with responses to intelligible speech (Scott et al., 2000;Scott et al., 2006;Narain et al., 2003). Whilst regions of posterio-lateral, but not anterior to primary auditory cortex, have been shown to be modulated by the position in which sounds are presented (van der Zwaag et al., 2011). Ahveninen et al. (2006) showed a double disassociation between spatial and phonetic processing, with anterolateral HG, anterior STG and posterior planum polare showing response adaptation to phonetic content, and the PT and posterior STG showing adaptation to location. In addition to a posterior "where" pathway, claims have also been made for other parallel or integrated posterior streams concerned with working memory and/or sensori-motor integration in humans (Rauschecker and Scott, 2009;Hickok and Poeppel, 2007).

1.3 Controversies surrounding the anterior "what" stream

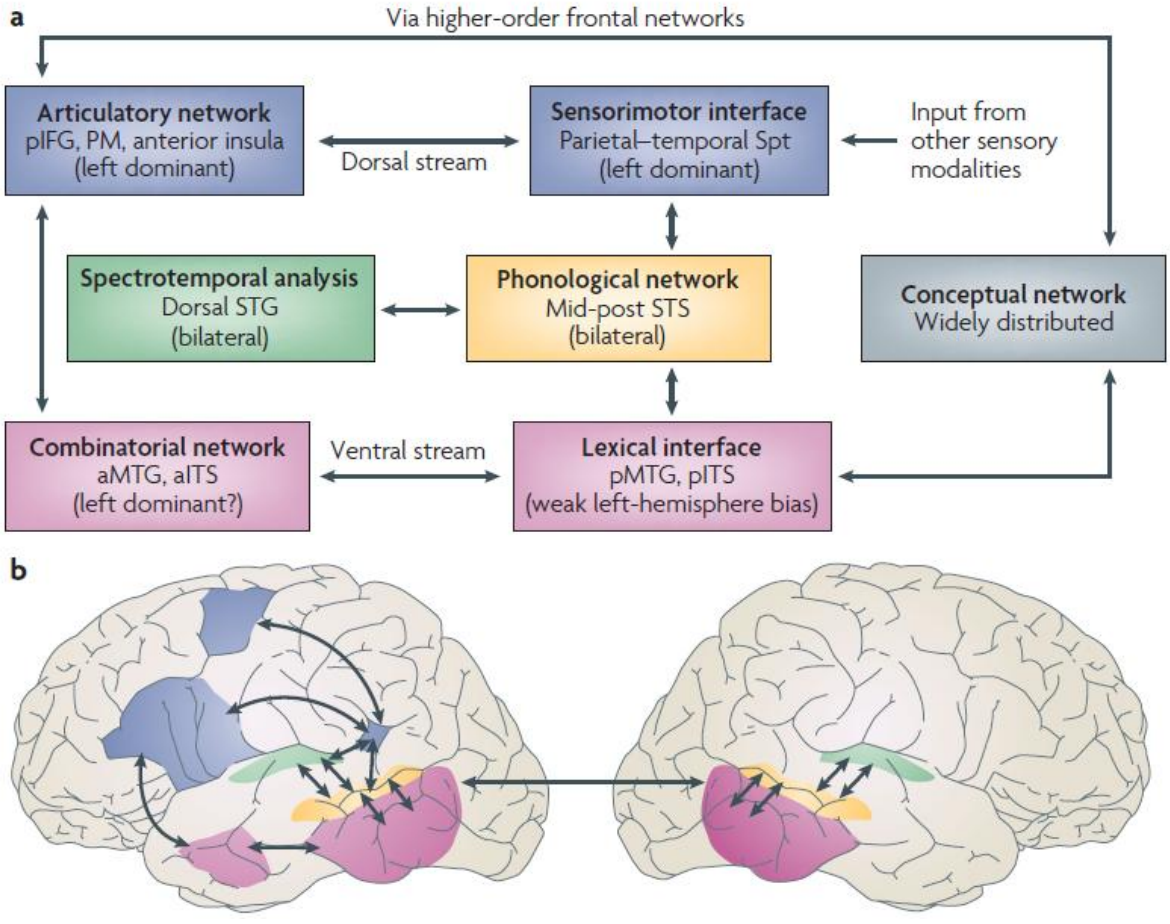
Whilst there is broad agreement concerning the existence of an anterior stream concerned with extracting meaning from speech, there is an ongoing controversy as to whether intelligible speech is first resolved in left anterior or bilateral posterior temporal fields (Hickok and Poeppel, 2007;Rauschecker and Scott, 2009).

Hickok and Poeppel (2007) conceptualise the anterior "what" stream as bilaterally represented. In their view, basic spectro-temporal processing occurs within HG, with the stream running posterio-laterally from HG to bilateral posterior STS, the suggested location at which intelligible phonemes are first extracted. The computational activities of the two hemispheres are said to differ and these differing roles are argued to reflect neuronal specialism for resolving information evolving over different time scales. The right hemisphere is suggested to show sensitivity to information encoded over longer time windows, and the left more sensitive to information encoded over short time windows (Poeppel, 2003) or with this information more bilaterally processed (Hickok and Poeppel, 2007) - see Zatorre and Belin (2001) for similar arguments. From bilateral posterior

STS, the stream is said to flow to the posterior middle and inferior temporal gyri (ITG) for mapping sound onto meaning, before flowing forward to the anterior temporal lobe for semantic and/or syntactic integration. The anterior temporal region is argued to be involved in integrating semantic knowledge across modalities, while the posterior region is devoted to the auditory modality. According to the Hickok and Poeppel model the posterior stream diverges from the anterior stream at bilateral posterior STS, and interactions between a region within the PT (referred to as Spt) and posterior frontal cortex and temporal structures support sensori-motor integration and working memory processes.

Figure 1.4 Hickok & Poeppel (2007).

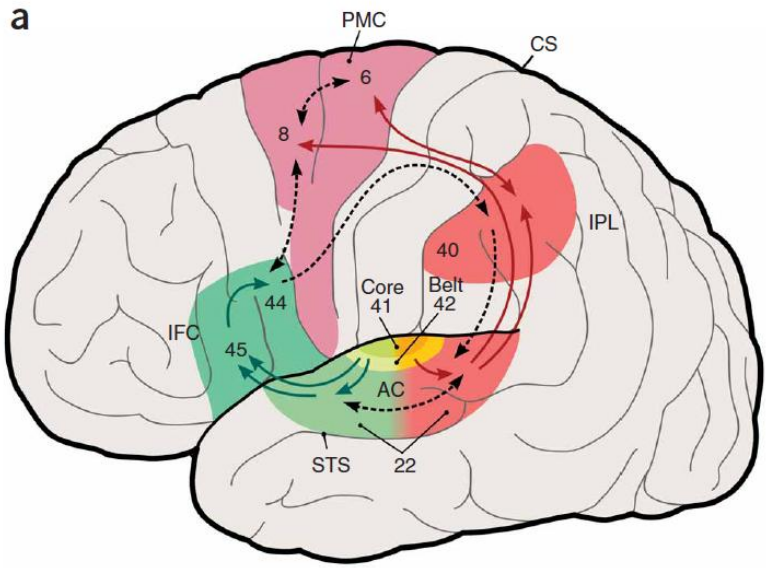
Reproduced with permission.



The opposing view, put forward by Scott and colleagues (Rauschecker and Scott, 2009;Scott and Wise, 2004), suggests that the anterior "what" stream is left lateralised. In a similar manner to Hickok and Poeppel, early cortical auditory areas are said to engage in low level acoustic processing. Beyond these early areas the two perspectives diverge with Scott and colleagues suggesting that instead of the anterior stream initially projecting posterior-laterally, it rather projects antero-laterally. Responses are argued to demonstrate a gradient of increasing sensitivity to acoustic complexity with distance from HG, and intelligible speech percepts emerge in the left anterior Superior Temporal Sulcus (STS), an area in which phonetic maps are suggested to be implemented. The resolution of intelligible speech in this anterior STS region allows representations to be ideally placed to interface with semantic representations stored in the "semantic hub" within the anterior temporal cortex, a region shown to atrophy in cases of semantic dementia (Patterson et al., 2007). By way of comparison with the Hickok and Poeppel model, whilst they suggest that bilateral responses to speech are driven by the acoustic properties of the signal, the Scott model advocates left lateralised responses driven by access to linguistic representations. Similar to the Hickok and Poeppel model, the posterior stream within this framework is suggested to support sensori-motor integration and working memory processes. Within this stream, the posterior STS is argued to play a specific role in representing transient representations of the sequence of sounds during speech perception, and connectivity between PT, frontal, parietal and auditory cortex is argued to mediate sensori-motor integration (Scott and Wise, 2004).

Figure 1.5 Rauschecker & Scott Model (2009).

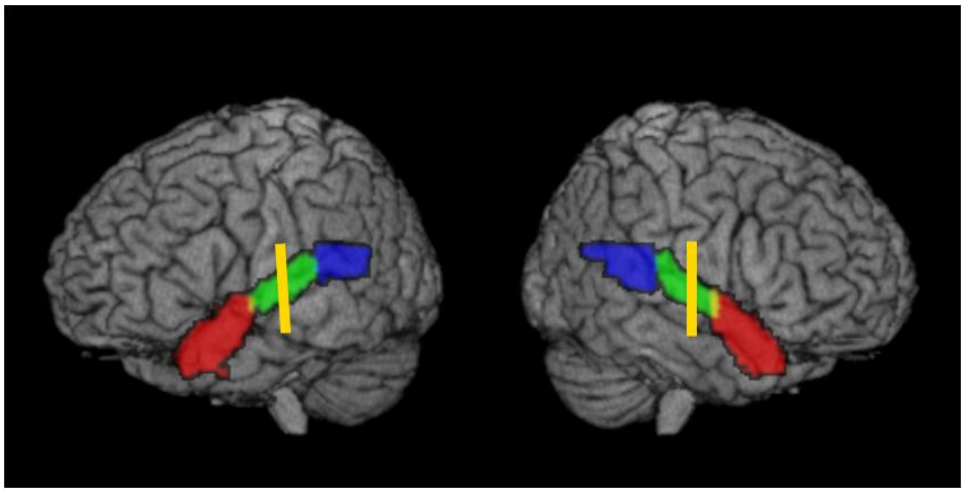
Reproduced with permission.



As is clear from the above description, the difference between the conceptualizations of the anterior stream is subtle. Indeed in respect of lateralisation the difference is one of degree. Thus while the Scott model argues for a left lateralised system, Hickok and Poeppel argue that intelligibility is resolved bilaterally while acknowledging that there might be a "weak left hemisphere bias" (Hickok and Poeppel, 2007) (p 395). Thus disagreement concerning lateralisation is more one of emphasis than absolutes, and it is unlikely that either group would argue for entirely equi-bilateral or entirely left lateralised processing. The anterior-posterior distinction is also subtle. Crucial to engaging in this debate is the distinction that one draws between anterior and posterior regions as there is no clear anatomical basis or consistent functional criterion for separating them. The majority of studies do not explicitly report the criterion used in labelling these regions (Scott et al., 2000;Obleser et al., 2007b), in recent times however in the light of developing controversies, the criterion has been explicitly reported (Okada et al., 2010) . Okada et al. (2010) used the position of HG as a landmark, with anterior STS defined as anterior to the anterior most point of HG, and posterior STS as posterior to the

posterior most point of HG, with mid STS defined as the region between these extremes. This proves however to be quite a stringent criterion with only the very most anterior (y>0, in the left) and posterior regions defined as such (y<-37) (as can be seen in Figure 1.4) and the functional relevance of a separate mid region that this approach creates is questionable.

Figure 1.6 Rendering of PAC as defined by Moroson et al. (2001) (green), region anterior to PAC (red), region posterior to PAC (blue). Yellow line marks the midpoint between the anterior and posterior most points of PAC.



As one of the concerns of this thesis is in delineating anterior versus posterior neural functions, some kind of working definition is required. The definition applied here was derived by identifying the anterior and posterior most coordinates of PAC defined in MNI space using the definitions of Morosan et al. (2001), and rather than introduce an arbitrary mid section, the mid-point of PAC is taken as the dividing line between anterior and posterior areas (see yellow line in Figure 1.4). This corresponds to roughly y=-17 in the left and y=-14 in the right. This criterion seemed to capture relatively well the distinction between anterior and posterior areas as described in reported activations in the functional imaging literature. It is acknowledged, however, that one might prefer to call the regions divided by this distinction mid-anterior and mid-posterior.

The following section discusses the prospective roles of anterior and posterior temporal cortex in resolving speech, and evaluates the evidence in support of left anterior versus posterior bilateral temporal cortex as playing a crucial role in resolving intelligible speech.

1.4 Identifying Regions which Respond to Speech: Evidence from Functional Neuroimaging

Identifying the regions specifically involved in processing intelligible speech has proved difficult. One might think that the simplest approach would be to contrast native speech with foreign speech. This approach is problematic, however, as languages differ along multiple dimensions including phonemic inventory, phonotactic and stress structure, all of which reduce the degree of experimental control. In preference many researchers have attempted to construct non-speech stimuli with similar acoustic properties to speech. Early studies contrasted speech to simple tones and noise bursts, both of which clearly lack the complexity of speech (Demonet et al., 1992;Zatorre et al., 1992).

A variety of more sophisticated non-speech stimuli have been used in recent times. Signal correlated noise, generated by multiplying the envelope of speech by a white noise, is well matched temporally but not spectrally to speech. Mummery et al. (1999) found activation which spread across anterior and posterior STS in the left hemisphere and mid to anterior STS in the right when contrasting speech with signal correlated noise. Uppenkamp et al. (2006) found bilateral activations spreading across anterior and posterior temporal cortices when they compared synthetically generated vowels to equivalent stimuli in which the formants varied in centre frequency randomly from one synthetic glottal cycle to the next (around every 10ms). Whilst these stimuli are well matched in terms of their overall level of acoustic complexity, the formant energy shifts abruptly between frequencies without the smooth transitions characteristic of speech.

Another approach has been to use sine wave speech, a stimulus in which single tones are synthesised to follow the formants of speech. These stimuli are unintelligible, sounding like strange whistles, until participants are informed that they can be understood as speech. A nice feature of these

stimuli is that the same stimulus is used in both the "speech" and "non-speech" mode which means they provide their own acoustic control. However, sine wave speech as it is often implemented does not contain harmonic or broadband formant structure and can be difficult to understand making it unclear whether the same neural processes are engaged when listening to sinewave speech as are engaged with natural speech. Using sine wave speech, Dehaene et al. (2005) found, when explicitly testing for differences in lateralisation, that a single cluster in the left supramarginal gyrus was more activated in the speech compared to the non-speech mode.

A number of studies have used reversed speech. Reversed speech controls for the acoustic complexity of speech as it contains all the same phonetic material, only presented in the opposite temporal order. However, amplitude modulations in speech are asymmetric and a great deal of relevant information is carried by the onset of sounds, time reversing can thus distort the information carried by onsets in an uncontrolled manner (Scott and Wise, 2004). Furthermore whilst reversing has profound effects on some speech sounds it leaves other more steady state sounds, such as fricatives and vowels, unaffected and thus entirely intelligible. Neuroimaging studies which have used reversed speech have either evidenced null results when contrasting words with reversed words (Binder et al., 2000) or left hemisphere activations spanning anterior and posterior STS and the left inferior frontal gyrus (Leff et al., 2008).

Arguably the most appropriate non-speech analogue used to date is rotated speech (Blesser, 1972). Speech rotation involves flipping the spectrum around a specified frequency typically around 2 kHz, with low frequencies becoming high, and high becoming low, followed by filtering to maintain the long term average spectrum of un-rotated speech. This has a similar temporal envelope to natural speech and preserves spectral complexity both in terms of overall spectral shape as well as harmonic and formant structure. Furthermore whilst it does not sound like natural speech, it does sound as if it is another albeit alien sounding language. Using Positron Emission Tomography (PET), Scott et al. (2000) found that clear speech activated the left anterior Superior Temporal Sulcus more greatly than rotated speech. The left posterior STS by contrast was shown to respond to the unintelligible phonetic

features preserved in rotated speech, consistent with a region more involved in complex acoustic processing than in responding to intelligible phonemes.

A later replication with functional Magnetic Resonance Imaging (fMRI) found a similar but not quite so clear cut result, finding both a left anterior and an additional left posterior STS activation to intelligible speech (Narain et al., 2003). More recent studies using rotated speech with larger numbers of subjects have sometimes additionally implicated right anterior and/or left posterior activations (Friederici et al., 2010;Awad et al., 2007;Spitsyna et al., 2006), albeit with the largest peaks found in the left anterior STS. This suggests that while the effect is strongest in left anterior STS there may also be weaker effects in the left posterior STS and the right anterior STS. It should be noted that contrary to the Hickok and Poeppel model, activation in the right posterior STS has been absent in all these studies. Indeed the most consistent factor across the studies that have contrasted speech with rotated speech is activation in the left anterior STS (Scott et al., 2000;Narain et al., 2003;Friederici et al., 2010;Obleser et al., 2007b;Liebenthal et al., 2005;Okada et al., 2010).

It has previously been argued that the intelligibility responses found in anterior areas in these studies are a function solely of prosodic or syntactic effects associated with using sentence level speech materials (Hickok and Poeppel, 2007). It is true that syntactic processing is sometimes associated with anterior areas, but equally it is noted that posterior regions have been too (Friederici et al., 2010). The fact that anterior regions have been implicated when simple consonant-vowel and consonant stimuli have been compared to their rotated equivalents argues against a purely syntactic/prosodic explanation (Liebenthal et al., 2005;Obleser et al., 2007b). Indeed the findings of a recent meta-analysis showed that responses to intelligible speech engage anterior areas regardless of the length of the speech stimuli, with responses becoming successively more anterior as stimuli increased from phoneme, to word to phrase level (DeWitt and Rauschecker, 2010). Furthermore if a prosodic account explained these findings one might imagine that the more salient prosodic structure of clear speech as compared to rotated speech would drive a more right rather than left lateralised response given the right hemisphere association with pitch processing (Zatorre and Gandour, 2008).

Another approach in identifying regions involved in speech perception has been to use adaptation paradigms to identify regions coding for the representation of a particular speech sound (Grill-Spector and Malach, 2001). In adaptation paradigms a repeated stimulus is played to subjects causing a successive reduction in activation in regions coding for that stimulus (habituation) and a recovery from habituation (dishabituation) on presentation of a different stimulus. With this technique it should be possible to identify regions responding to phonetic category change which should reflect regions involved specifically in speech rather than acoustic processing. One common finding from these studies has been that, alongside activations in temporal cortex, a region in inferior parietal cortex is often implicated in responding to phonetic change (Joanisse et al., 2007;Zevin and McCandliss, 2004) this is unexpected as this region lies in a region that responds to multiple modalities. It has since been identified as a behavioural orientating rather than purely phonetic response (Zevin et al., 2010), consistent with its implication in oddball detection tasks in a range of modalities (Downar et al., 2002).

Studies that have explored across category phonetic change have often identified a response in left posterior temporal cortex (Myers et al., 2009;Joanisse et al., 2007;Zevin and McCandliss, 2004;Blumstein et al., 2005) and the left Inferior Frontal Gyrus (IFG) (Blumstein et al., 2005;Myers et al., 2009). The IFG activations have been shown to reflect decision based rather than sensory responses (Myers et al., 2009). The left posterior temporal activations are at odds with an anterior view of speech perception. One factor common to these studies however is the fact they all used synthesized speech. This is in contrast to studies exploring across category phonetic perception that used naturally derived speech stimuli and implicated anterior regions (Obleser et al., 2007b;Obleser, 2010). What might underlie the recruitment of posterior regions specifically in the case of synthesized speech tokens? One explanation is that these stimuli are ”schematic”; often single acoustic cues have been used to signal phonetic contrasts, and as such the stimuli lack the complex co-varying acoustic structure that defines speech segments. These stimuli often sound highly unnatural and unlike speech that participants have routinely heard before. As a consequence it might be the case that participants have to tune low level acoustic processes or map to existing rich multi-

dimensional representations of speech sounds, i.e. a process that may involve a subtle form of perceptual learning.

Indeed a number of studies have associated activity in the left posterior STS with auditory perceptual learning. Leech et al. (2009) tacitly taught participants to categorize non-speech sounds and showed that the degree of increase in activation in the left posterior STS after training predicted success when subjects explicitly categorized the stimuli. Poldrack et al. (2001) compressed speech at different rates (from 15-60%) and found that the signal change in left posterior temporal cortex represented an inverted "u" shape, responding weakly at low and high levels of compression, and strongest at medium levels. This is consistent with a region that works hardest when speech is moderately distorted, and less so when it is either completely intelligible or unintelligible. Adank and Devlin (2010) investigated the time course involved in learning to understand temporally compressed speech. They showed that the left posterior STS evidenced a relative initial increase in signal to compressed compared to uncompressed speech, followed by a gradual adaptation as exposure continued and intelligibility increased. These studies suggest that posterior STS may be specifically recruited in learning new sounds and in the acoustic-phonetic processing required to resolve degraded speech. This might suggest that the posterior STS plays a greater relative role in the acoustic-phonetic processing required to resolve intelligible speech, rather than in responding to the resolved percept itself. Indeed Scott et al. (2006) manipulated intelligibility by increasing the number of bands in noise-vocoded speech (equivalent to increasing spectral detail); they showed that both anterior and posterior temporal cortex responded to increasing intelligibility, but posterior unlike anterior regions also demonstrated sensitivity to acoustic structure.

Posterior temporal cortex has associated with a number of complex auditory functions that are not specific to speech. The PT is an area which responds to hearing both speech and non-speech. It is suggested to function as a "computational hub" segregating and matching complex spectro-temporal patterns, likely playing an important role in auditory scene analysis in which individual sound sources are separated from mixtures, and localized in space (Griffiths and Warren, 2002). The left posterior STS has been shown to respond equally to speech and complex non-speech sounds (Scott et al., 2000)

and has been argued to be involved in the transient representation of the order of sounds during perception (Wise et al., 2001) likely implicating this region in short term working memory processes. Verbal short term memory has been argued to be an emergent property of the speech perceptual and productive systems rather than involving a dedicated anatomical structure (Buchsbaum and D'Esposito, 2008;Jacquemot and Scott, 2006). Posterior structures therefore have been shown to play an essential role in speech perception, but these functions have not always been shown to be specific to speech perception.

Another source of evidence for disassociating the function of anterior and posterior temporal cortex comes from the neuropsychological literature which is addressed in the following section.

1.5 Evidence from lesion and WADA studies

Since Carl Wernicke published his observations in 1874 on the association between damage to the left temporal lobe and impairments in speech comprehension, researchers have used neuropsychological evidence to localise the structures involved in speech comprehension. While functional imaging studies in the main are correlational rather than causal, evidence from lesion studies allows one to consider which regions might be necessary rather than just associated with a particular function. Unfortunately however, lesions rarely target discrete anatomical areas making drawing firm conclusions from lesion studies difficult. In addition the effects of lesions on brain connectivity are poorly understood, as are the mechanisms underlying the neural plasticity involved in recovery of function. This may be one reason behind the observation that lesions in multiple regions, including those beyond temporal cortex, affect the perception of speech (Blumstein et al., 1977).

Whilst damage to both anterior and posterior temporal cortex can cause impairments in speech perception (Dronkers et al., 2004), damage to posterior rather than the anterior temporal lobe seems to be most often associated with speech perceptual impairments (Hickok and Poeppel, 2007). This seems to argue against a primary role for anterior areas in resolving speech. There may however

be a simple explanation for this: infarcts confined to the anterior temporal lobe are rare as the artery supplying this region often arises proximally protecting it from emboli lodging at the more distal trifurcation of the middle cerebral artery (Crinion et al., 2006). Further, posterior temporal lesions have been shown to reduce the physiological response in anterior regions (Crinion et al., 2006). It seems highly probable therefore, especially given evidence in the macaque of dense connectivity between anterior and posterior temporal cortex within the STS, that posterior lesions could either impair speech perception directly or indirectly by reducing communication with anterior areas. Indeed a recent Dynamic Causal Modeling (DCM) analysis demonstrated that listening to intelligible speech as contrasted with reversed speech increased the strength of the connection between the anterior and posterior STS (Leff et al., 2008).

Proponents of bilateral phonetic processing have often argued that the fact that subjects perform much more poorly in speech perceptual tasks following bilateral rather than unilateral damage argues in favour of strongly bilateral speech perceptual processes (Hickok and Poeppel, 2007). It is an uncontroversial fact that damage to the left hemisphere has a much more profound effect on speech perceptual abilities than damage to the right, suggesting that there is not an equal contribution of the two hemispheres. Similar arguments, based on the observation that subjects are able to perform some simple speech perceptual tasks following incapacitation of the left hemisphere with WADA and temporalobectomy procedures, do not provide convincing evidence against a left hemisphere bias for speech perception. The demonstration that the right hemisphere can perform some simple tasks when the left is incapacitated is not evidence that the right hemisphere has the same level of speech perceptual expertise as the left, by way of analogy I can write with my left hand if my right is incapacitated but this does not mean that I can write equally well with either hand.

1.6 Conclusions and thesis outline

There is broad agreement that auditory perception proceeds hierarchically, engaging multiple streams of processing including an anterior stream specifically concerned with extracting meaning

from speech. Controversy exists as to the degree of lateralisation this pathway exhibits and the point at which intelligible speech emerges from within this stream. The patterns of lateralisation, or otherwise, in response to speech are argued to derive from either the acoustic properties of the signal or from the interface with linguistic representations. Evidence from functional imaging has implicated both the anterior and posterior STS in speech perception. Finding a suitable baseline capable of separating neural responses specific to speech as contrasted with acoustic complexity has proved difficult. Studies which have used complex non-speech baselines such as rotated speech have consistently implicated left anterior STS in responding to intelligible speech. The left posterior and right anterior STS have been associated, but less consistently, with responding to intelligible speech when complex baselines have been used. Posterior STS has been shown to be specifically involved in learning new sounds and in understanding degraded speech, consistent with a greater prospective role in acoustic-phonetic processing. Evidence from functional imaging suggesting left anterior STS to show the most consistent response to intelligible speech appear to contradict findings from lesion studies suggesting that damage to posterior regions are most often associated with impairments in speech perception. It is argued however that due to neural connectivity, lesions in posterior temporal cortex could impair the function either directly or indirectly by disrupting in the communication with anterior areas.

This thesis is principally concerned with addressing the controversies surrounding the anterior "what" pathway. It addresses two key questions:

1) Where are neural responses to intelligible, and intelligible but degraded speech, separated from responses to acoustic complexity?
2) Are the resulting patterns of lateralisation driven by the acoustic or linguistic properties of speech?

In Chapter 2 I describe the methods used in this thesis and explain the univariate and multivariate methods employed.

In Chapter 3 I address the findings of a recent study by Okada et al. (2010) which replicated the Scott et al. (2000) and Narain et al. (2003) studies. The Okada et al. study used univariate and multivariate analyses to argue for the importance of bilateral posterior rather than left anterior STS in resolving intelligible speech. I replicate the Scott et al. finding using the same techniques as Okada et al., demonstrating contrary to their findings the importance of left anterior STS in responding to intelligibility.

Then in Chapter 4 I explore responses to degraded speech which was derived from the first two formants of speech. I address whether lateralisation in response to intelligible speech is more likely to be driven by the acoustic or linguistic properties of speech. I demonstrate using multivariate methods a left lateralisation in resolving intelligible speech in the absence of any hemispheric preference for speech derived manipulations of amplitude and frequency.

In Chapter 5 I use Dynamic Causal Modelling with the same data to understand whether the observed left hemisphere preference for intelligible speech found in Chapter 4 is functionally relevant. I examine the connectivity between bilateral anterior and posterior temporal cortex and show that responses in the left hemisphere drive responses in the right. Further I identify a neural system which may represent the instantiation of the integration of higher level linguistic knowledge with lower level acoustic-phonetic processing.

In Chapter 6 I degrade speech by presenting concurrent distracting speech and non-speech sounds. I demonstrate that subjects who perform well at masking tasks tend to activate the lateral mid-posterior STG more than subjects who do not, and identify neural regions more activated by clear than masked speech. Finally the thesis will be brought together in a discussion that attempts to unify my findings.

Chapter 2 : METHODS

In this chapter a brief introduction to fMRI and univariate General Linear Modelling (GLM) is provided. This approach is contrasted with MultiVariate Pattern Analysis (MVPA) and Dynamic Causal Modelling (DCM).

2.1 Functional Magnetic Resonance Imaging – a brief introduction to the MR signal

In this thesis Functional Magnetic Resonance Imaging (fMRI) is used to investigate the neural basis of speech intelligibility. FMRI is a widely used neuroimaging technique used to measure changes in blood oxygenation and flow that occur in response to neural activity. Hydrogen atoms are abundant in the water molecules within brain tissue. As hydrogen atoms consist of single protons, thermal energy causes their atomic nuclei to spin. This spin generates an electric current on its surface and a small magnetic source. In the absence of a magnetic field the spin axes are orientated randomly. When placed in a large magnetic field the nuclei precess around an axis that is either parallel (low energy) or anti-parallel (high energy) to the field, with the majority aligned parallel. This generates a net magnetization in the direction of the field. When energy is emitted by an MR coil in the form of a radio-frequency pulse in a direction orthogonal to the magnetic field, some low energy nuclei absorb energy and change to a high energy state, aligning themselves anti-parallel to the direction of the field, changing the net magnetization. When excitation ceases, the excess spins at high energy return to a low energy state (parallel to the field) by releasing energy which is received by the coil. The time taken to return net magnetization to the low energy state is referred to as T1 recovery or longitudinal relaxation. T1 recovery is different for protons of different tissues. White matter has a very short T1 time, whilst cerebrospinal fluid has a very fast T1 time, and grey matter is intermediate. This property

allows T1 time to be exploited to construct contrast images capable of delineating different types of neural tissue; this is the basis for structural or so called T1 images.

Functional images are acquired using T2* contrast. After net magnetization is tipped into the transverse plane by the RF pulse, it is initially coherent, that is all the spins precess in phase. Overtime this coherence is lost, referred to as transverse relaxation. Coherence is lost via two processes: spin-spin interactions and local field inhomogeneities. These combined effects over time lead to signal loss referred to as T2* decay. When placed in a magnetic field oxygenated haemoglobin is diamagnetic, whilst deoxygenated haemoglobin is paramagnetic. Paramagnetic substances distort the surrounding magnetic field and as a consequence nearby protons will experience different field strengths and will precess at different frequencies, this results in more rapid decay of transverse magnetization and therefore a shorter T2*. The relative changes in the field strength in time and space are thus dependent on the ratio of oxygenated to deoxygenated haemoglobin. The fMRI signal referred to as the Blood Oxygenation Level Dependent (BOLD) response reflects these changes in the ratio of oxygenated to deoxygenated blood over time and space. The exact relationship between the BOLD signal and underlying neurophysiology is still relatively poorly understood. However, at a physiological level it is thought that cognitive stimulation induces local increases in neural activity that cause a small increase in oxygen consumption and energy metabolism. In response to this, changes in cerebral blood flow, cerebral blood volume and the cerebral metabolic rate of oxygen, cause an increase in blood oxygenation within the activated region, causing an accompanying increase in the BOLD signal.

The generic BOLD response has a characteristic shape, defined by an initial brief dip and a rise to peak at around 4-6s. This is followed by an undershoot, in which the signal drops below pre-stimulus levels, before finally returning to baseline around 20-30s after the initial response. Despite the above generic characterisation the BOLD response differs significantly both within and between individuals, and across regions within the brain. The fMRI signal is characterised by an enhanced spatial resolution over other commonly used imaging techniques, such as EEG and MEG, but a relatively poor temporal resolution owing to the sluggish nature of the BOLD response. As a

consequence fMRI is most often used to test hypotheses concerning where particular neural responses occur in the brain.

2.2 FMRI Experimental Design and Analysis

In a typical fMRI experiment participants are placed in an MRI scanner and presented with stimuli or asked to perform a task, if the occurrence of these events is correlated with an increase in BOLD response in a neural region, that region is usually assumed to be involved in the cognitive function assumed to be elicited by the event. As the brain is constantly active, measurements of brain activity are relative. To address this, the dominant methodological approach over the last thirty years has been to contrast the activity from one experimental condition with another, a so called "cognitive subtraction". The assumption being that the two conditions elicit identical responses in all but the cognitive function that the experimenter wishes to isolate, although for a critique of this approach see Price and Friston (1997). Other experimental approaches are also used; these include amongst others parametric designs in which a parameter is varied continuously to identify regions which show a response correlated with the manipulation, and adaptation designs which take advantage of the assumption that neurons show a reduction in successive responses to repeated stimulation.

During fMRI experiments stimuli are presented in runs; that is the scanner acquires data for a specific number of volumes followed by a short break when no data is acquired before scanning continues. In the case of auditory experiments one particular concern in experimental design is reducing the effects of the acoustic noise produced during the acquisition of brain volumes. This acoustic noise caused by the switching of the gradient coils can be reduced by so called sparse acquisition (Hall et al., 1999). In sparse acquisition single volumes are acquired followed by a delay, for example a volume might be acquired every 9 seconds for a duration of only 3 seconds. This is in contrast to continuous acquisition in which successive volumes are acquired without delay between volumes. By ensuring a delay between acquisitions sparse designs allow auditory stimuli to be

presented to subjects in the quiet between the volume acquisitions. Whilst this provides an advantage in allowing subjects to hear stimuli without interference, this comes at the cost of the acquisition of fewer brain images which can reduce statistical power. In this thesis both sparse and continuous data acquisition is used.

2.3 Preprocessing of fMRI data

fMRI data analysis involves a complex set of preprocessing steps. These include slice timing correction, realignment and unwarping, coregistration, normalization to a standard stereotactic space and smoothing. A short summary of these processes as they are conducted in the software package SPM (http://www.fil.ion.ucl.ac.uk/spm) is described in the following section. A whole brain volume usually consists of a number of slices. A single volume can take several seconds to acquire, typically around 2-4s, this can mean that a particular slice is acquired a number of seconds later than another. There are two main approaches for dealing with this issue: a slice timing correction can be applied which interpolates data in time so as to simulate instantaneous acquisition across the whole brain or a temporal derivative can be modelled within the statistical design to account for this variation. When the time taken to acquire a volume is short the need for slice timing is less, unfortunately slice timing is least effective when it is required the most; that is when the time to acquire a volume is longest. However, modelling the temporal derivative is also not without problems as in some instances it can reduce the power of the statistical model (Della-Maggiore et al., 2002). As slice timing is only appropriate when data is acquired continuously, slice timing has not been used in this thesis, except in the case of Chapter 6 in which continuous acquisition was used. Note that it is common practice in sparse acquisition to neither conduct slice timing or to model the temporal derivative provided the time to acquire a volume is relatively short.

Participants often move within the fMRI scanner, when they do so it means that a voxel sampled at a specific point in time and space, may not refer to the same voxel when sampled at

another time point. This reduces spatial accuracy and generates Type I error. In SPM, functional images are typically realigned using a rigid body transformation that uses least squares to minimize the difference between successive scans and a reference image. The transformation is then applied by resampling the data using interpolation. Nonlinear movement effects are further partialled out of analysis by the inclusion of movement parameters as nuisance variables in the statistical design. An additional unwarping step is often carried out in order to account for field inhomogeneities caused by air-tissue interfaces; these inhomogeneities reduce the intensity of the MR signal in these regions. This signal drop out distorts the shape of the functional images and interacts with subject movement making it more difficult to correct for subject movement, so called susceptibility-by-movement interactions. All analyses in this thesis use unwarping to correct for these effects as this thesis specifically addresses the function of the anterior temporal lobes which have been shown to suffer from signal drop out (Devlin et al., 2000).

A coregistration step is conducted to align the structural, T1 image, to the mean realigned functional, T2* image. One difficulty in achieving this task, is that the two types of image are from different modalities, that is one is T1 weighted and the other T2 weighted. As a consequence coregistration cannot be achieved by minimizing the sum of squared differences between the images as the same intensities might signify different types of tissue in the two types of image. To address this, coregistration in SPM uses a mutual information technique which allows dependencies between the intensities of the two images, i.e. a high intensity in A predicts a low intensity in B, to be exploited to maximize the fit between the images. Images are normalized to a reference space, to allow neural responses of different subjects to be considered and to allow results to be compared across studies. In this thesis normalization was conducted by segmenting the T1 image to estimate the probability of tissue types at each voxel and using the parameters from this segmentation to transform the functional images to MNI space. Finally, smoothing is applied to functional images to maximize signal to noise ratio, reduce intersubject variability, ensure that noise is Gaussian distributed to facilitate correction for multiple comparisons. Smoothing makes each voxel a weighted sum of its neighbours by

applying a 3d Gaussian of a specified FWHM at each voxel. In this thesis a smoothing kernel of 8mm FWHM has been used, this is a commonly used intermediate level of smoothing for group analyses.

2.4 Univariate General Linear Modelling

Following preprocessing, statistical tests are conducted on the grey scale intensity values of the functional images within discrete volumetric units (voxels). Typically each voxel is analysed in isolation of all others, termed a mass univariate GLM approach. The matrix form of which is presented below:

$$Y = X\beta + e$$

where Y =column vector of observations, β = column vector of beta coefficients, X = explanatory variables and e=error term

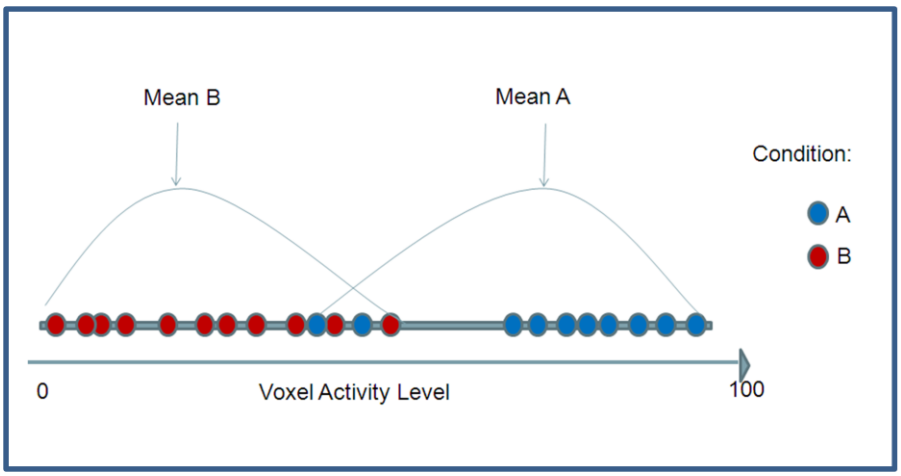
The parameters of the above model are estimated, using ordinary least squares, such that the explanatory variables (a design matrix with a row per observation and a column per explanatory variable) predict the observed time series at a specific voxel as closely as possible. Explanatory variables include both effects of interest and nuisance variables such as movement parameters. The residuals of the fit form the error component. Existing knowledge about the shape of the BOLD response is incorporated into the model by convolving the predicted time series with a canonical hemodynamic response function. At each voxel a parameter value, which represents the slope of the regression line for that parameter, and a measure of the estimated variance is calculated. Null hypotheses are tested to assess whether the parameters of the model at each voxel are different to 0 or different to each other. Thousands of statistical tests are conducted at each individual voxel, with a correction usually made to reduce Type I error. The resulting statistical parametric map shows voxels in which the null hypothesis can be rejected at a specified level of confidence.

In order to test hypotheses about groups of subjects, images representing the contrast estimate for each subject are taken forward to the second level, and a new design matrix constructed with each row representing a single subject. If the subjects from whom the data is acquired are drawn from a random population, then as the contrast estimates take account of individual variability between subjects, inferences can be generalized beyond the individuals in the experiment to the population from which they are sampled. All analyses in this thesis use this random effects summary statistic approach.

2.5 Multivariate Pattern Analysis

At a basic level a "cognitive subtraction" (a univariate t-test comparing two conditions) can be thought of as a one dimensional analysis that asks whether the mean of samples at a single voxel from one condition are greater than another (whilst also taking into account the variability in the measurements), that is, it asks whether the conditions differ along a single magnitude dimension (see Figure 2.1).

Figure 2.1: Representation of a t-test as a one dimensional analysis.

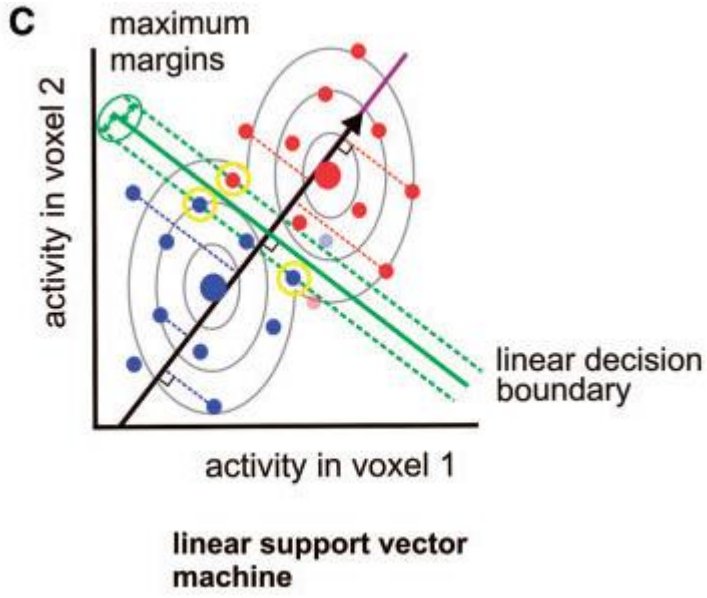


In recent years machine learning approaches to fMRI analysis have become an increasingly popular alternative and/or supplement to traditional univariate GLM. Typically in pattern classification, an algorithm is shown a subset of training data and is provided with the associated label that identifies the experimental condition to which it belongs, i.e. scan 1=condition A, scan 2=condition B etc., it then attempts, by considering the pattern of activity across multiple voxels, to learn a function that successfully maps the brain images to their associated labels. A subset of data, which was withheld during the training phase, is then used to assess how well the learnt function performs in classifying the unseen test set – a measure of the generalization of learning. If the classifier performs at a level above chance then this is taken as evidence that there was information in the brain images capable of distinguishing between the conditions.

In the case of linear discriminant methods, which are used in this thesis, the classifier learns a discriminating hyperplane that separates brain volumes representing two or more conditions, in a multidimensional space with as many dimensions as voxels. Thus by way of analogy with univariate analysis, rather than distinguishing between conditions on a single dimension (see Figure 2.1), multivariate pattern analysis makes use of multiple dimensions to distinguish between conditions. A simple two voxel classification example is illustrated in Figure 2.2, for each data example (a brain volume corresponding to an experimental condition) the activity level in one voxel is plotted relative to the other, and a separating boundary constructed which separates the conditions within this space. In reality classifiers often use hundreds or thousands of voxels/dimensions, rather than the two voxel examples shown here.

Figure 2.2 Classification in a two voxel space. Each circle represents the intensity value at two voxels from scans belong to condition 1 (blue) and condition 2 (red).

Reproduced with permission: Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRIan introductory guide. Social Cognitive and Affective Neuroscience 4:101-109.



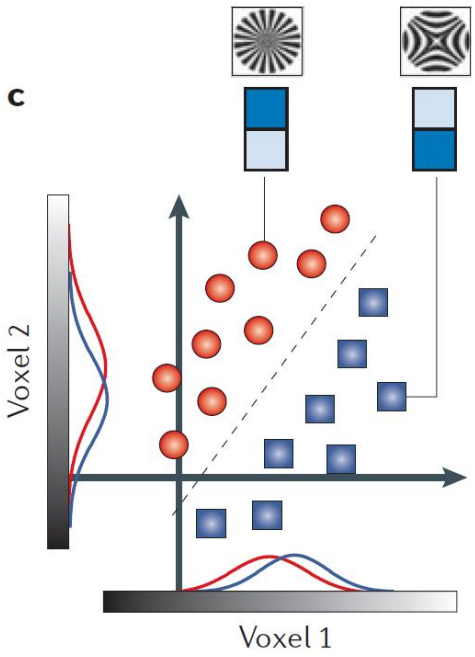
Pattern classification methods offer a number of advantages, both theoretical and practical, over traditional univariate approaches. fMRI data is by essence highly dimensional as recordings are made at hundreds of thousands of locations. The implicit assumption in univariate analysis is that activation in one voxel is independent of all others. This assumption is clearly not true. Each voxel spans tens of thousands of neurons and neural firing is highly correlated. It is thus highly likely that an important feature of neural coding is the co-firing of neurons across multiple locations.

Classification methods have often been shown to provide greater sensitivity to experimental effects compared to standard univariate GLM. Indeed a number of studies have shown successful classification when univariate analysis has failed (Formisano et al., 2008;Obleser, 2010) and classifiers have been used to address experimental questions previously thought to be beyond the resolution of fMRI analyses (Kamitani and Tong, 2005). One source of this additional sensitivity

derives from the integration of information across multiple voxels. From an information perspective by considering the interaction between voxels, the amount of information in the analysis is increased from n to n^2 voxels (O'Toole et al., 2007). Voxels that carry no information about an experiment when analysed in isolation can be shown to do so when analysed alongside another (Guyon et al., 2002;Haynes and Rees, 2006). Figure 2.3 provides an example of this in a two voxel example, the means and distributions of samples from the two conditions largely overlap at each voxel, and thus neither voxel would be "significantly activated" in a univariate analysis. However, it is possible to see that by considering their co-activation the conditions can be separated, demonstrating that they contain multivariate information.

Figure 2.3 Example of how voxels which independently fail to separate experimental conditions can do so when jointly analysed.

Reproduced with permission: Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7:523-534.

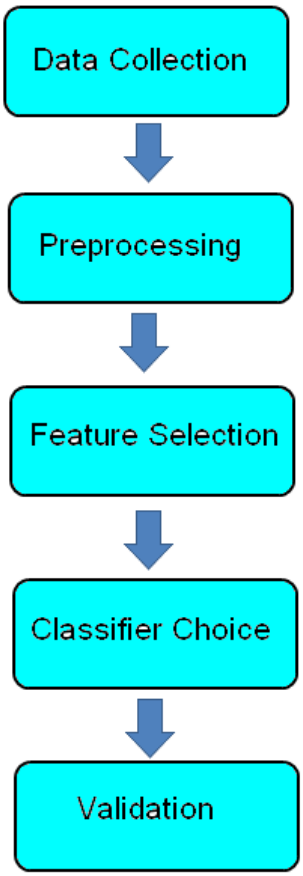


Pattern classification also benefits from not needing to make correction for multiple comparisons at each voxel. Traditional univariate analyses use conventional inferential statistics on a mass scale, often requiring severe correction for the thousands of tests conducted, with a resulting reduction in sensitivity. Related to this, univariate data is often smoothed to meet the assumptions of

random field theory, necessary to facilitate correction for multiple comparisons, and more generally to improve signal to noise ratio. Classification does not require smoothing as voxel wise correction is not required allowing analysis at a finer degree of spatial resolution.

There are a number of processing steps that are typically carried out in multivariate pattern analysis (see Figure 2.4). In the next section these steps will be outlined and some of the considerations discussed from the perspective of auditory research.

Figure 2.4 Typical pattern classification workflow.

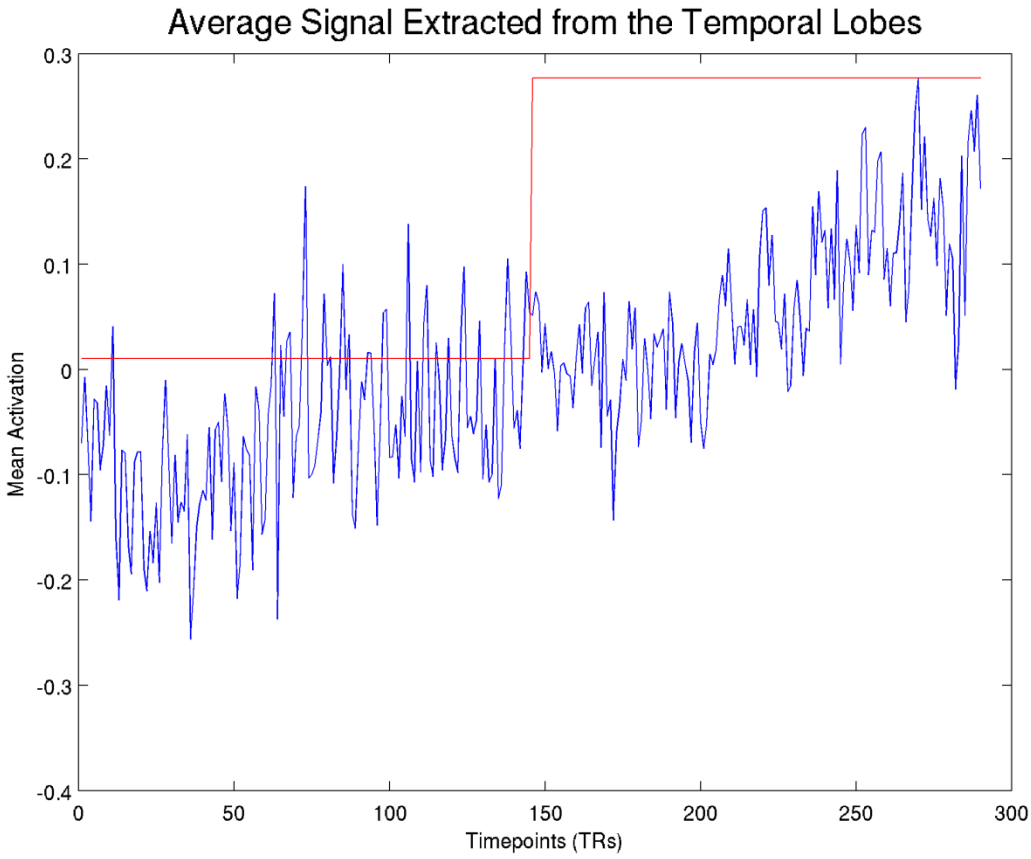


2.6 Classification workflow – data collection

Pattern analysis requires enough data to be collected to allow adequate numbers of training and test examples, and so that these can be kept separate so as to estimate the true accuracy of the classifier (Kriegeskorte et al., 2009). Training and test sets need to be kept separate so as to ensure that we can prove that learning is not just specific to the training set and can generalize to unseen examples, this ensures that we do not "overfit" the data, i.e. that we do not learn the noise structure to the detriment of the signal. Typically data is either split into the same number of training and test examples usually separated by runs or by splitting data to have a greater ratio of training to test data (e.g. 2/3). There is a trade off to be made between providing the classifier with enough data to learn stable mappings, and having enough test data to gain accurate estimates of the classifiers accuracy. In univariate analysis it is in the main better to collect as many volumes as possible within the limits of participant fatigue, as this leads to a more reliable averaged response. In pattern analysis by contrast it is sometimes better to collect a smaller number of less noisy examples, the reason being that a small number of "bad" training examples can have a significant effect on the classification boundary - a problem more pronounced with some classifiers than others. If the training examples are noisy, trials can be averaged to improve signal to noise ratio (Mourao-Miranda et al., 2006).

Noise effects such as scanner drift can have a strong influence on fMRI data. As a consequence it is often recommended that training and test data sets are partitioned by run to ensure that the classifier is unable to exploit any noise effects associated with the runs to improve classification accuracy. It is thus useful to have at least two runs of data so that training and testing is possible on separate runs. See Figure 2.5 for an example of scanner drift observed in an experiment conducted in this thesis.

Figure 2.5 Data acquired from the temporal lobes. The red line indicates different runs.



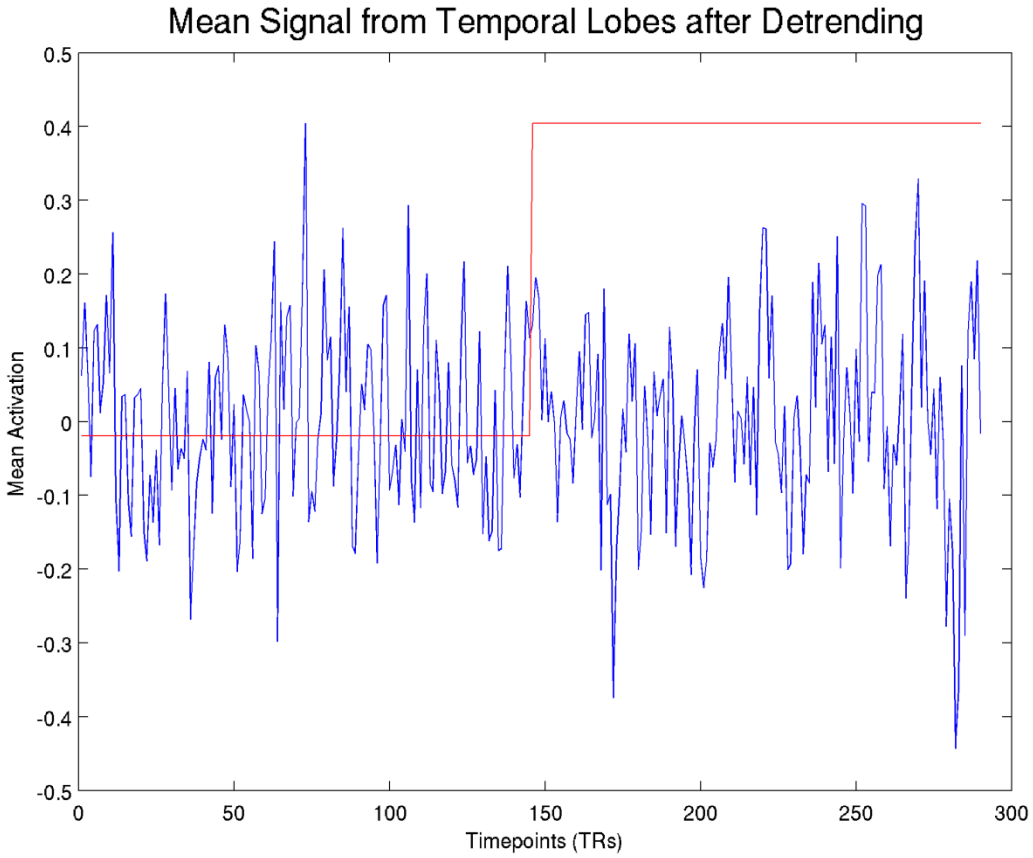
In auditory fMRI experiments, data is often acquired using sparse acquisition to reduce the effects of the interfering noise from the scanner. One advantage of this method is that the long gap between volume acquisitions ensures that there is little autocorrelation between scans making it less likely that the classifier is taking advantage of this artefact in classifying images. If autocorrelation is expected, it is possible to use permutation testing to account for these effects on classification performance (Golland and Fischl, 2003). As classification has been conducted on sparse data in this thesis, autocorrelation is not an issue, allowing a classical statistical approach.

2.7 Pre-processing for classification

One must decide which of the previously described fMRI pre-processing methods to use to prepare the data for classification and the type of image to use in the analysis. Etzel et al. (2011) attempted to identify optimal pre-processing steps for classification and concluded that the optimal methodological choices are likely to be individual to each data set making it difficult to suggest one optimal pre-processing pipeline. In this thesis data has been realigned and unwarped, and z-scored across stimuli by run at each voxel. Z scoring is achieved by subtracting the mean value across stimuli from the response in each voxel to all stimuli within a run, then dividing the resulting value by the standard deviation. This removes baseline shifts in amplitude across runs and ensures similar scaling across voxels. An alternative is to z-score by considering the response of all voxels to a given stimulus and normalizing in that manner. Misaki et al. (2010) found these two approaches yielded equivalent results. Data was further detrended by run to remove linear and quadratic trends caused by scanner drift (see Figure 2.6 for an example of data post z-scoring and detrending). Data in this thesis was not smoothed so as to take advantage of the increased spatial resolution of classification methods, evidence suggests however that there is often a moderate classification gain from smoothing (Etzel et al., 2011).

In this thesis classification was conducted on "raw" functional images (albeit after realigning and unwarping), it should be noted however that successful classification can also be achieved using the beta images in which responses are modelled for single trials (Formisano et al., 2008) or when all trials are modelled together, or using SPMt maps (Obleser, 2010). Using SPMt rather than beta maps has been shown to yield higher classification, especially when using support vector machines (Misaki et al., 2010). One advantage of working from beta images or SPMt maps is that classification is conducted on data that has been fit to a model which incorporates the shape of the hemodynamic response. However, when working with sparse data a single volume is captured at estimated peak of the BOLD response and so this does not confer the same advantage.

Figure 2.6 The same data as shown in Figure 2.5 following z-scoring and detrending.



2.8 Feature selection for classification

Feature selection, that is deciding which voxels to include in the analysis, is an important step in classification. Classifiers often perform poorly with too many voxels. A key concern in feature selection is to ensure that the data used to define the reduced feature set is independent from the test data, if the classifier is allowed to "peak" the test data in feature selection it will be optimistically biased with an increased chance of spurious classification (Kriegeskorte et al., 2009). A broad distinction in feature selection methods can be drawn between filter and wrapper methods. Filter methods select a subset of features by ranking them using a method independent of the classification method, for example by conducting a univariate test to select activated voxels for further

classification. Wrapper methods by contrast use the same classification method to rank and remove features as is used in classification; this allows the selective removal of voxels with low predictive power. Feature selection approaches used previously in fMRI include dimension reduction using principle component analysis (Mourao-Miranda et al., 2005), univariate analysis to create functionally defined regions of interest (Okada et al., 2010), masking with anatomically defined regions of interest (Etzel et al., 2009), and wrapper methods such as Recursive Feature Elimination (RFE) (Formisano et al., 2008). In this thesis we have adopted two different approaches: anatomical regions of interest and RFE.

Anatomical ROIs are often defined based on gyral anatomy. When using ROIs derived from gyral anatomy, it is best to define these individually in each subject (Poldrack, 2007). One approach is to manually define regions in each subject's native space based on agreed anatomical definitions. In Chapter 3 this approach was adopted to define Heschl's Gyrus (HG) using the definitions of Penhune et al. (1996). An alternative but less recommended approach is to use ROIs defined on a single subject atlas in stereotactic space, this is less preferable as normalization is imperfect in matching brains across subjects. In Chapter 3, in the case of larger anatomical areas such as the STG, ROIs defined in MNI space on the single subject brain, were transformed into the subjects' native space. An alternative approach adopted in Chapter 4 is to use an automated parcellation method (FreeSurfer) to define subject specific ROIs (Destrieux et al., 2010). Freesurfer is an automated approach that uses the individual anatomy of each subject, to probabilistically map regions of interest by reference to a database of previously defined manual parcellations from multiple subjects. This has the advantage of being probabilistic, and thus accounting for variability in anatomy between subjects, and taking into account the individual gyral anatomy of each subject.

Often when functional ROIs are used as a feature reduction method, a standard univariate statistical test is conducted on a subset of data from the main experiment or a separate functional localizer scan is used to identify activated voxels (see Okada et al. (2010) for an example of the use of both methods). One criticism of this approach is that the mask created by this test restricts the subsequent multivariate analysis only to those voxels showing an initial univariate response. A

different approach is to use a multivariate method to define a subset of voxels for subsequent classification; RFE takes this approach. In RFE with Support Vector Machines (SVM) a classifier is trained using a set of voxels. The voxels are then ranked by the magnitude of their associated weight; the weights reflect the importance of each voxel to defining the classification boundary. A specified number of voxels with the smallest weights are then removed and the process starts again, successively pruning away voxels to a subset of best performing features. This method is used descriptively in Chapter 4 to demonstrate the sensitivity of SVMs using large anatomical regions of interest.

2.8 Classifier Selection

A broad distinction can be made between linear and non-linear classifiers; the latter allow much more complex boundaries to be constructed. Typically researchers have tended to use linear in preference to nonlinear classifiers. This is for two reasons, firstly it is more difficult to interpret the importance that each individual voxel plays in defining the classification boundary in the case of non-linear classifiers, and secondly they have tended to be outperformed by their linear counter parts (Misaki et al., 2010). This is likely to arise from overfitting of the training data, which occurs as non-linear classifiers are able to use a larger number of parameters in model fitting and construct more complex decision boundaries. A further broad distinction can be made between generative and discriminative classifiers. Discriminant classifiers attempt to directly learn a given prediction function from training data by learning the parameters of the function, whilst generative classifiers learn a statistical model that could be used to generate an example from each class. Some of the previous classifiers used with neuroimaging data to date include gaussian naïve bayes (Mitchell et al., 2004), logistic regression (Yamashita et al., 2008), SVMs (Mourao-Miranda et al., 2005), Linear Discriminant Analysis (LDA) (Carlson et al., 2003) and pattern-correlation classifiers (Haxby et al., 2001). The essential difference between classifiers is the way in which they define the boundary between the experimental conditions.

The pattern correlation classifier is perhaps the simplest classifier; it works by classifying patterns according to the strength of the correlation coefficient between an example and category exemplars. These exemplars are the average response pattern estimated from the training data for each category, with the test pattern assigned to a category based on the category that it is most correlated with. SVMs by comparison are computationally more expensive, but have become arguably the most popular classifier as they tend to outperform or perform as well as the best performing classifiers when multiple classifiers have been compared (Mourao-Miranda et al., 2005;Mitchell et al., 2004;Misaki et al., 2010).

SVMs define a hyperplane within the multidimensional voxel space that maximizes the distance between the most similar examples from each experimental condition. This can conceptualised by imagining a decision boundary in a two voxel space that exactly separates two conditions, then widening the margin equally on each side of the boundary while adjusting the angle and position of the boundary, until the margin cannot be widened any further without including one of the training examples (Mur et al., 2009). The points on this margin, i.e. the most similar examples of each condition, are referred to as the support vectors. A parameter, referred to as C, can be adjusted to allow some misclassifications to occur such that a small number of examples are allowed to fall within the margin or on the opposite side of the decision boundary; this is referred to as a soft margin and occurs with small values for C. A hard margin is so defined when the C value approaches infinity and in this circumstance the two classes have to be exactly linearly separable for a solution to be found.

The location of the hyperplane is defined by a weight vector, which is orthogonal to the decision boundary, and a parameter that shifts it to its best location, known as the bias. The weight vector is the direction in the data of maximum discrimination. It constitutes a set of values that weight each voxels' contribution to a function that can be used to predict the identity of new data examples. The fundamental task in training a SVM is learning the values of the weight vector and bias. The success of the function is evaluated by assessing its accuracy in predicting the identity of unseen examples. Computing the weighted sum of voxel responses is equivalent to projecting the

data examples onto a linear discriminant dimension (a line in multivariate space) with a threshold - the location of the decision boundary on that dimension - used to assign examples to each condition. The function takes the following form:

$$f(x)=w^{T}x+b$$

where w=transpose of the weight vector with n values, x=data example with n voxels and b=bias

Given a positive and a negative class: +1 =condition A, -1 = condition B, the identity of test data is assigned to condition A if $f(x) > 0$, and as condition B if $f(x) < 0$. By examining the values of the weight vector we can gain an intuition into how each voxel contributes to classification. The weight vector represents the weighted average of the support vectors. Voxels that receive a large weight contribute more to the decision boundary than voxels that receive a small weight. Furthermore, voxels receiving a positive weight represent voxels in which there was a relative increase in BOLD signal to condition A in the support vectors, whereas those receiving a negative weight show a relative increase to condition B (Mourao-Miranda et al., 2005).

The defining feature of SVMs is that they only use a subset of the available data to define the classification boundary – the support vectors. This means that both the boundary and weights are defined by the examples in the data that are most similar across the two conditions. Indeed all data examples could be removed from the analysis except the support vectors without changing the solution. This is in contrast to methods such as LDA in which all data examples are used to construct the boundary with the removal of any example changing the solution. Thus one criticism that is often made of SVMs is that the solutions are constructed from the fringes rather than the centres of the distributions of the data from each condition. In the case of this thesis this is viewed as an advantage, as non-speech stimuli have been constructed so as to simulate speech as closely as possible without

inducing a speech percept, thus solutions defined by similarity are an advantage. The fact that the solution is defined by the support vectors alone confers an advantage in reducing overfitting and increasing generalization between training and test sets. By only using the support vectors all other non-informative patterns in the data are essentially given zero weight, which effectively weighs down noisy features that are highly correlated with each other (Pereira et al., 2009). SVMs thus work well with large numbers of voxels because they minimize classification error whilst taking into account model complexity (Sato et al., 2009). Indeed a number of studies have shown that SVMs are relatively robust regardless of the number of voxels used (Ku et al., 2008;Cox and Savoy, 2003), and previous studies have shown equivalent high levels of performance regardless of whether three hundred or three thousand voxels have been used (Misaki et al., 2010).

An additional technique used in this thesis is searchlight analysis. The searchlight technique is not specific to a type of classifier. The searchlight method allows the user to identify locally represented information in an unbiased manner. In a searchlight analysis the classification of each voxel and its immediate surrounding neighbours is considered in turn within an anatomical region. The voxel at the centre of the neighbourhood is assigned the classification performance of its local neighbourhood. Whilst this is a useful technique two factors need to be taken account in its use. Firstly, this technique can only discern the representation of local information and cannot be used to understand how information might be integrated over larger areas. Secondly, as it is difficult to meaningfully extract and summarise the weight vector from these overlapping neighbourhoods it is difficult to understand how classification is being achieved within these discriminative patches. In this thesis, the searchlight method has been used in combination with a more global region of interest based approach to understand how information is represented at different spatial scales.

2.9 Validation

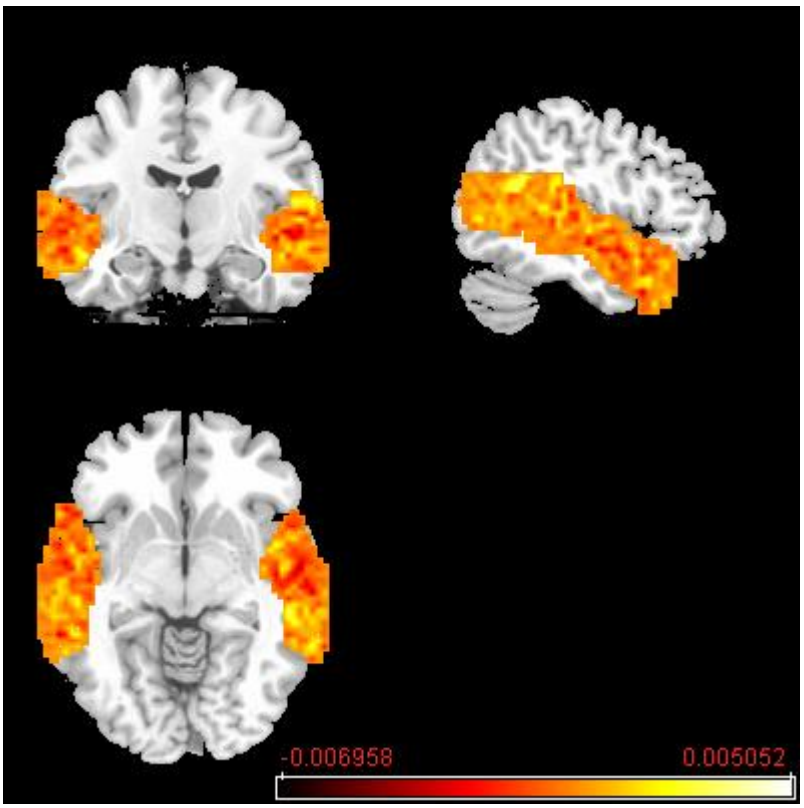
As stated previously successful classification is validated by the degree of successful generalisation in learning that transfers from a training to a test data set. Typically k-fold cross validation is used to validate the success of the algorithm; this ensures a maximum number of training examples whilst ensuring that training and test data are kept separate. This is achieved by dividing the data into k independent data subsets, with all but one subset used for training and the remainder kept for testing, then repeating the procedure until each subset has been used in testing once; typically performance is then averaged across the folds to attain a summary estimate of the classifiers accuracy. The most extreme form of this approach is referred to as "leave one out", in which successive single data examples are left out for testing whilst training is conducted on all the remaining examples. This has the advantage that the maximum number of examples can be used in training the classifier. It is a computationally demanding approach that has been shown to result in higher classification accuracy than the leave run out approach (Misaki et al., 2010). However, this increased level of performance may arise from the classifier exploiting within-run noise effects. It is for this reason that a more conservative approach leave one run out approach has been adopted.

Cross validation can be conducted either within or across subject. In the within subject case, a subset of each subjects' data is used for training and testing, and each subject then contributes a single classification score (usually averaged across folds) to a population of scores from the group. In this instance every subject is likely to use a different discriminative pattern. In the across subject case, a single subject is usually left out as a test example and the data from the remaining n-1 subjects is used to train the classifier. Thus in this instance, the discriminative pattern is defined at any one time by n-1 subjects, and is thus more representative of the group of subject. In across subject cross validation individual subjects who classify poorly can have a significant effect on classification performance, as that subjects data is used n-1 times in training the classifier. It is the authors

experience that often one or two subjects from any group of subjects classify poorly. It is for this reason that a within subject approach has been taken.

After cross validation and the acquisition of a population of scores from the subjects in the experiment, it is possible to ask whether these scores are significantly different to chance level, a 50% accuracy level in the case of two conditions, or whether for example the scores arising from one ROI are significantly different to another. In this thesis non-parametric statistical tests have been used as they make fewer assumptions and are more robust to the effects of outliers - an approach recommended by Demsar (2006). From a classification analysis, it is possible to examine both classifier accuracy and the weight vector. Whilst it is easy to summarize classification accuracy for readers, it is less easy to summarize the weight vector. Images of the weight vector can be hard to interpret as every voxel receives a weight which can be either positive or negative (see Figure 2.7 for weights extracted from data in this thesis for a single slice). Furthermore in a within subjects cross validation, each subject has a different weight vector for each fold of the data, making it harder to summarize the weights across subjects. In this thesis weight maps have been created by thresholding and showing the degree of concordance across subjects so as to make interpretation easier for the reader.

Figure 2.7 Classifier weight vector for a single slice in a single subject.



One final point to be made about interpreting the results of classification analysis is that classification can be thought of as the more general case of subtractive univariate t-tests. That is, a univariate pattern in which the mean for condition A is greater than condition B at each individual voxel, is just one of many types of discriminative pattern that multivariate pattern analysis can discriminate. Indeed classification can also identify the opposite pattern that is greater activity for condition B than A at each voxel at individual voxels, or more complex discriminative relationships involving the interaction between voxels and their relative activity levels. It is thus important to try and understand how classification is achieved as different discriminative patterns constrain the interpretation of what it means to have successful classification within a region.

Classification methods to date have been used to ask both traditional questions concerning the spatial localization of cognitive functions (Okada et al., 2010), and also to address broader questions

concerning how neural information is represented (Haxby et al., 2001). Indeed a number of studies which have used pattern based classification analysis have found evidence to suggest that the neural processing of information may be more spatially distributed than previously thought (Formisano et al., 2008;Staeren et al., 2009;Haxby et al., 2001;Obleser, 2010). Another multivariate technique that is well placed to ask questions concerning how as well as where information is processed is DCM. DCM is also used within this thesis; the following sections provide a short description of the method.

2.10 Dynamic Causal Modelling

The fundamental aim of DCM is to apply a mechanistic approach to understand how neural regions interact with one another. DCMs are both dynamic and causal, in the sense that they allow the modelling of how dynamics in one region cause changes in the dynamics of another, and how these interactions are modulated by experimental manipulations (Stephan et al., 2010). This is achieved via a biologically plausible model that describes the interactions between neural activity in different regions over time, and a hemodynamic model that translates the neuronal states to a predicted BOLD signal. The parameters of the neuronal hidden state model can be estimated by perturbing the system with a set of known inputs, the events in an experiment, and observing the output, the BOLD response, and then comparing the degree to which the response predicted by the neuronal state model fits the observed data.

The hidden neuronal state equation can be written as:

$$\dot{z} = (A + \sum u_j B^j) z + Cu$$

The hidden state, $\dot{z}$, reflects changing synaptic activity in a region over time, with the state of the region a function (1) of the current state of that region, z, (2) perturbing inputs to the system (if

they are specified as entering at this region) specified in the C matrix, (3) the influence of other regions, specified in the A matrix, reflecting the coupling between regions to all events, (4) context dependent changes, the modulation of the A matrix by a subset of events specified in the B^j matrix. The parameters A B^j C are coupling matrices that can be estimated via Bayesian inversion, with the strength of the coupling between regions measured as a rate of change constant (measured in Hertz) that reflect how the rate of change in one region affects the rate of change in another. The result of integrating the above neuronal model is the extraction of a time series of predicted neural activity, which is in turn passed through a hemodynamic state model which translates the neural time series to a BOLD time series; this model describes how neuronal activity induces changes in vasodilation, blood flow, volume, deoxyhemoglobin, and eventually the BOLD signal (Stephan et al., 2007).

DCM uses a Bayesian framework for estimating the parameters of both the neuronal and hemodynamic state models, such that each parameter is represented by a Gaussian probability distribution and a prior mean value and variance. As these models are furnished with a large number of parameters, these priors can be used to constrain the parameters and reduce overfitting. In the case of the neuronal state model the priors for the parameters are referred to as shrinkage priors; coupling parameters are assigned zero mean priors ensuring that they will be estimated as zero in the absence of contrary evidence. The priors for the hemodynamic model by contrast reflect knowledge concerning the range of values which should be expected based on prior experimentation (empirical priors); additional principled priors can be used to constrain parameters so that they cannot be negative. Each prior has an associated variance reflecting the level of confidence in the assigned mean prior; priors with tight variances are less likely to change during the estimation procedure. Using a Bayesian approach the posterior probability for the parameter values are estimated via inversion, such that the parameters are adjusted to maximize the similarity between the data predicted by the model and the observed BOLD response (whilst also taking into account the need to constrain model complexity), with the parameters updated iteratively in the light of the data until convergence, the point at which changing the parameter fails to increase the probability further.

In DCM hypotheses can be investigated at the level of overall model structure (finding a best model) or at the level of the individual parameters (Stephan et al., 2010). When examining model

structure, a number of models are generated that manipulate aspects of model structure that embody hypotheses concerning the functional relationship between regions, for example the presence/absence of particular connections in the A matrix. The parameters of these different models are then estimated, and a decision is made as to which predicted model most closely fits the observed data (within the space of models tested). Models are contrasted by comparing their model evidences; the probability of getting the observed data given the model or rather their "free energy" (F value) which is an approximation to the model evidence (Stephan et al., 2007). This can be achieved by calculating the ratio of model evidences to produce a Bayes factor, as the F values represent the log of the model evidence, the log Bayes factor can be attained by subtracting the F values. A difference in $F > 3$ is strong evidence in favour of a model (Kass and Raftery, 1995). The process of choosing a "best model" is referred to as Bayesian Model Selection (BMS).

At the group level DCM can be conducted as a fixed or a random effects analysis. A fixed effects analysis is warranted when investigating basic physiological responses that are unlikely to vary significantly between subjects, whereas random effects analyses are suggested in the instance where complex cognitive functions are investigated (Stephan et al., 2010). In fixed effects BMS the F values of models are subtracted for each subject and their differences summed to generate a group Bayes factor, with the same evidence criterion used as was previously described. In random effects analysis, which is conducted in this thesis, the probability that the data of a random subject was generated by a specific model (referred to as the expected probability) and the probability that a specific model is more likely than the other models in the model space can be calculated (the exceedance probability). The exceedance probabilities sum to one, if the probability for any single model is greater than 0.95 this is strong evidence of a winning model.

If a winning model is identified by BMS it is common practice to conduct further inferences on the parameters of the winning model to establish whether they are significantly greater than zero, this is simply done by extracting the parameters for each subject and conducting one sample t-tests to establish whether the population of scores are significantly different to zero. Note that it is not necessarily the case that every parameter within a winning model is significant.

In some circumstances it might be of interest to ask a question concerning a particular aspect of model structure, such as whether forward or backward connections are more important. In this circumstance it is possible to partition the model space into families based on structural features and conduct model selection upon these families rather than individual models (Penny et al., 2010). This is a particularly useful technique if there is no clear winning model. The chances of finding a clear winning model can be less when there are a large number of models and when random effects analyses are conducted, particularly if a number of models share an important feature. One approach to deal with this problem is to partition the models into families based on shared structural features and then to conduct Bayesian Model Averaging (BMA) on the winning models. This so called family level inference approach is conducted in this thesis in Chapter 4.

In BMA after model estimation the expected probability of each model for each subject is calculated. The estimated connection strengths are sampled repeatedly for the models according to the expected probabilities for each subject, with more probable models sampled more frequently. This is conducted subject by subject and finally the parameters are averaged across subjects giving a distribution of values for each parameter. The distribution of values can be tested to see if they are greater than zero.

2.11 Data acknowledgment and statement of publications

I am indebted to Jeong Kyong for designing the experiment and collecting the data which is used in Chapter 3. I conducted all the analyses and interpretation presented herein. The resulting manuscript is currently under revision at Cerebral Cortex: *The pathway for Intelligible Speech: A reply to Okada et al. (2010).* I am a joint first author on this paper. I helped to collect, contributed to the experimental design and conducted analysis of the data that constitutes Chapters 4 & 5 of this thesis. The manuscript that represents work from Chapter 4 is currently under revision at the Journal of Cognitive Neuroscience: *Left-dominant decoding of speech intelligibility: Evidence from*

univariate and multivariate analyses of functional imaging data. I am a joint first author on this paper with Carolyn McGettigan who also helped to design, analyse and collect data for this study. Poonam Shah and Zarinah Agnew also contributed to data collection. I was responsible for the experimental design and data collection for the study that is included in Chapter 6. Zarinah Agnew and Carolyn McGettigan assisted in data collection for this study. Research arising from the literature review in this thesis contributed to the following paper: Scott SK, Evans S (2010) *Categorizing speech.* Nature Neuroscience 13:1304-1306.

Chapter 3 : EXPERIMENT 1

3.1 CHAPTER SUMMARY

Okada and colleagues (2010) recently published a replication of Scott et al. (2000), a study which examined neural responses to speech intelligibility. In the original study neural responses to intelligible speech, as contrasted with acoustic complexity, were identified in the left anterior STS. Okada et al. used equivalent stimuli and both univariate and multivariate analyses to argue against the importance of left anterior temporal cortex and instead for the role of bilateral posterior regions in resolving intelligible speech. In this chapter a similar re-analysis using univariate and multivariate methods is conducted on data derived from a replication of Scott et al. (2000) with the aim of interrogating the prospective roles of bilateral anterior and posterior temporal cortex.

3.2 INTRODUCTION

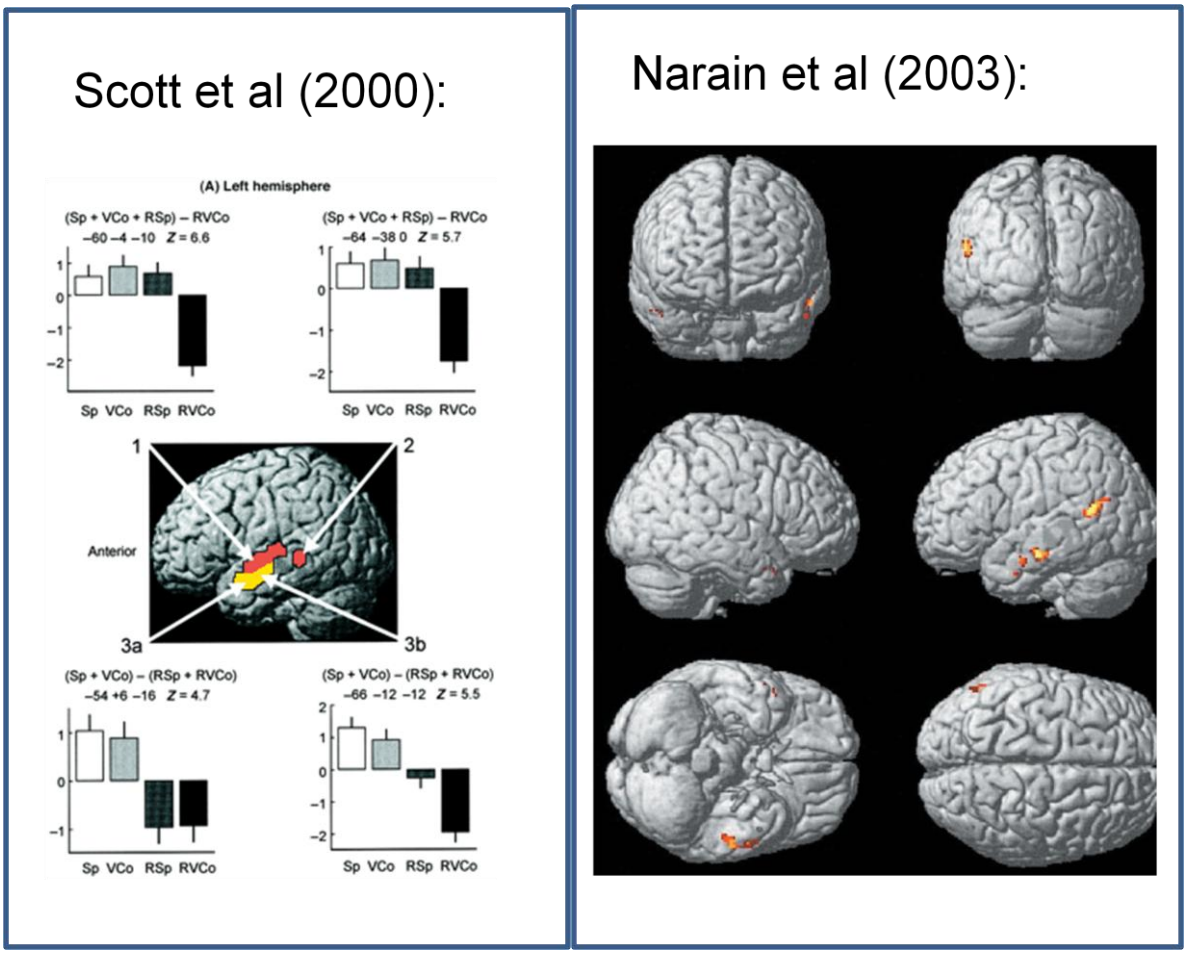
Many functional imaging studies have attempted to isolate neural regions that are sensitive to intelligible speech as compared to those regions which respond to acoustic complexity in the absence of intelligibility. The selection of a suitable baseline comparison condition has often proved difficult due to the inherent acoustic complexity of the speech signal. Using rotated speech (Blesser, 1972), which is well matched to speech in both spectral and amplitude variation, it has been shown that the left anterior superior temporal sulcus (STS) responds preferentially to intelligible speech (Narain et al., 2003;Scott et al., 2000). These studies looked for the commonality in neural response to two kinds of intelligible speech which differed in surface acoustic structure: clear speech (clear) and noise-vocoded speech (NV) (Shannon et al., 1995), as contrasted with two unintelligible sounds: rotated (rot) and rotated-noise-vocoded speech (rotNV). These studies employed traditional univariate

statistical analyses, the GLM, to predict the time series at each individual voxel and showed regions evidencing a relative increase in BOLD signal to intelligible as contrasted with unintelligible sounds.

The original Scott et al. (2000) PET study showed, using contrast estimate plots, that neural responses became increasingly invariant to the surface acoustic structure of the speech signal and more sensitive to differences in intelligibility as responses progressed antero-laterally towards the temporal pole. Left posterior STS and a mid STG region in the left hemisphere exhibited an increased response to stimuli with any phonetic content regardless of its intelligibility (see the red region and response plots 1 + 2 in Figure 3.1 left). Within regions which showed a main effect of intelligibility, the mid STS showed an increase in response to the intelligible conditions and a greater degree of differentiation within stimulus type (region in yellow and plot 3b) compared to the most anterior STS region which showed a strong intelligibility effect but less differentiation within conditions (plot 3a). Note that the plot in 3a is most indicative of an archetypical intelligibility response; showing a large equivalent increase to both intelligible conditions, and reduced a equivalent decrease in signal to both the unintelligible conditions. The Narain et al. (2003) study, an fMRI replication of Scott et al., conducted the global null conjunction of the two intelligibility subtractions [clear – rot] and [NV - rotNV], and found both left anterior and posterior STS activations (figure 3.1 right). Thus whilst both studies reported left anterior STS activations, the Narain et al. study also implicated an addition left lateralised posterior activation.

Figure 3.1 The results from Scott et al. (2000) (left) and Narain et al. 2003 (right). Note the key to the abbreviation on the left: Sp=clear speech, VCo=noise-vocoding, RSp=rotated speech and RVCo=rotated noise-vocoded.

Reproduced with permission.

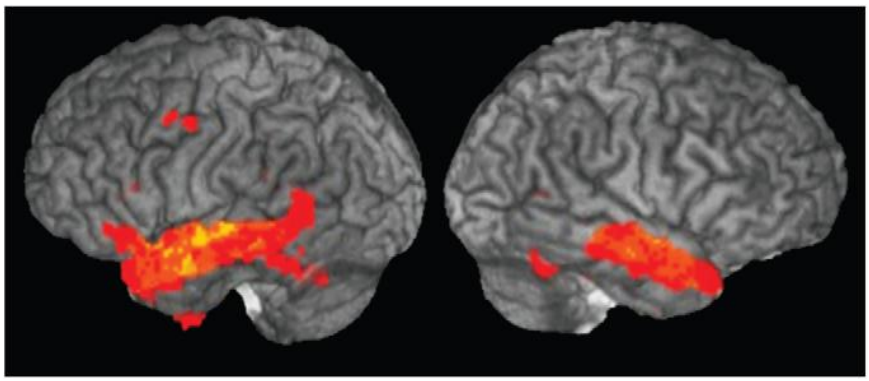


These studies have proved influential, with the Scott et al. study at the time of writing having been cited 407 times (ISI world of knowledge, 15th July 2011). A recent study in Cerebral Cortex [Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G. 2010. Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. Cereb Cortex 20: 2486-2495] replicated the Scott et al. (2000) methodology with fMRI. They conducted a univariate analysis that showed widespread bilateral activation spanning anterior and posterior temporal cortex to the average of clear speech and noise-

vocoded speech (a main effect of intelligibility), relative to their unintelligible rotated equivalents (see Figure 3.2).

Figure 3.2 Univariate analysis from Okada et al. (2010). The average of the two intelligible conditions subtracted from the unintelligible conditions: [clear + NV] – [rot + rotNV].

Reproduced with permission.



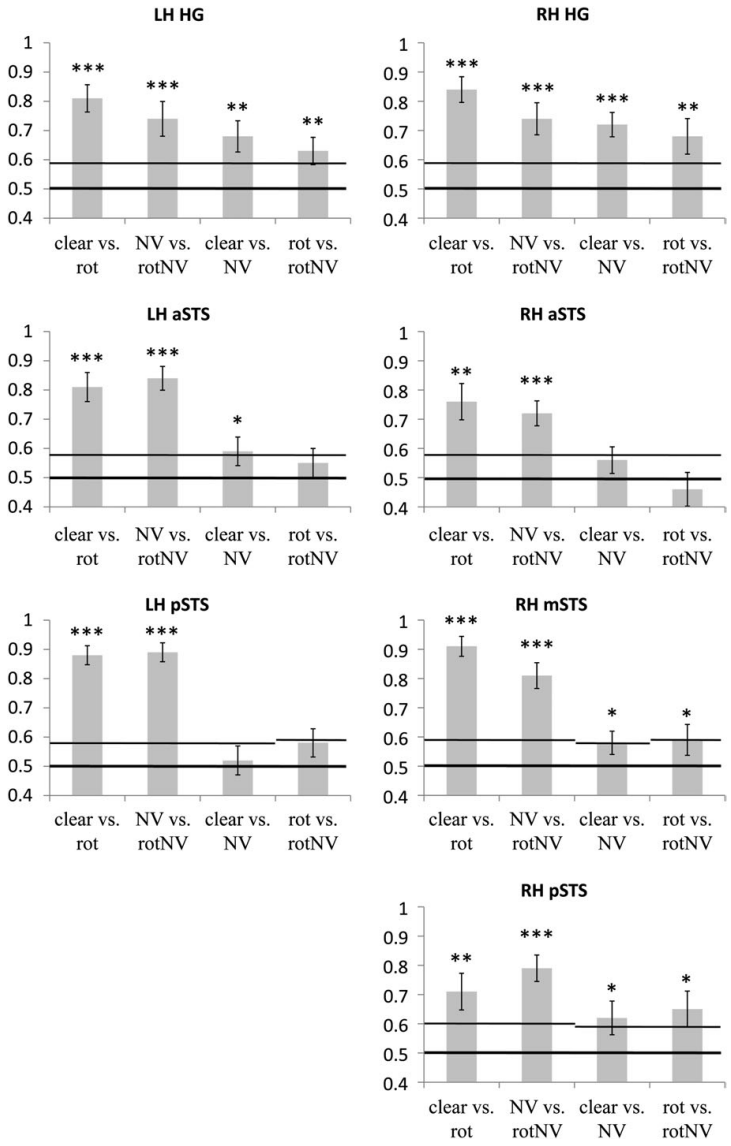
Further to this, the authors conducted a multivariate pattern analysis using a SVM within ROIs in HG and the STS. HG was defined individually in each subject using a localiser scan of [noise - rest], whilst regions in the STS were defined with a subset of data using the contrast of [clear - rot]. In the case of the STS ROIs, data was extracted from 7^3 voxel cubes at peaks defined as anterior, posterior or mid, dependent on where they fell within the STS. This distinction was derived by defining within each subject, the region more anterior than the anterior most point of Heschl's Gyrus as anterior STS, posterior to the posterior most point as posterior STS, and a mid region representing the region between these points. As they failed to find consistent activation in the left mid region they excluded it, leaving only anterior and posterior ROIs in the left but anterior, mid and posterior ROIs in the right.

They then conducted pair wise classifications of the most closely acoustically controlled intelligible/unintelligible pairings: [Clear vs. rot] and [NV vs. rotNV], in addition to classifications

that they argued differed predominately on an "acoustic basis": [Clear vs. NV] and [rot vs. rotNV]. This was despite the fact that noise-vocoded speech differs in intelligibility to clear speech (Scott et al., 2000). They argued that a region principally involved in resolving intelligibility should maximally separate stimuli that differ in intelligibility, whilst performing poorly at separating stimuli that differ on an "acoustic" basis.

HG bilaterally was shown to be able to separate all the intelligibility and acoustic classifications at a level greater than chance (see Figure 3.3). The regions of the STS were able to separate intelligible from unintelligible speech at a level greater than chance, but had a variable ability in separating the "acoustic" contrasts. As the left anterior STS region was able to distinguish between clear speech and noise-vocoded speech at a level just greater than chance, they argued that it made it an unlikely candidate region for resolving intelligible speech. It should be noted however that both right mid and posterior STS also separated the acoustic classifications at a level greater than chance, with both right anterior and left posterior STS the only regions to conform to the criterion they suggest for an intelligibility selective region. Note therefore that Okada et al.'s results are not clear cut in supporting either bilateral posterior or left anterior STS as the key region/s involved in resolving intelligible speech as per their adopted criterion.

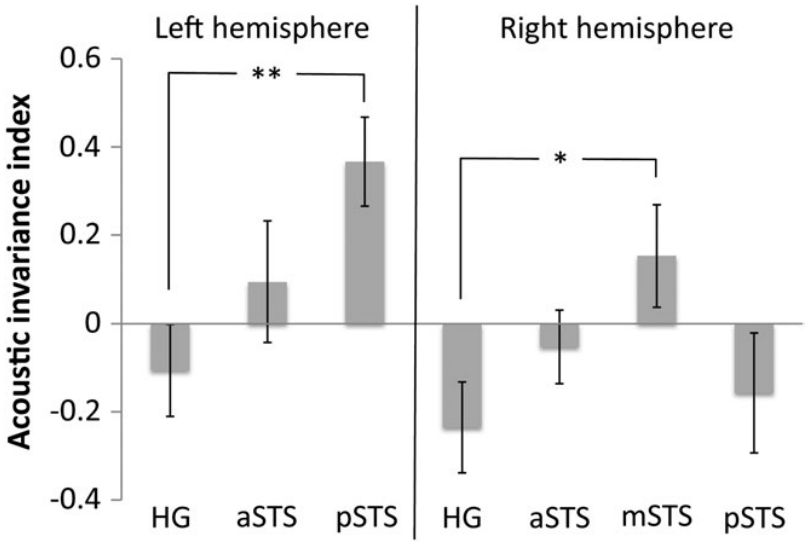
Figure 3.3 MVPA classifications from Okada et al. (2010).



Rather than statistically compare the raw classification scores of the different regions directly they calculated an "acoustic invariance index". This index transforms the classification scores so as to make intelligibility performance relative to acoustic performance. This was achieved by taking the sum of the 2 intelligibility classification scores and subtracting the 2 acoustic classification scores, and then subtracting the sum of the absolute values of the acoustic effects. With this they demonstrated that left HG and left pSTS, and right HG and right midSTS differed significantly (see Figure 3.4), the significance of this being that HG should show the greatest sensitivity to acoustic distinctions. Note however that anterior and posterior STS did not differ significantly from each other

in either hemisphere, and that they failed to demonstrate a difference between HG and right posterior STS.

Figure 3.4 The results of the acoustic invariance index in Okada et al. (2010).



To summarise Okada et al.'s findings, they found widespread bilateral activation in anterior and posterior temporal cortex to the "main effect of intelligibility". Using multivariate pattern analysis they demonstrated that both anterior and posterior temporal cortex bilaterally could separate intelligible speech from unintelligible sounds at a level greater than chance. The anterior STS additionally separated an "acoustic" contrast of clear speech from noise vocoded speech, which they argue is a profile incompatible with a region exclusively involved in resolving intelligible speech. By using a metric that transforms accuracy on the intelligibility contrasts to be relative to the "acoustic" contrasts, they demonstrated that left but not bilateral posterior STS showed a profile of response significantly different to HG.

There were a number of omissions in the analyses conducted by Okada et al. Firstly, by presenting only the main effects of intelligibility in their univariate analysis, the authors did not make

use of the factorial design of the study in their analysis. As a result they could not identify whether there were regions showing an interaction between rotation and vocoding, which might indicate areas where the response to the two intelligibility simple effects were not equivalent. Also by only examining classification within 7^3 voxel cubes they could not comment on how information might be integrated across anatomical regions. Finally as they failed to examine the weight vector of their classifiers, which quantifies the relative contribution of each voxel to defining the classification boundary, they could gain no insight into how classification was achieved. In the following chapter univariate and multivariate pattern analyses are conducted on data acquired from a replication of Scott et al (2000) to address these methodological issues. A factorial univariate analysis is conducted and classification is carried out within local discriminative patches and within and across anatomical regions; by extracting the classifier weights an attempt is also made to understand how information is classified in different neural regions.

3.3 METHOD

Participants

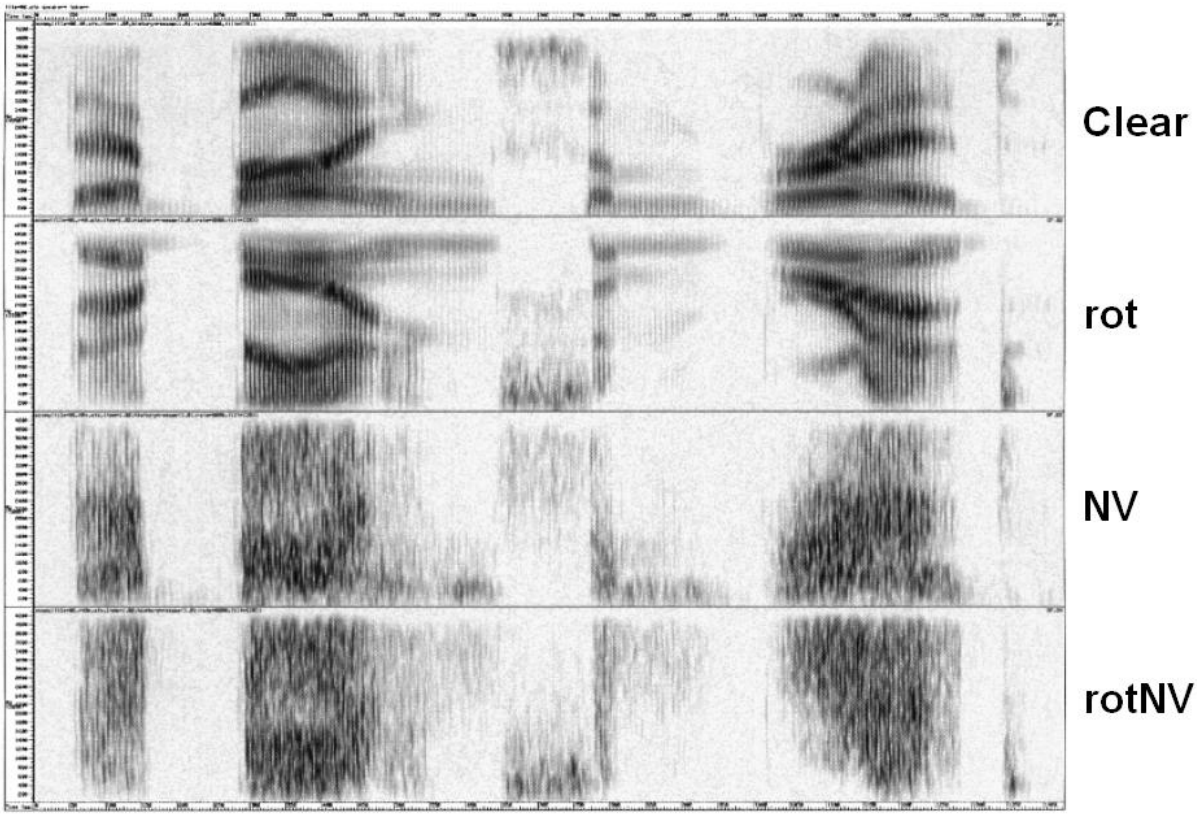
Twelve right handed subjects participated in the experiment (aged 18-38, mean age 25, 3 male). All subjects were native speakers of English with no known hearing or language impairments. All subjects gave informed consent and the experiment was performed with the approval of the local ethics committee of the Hammersmith Hospital.

Stimuli

All stimuli were drawn from low pass filtered (3.8 kHz) digital representations of the Bamford-Kowel-Bench (BKB) sentence corpus (Bench et al., 1979). There were four stimulus conditions: natural speech (clear), noise-vocoded (NV), spectrally-rotated (rot) and rotated-noise-vocoded speech (rotNV). The rotation of speech is achieved by inverting the frequency spectrum around 2 kHz using a simple modulation technique; this retains spectral and temporal complexity but makes the speech unintelligible (Blesser, 1972). It has been described previously as sounding like an alien speaking your language with different articulators (Blesser, 1972). A filter was used to give the rotated speech approximately the same long-term average spectrum as the original un-rotated speech using measurements derived from Byrne et al. (1994). Noise-vocoding involves passing the speech signal through a filter bank (in this case 6 filters) to extract the time-varying envelopes associated with the energy in each spectral channel. The extracted envelopes are then multiplied by white noise and combined after re-filtering (Shannon et al., 1995). This retains the amplitude envelope cues within specified spectral bands but removes spectral detail. With six bands, the speech can be understood with a small amount of training but sounds like a harsh whisper with only a weak sense of pitch. Subjects underwent a short training, as described in Scott et al. (2000), to ensure that they understood the NV speech. The combination of vocoding and rotation sounds like intermittent static noise with weak pitch changes. It does not contain recognisable phonetic content and is not intelligible or recognisable as speech. See Figure 3.5 for spectrograms of example stimuli.

The clear and NV conditions are both intelligible, whilst the rot and rotNV conditions are both unintelligible. The distinction between the two intelligible, clear and NV, and between the two unintelligible conditions, rot and rotNV, represents a difference in “acoustics” as defined by the schema of Okada et al.

Figure 3.5 Spectrograms of the stimuli.



Functional Neuroimaging

Subjects were scanned on a Philips (Philips Medical Systems, Best, The Netherlands) Intera 3.0 Tesla MRI scanner using Nova Dual gradients, a phased array head coil and sensitivity encoding (SENSE) with an underlying sampling factor of 2. Functional MRI images were acquired using a T2*-weighted gradient echo planar imaging sequence which covered the whole-brain (repetition time: 10s, acquisition time: 2s, TE: 30ms, flip angle: 90º). Thirty-two axial slices with a slice thickness of 3.25mm and interslice gap of 0.75mm were acquired (resolution: 2.19 x 2.19 x 4.00mm; field of view 280 x 224 x 128mm). Quadratic shim gradients were used to correct for magnetic field inhomogeneities. T1 images were acquired for all subjects (resolution=1.20 x 0.93 x 0.93mm).

Participants listened to the sounds delivered via an MR-compatible binaural headphone set (MR confon Gmbh, Magdeburg, Germany). All the stimuli were presented using E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA, USA) installed on an MR interfacing IFIS-SA system (Invivo Corporation, Orlando, FL, USA).

Data were acquired using sparse acquisition which ensured that the stimuli were presented in silence (Hall et al., 1999). Stimuli were presented during a 7.5s MR silent period which was followed by a 2s image acquisition and a 0.5s silence. Two runs of data were acquired, with each run consisting of 24 trials of each condition presented in a pseudo-randomised order (96 trials/volumes per run). A total of 192 trials/volumes were acquired for each subject. Each trial comprised three randomly selected sentences, with each sentence less than 2s in duration. Subjects listened passively to the sentences in the scanner and were instructed to try and understand each sentence.

Data analysis

Univariate Analysis

Data were analysed using Statistical Parametric Mapping (SPM8; http://www.fil.ion.ucl.ac.uk/spm/). Scans were realigned, un-warped and spatially normalised using the parameters arising from the segmentation of each participant's T1-weighted image, and smoothed using an isotropic Gaussian kernel of 8 mm full-width at half maximum. A first order Finite Impulse Response filter with a window length equal to the time taken to acquire a single volume, effectively a box car function, was used to model the hemodynamic response (de Zubicaray et al., 2007). A high pass filter with a time constant of 128s was applied to remove low frequency noise.

The four stimulus conditions (and 6 movement regressors of no interest) were entered into a general linear model at the first level. The first level con images of each condition were entered into a factorial within subjects ANOVA with the factors: vocoding (acoustic manipulation) and rotation (intelligibility manipulation), at the second level. All statistical maps were False Discovery Rate (FDR) corrected at $p<0.05$, unless otherwise stated. No minimum cluster extent was imposed on the statistical maps, however for the sake of brevity only cluster extents above 10 are reported in tables detailing activations.

Multivariate Pattern Analysis.

Classification was conducted on unsmoothed images in each subject's native space. Functional images were un-warped and realigned to the first acquired volume using SPM8. Training/test examples were constructed from single volumes. Linear and quadratic trends were removed and the data z-scored within each run. The data were separated into training and test sets by run to ensure that training data did not influence testing (Kriegeskorte et al., 2009).

Two different multivariate analysis approaches were adopted. In the first instance a searchlight analysis was conducted which examined the distribution of local information within the temporal lobes (Kriegeskorte et al., 2006). In a searchlight analysis the classification of each voxel and its immediate surrounding neighbours is considered for all the voxels within an anatomical region. The voxel at the centre of the neighbourhood is assigned the classification performance of its local neighbourhood. The Searchmight toolbox was used for this analysis (http://minerva.csbmb.princeton.edu/searchmight). Classifications were conducted using a linear support vector machine (Libsvm) in 7^3 voxel cubes centred at each voxel within a bilateral temporal lobe mask. The C parameter controls the trade off between maximising the size of the margin (the distance between the support vectors and the boundary) and the number of misclassified data points. This was set to be equal to the default: 1/number of features to allow some misclassifications to occur.

A hard margin, a C value approaching infinity, ensures that no examples are misclassified in the training set. This cannot be used in circumstances when the conditions are not linearly separable as was the case using these small regions of interest. Local information analyses were conducted to replicate all the pair wise classifications conducted by Okada et al. The searchlight procedure was carried out in each subjects' native space. After cross validating classifier performance using each held out run (Kriegeskorte et al., 2009), the probability of obtaining the given classification result at each voxel was assessed against a binomial distribution. An FDR correction of $p<0.01$ was applied to control for false positives. These FDR corrected binary maps were transformed from native space to MNI space, using the parameters acquired from segmentation, to allow comparison across subjects.

In addition to a local analysis, a more global analysis was conducted which assessed the classification performance of whole discrete anatomical regions. This allows understanding of how information might be integrated within and across whole anatomical structures, rather than within small discriminative patches. For the global analyses, the linear Support Vector Machine (SVM) from the Spider toolbox (http://www.kyb.tuebingen.mpg.de/bs/people/spider/) with the Andre optimization and a hard margin was used. Note a hard margin could be used in this instance due to the larger regions of interest which provides a more complex feature space that always makes it possible to separate the conditions. An intelligibility contrast was created by constructing an intelligible class from volumes of clear and NV stimuli, and an unintelligible class from volumes of rot and rotNV stimuli. By using a hard margin and collapsing the two intelligible and two unintelligible conditions, the SVM must find a separating hyperplane that exactly separates Clear and NV from the rot and rotNV conditions, this can be thought to be the conceptual equivalent of finding the conjunction null. The first classifier was trained on the first run and tested on the second, and vice versa for the second classifier. The "true" accuracy of classification was estimated by averaging the performance across the two classifiers for each subject.

Anatomical Regions of Interest (ROIs) were constructed for each subject to reduce the number of voxels in the analysis. ROIs for Heschl's Gyri were hand drawn on T1 structural images that had been coregistered to the mean functional image of each subject using the definitions of

Penhune et al. (1996) and the Anatomist software (http://brainvisa.info/). For larger anatomical areas, ROIs from the AAL ROI library were used, which had previously been defined by hand on a brain matched to the MNI/ICBM template using the definitions of Tzourio-Mazoyer et al. (2002), available via the Marsbar toolbox (Brett et al., 2002). These ROIs were transformed into the native space of each subject via the inverse normalisation parameters identified via segmentation. ROIs included: Superior Temporal Gyrus (STG), Middle Temporal Gyrus (MTG) and Inferior Temporal Gyrus (ITG) of the left and right, and combined hemispheres. An ROI was also constructed consisting of the concatenation of all the bilateral temporal and Heschl's gyri. An ROI in the visual cortex, the Inferior Occipital Gyrus (IOG), was also included as a control region. Note that the searchlight procedure was only conducted using the ROI which included all the bilateral temporal and HG.

3.4 RESULTS

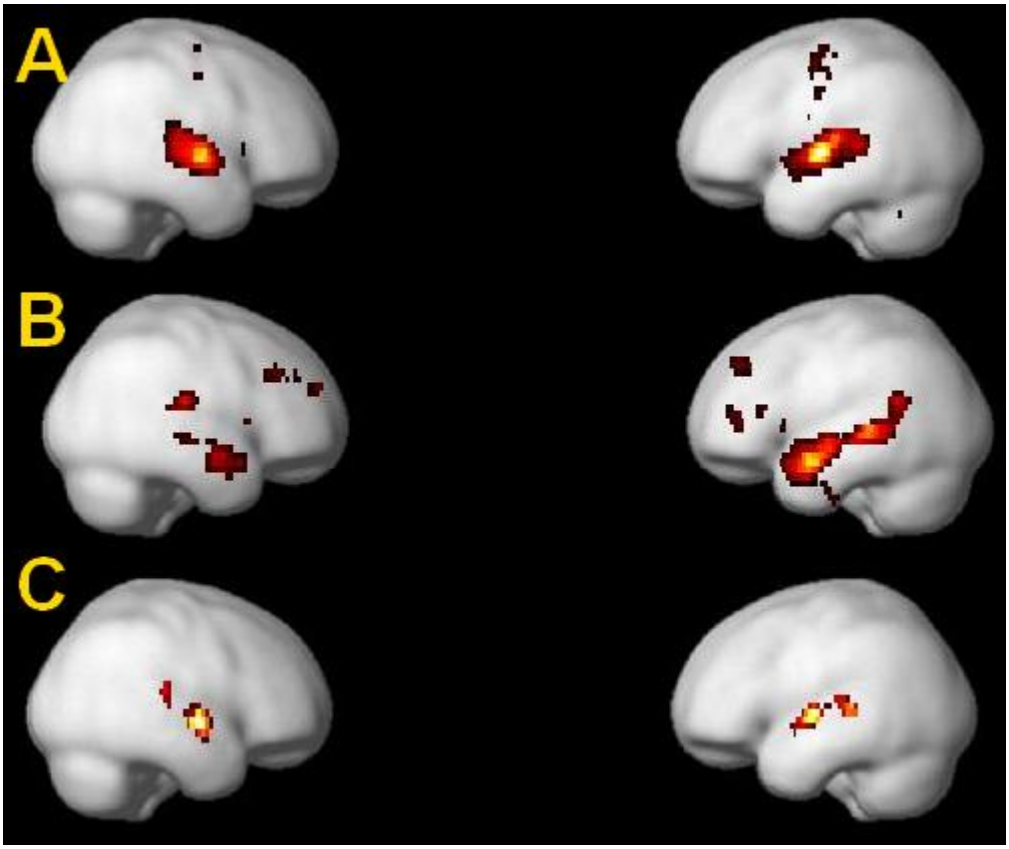
Univariate Analysis

Okada et al. (2010) only reported the main effect of intelligibility: [Clear +NV]-[rot+rotNV]. In contrast, here a factorial analysis was conducted to allow the examination of the main effects of rotation (intelligibility manipulation) and vocoding (acoustic manipulation), and their interaction. For the main effect of vocoding, [Clear + Rot]-[NV+rotNV], clusters of activation were focused predominantly within the temporal lobes bilaterally, with activation spreading within HG and across both the superior and middle temporal gyrus, and additional small clusters were identified within the supplementary motor area (SMA) and pre and post central gyri (see Figure 3.6A). The largest peak level activations were found in bilateral HG, and the superior and middle temporal gyri. The observed strong bilateral activations are in contrast to the right lateralised activations for the same contrast found in Scott et al. (2000). In Scott et al. this contrast was reported as representing a difference between the presence versus absence of pitch. It should be noted however that noise-vocoding in this

instance is not totally absent of a pitch percept as regularities in the amplitude envelopes across and within frequency channels provide a weak sense of pitch. In addition rotation creates a slightly unnatural pitch percept as it maintains the regular spacing of harmonics but changes their absolute frequencies. Narain et al. (2003) did not report the main effect of vocoding due to the above described confounds; whilst the relevance of the pitch contrast have diminished, the relevance of the intelligibility contrasts have endured, it is for this reason that the main focus of this chapter is upon intelligibility. The described confounds, combined with the increased statistical power of this study compared to the original PET study, may explain why the pattern of right lateralisation was not replicated.

The main effect of intelligibility, [Clear + Rot]-[NV+rotNV], was associated with clusters of activation which spread across both anterior and posterior superior temporal cortex bilaterally extending into the supramarginal gyrus on the right, with additional small clusters of activity found within bilateral anterior cingulate and prefrontal cortex. The largest peak level activations were found bilaterally within anterior and posterior STS, with the largest peak in the left anterior STS (see Figure 3.6B). Note the similarity in the activation pattern within the temporal lobes between this image and the statistical map presented by Okada et al. for the same contrast (Figure 3.2). Only 0.7% of the cluster within left STG was observed to fall within primary auditory cortex, with no activation falling within primary auditory cortex in the right.

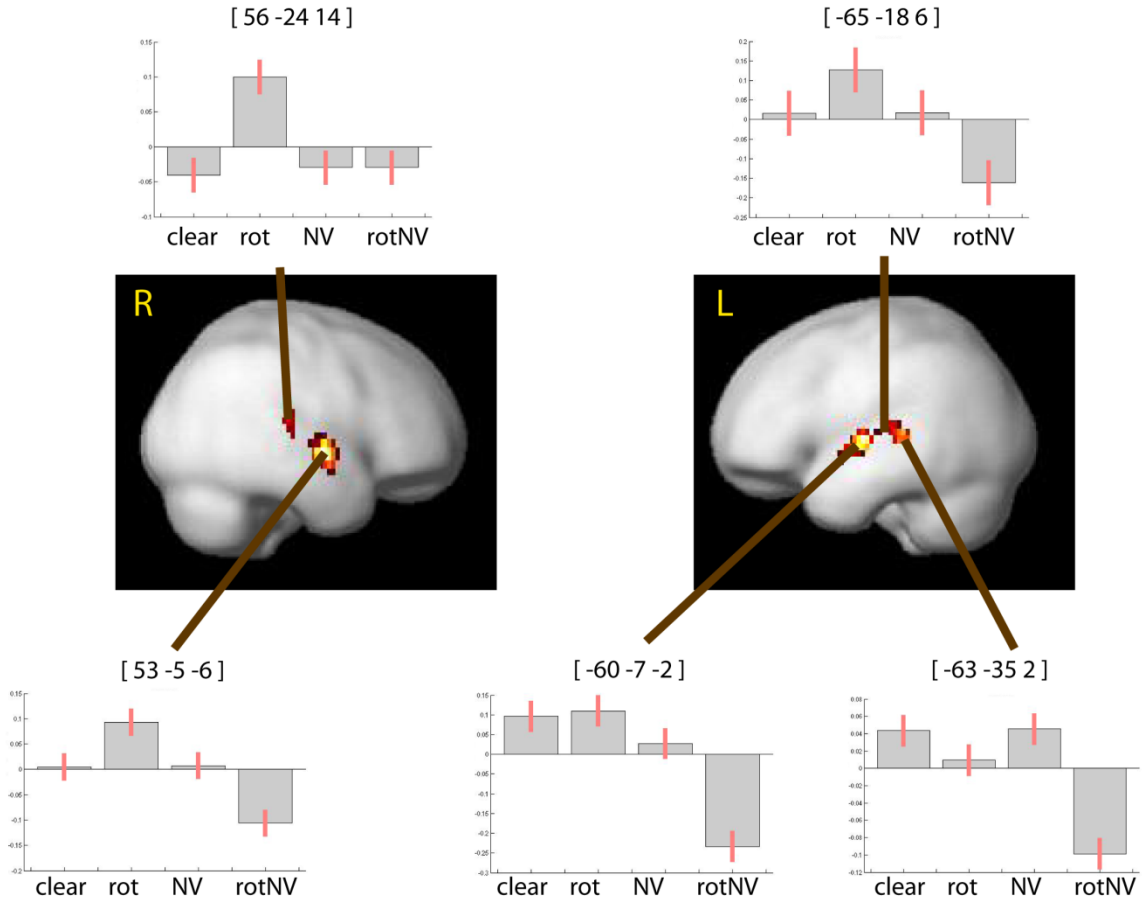
Figure 3.6 From the top: main effects of vocoding (A), rotation (B) and the interaction between vocoding and rotation (C).



When the f-test of the interaction between vocoding and rotation was examined, clusters of activation were found predominantly within bilateral mid-posterior temporal cortex extending into the STS both posteriorly and more anteriorly (see Figure 3.6C). Peak level activations were found in left mid-anterior STG, PT and posterior STS, and right mid-anterior STG and PT. For the sake of completeness response plots for all the peaks are shown see Figure 3.7. In the left hemisphere the plots were characterised by similar responses to clear, rot and NV and a relative deactivation to rotNV. Note that the confidence intervals in the left posterior STS [-63 -35 2] overlapped between clear, rot and NV, but not with rotNV. In the right hemisphere both peaks showed the greatest response to rot, with the right PT peak [56 -24 14] showing equivalent reduced responses to clear, NV, rotNV, and the right mid-anterior region [53 -5 -6] showing a similar profile but registering an intermediate response to NV and clear speech.

Figure 3.7 Response plots (including the 95% confidence interval) showing the parameter estimates at the peak level activations for the interaction between rotation and vocoding.

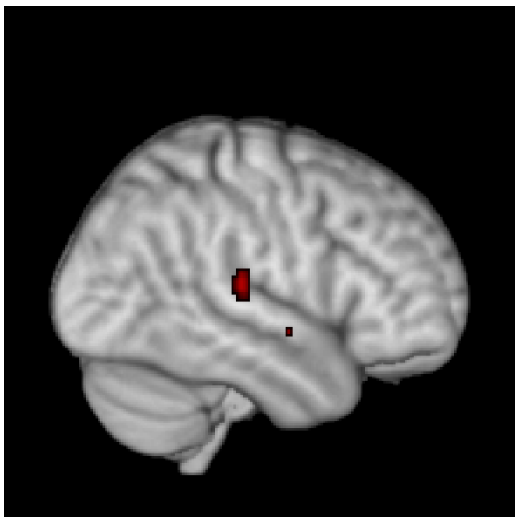
Note that as there was no implicit baseline, plots are relative to the mean parameter value across conditions rather than to a baseline.



To explore the nature of these interactions in more detail, the interaction was masked inclusively with the simple intelligibility effects at the same corrected threshold. This showed that a region in right PT and right mid STG responded more strongly to rotated speech than any other condition (see Figure 3.8). When the interaction in left posterior STS was specifically examined, no

significant difference in any direction was shown between clear, rot or NV, but there was a relative increase to all those conditions relative to rotNV.

Figure 3.8 Regions within the interaction responding more to rot than any other condition.



Having demonstrated a significant interaction between the factors, the simple effects of the intelligibility subtractions [clear - rot] and [NV - rotNV] were examined (see Table 3.1). As suggested by the interaction, the simple intelligibility effects generated very different statistical maps: [clear - rot] activated a region exclusively within the left anterior STS (Figure 3.9A), while the [NV - rotNV] activated large clusters extending along much of the length of the STS bilaterally, with a large cluster also centred in the left Inferior Frontal Gyrus (Figure 3.9B). The largest peaks were identified within left anterior STS for both contrasts.

A conjunction analysis was carried out to isolate activations common to the two individual intelligibility subtractions. It has been noted that there has been confusion in the past concerning the interpretation of conjunction analyses (Nichols et al., 2005). For the sake of clarity, there are two types of conjunction, the conjunction null and the global null conjunction (Nichols et al., 2005;Friston et al., 2005). The statistical maps that result from the conjunction null implicate voxels that survive a

specified threshold across all the individual subtractions that make up the conjunction. In contrast, the global null conjunction, displays voxels that show effects that are in a similar direction but that are not necessarily individually significant across all the subtractions that constitute the conjunction at the specified threshold. Therefore in the case of the global null conjunction it should be noted that a significant conjunction does not mean that all the contrasts were individually significant (i.e., a conjunction of significance). It rather means that the contrasts were consistently high and jointly significant. This is equivalent to inferring that one or more effects were present (Friston et al., 2005).

The only voxels that survived the conjunction null of the two individual intelligibility subtractions were found in the left anterior STS exclusively. Note that this region was non-overlapping with the region showing a significant interaction between rotation and vocoding. This statistical analysis reflects the concept of the original Scott et al. (2000) design, which used more than one intelligibility subtraction in an attempt to isolate a more invariant intelligibility response. The statistical map for the conjunction null of the two intelligible conditions is identical to the [clear - rot] subtraction alone (Figure 3.9A). This reflects the fact that the activation in the [clear - rot] subtraction is present in the [NV - rotNV] subtraction but not vice versa. Note for the sake of completeness, the conjunction null of all the four possible intelligibility subtractions, i.e. additionally inclusive of [Clear - rotNV] and [NV - rot], also activates the left anterior STS exclusively (not shown).

In Narain et al. (2003), both anterior and posterior peaks were identified using the global null conjunction rather than conjunction null analysis. It is possible that the posterior peaks identified in Narain et al. (2003) were driven by the effects of the [NV-rotNV] contrast – as also seems to be the case in this study. A global null conjunction reveals a similar pattern of activation to that found in Narain et al., with the activation cluster spreading across both anterior and posterior STS in the left, in addition to activation in the anterior inferior temporal gyrus and the right anterior STS which was not shown in Narain et al (Figure 3.9C).

Figure 3.9 Top to bottom: Simple intelligibility effects (A) clear - rot (B) NV - rotNV (C) the global null conjunction of the two simple effects.

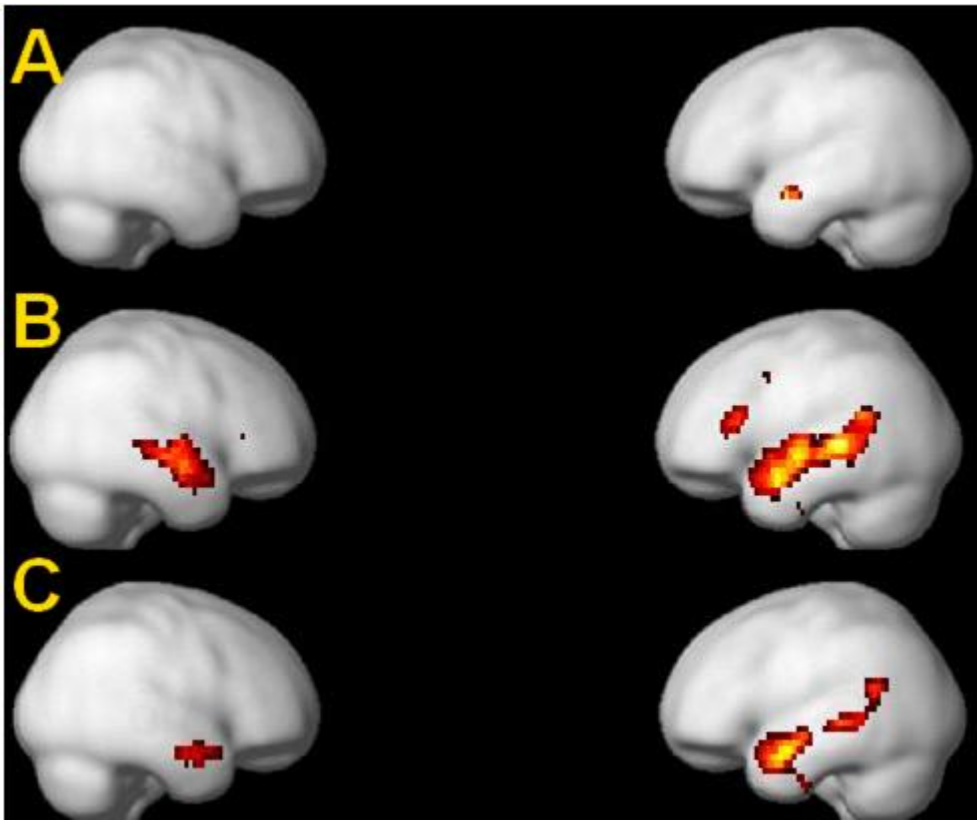
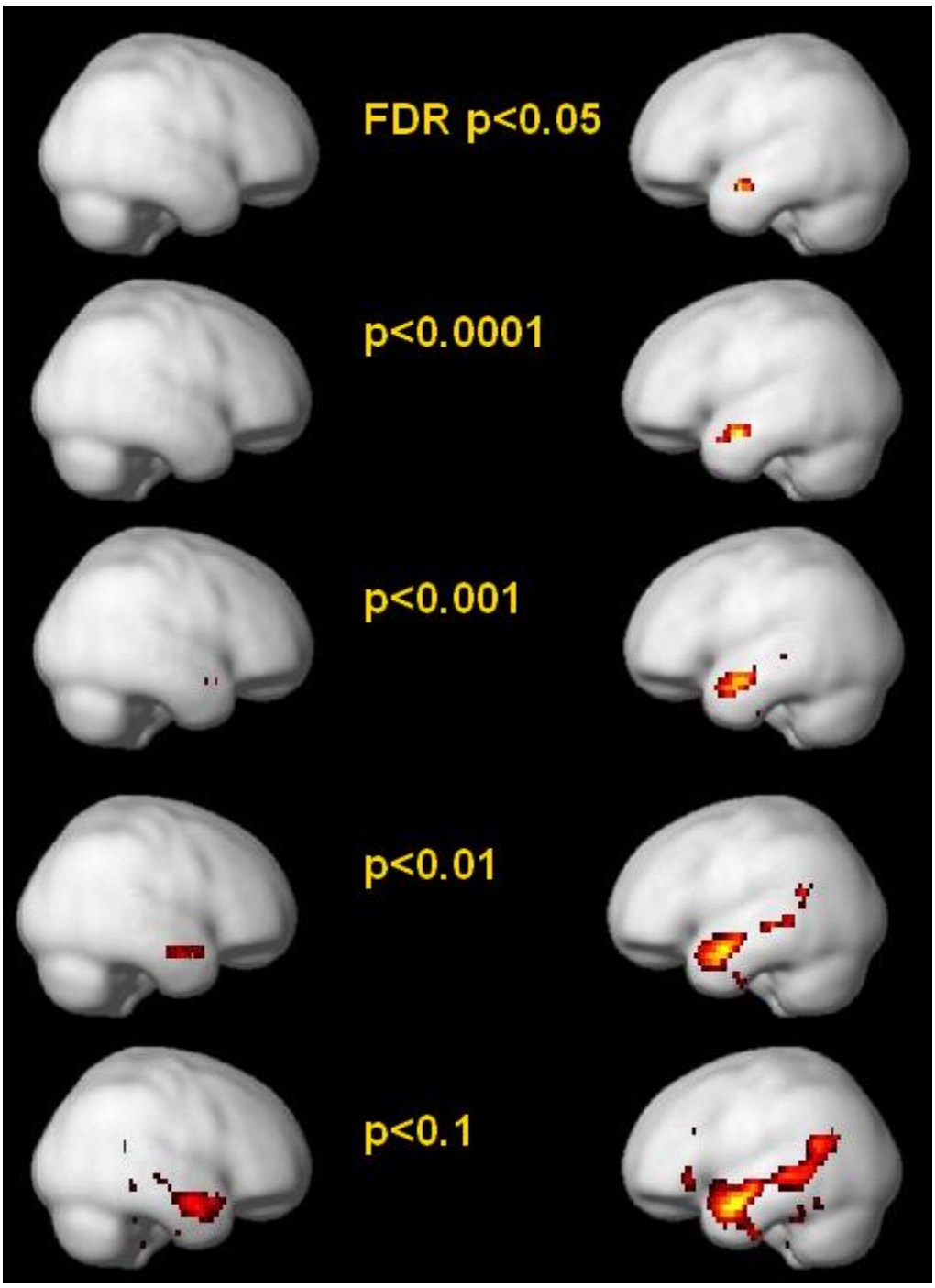


Table 3.1 Peak level activations, FDR p<0.05, voxel extent > 10.

Location	MNI X	y	Z	Extent	Z
[Clear - Rot]					
Left mid-anterior STS	-52	-5	-18	16	4.55
Left anterior STS	-58	2	-18		4.46
[NV - rotNV]					
Left anterior STS	-58	2	-14	1058	6.83
Left mid-posterior STS	-63	-35	2		6.46
Left mid-anterior STG	-60	-7	-6		6.13
Right anterior STS	60	0	-14	455	5.77
Right anterior STS	53	8	-18		4.78
Right mid-posterior STS	56	-18	-6		4.45
Left Inferior Frontal Gyrus	-50	28	14	96	4.42

Statistical maps for the conjunction null of the simple intelligibility effects were produced showing a range of statistical thresholds to understand whether the described effects were merely a thresholding effect (see Figure 3.10). This showed that posterior and bilateral responses did begin to emerge at more liberal thresholds, albeit the threshold had to be reduced significantly to evidence substantial activity in those regions (e.g. $p<0.01$).

Figure 3.10 The conjunction null of the simple intelligibility contrasts at a range of statistical thresholds.



To explore the data further, a contrast examining the main effect of intelligibility masking out the interaction at p<0.05 was conducted (see Figure 3.11). Again the largest peak was found in anterior STG and STS and additional activation was found in right anterior and left posterior STS.

Figure 3.11 The main effect of intelligibility masking out the interaction at p<0.05.

Multivariate Pattern Analysis

Local Information

In the local information maps each voxel represents the classification accuracy of a small cube of data centred at that voxel. Figure 3.12 shows voxels surviving FDR correction at a level of $p<0.01$, with the colour bar representing the concordance between subjects in implicating the same voxel/neighbourhood at that corrected level. Classifications were conducted for the two intelligibility contrasts: clear vs. rot (Figure 3.12A) and NV vs. rotNV (Figure 3.12B), and for the two "acoustic" contrasts: clear vs. NV (Figure 3.12C) and rot vs. rotNV (Figure 3.12D), to partially replicate the analyses of Okada et al.

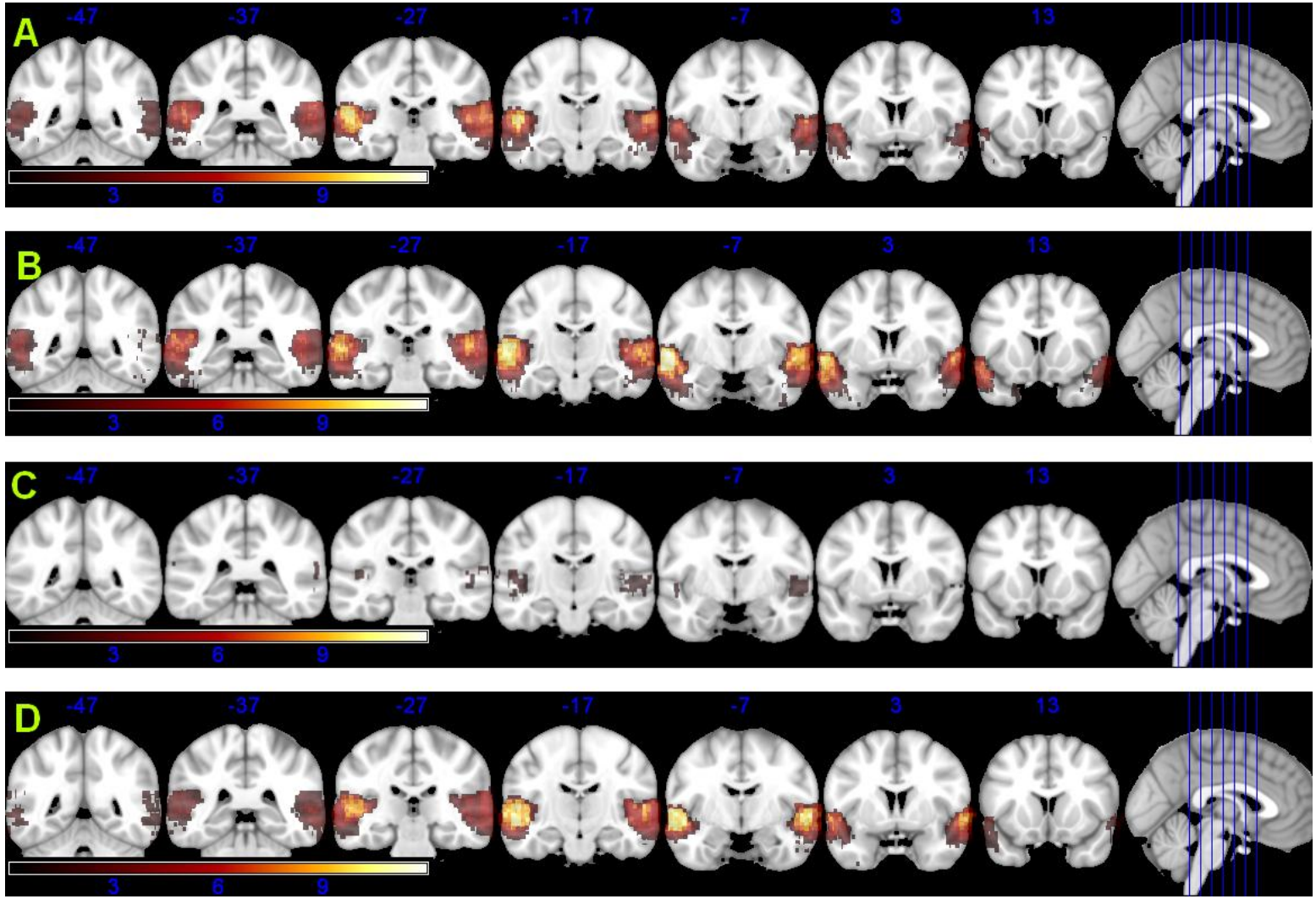
For the two intelligibility contrasts, the analyses demonstrated that local information was located predominantly bilaterally in both anterior and posterior superior temporal cortex, and also extending into HG. Visual inspection of these maps, suggests that there seemed to be a slightly greater concordance between subjects implicating mid-posterior regions for successful clear vs. rot

classifications, whereas the balance seemed to be more equal between anterior and posterior regions in the NV from rotNV classifications.

For the two acoustic contrasts, there was a larger number of successful voxel neighbourhoods and greater concordance between subjects in the NV vs. rotNV as contrasted with the Clear vs. NV classifications. In both sets of classifications neighbourhoods in HG showed above chance performance across subjects. For the Clear vs. NV, neighbourhoods in left HG and mid STG/STS, and in the right HG and mid-posterior STG/STS, showed above chance performance, although the same voxel neighbourhood was not implicated in more than three subjects. For the NV vs. rotNV the most discriminative region included and extending beyond HG to both anterior and posterior temporal regions.

Note that these local information analyses only take into account classification performance, and do not include information about the weight vector of the classification function.

Figure 3.12 Searchlight classifications of (A) clear vs. rot (B) NV vs. rot (C) clear vs. NV (D) rot vs. rotNV. Colour bar represents the number of subjects implicating the same voxel/neighbourhood as classifying at an above chance level.



Global Information

Classifications were then conducted using all the voxels within discrete anatomical regions; this allowed the examination of how information is integrated within whole anatomical structures rather than just looking at the performance of small discriminative patches. Firstly the ability of each anatomical structure to separate intelligible speech from unintelligible sounds was examined. Then by extracting the weight vector from a classification which included the whole of the bilateral temporal lobes (and HG); the relative importance of each voxel in separating intelligible speech from unintelligible sounds was quantified. By examining the sign of these weights it was possible to describe which cortical areas were likely to be coding for intelligible as contrasted with unintelligible sounds.

Bilateral ROIs were used to assess the relative contribution of auditory and auditory association cortices in separating intelligible from unintelligible sounds. A one-tailed Wilcoxon signed rank test with a Bonferroni adjusted threshold of $p<0.01$ (correcting for 5 tests), demonstrated that HG and all temporal ROIs could separate intelligible speech from unintelligible sounds at levels greater than chance (see Figure 3.13A). As expected the control ROI (the IOG) did not perform significantly better than chance ($p > 0.05$).

The Friedman omnibus test demonstrated that there were significant differences between the performance of the different ROIs ($F(4,44)=45.35$, $p<0.001$). The Nemenyi test, a non-parametric post hoc test similar to the Tukey test for ANOVA, was used to carry out follow-up pair-wise comparisons (see Demsar, 2006). In this test the ROIs are ranked within each subject with the best performing ROI, e.g. STG, ranked first. The average rank for each ROI is then calculated by averaging the ranks of each ROI across subjects. Two classifiers are deemed to perform significantly differently to each other if the average rank for each ROI differs by at least the critical difference based on the Studentized range statistic divided by $\sqrt{2}$. This analysis showed that the difference between the

average rank of the STG compared to all other regions except the MTG exceeded the critical difference of 1.761, suggesting that the STG and MTG performed at similar levels, and were the most informative regions in separating intelligible speech from unintelligible sounds.

In order to quantify the amount of information within each hemisphere, left and right ROIs were contrasted (see box plots in Figure 3.13B). Paired Wilcoxon signed rank tests, at a Bonferroni adjusted $p<0.013$ (correcting for four tests), demonstrated that the left MTG performed significantly better than the right ($w=3.5$, $df=11$, $p=0.003$). There was no significant difference between the left and right STG ($w=30.5$,$df=11$,$p=0.532$), while HG ($w=14$,$df=11$,$p=0.050$) and ITG ($w=9$,$df=11$,$p=0.015$) showed a trend towards significance, with left performing better than right, but did not survive correction for multiple comparisons. In the univariate analysis the peak level activations were located mainly within the STS, the sulcus which separates the STG and MTG. The greater performance of the left MTG compared to the right likely reflects the fact that the STS is often included within the MTG rather than the STG ROI in the AAL parcellation (see Figure 3.14).

Figure 3.13 (A) Combined left-right ROIs (B) Left vs. right ROIs.

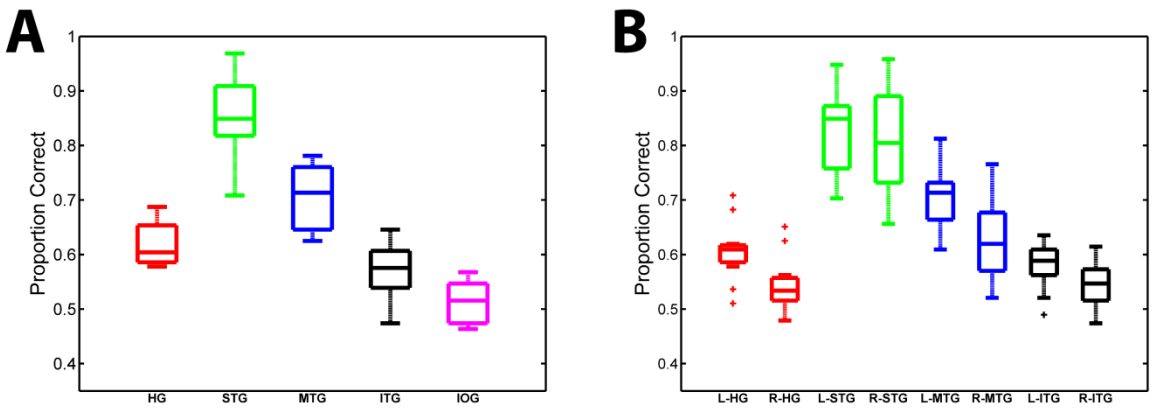
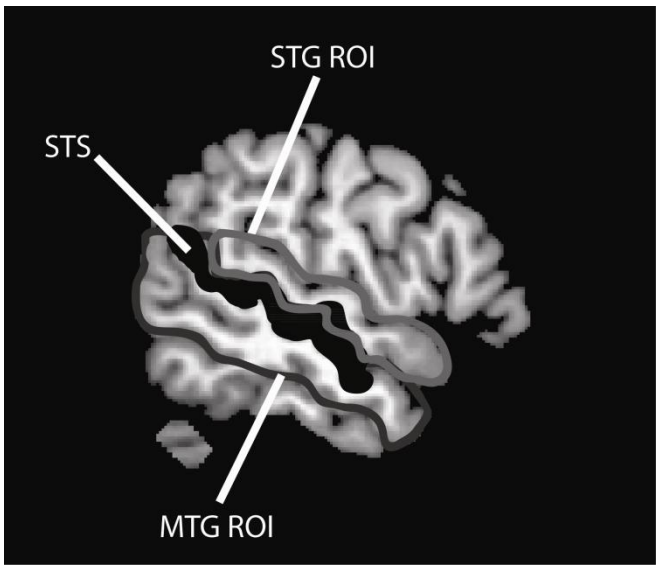


Figure 3.14 Illustration of the position of the STS in relation to STG and MTG.

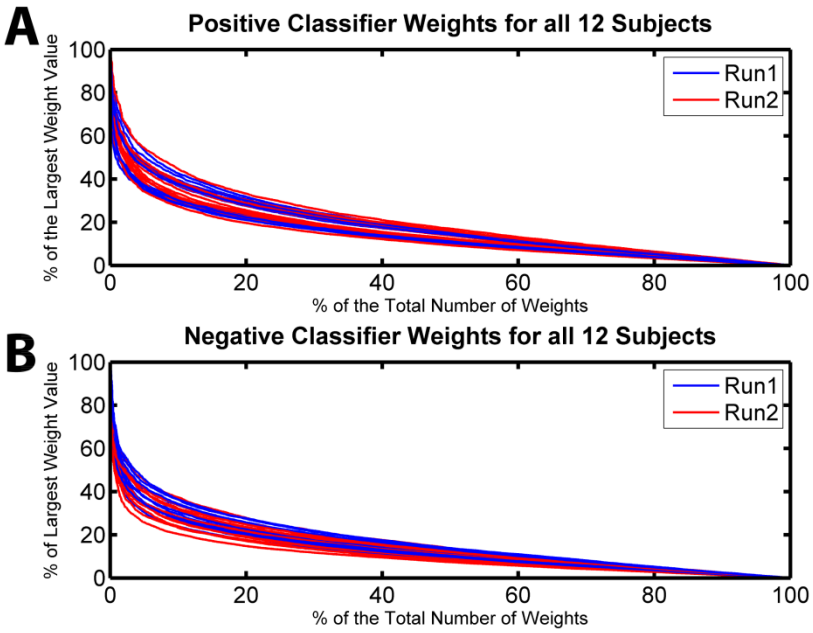


To summarise, the STG and MTG performed at an equivalent level and performed better than all other regions, with the MTG showing a left hemisphere preference, likely reflecting the fact that the STS was often included within the MTG ROI.

It has been noted that the direct comparison of classification scores from anatomical ROIs of different sizes may be problematic as larger ROIs may, on the one hand, contain more classifiable information by virtue of their size, and on the other, the presence of many irrelevant voxels in larger ROIs may make it harder to find a discriminative pattern (Etzel et al., 2009). A final ROI was therefore constructed which included all bilateral temporal and Heschl's gyri. This ROI performed very successfully, separating the conditions correctly 81% of the time (group median). The classifier weight vector from this large ROI was extracted to quantify the relative contribution of all voxels in the temporal lobes (and HG) in separating intelligible from unintelligible sounds. This allowed us to gather converging evidence to support our previous findings and to understand how the voxels across anatomical regions were contributing to classification.

In order to understand how the magnitude of the weights varied, classifier weights were separated into positive and negative weights for classifiers trained on run 1 and run 2, and sorted in magnitude. Large positive weights indicate voxels important to defining the hyperplane that show a relative increase in signal to intelligible speech, whilst large negative weights represent a relative increase to unintelligible sounds. To make the weights comparable across participants and runs, each weight value was expressed as a percentage of the largest overall weight value and plotted against the weight's ranked magnitude expressed as a percentage of the total number of weights. This plot demonstrated that the relative size of the weights decayed in an exponential fashion, with weights ranking in the top 30% accounting for the majority (around 80%) of the range in weight magnitude (see Figure 3.15A & B).

Figure 3.15 (A) Positive classifier weights (B) Negative classifier weights.

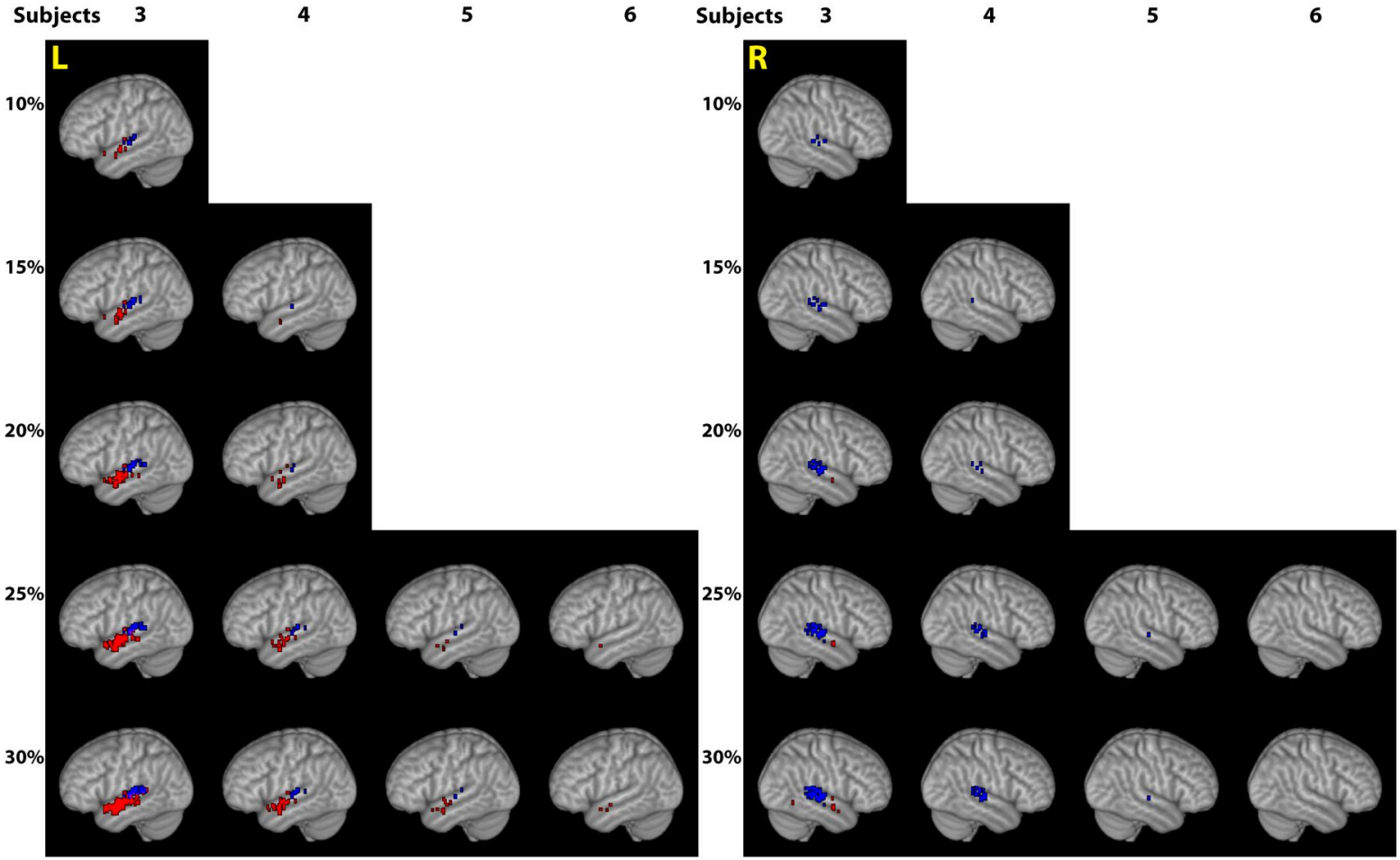


Weights were assigned to percentage bands based on their magnitude, i.e. the largest 10, 15, 20, 25 and 30% weights. For a voxel to be considered for visualisation, the voxel weights had to be of

a consistent magnitude across the two runs. Therefore only voxels implicated in the same percentage band across both the runs were considered, i.e. voxels in the top 10%, 15% etc. in both runs. This prevented voxels being included which were of a small magnitude in one run and a large magnitude in another, which would likely reflect noise in the data. Voxels were then transformed into the shared MNI space using the segmentation parameters acquired from the T1 segmentation to allow comparison across subjects.

In order to visualise the weights whilst reducing arbitrary thresholding effects, different percentage weight bands up to and including 30% were displayed as a function of how many subjects also implicated the same voxels within that weight band. These images clearly demonstrate that when subject agreement and the specificity of the weight banding (with the emphasis on the most important weights) are played off against one another, weights characteristic of intelligibility (positive weights) are located in left anterior temporal cortex (see Figure 3.16). In contrast, negative weights, that is voxels characteristic of the unintelligible sounds are located in earlier auditory cortex predominantly within bilateral mid-posterior STG and PT.

Figure 3.16 Classifier Weights in the left and right hemisphere as a function of relative importance (percentage band) and number of subjects implicating the same voxel (subject consistency). Red=intelligible. Blue=unintelligible.



3.5 DISCUSSION

Three different analyses were conducted in this study: a univariate analysis, and a "local" and "global" multivariate analysis. In the univariate analysis, the main effect of intelligibility replicated the results of Okada et al. in showing bilateral activation spreading across anterior and posterior temporal cortex. Note as both studies used the same statistical threshold this suggests that the studies have roughly equivalent statistical power. Activation was identified in bilateral mid-posterior superior temporal cortex for the interaction between rotation and vocoding. As suggested by the presence of a significant interaction, the simple intelligibility effects activated very different regions, with the [NV- rotNV] contrast activating anterior and posterior temporal cortex bilaterally, in contrast to the [clear - rot] contrast which exclusively activated the left anterior STS. The conjunction null of these two simple effects exclusively activated the left anterior STS. In contrast, the global null conjunction was associated with activation in both left anterior and posterior STS, suggesting that the activation in the left posterior STS in both this and Narain et al. (2003) study were driven mainly by the response to [NV - rotNV].

The local information analysis, replicated Okada et al. in demonstrating that local neighbourhoods of voxels in both anterior and posterior temporal cortex, and HG bilaterally were informative in separating intelligible from unintelligible speech. There was a slightly greater consistency across subjects in implicating mid-posterior regions compared to anterior regions in classifications of clear from rot; this was in contrast to classifications of NV from rotNV which were more equally balanced between anterior and posterior regions. In the case of the "acoustic" contrasts, above chance classification was again bilateral, concentrated in HG and in the case of rot from rotNV additionally distributed widely across anterior and posterior superior and middle temporal cortex.

When "global" classifications were conducted, which investigated how information was integrated across whole anatomical regions, the STG and MTG were shown to be most successful in

classifying intelligible from unintelligible sounds. Classification exhibited a slight leftward advantage but only in the MTG, the ROI in which the STS was often shown to fall in the anatomical parcellation used. The weight vector was extracted from classifications that used the entire temporal cortex bilaterally. By examining the weight vector it was possible to quantify the relative importance of all voxels within this region in separating intelligible from unintelligible speech. The positive weights, those showing a relative increase in BOLD signal to intelligible speech were predominately found in left anterior temporal cortex, whilst negative weights, those showing a relative increase to unintelligible sounds were found in earlier auditory cortex predominantly within bilateral mid-posterior lateral STG and PT.

Synthesising the results across these different analyses, our findings corroborate the importance of left anterior temporal cortex in processing intelligible speech (Scott et al., 2000;Narain et al., 2003;Scott et al., 2006) and also clarify the reasons why an additional posterior activation was identified in Narain et al. (2003). The univariate analysis showed that the left anterior STS was the only region implicated in responding to both simple intelligibility effects at a corrected threshold. Local multivariate information analyses indicated that neighbourhoods of voxels in both anterior and posterior temporal cortex bilaterally were capable of distinguishing between intelligible and unintelligible sounds, however when the classifier had access to information across the entire bilateral temporal lobes, the response in the left anterior temporal cortex was most important in coding for intelligible speech as contrasted with unintelligible sounds.

Mid-posterior bilateral regions by contrast were implicated in the interaction between vocoding and rotated speech. In the left posterior STS the profile was characterised by a similar increased response to clear, rot and NV and a relative deactivation to rotNV. This mirrors the finding of Scott et al. (see Figure 3.1 plot 2) that showed that left posterior STS responded to the phonetic features in rotated speech regardless of their intelligibility. Rotated speech contains phonetic features such as presence/absence of voicing and formant transitions, but these do not lead to the percept of an intelligible phoneme. Rotated-noise-vocoded speech in contrast does not contain recognisable phonetic features perhaps explaining the relative decrease in activation to this condition. As such the

left posterior STS may be relatively more tuned to acoustic-phonetic functions supporting the resolving of intelligible speech (the process) than responding to the fully resolved intelligible percept (the product).

Regions of the right temporal lobe showed an increased response to rotated speech over all other conditions including clear speech. It is difficult to understand why a region would respond most selectively to rotated speech. Interpreted within the context of predictive coding these regions might be associated with sending error signal to higher cortical areas. In predictive coding higher levels of the cortical hierarchy feedback sensory predictions to lower levels of the cortical hierarchy, with lower levels sending forward error signal concerning the mismatch between prediction and experience (Friston and Kiebel, 2009). One speculative suggestion might be that these regions respond most strongly to rotated speech because it is "speech like" whilst still being unintelligible, and may therefore generate the greatest mismatch with expectation.

Whilst activation in bilateral posterior regions was not shown in the conjunction null of the two simple intelligibility effects at a corrected level, it is impossible to rule out that we might have failed to see activation there because of a lack of statistical power. Different statistical contrasts have different associated effect sizes and as a result it is common practice for researchers to sometimes reduce statistical thresholds for more "subtle" contrasts. The conjunction null analysis was shown at a range of thresholds; substantial activations in posterior and bilateral temporal cortex only begin to emerge at an uncorrected threshold of $p<0.01$, and only in the right anterior and left posterior STS and not the right posterior STS. This is in contrast to the robust effects demonstrated using an FDR $p<0.05$ threshold for all other contrasts, including the other simple intelligibility effects.

Some studies have demonstrated, in addition to a consistent activation in left anterior STS, activations also in left posterior STS (Spitsyna et al., 2006) and right anterior STS (Friederici et al., 2010) or both those regions (Awad et al., 2007), when contrasting speech to complex non-speech baselines. In this study while the most discriminative voxels coding for intelligible speech were found in anterior regions, less discriminative voxels coding for intelligible speech were also found in

posterior and bilateral regions, albeit only when the criterion was reduced to show voxels which didn't contribute so highly to classification (not shown). Also when the main effect of intelligibility was shown whilst exclusively masking out the interaction, activation was also found in right anterior and left posterior STS. It is clear as more studies are published examining speech intelligibility that the anterior vs. posterior, or unilateral vs. bilateral debate, is one of degree rather than absolutes. Here evidence is found that the left anterior STS has the largest effect size, suggesting that it is the key intelligibility region, but not excluding the possibility that other regions may play some contributory role. Indeed in the macaque the bilateral anterior and posterior temporal cortex have been shown to be densely connected suggesting a likely functional interdependence between the regions (Pandya et al., 1969;Seltzer and Pandya, 1989). One possible suggestion is that the left anterior STS operates as the key component in a flexible network of regions involved in processing intelligible speech, with the relative contributions of other regions modulated by task and auditory stimulus.

This study differed from Okada et al. in a number of aspects of design and analysis which may explain some of the differences in the findings of the two studies. Okada et al. (2010) required subjects to indicate with a button press when the stimuli were intelligible, an active task which differs from the passive listening paradigm used in this and other studies of intelligibility (Scott et al., 2000;Scott et al., 2006;Narain et al., 2003). They also presented their stimuli in a background of continuous scanner noise, rather than using a sparse design in which the stimuli are presented in silence. Scanner noise is characterised by intense spectral peaks at the switching periodicity of the coils, and by harmonics within the range crucial for speech intelligibility (Hall et al., 1999). Excessive scanner noise may saturate the response in auditory regions, increase attentional load and engage top down compensatory processing strategies (Hall et al., 1999;Edmister et al., 1999;Schmidt et al., 2008;Tomasi et al., 2005).

Okada et al. (2010) only presented the main effect of intelligibility, in which they found widespread activation across anterior and posterior temporal fields (with peaks in anterior regions). They did not carry out a factorial analysis and so were unable to identify whether a significant

interaction existed between rotation and vocoding which would have informed them of the importance of interpreting the simple intelligibility effects.

The pattern analyses in the two studies also differed significantly. Okada et al. only examined "local" classification accuracy within voxel cubes extracted from a single specific location in anterior and posterior STS in each subject. In this study two classification approaches were adopted one to explore local information and the other to understand how information was integrated within and across discrete anatomical structures. When single cubes of data are extracted for analyses as was the case in Okada et al., it places a great emphasis on having found the right location, shape and size of area in which to look for a discriminative pattern. In our local analysis a searchlight procedure was used to address the difficulty of finding the right location. By additionally using whole anatomical ROIs it was possible to conduct classifications in which the algorithm was not constrained to local neighbourhoods, allowing understanding of how information might be integrated across wider areas. Further, unlike Okada et al., the weight vector was examined in order to understand how classification was achieved. When classification accuracy is interpreted without examining the classifier weights, as was the case in Okada et al., no insight is possible into how the classifier performs its discrimination. A high classification performance only indicates that there is a pattern of activation which can be successfully learnt that separates the conditions. Successful classification could reflect a number of different discriminative patterns, some of which are incompatible with the interpretation that a region is coding a response to intelligible speech. Indeed it would be possible to achieve a high level of classification if there was a relative increase in signal for unintelligible sounds compared to intelligible speech, and in fact good classification could still be achieved if there was no signal at all for intelligible speech, provided there was reliable signal for the unintelligible sounds – both scenarios would be difficult to reconcile with an interpretation that those voxels were coding for intelligible speech.

Okada et al. found that the left anterior STS was also able to distinguish the so called "acoustic contrast" of clear speech from noise-vocoded speech. The possibility that anterior STS may have been able to separate these conditions on the basis of intelligibility rather than acoustics was not

discussed in the Okada et al. study, even though noise-vocoded speech is harder to understand and less familiar to participants than clear speech. This would have been likely to have been exacerbated by the presence of continuous scanner noise which would make the noise-vocoded speech disproportionately harder to understand compared to the clear speech. This finding was not replicated in this study using the searchlight method; successful neighbourhoods in the STS were sparse and confined to the left mid and right mid-posterior STS. This might be because sparse acquisition was used in this study which would not have exacerbated the disproportionate difference in intelligibility between the two conditions in the same way.

Large numbers of above chance neighbourhoods were found with high concordance between subjects along the STG, MTG and STS bilaterally for the rot vs. rotNV contrast. A very similar area was shown to be significantly activated by this contrast in the univariate analysis (not shown), corroborating this result. This is unexpected considering that Okada et al. found above chance performance only in right mid STS. However as both anterior and posterior STS have been shown to respond strongly to complex non-speech stimuli it is not surprising (Hall et al., 2002). As previously stated the clear vs NV and rot vs rotNV contrasts differ along multiple complex acoustic dimensions and are also confounded by a difference in intelligibility in the case of the case of clear vs NV. This is why it was decided not to fully replicate Okada et al.'s approach in calculating an acoustic invariance index. In the light of our searchlight results it also seems incorrect to equate these "acoustic contrasts" as occurs implicitly when calculating the invariance index as they clearly engage very different regions, a similar logic applies to the intelligibility contrasts.

This study replicates Okada et al. in showing that there was also a discriminative pattern of responses within HG and in the ITG capable of separating intelligible from unintelligible sounds. Neither of the regions are usually associated with intelligibility responses, which demonstrates the additional sensitivity of pattern analysis (Mur et al., 2009;Haynes and Rees, 2006). The finding of a discriminative response in HG could be due to simple acoustic differences between speech and rotated speech; for example, some fricatives when rotated around 2 kHz manifest as a low frequency noise, something not usually found in natural speech. Alternatively this discrimination might result from

some higher level auditory process within Heschl's Gyrus (King and Nelken, 2009) or from feedback from later auditory association areas. A more speculative possibility is that there may be a gradient of increasing sensitivity to intelligibility that begins in much lower level cortices than previously thought. Whichever of these suggestions is correct, these results provoke interesting future research questions into the role of HG especially in regard to whether HG has a simple passive sensory role, or a more complex constructive higher level function in speech perception.

3.6 CHAPTER CONCLUSION

In summary, this study replicates the findings of the original Scott et al. study in emphasising the importance of left anterior STS in resolving intelligible speech. It also clarifies the findings of Narain et al. (2003) by suggesting that the left posterior STS activation identified in that study was likely to be driven by the response to the difference between NV and rotNV. Using multivariate pattern analysis the findings of Okada et al. were replicated in showing that local information in both anterior and posterior temporal cortex could be used to separate intelligible speech at a level greater than chance. However crucially, it was shown that when the classifier was able to integrate information across discrete anatomical regions, the response in the left anterior STS was most characteristic of a response to intelligible speech. It was speculated that posterior temporal regions, although not playing a major role in this study, may play a role alongside anterior regions in resolving intelligible speech as part of a flexible intelligibility network.

The finding that HG could discriminate between the intelligible and unintelligible conditions, and the fact that some regions responded more strongly to rotated than to clear speech may suggest that rotated speech is not the most ideal unintelligible baseline. The next chapter used an alternative, arguably more closely controlled, non-speech baseline to identify responses to intelligible but degraded speech, and to ask whether any resulting patterns of lateralisation were driven by acoustic or linguistic factors.

Chapter 4 : EXPERIMENT 2

4.1 CHAPTER SUMMARY

In this chapter the question is addressed as to whether the relative left lateralisation for intelligible speech is driven by its linguistic or acoustic features. In this study, unintelligible stimuli were generated in which speech-derived modulations of formant frequency and amplitude were absent, applied singly or in combination. Furthermore to assess responses to intelligibility, two dually modulated conditions – an unintelligible condition in which spectral and amplitude modulations came from two different sentences - and a condition with matching spectral and temporal modulations that listeners could understand after a small amount of training were generated. Univariate and multivariate analyses are used to characterise neural responses to these stimuli and draw out, where present, subtle hemispheric biases.

4.2 INTRODUCTION

It has been proposed that the hemispheres of the brain show a preference for processing different acoustic features, and that these preferences might drive the lateralisation or otherwise of speech processing. The Asymmetric Sampling in Time (AST) hypothesis posits that the two hemispheres sample the speech stream at different rates, with populations of neurons in the left hemisphere tuned to information encoded over short time windows (~20-40ms) , and the right hemisphere to longer windows (~150-250ms) (Poeppel, 2003). As information supporting speech intelligibility evolves over a range of time scales, this has been argued to drive a bilateral response to speech (Poeppel and Hickok, 2004). The strong view claims that the left hemisphere is specialized for resolving fast changes such as formant transitions, and the right hemisphere for slower change such as information about syllabic and intonation structure, although a more moderate view has also been

proposed (Hickok and Poeppel, 2004). A similar but subtly different hypothesis suggests that a trade-off between temporal and spectral processing exists, with the left hemisphere specialized for processing temporal and the right for spectral information (Zatorre and Belin, 2001). These theories stand in contrast to hypotheses which suggest a left hemisphere specialization for speech driven by access to linguistic representations rather than acoustics (Scott et al., 2000;Wolmetz et al., 2011).

There have been a number of studies that have directly tested for hemispheric lateralisation to specific acoustic features, these have often shown a left hemisphere advantage for fast temporal and/or a right hemisphere advantage for a response to spectral or slowly evolving information (Boemio et al., 2005;Schonwiesner, 2005;Zatorre and Belin, 2001;Obleser et al., 2008;Warrier et al., 2009). These effects have been demonstrated mainly in Heschl's Gyrus and/or anterior-lateral STG/STS (Schonwiesner, 2005;Warrier et al., 2009;Zatorre and Belin, 2001;Obleser et al., 2008). Without exception these differential effects have been rather subtle, with whole brain analyses showing robust bilateral effects, and only post-hoc regions of interest analyses supporting any hemispheric preference. Further whilst the findings of a preferential response in the right hemisphere to spectral or slowly evolving information has been shown relatively consistently, a left hemisphere advantage for rapid changes has proved more elusive (Boemio et al., 2005).

The theory of a left hemisphere preference for rapid temporal processing is not a new one, and is derived from an old finding that showed that patients with aphasia and left hemisphere lesions performed poorly on temporal order judgments (Efron, 1963), although note in this study that the patients with receptive aphasia actually performed worse on visual compared to auditory temporal judgments. Subsequent studies of developmental language disorders also sought to explain language deficits in terms of an impairment in rapid temporal processing (Tallal, 1980;Tallal and Piercy, 1973). These findings, alongside the observation of greater white matter volume in the left compared to the right hemisphere in primary auditory areas (Penhune et al., 1996), have been influential historically in progressing the rapid left hemisphere theory. The theory that rapid auditory processing deficits cause language impairment is difficult to reconcile with recent evidence showing that impairments are not restricted to short time intervals and have been shown to manifest in some cases exclusively at long

intervals (Rosen, 2003). In addition the observation that primary auditory cortex cannot be defined reliably using MRI has brought into question previous demonstrations of white matter asymmetry in this region (Rademacher et al., 2001). However despite these findings the rapid left hemisphere theory has continued to gain traction within the neuroscience community.

There are a number of methodological issues in the existing body of literature exploring the auditory basis of hemispheric lateralisation. Whilst there have been a number of demonstrations of subtle hemispheric preferences, these experiments have in the main used simple non-speech stimuli (Boemio et al., 2005;Zatorre and Belin, 2001;Warrier et al., 2009;Jamison et al., 2006). For example Zatorre et al. (2001) and Jamison et al. (2006) used pure tones as the basis for spectral and temporal manipulations. Extending the interpretation of results from studies such as these to the processes underlying speech processing is problematic. Whilst intuitively it seems logical to extrapolate from auditory to speech perception, the reality is rather different. For example, it has been show that impairments in low level auditory perceptual tasks are not necessarily associated with deficits in speech perception suggesting that the relationship between auditory and speech perception is complex and to some extent dissociable (Ramus et al., 2003). Obleser et al. (2008) adopted a different approach, rather than use simple non-speech stimuli, they degraded the temporal and spectral information in speech, unfortunately however this also leads to a concomitant modulation in intelligibility, which in itself has often been shown to be left lateralised (Scott et al., 2000;Narain et al., 2003;Scott et al., 2006). To date few studies have attempted to explore patterns of auditory lateralisation using unintelligible stimuli with acoustic manipulations derived directly from speech. This would allow firmer conclusions to be drawn between any observed lateralisation effects and speech perception.

Two important methodological issues exist in the analyses that have been conducted in this area. Firstly non independent data has sometimes been used to define regions of interest in which subsequent statistical tests were conducted (Zatorre and Belin, 2001), so called double dipping (Kriegeskorte et al., 2009). Secondly data of an arbitrary shape and size has often been extracted on which to conduct further statistical tests (Schonwiesner, 2005;Jamison et al., 2006). The size and

shape of these spheres could have an influence on the resulting analyses especially where differences between conditions are subtle. A more principled approach would be to use apriori defined anatomically defined masks or use a subset of independent functional data to define ROIs. A third issue derives from the statistical hypotheses explored in these studies. The strongest evidence in demonstrating, for example, a relative hemispheric preference for spectral as contrasted with temporal processing would be to show firstly, that a right hemisphere region responds significantly to spectral modulation (show that the response is significantly different to 0), and secondly to show that it responds more to spectral than to temporal modulation (show that the difference between the response to temporal and spectral modulation is greater than 0). A number of studies do not take this approach and rather demonstrate that the response is different to 0, without showing that the difference between the responses is greater than 0. The conclusions that can be drawn from these studies is weaker as a result (Zatorre and Belin, 2001;Schonwiesner, 2005).

To address these perceived gaps in the existing literature, unintelligible stimuli were generated in which speech-derived modulations of formant frequency and amplitude were absent, applied singly or in combination. Furthermore to assess responses to intelligibility, two dually modulated conditions were generated – an unintelligible condition in which spectral and amplitude modulations came from two different sentences - and a condition with matching spectral and temporal modulations that listeners could understand after a small amount of training. Pattern classification using SVMs was conducted within anatomical regions of interest to tease apart subtle hemispheric preferences, whilst avoiding double dipping and the arbitrary definition of regions of interest. A number of complimentary empirical and descriptive methods were conducted to demonstrate the presence or absence of relative hemispheric preferences to manipulations of amplitude, spectrum or intelligibility, crucially it was determined whether regions showed a direct relative difference in preference for one type of acoustic modulation over another.

4.3 METHOD

Participants

20 right-handed speakers of English (10 female, aged 18-40 years) took part in the study. All participants reported normal hearing and no history of speech and language difficulties or neurological problems. All subjects gave informed consent in accord with the approval of the UCL Department of Psychology Ethics Committee.

Stimuli

All stimuli were based on sine wave speech in which pure tones are synthesized to follow the formants of speech (Remez et al., 1981). Typically when subjects listen naively to sinewave speech they report hearing the stimuli as "whistles", but it becomes largely intelligible when they are informed that is speech – listening in so called "speech mode". The stimuli were derived from a set of 336 semantically and syntactically simple sentences known as the Bamford-Kowel-Bench (BKB) sentences (Bench et al., 1979). These were recorded in an anechoic chamber by an adult male speaker of Southern British English at a sampling rate of 11.025 kHz with 16 bit quantization.

A semi-automatic procedure was used to track the frequencies and amplitudes of the first two formants of speech every 10ms. The construction of stimulus conditions followed a 2x2 factorial design with spectral and amplitude modulation as factors (with modulation type present or absent). In order to provide formant tracks that varied continuously over the entire utterance (e.g. such that they persisted through consonantal closures), the formant tracks were interpolated over silent periods using piecewise-cubic Hermite interpolation in log frequency and linear time. Static formant tracks were set to the median frequencies of the measured formant tracks, separately for each formant track.

Similarly, static amplitude values were obtained from the median of the measured amplitude values larger than zero.

Five stimulus conditions were created where S and A correspond to Spectral and Amplitude modulation, respectively. The character ‘o’ indicates a steady/fixed state while ‘mod’ indicates a dynamic/modulated state.

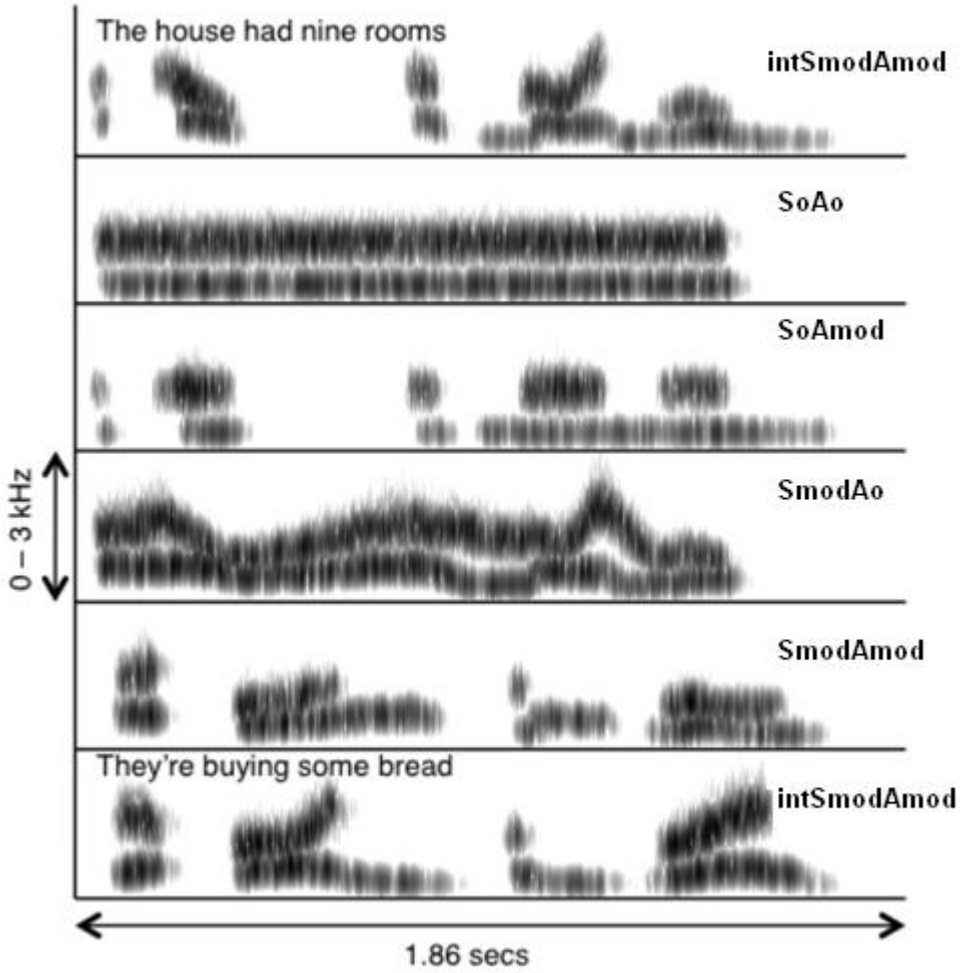
(1) SoAo, steady state formant tracks with fixed amplitude.
(2) SoAmod, steady state formant track with dynamic amplitude variation.
(3) SmodAo, dynamic frequency variation with fixed amplitude.
(4) SmodAmod, dynamic frequency and amplitude variation but each from a different sentence, making the signal unintelligible. Linear time scaling of the amplitude contours was performed as required to account for the different durations of the two utterances.
(5) intSmodAmod, an intelligible condition, with dynamic frequency and amplitude variation taken from the same original sentence. These were created in the same way as (1)-(4) but with less extensive hand correction (the interpolations for the unintelligible conditions were particularly vulnerable to small errors in formant estimation).

Each stimulus was further noise-vocoded (Shannon et al., 1995), to enhance auditory coherence. For each item, the input waveform was passed through a bank of 27 analysis filters (each a 6th-order Butterworth) with frequency responses crossing 3 dB down from the pass-band peak. Envelope extraction at the output of each analysis filter was done using full-wave rectification and 2nd-order Butterworth low-pass filtering at 30 Hz. The envelopes were then multiplied by a white noise, and each filtered by a 6th-order Butterworth IIR output filter identical to the analysis filter. The Root Mean Square (RMS) level from each output filter was set to be equal to the RMS level of the original analysis outputs. Finally, the modulated outputs were summed together. The cross-over frequencies for both filter banks (over the frequency range 70-5000 Hz) were calculated using an equation relating

position on the basilar membrane to its best frequency (Greenwood, 1990). Figure 4.1 shows example spectrograms from each of the five conditions.

The intelligibility of the modulated stimuli (i.e. excluding the SoAo condition) was tested in 13 adult listeners by Rosen et al. (in revision), using 10 items from each condition. The mean intelligibility scores were 61%, 6%, 3% and 3% keywords correct for the intSmodAmod, SmodAmod, SmodAo, and SoAmod conditions, respectively.

Figure 4.1 Spectrograms of example sentences from the five conditions.



Behavioural Testing

A behavioural pre-test was conducted to familiarise and train participants to understand the $_{int}S_{mod}A_{mod}$ speech to ensure that they were listening in'speech mode' during the scanning session (Dehaene et al., 2005). Each sentence was played to the participant and they were asked to repeat back what they had heard, the number of key words correct was recorded. If the subject identified all three words, they were provided with positive feedback and the next sentence was played. If the participant reported less than 3 words correct, the sentence was verbally repeated and played again. This process was continued until the participant correctly repeated all the key words in five consecutive sentences, or until 98 sentences were presented.

A post-test was conducted using 80 sentences from the $_{int}S_{mod}A_{mod}$ condition, half of which had been presented in the scanner and half of which were novel exemplars from the same condition. After each sentence, participants were again asked to repeat back the sentences and the number of correct key words was recorded.

fMRI experiment

Functional imaging data were acquired on a Siemens Avanto 1.5 Tesla scanner (Siemens AG, Erlangen, Germany) with a 32-channel birdcage headcoil. Two runs of functional images were collected (TR = 9 seconds, TA = 3 seconds, TE = 50 ms, flip angle 90 degrees, 35 axial slices, 3mm × 3mm × 3mm in-plane resolution). A sparse-sampling routine was employed (Hall et al., 1999), in which two stimuli from the same condition were presented sequentially during the silent period, with the onset of the first stimulus presented 5.3 seconds (with jittering of +/- 500 ms) before acquisition of the next scan commenced.

In the scanner, the auditory stimuli were delivered via air-conduction headphones (Etymotic Inc., Elk Grove Village, IL, USA). In each functional run, the participant heard 50 stimuli from each of the five auditory conditions (2 stimuli per trial). Participants were instructed to listen carefully to all the stimuli. They were told that they would hear some examples of the same sort used in the training phase, which they should try to understand. The order of presentation of stimuli from the different conditions was pseudorandomised to allow a relatively even distribution of the conditions across the run without any predictable ordering effects. A silent baseline was included in the form of four silent mini-blocks in each functional run, each comprising five silent trials. After the functional run was complete, a high-resolution T1-weighted anatomical image was acquired (HIRes MP-RAGE, 160 sagittal slices, voxel size = 1 mm^3).

Univariate Analysis

Data were preprocessed and analyzed in SPM8. Functional images were realigned and unwarped, coregistered and normalised using parameters obtained from the segmentation of the anatomical image, and smoothed using a Gaussian kernel of 8mm FWHM. A first order Finite Impulse Response filter with a window length equal to the time taken to acquire a single volume, effectively a box car function, was used to model the hemodynamic response (de Zubicaray et al., 2007). A high pass filter with a time constant of 128s was applied to remove low frequency noise. Each condition was modelled as a separate regressor in a generalised linear model (GLM). Six movement parameters were included as regressors of no interest.

At the first level, con images were created by the comparison of each auditory condition against an implicit silent baseline. The con images for the four unintelligible conditions (excluding the intelligible condition) were entered into a random-effects 2 × 2 within subject ANOVA with factors Spectral Modulation (present or absent) and Amplitude Modulation (present or absent). This allowed examination of main effects of spectral and amplitude modulation and their interaction. An

intelligibility contrast was created, by generating con images of [intSmodAmod - SmodAmod] at the first level and entering these images into a one sample t-test at the second level.

All statistical images are shown at $p < .05$ FDR corrected with no cluster extent. Tables are restricted to clusters greater than 40 voxels for the sake of brevity. Anatomical localization of activations was guided by reference to the SPM anatomy toolbox (http://www.fz-juelich.de/inm/inm-1/DE/Forschung/_docs/SPMAnantomyToolbox/SPMAnantomyToolbox_node.html).

Multivariate Analysis

Functional images were unwarped and realigned to the first acquired volume. Training and test examples from each condition were constructed from single volumes. The data were separated into training and test sets by functional run, to ensure that training data did not influence testing (Kriegeskorte et al., 2009). Linear and quadratic trends were removed and the data z-scored within each run. A linear Support Vector Machine (SVM), using a hard margin and the Andre optimization, from the Spider toolbox (http://www.kyb.tuebingen.mpg.de/bs/people/spider/) was used to train and validate models. The first classifier was trained on the first run and tested on the second, and vice versa for the second classifier. Accuracy was estimated by averaging performance across the two classifiers for each participant.

The classifications were performed using a number of subject-specific, anatomically defined ROIs. The Freesurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) was used to perform cortical reconstruction and volumetric segmentation via an automated cortical parcellation of individual participants' T1 images. Two participants were excluded from the multivariate analyses due to unsuccessful cortical parcellation. Subject-specific, left and right hemisphere ROIs for Heschl's Gyrus (HG), MTG STG and the STS (Temporal) and the combined left and right Inferior Occipital Gyrus (IOG) were created. These anatomical regions were included based on *a priori*

hypotheses about the key sites for intelligibility and acoustic processing of speech (Davis and Johnsrude, 2003;Eisner et al., 2010;Obleser et al., 2008;Scott et al., 2000), and hence not contingent on the results of the univariate analyses. Recursive Feature Elimination (RFE) (Guyon et al., 2002) using the Spider Toolbox and SVM was conducted as an additional exploratory feature selection step. RFE using SVM works by ranking voxels according to their associated weight and removing a specified number of voxels, with the procedure conducted iteratively to successively prune away irrelevant voxels.

4.4 RESULTS

Behavioural Results

In the pretest, a stringent criterion of 5 consecutive correct responses (with 100% accuracy on keyword report) was used to ensure a thorough training on the ${}_{int}S_{mod}A_{mod}$ condition before the scan. For those participants who reached this criterion in the pre-test, the mean number of trials to criterion was 46.6 (SD 17.7). Four of the 20 fMRI participants did not reach this criterion within the list of 98 pre-test items. However, as all participants achieved 3 consecutive correct responses within an average of only 23.5 trials (SD 16.4), with six participants reaching this threshold within the first 6 items, it was clear that all participants would be able to understand a sufficient proportion of the ${}_{int}S_{mod}A_{mod}$ items in the scanner to support our planned intelligibility contrasts.

In the post test, the average accuracy across the whole post-test item set (calculated as the percentage of keywords correctly reported) was 67.2% (SD 8.5%), representing a mean improvement of 4.6% (SD 5.7%) on pre-test scores (mean 62.4%, SD 6.4%). This improvement was statistically significant ($t(19) = 3.615$, $p < .01$). There was no difference in accuracy between the old (67.0%) and new (67.5%) items ($p > .05$).

fMRI Results

Univariate Analysis

The main effect of amplitude modulation gave rise to clusters of activity predominantly focused in bilateral STG and HG (see Figure 4.2A). The main effect of spectral modulation gave rise to more widespread activation, again concentrated predominantly in bilateral STG, there was an additional large cluster found in the left precentral gyrus (see Figure 4.2B). Significant peak level activations for the main effects are listed in Table 4.1. Note that the spatial extents of activation were fairly similar in each hemisphere for both contrasts.

Figure 4.2 Univariate analyses (A) Main effects of amplitude (B) Main effect of spectrum.

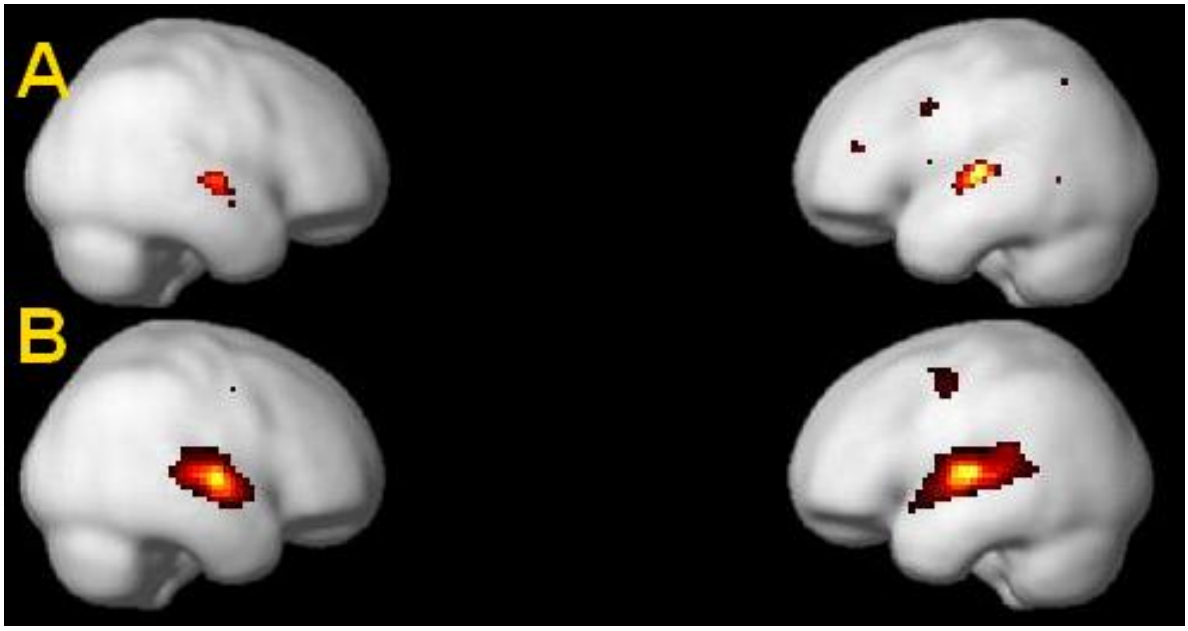


Table 4.1 Peak level activations for the main effects and interactions, FDR p<0.05, cluster extent > 40.

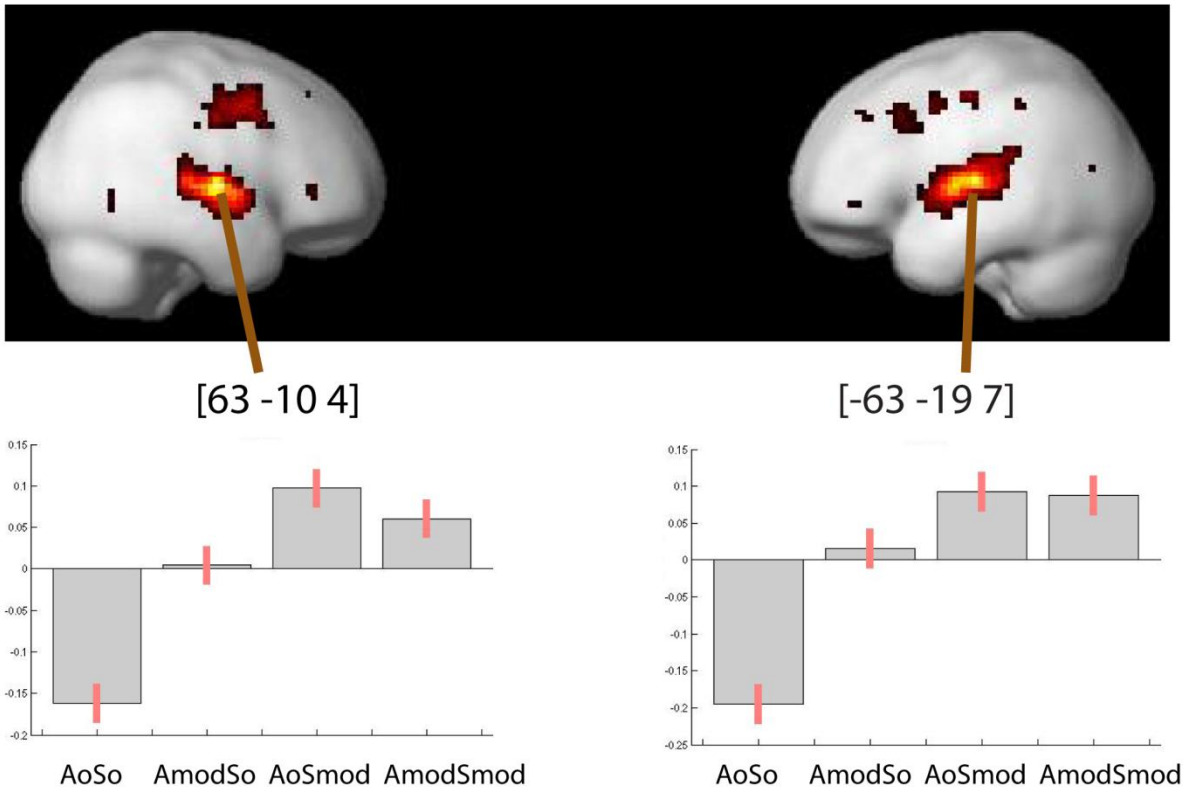
Contrast	Extent	Region	Coordinates			z
			X	y	z	
Amplitude	86	Left STG	-60	-19	4	6.12
	48	Right STG	63	-10	1	4.72
Spectrum	554	Left STG	-60	-13	7	Inf
			-51	-37	16	4.09
	395	Right STG	63	-7	4	6.86
	41	Left precentral gyrus	-54	-4	49	4.15
Interaction	425	Right STG	63	-10	4	6.64
		Right Insula	42	-1	-11	3.62
	459	Left STG	-63	-19	7	6.26
		Left STG	-51	-7	4	5.96
		Left STG	-54	-37	19	3.66
	187	Right precentral Gyrus	57	8	46	4.89
		Right postcentral Gyrus	51	-10	37	3.96
		Right postcentral Gyrus	63	-16	43	3.04
	46	Left anterior cingulate	-6	20	31	4.38
	42	Left middle frontal gyrus	-33	17	37	3.88
		Left middle frontal gyrus	-30	23	31	3.33

An f-test examining the interaction of spectral and amplitude modulation was used to test for additive responses to the intelligibility relevant acoustic modulations (i.e. greater activation for the dually modulated SmodAmod condition than the sum of the responses to the singly modulated SoAmod and SmodAo conditions). This contrast gave rise to activations in bilateral STG (with the cluster spreading into the insula in the right) in addition to activations in the left anterior cingulate,

middle frontal gyrus (MFG) and the right pre- and postcentral gyrus. When contrast estimate plots were examined it was clear that there were no peaks demonstrating an additive effect of the modulations. Plots from the main peaks are shown in Figure 4.3, and provide an example of sub-additivity of the two factors i.e. that the difference in signal between SoAo and the singly modulated conditions (SoAmod or SmodAo) was larger than that between those singly modulated conditions and the SmodAmod condition. See Table 4.1 for peak activations.

Figure 4.3 Interaction between amplitude and spectrum including plots of effect size from the largest peaks in each hemisphere.

Plots are relative to mean parameter value across conditions rather than baseline



The intelligibility subtraction [$_{int}S_{mod}A_{mod}$ - $S_{mod}A_{mod}$] gave rise to significant activation in bilateral STS and STG, with the peak voxel in left mid-anterior STG and a larger relative cluster

extent in the left hemisphere. Clusters of activation also extended bilaterally within the insula, IFG and middle frontal gyrus, supplementary motor area and precentral gyrus (see Figure 4.4, significant peak level activations, > 40 cluster extent, are listed in Table 4.2).

Figure 4.4 Intelligible - Unintelligible: [intSmodAmod - SmodAmod].

Table 4.2 Peak level activations Intelligible > Unintelligible [intSmodAmod - SmodAmod], FDR p<0.05, cluster extent > 40.

Contrast	Extent	Region	Coordinates			Z
			X	y	z	
Intelligible > Unintelligible	2191	Left mid-anterior STG	-57	-13	1	6.08
		Left mid-posterior STS	-63	-34	7	5.67
		Left mid-anterior STG	-54	-7	-5	5.65
	1027	Right mid-anterior STG	60	-13	-5	5.63
		Right mid-anterior STS	51	-16	-8	4.96
		Right temporal pole	60	11	-11	4.66
	113	Left SMA	-6	11	58	5.10
		Left SMA	-6	23	52	3.24
	87	Right precentral gyrus	51	2	49	4.14
		Right precentral gyrus	54	-7	40	3.01
	93	Right IFG	54	23	16	3.79
		Right IFG	42	14	13	3.18

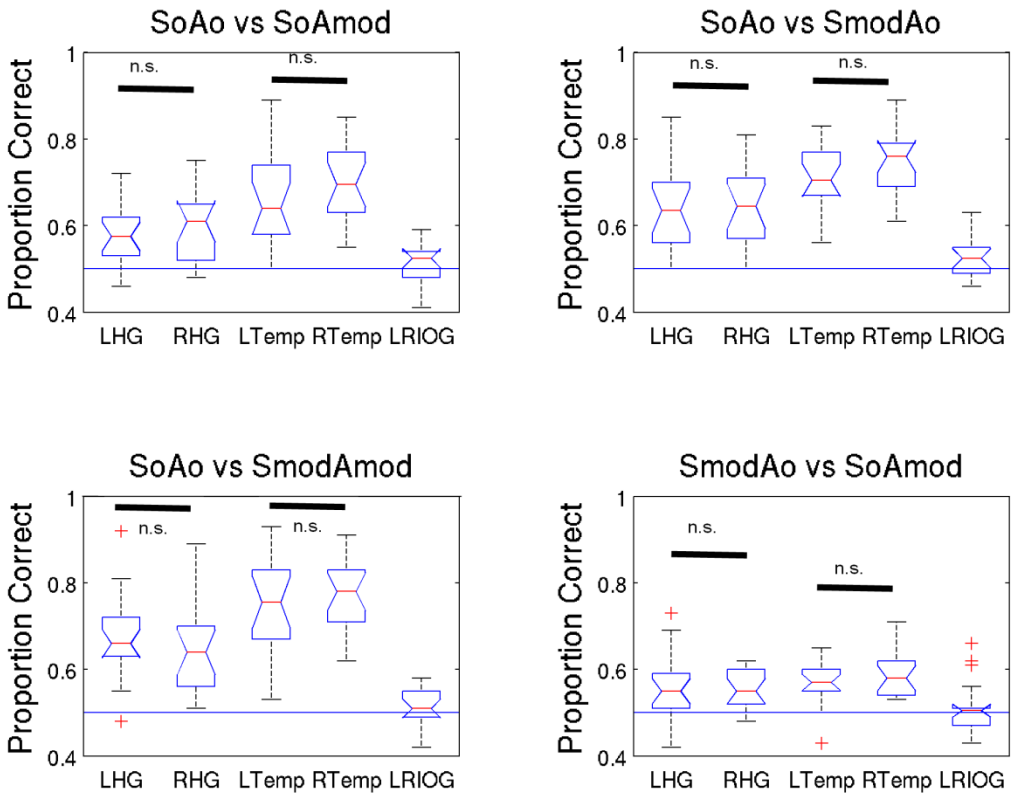
Multivariate Analysis

The average number of voxels across subjects in each ROI were as follows LHG (77), RHG (56), LTemp (1406) and RTemp (1427) and LRIOG (392). There was no significant difference between the number of voxels in the left vs. right temporal ROIs ($w=66, df=17, p > 0.05$), but there were significantly more voxels in left as contrasted with right HG ($w=3, df=17, p<0.001$), as identified by repeated measures two-sided wilcoxon signed rank tests.

The support vector machine was trained and tested on four acoustic classifications: SoAo vs. SoAmod, SoAo vs. SmodAo, SoAo vs. SmodAmod, SoAmod vs. a SmodAo, and an additional intelligibility classification: intSmodAmod vs. SmodAmod. Performance in each classification was tested against a chance performance of 0.5, using a one-sided Wilcoxon signed rank test, with a corrected significance level of $p < .01$ (for the five ROIs tested in each classification). The control ROI, the IOG, performed no better than chance on all classifications (inclusive of both the acoustic and intelligibility classifications).

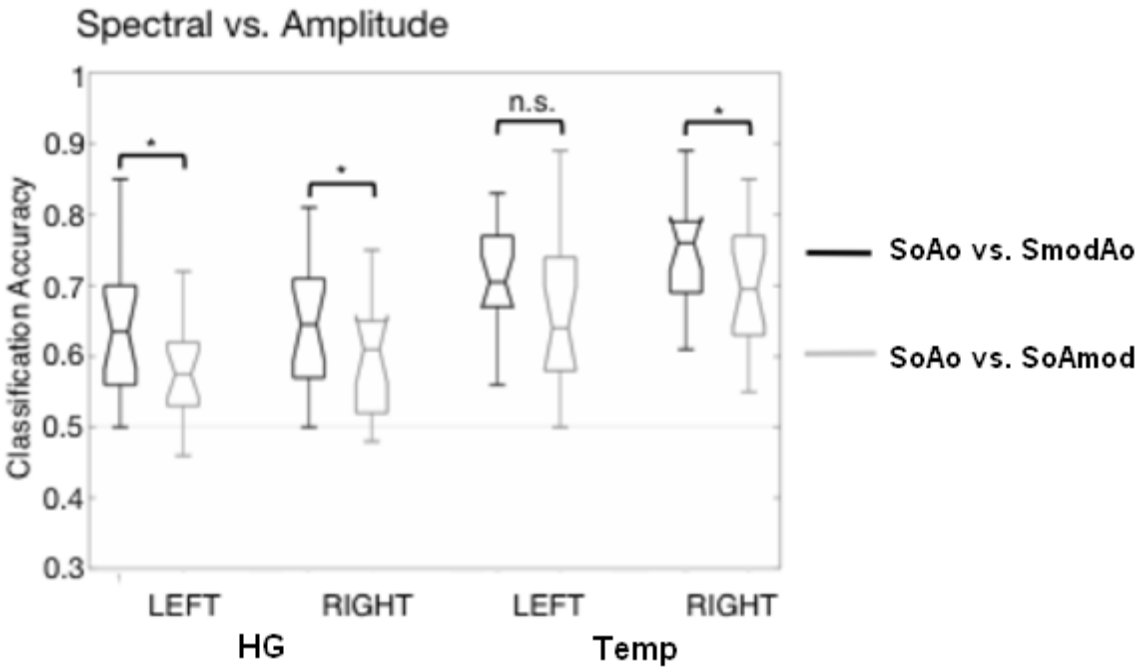
Both left and right Temporal and HG ROIs, performed significantly better than chance for all acoustic classifications (see Figure 4.5). When the classification performance of left versus right hemisphere ROIs were directly contrasted, there was no significant difference between the left and right for any of the acoustic classifications in all ROI pairs ($p > .025$, significance level corrected for 2 left-right comparisons for each contrast for a repeated measures two-sided Wilcoxon signed rank test).

Figure 4.5 Box plots of classification scores for the group of subjects in each ROI for the acoustic contrasts.



In order to compare classification performance for spectral versus temporal modulations within each hemisphere, scores from the SoAo vs. SoAmod and SoAo vs. SmodAo classifications were directly compared (see Figure 4.6) within left HG, right HG, left temporal and right temporal ROIs (paired, two-sided Wilcoxon signed rank tests corrected for four tests, $p < 0.013$). This showed that the classification of spectral modulations was significantly more accurate than the classification of amplitude modulations in left HG ($w = 16$, $df=17$, $p = .004$), right HG ($w = 31$, $df=17$, $p = .004$) and right STG+MTG ($w = 23.5$, $df=17$, $p = .012$). The difference in the left temporal ROI was significant at an uncorrected alpha of .05 ($w = 31$, $df=17$, $p = .018$). This reflected the fact that there was in general a higher level of classification for spectral compared to amplitude modulation irrespective of hemisphere.

Figure 4.6 Comparison of classifier scores from the SoAo vs. SoAmod and SoAo vs. SmodAo in the left and right HG and Temporal ROIs.

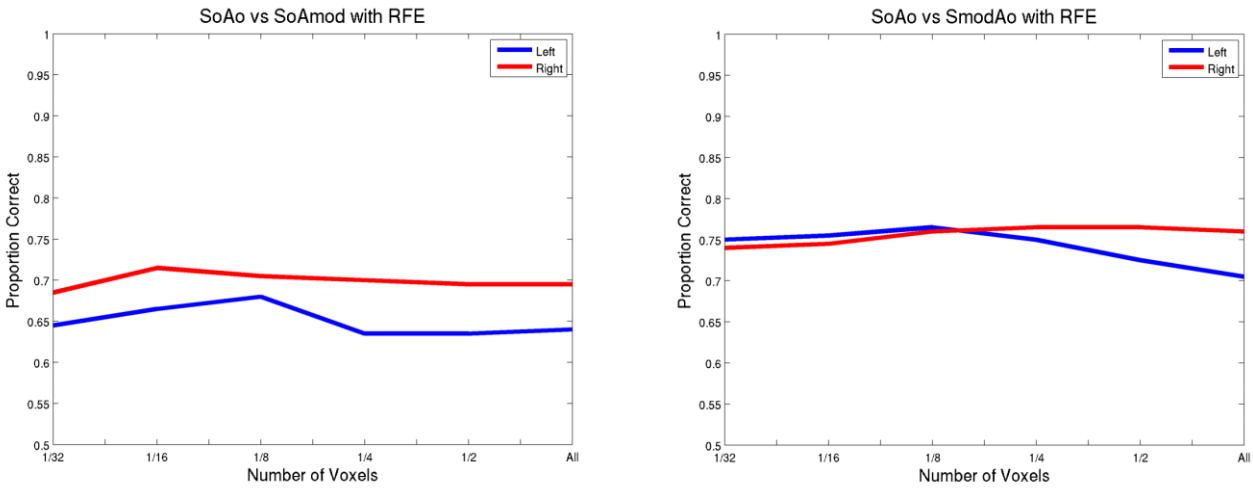


It is conceivable that the SVM failed to demonstrate differences in accuracy for the left vs. right hemisphere ROIs due to the large number of voxels in the Temporal ROI, indeed SVMs can perform poorly with too many irrelevant features. In order to explore this possibility classification was rerun for SoAo vs. SoAmod and SoAo vs. SmodAo using Recursive Feature Elimination (RFE) and SVM. Classifications were run first with all voxels in either the left or right temporal ROI, then features were reduced by successively halving the number of voxels so as to constitute 1/2, then 1/4, 1/8, 1/16, 1/32, the original number of voxels, this corresponded to using (an average of across subjects of) 1406, 702, 350, 174, 87, 43 voxels in the left hemisphere and in the right to 1426, 712, 356, 177, 88, 44 voxels. RFE was conducted on one run with the performance of the selected voxels validated by testing on the other run, and vice versa for each run, with the performance of the two runs averaged. Note that ordinarily feature selection would be carried out on a subset of data separate

to the training set; in this case the use of RFE was an exploratory descriptive technique rather than designed to estimate the true accuracy of the classifier.

There was no significant difference in classification performance using the highest performing set of reduced features compared to using the whole anatomical region, in the left for amplitude modulation, 1/8~174voxels ($w=49, df=17, p > 0.05$), or the right 1/16~88 voxels ($w=55.5, df=17, p > 0.05$). In the left for the spectral modulation 1/8~174 voxels ($w=32.5, df=17, p > 0.05$) or the right 1/2~712voxels ($w=39, df=17, p > 0.05$). See Figure 4.7 for plots showing how classification performance was modulated by using a smaller numbers of "optimised" features. This procedure suggests that the large number of voxels in the Temporal ROI was not detrimental to showing an effect of sensitivity between hemispheres for the two types of modulation.

Figure 4.7 RFE classifications using different numbers of voxels for the SoAo vs. SoAmod and SoAo vs. SmodAo contrasts in the left and right hemisphere using a 1/2, 1/4, 1/8 etc.. the original number of voxels.



The top 30% of positive and negative weights were extracted from the acoustic contrast: SmodAo vs. SoAmod , using the separate bilateral temporal (inclusive of STG,STS and MTG) and HG masks. Both classifications, assessed by a one tailed signed rank Wilcoxon test, were

significantly different to chance: bilateral temporal ($w=0$, $df=17$, $p<0.001$) and HG ($w=27.5$, $df=17$, $p<0.01$). Only voxels that belonged in the top 30% for both cross validated runs were extracted to reduce noisy voxel weights. By examining the classifier weights it was possible to ascertain which voxels contributed most to the classification and whether these voxels exhibited a relative increase in signal to spectral or amplitude modulation. Weights were simultaneously visualized for both classifications in native space (see Figure 4.8 for weights in three representative participants: 's3','s4','s12'). Negative weights, shown in red representing an increase in signal to SmodAo in the support vectors, and positive weights shown in blue representing an increase to So Amod, seemed to be well distributed within and between the hemispheres suggesting a lack of hemispheric preference for modulation type. This was confirmed by counting the number of positive and negative weights within each hemisphere; the number of weights of each type were compared within hemisphere using a two tailed signed rank test, there was shown to be no significant difference in the proportion of the two types of weight for the HG and Temporal ROI ($p > 0.05$) (considered separately), see Figure 4.9.

Figure 4.8 Classifier weights shown in native space for three representative subjects: S3, S4, S12, for the acoustic classification: SmodAo vs. SoAmod. Red voxels= SmodAo and blue= SoAmod

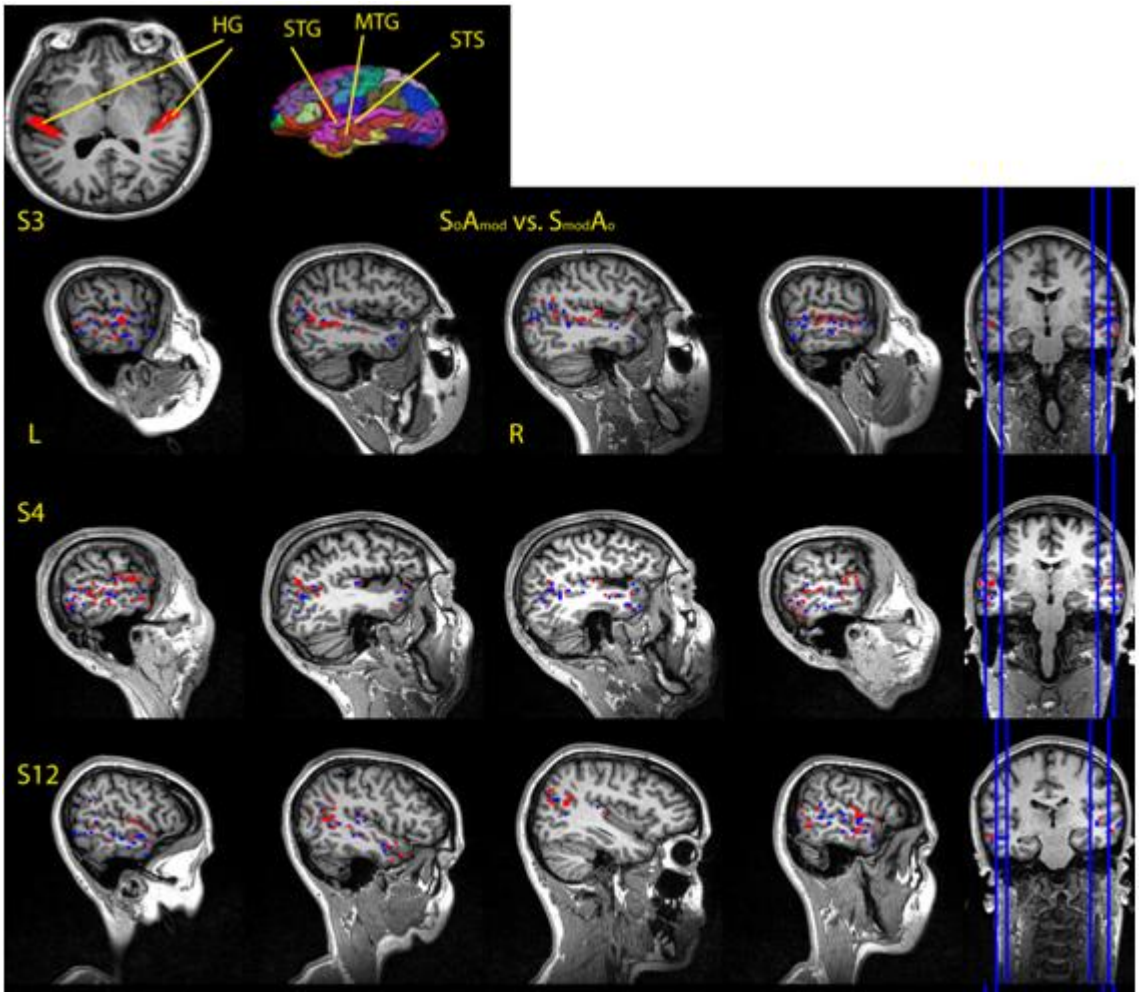
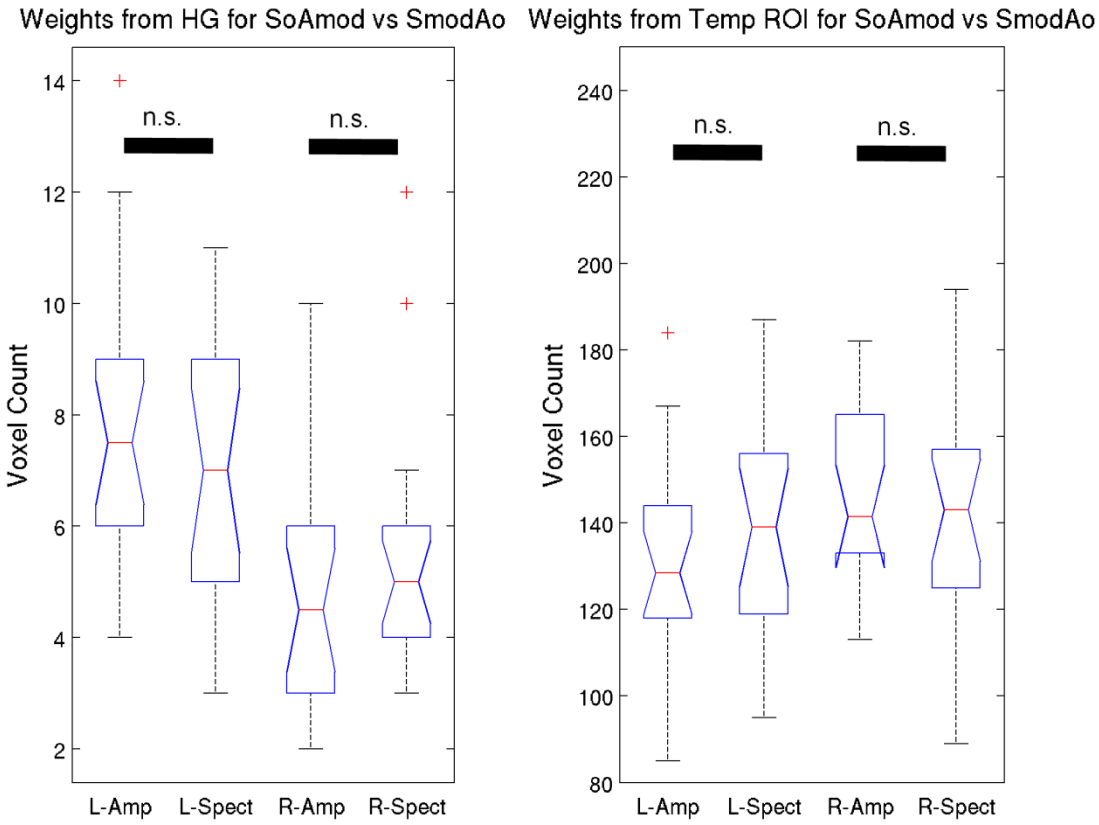


Figure 4.9 Voxel counts for weights characterising amplitude and spectral modulations. Results show voxel counts for the group of subjects.



For the intelligibility classification, SmodAmod vs intSmodAmod, the left and right temporal ROIs were shown to perform at a level above chance, but left and right HG did not (corrected for five tests, $p<0.01$, one-sided Wilcoxon against chance level of 0.5) (see Figure 4.10, left). The comparison of left versus right ROIs for this contrast showed that performance was equivalent between hemispheres for HG, but was significantly greater in the left temporal ROI compared to its right-hemisphere homologue ($w = 22.5$, $p = .011$; paired, two-sided Wilcoxon signed rank tests, corrected significance level of $p < .025$). The top 30% positive and negative classifier weights were extracted from this intelligibility contrast (using the same approach as for the acoustic classification) using a bilateral temporal mask (inclusive of STG,STS and MTG). Classification using this ROI was highly significant ($w=0,df=17,p<0.001$). Classifier weights were extracted within native space, three representative subjects 's3','s4','s12' are shown in Figure 4.11, red weights for intelligible and blue

for unintelligible sounds. Classifier weights characteristic of an increase in signal to intelligible and unintelligible sounds (in the support vectors) were well distributed within and across both hemispheres, when positive and negative weights were counted and compared, there were significantly larger numbers of voxels characterizing a response to intelligible speech in the left (w=7,df=17, p<0.001) and unintelligible sounds in the right respectively (w=31,df=17,p=0.0176; wilcoxon two tailed sign rank corrected for two tests) (see Figure 4.10, right).

Figure 4.10 Classification of SmodAmod vs intSmodAmod (left) and voxel counts for weights for the same classification (right).

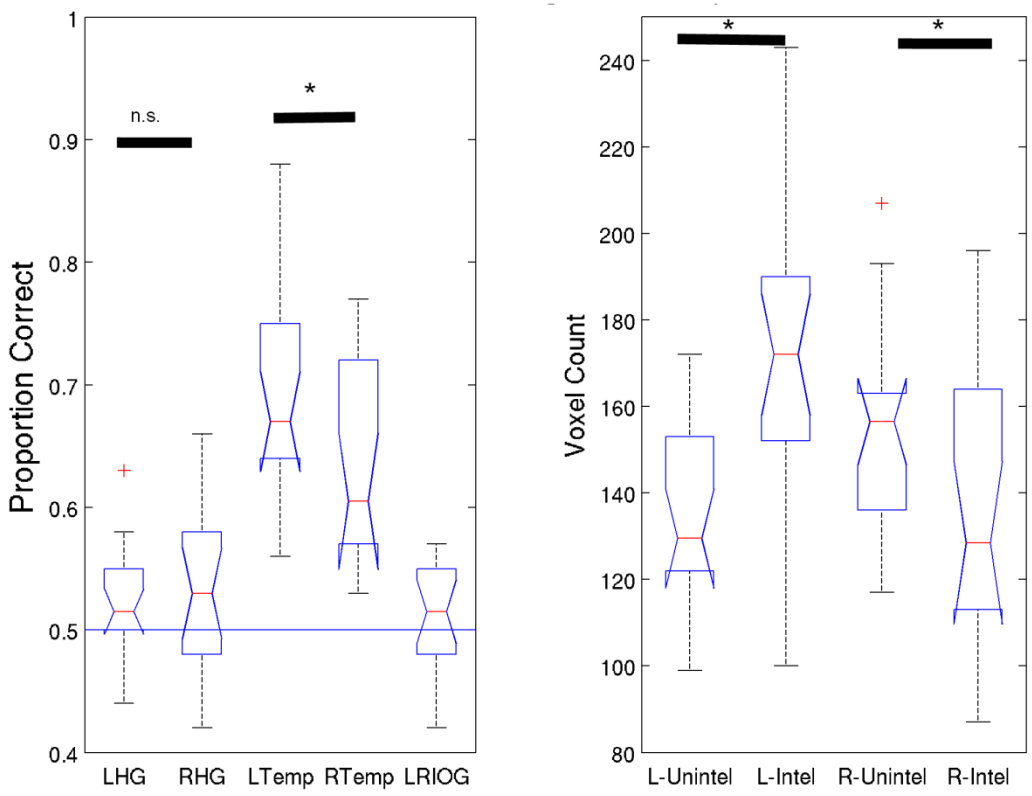
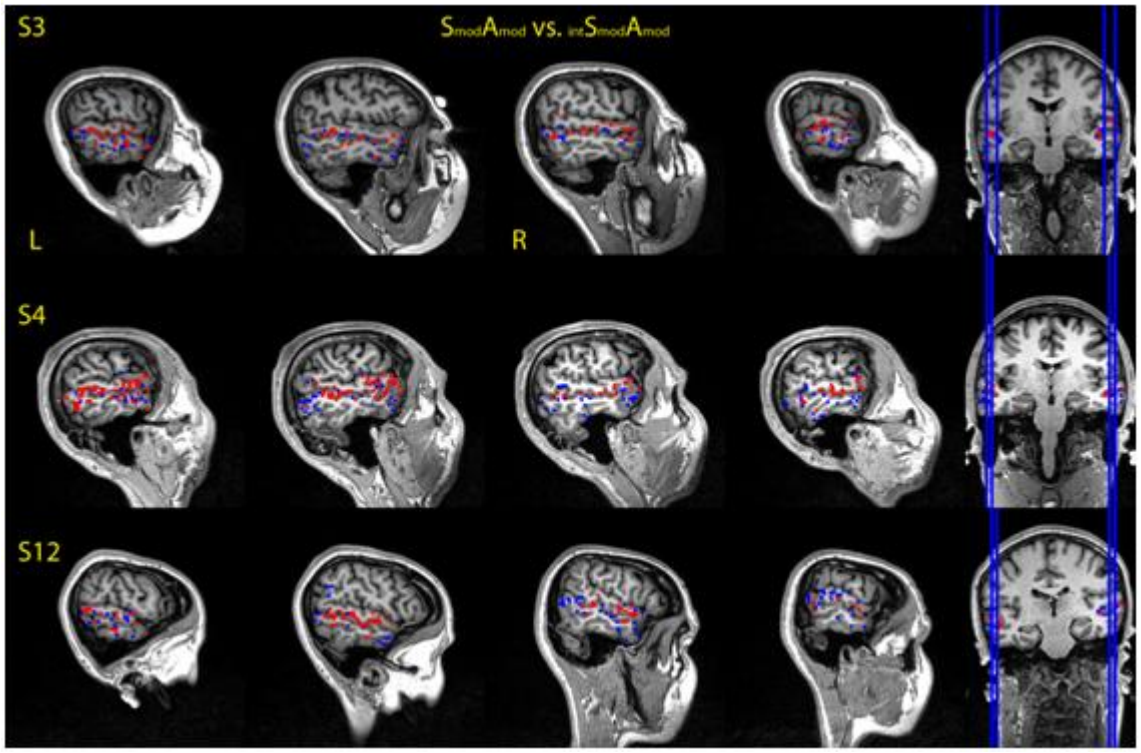


Figure 4.11 Classifier weights for the intelligibility contrast: SmodAmod vs intSmodAmod. Red voxels= intSmodAmod and blue=SmodAmod.



4.5 DISCUSSION

A significant left hemisphere preference was identified for processing intelligible speech, whilst no evidence was found for an expected left lateralisation for modulations of amplitude and right lateralisation for spectrum. The univariate analysis showed robust bilateral activation in the STG for the main effects of spectral and temporal modulation with similar cluster extents found in each hemisphere. This is in accord with previous studies that have demonstrated bilateral responses to spectral and amplitude modulation (Hall et al., 2002;Hart et al., 2003). The interaction between the spectral and temporal factors, showed a non-linear subadditivity; that is the difference in signal between SoAo and the singly modulated conditions (SoAmod or SmodAo) was larger than that between those singly modulated conditions and the SmodAmod condition.

A multivariate analysis was conducted within anatomical ROIs. This showed that HG and the Temporal ROI (inclusive of STG, MTG & STS) within each hemisphere could separate all the acoustic classifications at a level greater than chance. When the performance of the left hemisphere was directly contrasted with the right for each of these classifications, there was no significant difference shown between the hemispheres. However, when classifications of the singly modulated conditions versus the no-modulation condition were compared within hemisphere, it appeared that both hemispheres showed a subtle preference for correctly classifying spectral over amplitude modulation. When classifier weights were extracted from the classification of SmodAo vs. SoAmod it was shown that there was no difference in the relative quantity of the two types of weight within each hemisphere.

Why was no relative hemispheric preference found for the two types of modulation? It is unlikely that the explanation derives from a lack of sensitivity in the analysis. A number of facts point to this conclusion. Firstly using RFE it was shown that similar levels of performance was achieved using a smaller number of relevant voxels, as was achieved using whole anatomical regions, suggesting that the size of the ROI did not prohibit finding an effect. Indeed whilst SVM performance can be negatively affected by too many irrelevant features, their defining characteristic is the ease with which they deal with large feature spaces. Secondly, whilst a relative hemispheric preference was not shown when directly contrasting the hemispheres, a relative improvement in classification to spectral as compared to amplitude modulation was shown, suggesting some degree of sensitivity to the difference between the acoustic conditions. Thirdly, all four possible acoustic classifications were examined and a range of converging techniques were used, including plotting classifier weights, to identify any subtle preferences if they were present. Whilst the absence of evidence is not the same as evidence of absence; the fact that the lateralisation patterns were not shown after an exhaustive search argues against a lack of sensitivity and in favour of a lack of an underlying effect.

One might take the observation that an increased ability to separate spectral from amplitude modulations in the right hemisphere provides partial support for a right hemisphere preference for

spectral information. Indeed a right hemisphere preference for spectral and slowly evolving information has been more reproducible than a left hemisphere preference for temporal information (Boemio et al., 2005;Kumar et al., 2007;Belin et al., 1998). The fact however that this preference was not selective for the right hemisphere, suggests that a more general explanation might be warranted. One very simple explanation might be that the spectral modulation was just more attentionally arresting than the amplitude modulation. Indeed the dynamic range of the two types of modulation were not equated, so there could have been a greater relative degree of modulation in frequency compared to amplitude, if this was the case it is unclear how one might go about attempting to equate the two types of modulation either physically or perceptually.

There are a number of methodological differences between this and previous studies of hemispheric lateralisation which may help to explain the discrepant findings. Whilst anatomical masks were used in this study, functionally defined ROIs have often been used to address hemispheric differences. Unfortunately the same data has often been used to define regions of interest (of arbitrary size and shape) on which further statistical tests have been conducted, likely inflating the false positive rate. One reason functionally defined regions have been used is that large anatomical masks, such as the STG, can show poor sensitivity to subtle experimental effects when assessed by extracting mean signal, especially if only a small number of voxels are activated within a large region (Poldrack, 2007). It is argued here that the use of SVMs represent a significant methodological step forward in addressing issues of lateralisation, as large anatomical ROIs can be used whilst still maintaining great sensitivity to experimental effects. Furthermore the use of classification accuracy as a metric avoids thresholding effects associated with voxel counting and preserves functional-anatomical differences between the hemispheres which can be distorted when flipping the left-right orientation of statistical maps – approaches which have been adopted to address lateralisation in the past (Josse et al., 2008;Vouloumanos et al., 2001).

It seems very likely that the differences between the stimuli in this and previous studies might contribute to our failure to replicate previous work in this area. The strongest evidence in favour of an acoustic basis of speech lateralisation would be in the demonstration that the kinds of acoustic

modulations actually present in the speech signal drive each hemisphere preferentially. Indeed it is unclear how much can be learnt from extrapolating from simple non-speech stimuli, such as tones, to the processes underlying speech perception. This study is one of the first that has attempted to demonstrate auditory hemispheric lateralisation to modulations directly derived from speech.

It should be noted that the acoustic manipulations in this study are not entirely orthogonal. Indeed as is observed by Zatorre et al. (2001), the representation of temporal and spectral dimensions as dichotomous is false, as there is always a trade-off between temporal and spectral information, for example, a tone has to have an infinite duration to constitute a truly singular spectral component. At its most basic level, modulation in formant frequency must also necessitate modulation in amplitude because as the centre frequency of the formants change with time there is a concomitant increase and decrease in amplitude of the response within the corresponding auditory filters within the periphery. It is difficult to think how it would be possible to avoid this confound. Furthermore in the case of our stimuli, the singly amplitude modulated condition also involves a small concomitant modulation in spectrum; the amplitude value is defined every 10ms independently for each formant, as a result the instantaneous spectrum constantly fluctuates by small amounts, reflecting the relative change in intensity of the formants that are fixed in centre frequency. The fact that the acoustic modulations are not entirely orthogonal might explain why there was a subadditivity in the presence of the combination of the two types of modulation. It may also have been reflected in the reduced relative classification of the SoAmod vs SmodAo compared to the other acoustic contrasts. It is difficult however to conceive of how one might generate similar manipulations derived from speech without these kinds of concomitant confounds, hence the pattern of results in this study reflect the degree to which modulation of spectrum and amplitude can be orthogonalised in the context of speech rather than failings in stimuli design.

The univariate intelligibility contrast activated bilateral anterior and posterior temporal cortex and extensive regions of bilateral frontal cortex. Larger cluster extents were found in the left temporal cortex with the largest overall peak found in mid-anterior STG. Follow up classifications showed that the temporal ROIs but not HG within each hemisphere were able to separate intelligible from

unintelligible sounds at a level greater chance. The left temporal ROI (STG, MTG & STS) performed significantly better than the right. When the classifier weights were inspected voxels important in the classification which showed a relative increase to intelligible speech and unintelligible sounds, were well distributed both within and across hemisphere, however there was a significantly larger number of weights loading on intelligible as compared to unintelligible sounds in the left hemisphere, with the opposite true for the right hemisphere.

The finding of an increased ability to separate intelligible from unintelligible sounds in the left hemisphere, in addition to the significantly larger number of highly discriminative weights characterizing an increase in signal to intelligible sounds in the left hemisphere, provides converging evidence in support of a left dominant response to intelligible speech, replicating previous studies which showed an increase in activity in the left hemisphere to intelligible speech (Scott et al., 2000;Narain et al., 2003). The absence of evidence in favour of an auditory basis for hemispheric lateralisation, in combination with evidence in support of a left hemisphere preference for intelligible speech argues in favour of a left hemisphere lateralisation for speech, driven by access to linguistic representations rather than acoustic features of the speech signal. It is acknowledged however that further work is required to confirm this result, as it is acknowledged that it is difficult to argue in favour of a null result.

A nice feature of the current stimuli is that the intelligible and unintelligible tokens sound very similar, such that one wouldn't be able to easily report which stimuli came from the intelligible and which from the unintelligible category, excepting the difference in intelligibility. This makes it likely that subjects attended to both kinds of stimuli similarly, suggesting that the left hemisphere lateralisation was not purely driven by additional attention directed to the intelligible stimuli. One criticism that can be made of rotated speech is that, whilst it has many speech like qualities, it is very easy to tell that it is not speech; as a consequence there is a chance that subjects might not engage so attentively when they listen to rotated speech as compared to intelligible speech. This potential confound is not true of the stimuli used in this experiment.

The univariate intelligibility effect in this study, was very different to that observed in Chapter 3, in which only the left anterior STS was activated by the conjunction null of the two intelligibility contrasts. In this study activation was bilateral spreading across anterior and posterior superior temporal cortex bilaterally, albeit with peaks found in mid-anterior left STS/STG. The difference between the activity patterns in the two studies is likely to be explained by differences in the stimuli. In Chapter 3 a response to an entirely intelligible stimulus was contrasted with an entirely unintelligible one (clear - rot, acted to constrain the resulting conjunction null map to left anterior STS), whilst in this study the intelligible stimulus was only partially intelligible (subjects reported around ~65% key words correct). As a consequence a number of additional cognitive processes might have been engaged, including increased auditory and working memory demands. In addition subjects showed a small but significant increase in behavioural performance between pre and post scanning suggestive of some degree of perceptual learning. The additional bilateral and posterior temporal activations demonstrated in this study might then be attributed to these factors.

Extensive activation was found in the frontal cortex, including activations within the insular, inferior frontal gyrus, SMA and precentral gyrus. Unfortunately the design of this study does not make it possible to differentiate the function of the responses in these regions. The activation of both anterior and posterior frontal cortex argues for the engagement of both the posterior "how" and the anterior "what" stream. Indeed it is possible to imagine that the processing of this novel degraded speech might require the matching of existing motor representations of speech to the auditory signal, likely facilitated by the "how" stream which integrates perception with action (Rauschecker & Scott, 2009). Responses in prefrontal cortex have been shown in a number of studies involving degraded speech comprehension as a correlate of increased comprehension or perceptual learning (Adank and Devlin, 2010;Davis and Johnsrude, 2003;Eisner et al., 2010). Eisner et al. (2010) related activation in posterior parts of the left inferior frontal gyrus to variability in working memory capacity, and suggested that working memory processes may mediate perceptual learning of noise-vocoded speech. Further, the anterior insular has been associated with motor planning in speech production (Dronkers, 1996). Thus a number of candidate functions for the frontal activations can be suggested including

working memory/perceptual learning, sub vocal articulatory strategies and/or the recruitment of motor representations to aid in comprehension in the case of effortful speech perception.

The classifier weight vector maps from this study and Chapter 3 were very different, with the map in Chapter 3 showing very focal discriminative responses to intelligible and unintelligible sounds respectively. In this study there wasn't the same focal distribution of weights; the weights for the two conditions were well distributed within and between the hemispheres, albeit with a statistically significant difference in the relative proportions of weights coding for intelligible as compared to unintelligible weights in each hemisphere. As such the type of pattern the classifier was exploiting seems to be different between the studies. In Chapter 3 the classifier exploited the "univariate pattern" seemingly exploiting differences in the amplitude of response between the conditions. In the univariate analysis in this study strong bilateral activations were found across most of the superior temporal gyri for the intelligible compared to the unintelligible condition. It seems likely, supported by the observation that the weights for each class were relatively well distributed, that the discriminative pattern in this chapter reflected a more "multivariate pattern" in which the overall amplitude between conditions at each voxel was of less importance, and rather the relative pattern of increases and decreases in signal across multiple voxels was discriminative.

The response in HG was able to successfully separate all of the acoustic contrasts at a level greater than chance, but not the intelligibility contrast. This was in contrast to the findings of Chapter 3 which showed that HG was able to separate the intelligibility contrast. The classification performance within HG in Chapter 3 was relatively low and the classification results in this study generally were slightly lower than in Chapter 3; facts which should constrain our interpretation of this finding. However if emphasis were placed on the null result, then this could suggest that these stimuli provide a better acoustic control for intelligibility than rotated speech. It would also support the suggestion that participants would experience the most closely controlled intelligible-unintelligible conditions similarly barring their intelligibility. And it could be interpreted as providing contradiction to the suggestion that sensitivity to intelligibility begins as early as HG as was suggested in Chapter 3.

4.6 CHAPTER CONCLUSION

In this chapter no evidence was found to suggest that speech derived manipulations of amplitude or spectrum selectively drive the left or the right hemisphere. Evidence was found however that responses to intelligible speech were preferentially processed in the left hemisphere. Evidence for these claims were derived from a pattern classification which identified that the learnable response in the left hemisphere was more informative than the right in successfully separating intelligible from unintelligible speech. Examining the voxel weights we showed that voxels in the left hemisphere contributed most to coding for intelligible speech. Whilst these findings are interesting, it remains to be seen whether this observed hemispheric lateralisation has a functional relevance. The next chapter uses DCM to understand how regions in the bilateral temporal lobes interact with one another in resolving intelligible speech, with the hope of understanding whether the observed hemispheric lateralisation is functionally relevant.

Chapter 5 : EXPERIMENT 3

5.1 CHAPTER SUMMARY

Having shown in the previous chapter a left hemisphere preference for intelligible speech, using a locationist approach, the following chapter employs a systems based approach, to understand whether bilateral anterior and posterior temporal cortex are functionally connected and if so how these regions might interact with one another.

5.2 INTRODUCTION

In the previous chapter no evidence in support of an auditory basis for speech lateralisation was found. In contrast, evidence was found in support of a subtle left hemisphere preference for the processing of intelligible speech. In the absence of evidence in support of auditory lateralisation and the observed lateralisation for intelligible speech, it was argued that lateralisation for speech is more likely to be driven by the interface with linguistic representations than by acoustic sensitivities. These conclusions were derived from data acquired while subjects listened to degraded but mostly intelligible speech; data in which subjects showed a small but significant improvement in comprehension of the stimuli during the course of the scanning session. In the current study DCM was conducted with this same data set to understand the nature of the interaction between regions of the bilateral temporal lobes when subjects listen to degraded but intelligible speech.

The speech perceptual system is relatively robust to degradations of the speech signal and to competition from other sound sources in the environment. It is this robustness that helps us to understand speech on mobile phones and tannoy systems, and in reverberant rooms and crowded parties. One mechanism likely to support speech perception, especially when signals are degraded or masked by other sounds, is the integration of lower level acoustic with higher level linguistic

information (Davis and Johnsrude, 2007). Indeed there have been numerous demonstrations of the influence of higher level linguistic information on the perception of speech. For example, listeners are better at understanding sentences rather than single words presented in noise (Miller et al., 1951), are unaware if a phoneme within a word is replaced by a tone or cough (Warren, 1970) and show a preference for categorizing ambiguous speech sounds to form words rather than non-words (Ganong, 1980). The exact mechanism by which higher level information is integrated in veridical speech perception is the subject of controversy, with some researchers suggesting that interactive mechanisms directly influence pre-lexical acoustic processing and others suggesting that pre-lexical processing remains autonomous until integrated with higher level information at a later stage (Norris et al., 2000;McClelland et al., 2006).

Whilst the exact mechanism by which higher and lower level information is integrated in veridical speech perception is the subject of much controversy, it is less controversial to suggest that low level acoustic processes can be directly influenced by higher level information in the process of learning to understand degraded speech or new speech sounds. A number of behavioural studies have found evidence in support of this (Davis et al., 2005;Norris et al., 2003). For example, Norris et al. (2003) demonstrated that subjects use lexical knowledge to disambiguate newly encountered speech sounds. They exposed listeners to an ambiguous sound between [f] and [s]; for one group this sound appeared in contexts where its interpretation as [f] but not [s] was consistent with a real word, with the opposite true of a different group; on subsequent exposure to a continuum of sounds ranging from [f] to [s], the phoneme boundary for each group was shifted towards the sound for which the interpretation was consistent with the lexical context suggesting a lexical bias in the learning of newly encountered speech sounds. Despite behavioural evidence for the retuning of lower level perceptual processes by higher level linguistic information, there has been little clear evidence of how this might be instantiated at a neural level.

There have been comparatively few studies exploring the neural basis of degraded speech comprehension. Studies conducted in this area have tended to implicate a network of regions including inferior frontal, anterior/posterior temporal and/or inferior parietal regions (Obleser et al.,

2007a;Scott et al., 2009;Scott et al., 2004;Obleser and Kotz, 2010;Eisner et al., 2010;Adank and Devlin, 2010;Poldrack et al., 2001). Adank and Devlin (2010) investigated the time course involved in perceptual learning. They compared neural adaptation responses to time compressed speech. Four blocks of compressed and non-compressed speech were presented to the subjects. Regions responding to adaptation were found in bilateral posterior STS, left ventral premotor and left anterior STG. The profile of the response in each hemisphere was notably different. Whilst left hemisphere regions did not show adaptation to normal speech, they did so for compressed speech, showing an increased response to the first block of compression and returning to the level elicited by normal speech by block three. In the right hemisphere, regions adapted to both normal and compressed speech, with the response in posterior STS but not right anterior STG returning to the level elicited by normal speech. The authors argued that this right hemisphere adaptation is best explained by a more acoustically driven response. Evidence for which derives from the observation that right hemisphere regions showed an adaptation to non-compressed sentences likely driven by acoustic adaptation to hearing speech during continuous scanning, and the failure of the compression adaptation response to return to baseline even after extended exposure. The left hemisphere by contrast did return to baseline speech levels after repeated exposure arguably reflecting increased comprehension of the sentences. This profile is consistent with regions that work hard to resolve the intelligibility of speech and show a decreased response once this is achieved.

Poldrack et al. (2001) also examined neural responses to time compressed speech by examining responses to different compression rates (60, 45, 30 and 15% of the original length). In a sentence verification task behavioural performance ranged from 53% at the highest level of compression, to 83% at the lowest level (chance level = 50%). Using contrast weights they looked for regions showing either an increase or decrease in signal as compression rate increased or a convex pattern, i.e. reduced signal at the highest and lowest rates with greatest response at intermediate levels. A region showing a convex pattern is characterised by working hardest at moderate levels of compression, and less so when stimuli are mostly intelligible or unintelligible (although it must be

noted that subjects did not show a behavioural difference between the 45 and 60% rates, weakening this argument). The left posterior STS and IFG showed this profile.

A number of studies have explored the interaction between levels of signal degradation and linguistic factors such as semantic facilitation. These studies showed that the greatest facilitation from semantic information occurs at intermediate levels of signal degradation, suggesting that semantic information is recruited to aid comprehension less when the signal is too degraded for linguistic information to be of use or when it is not required to further aid comprehension (Obleser et al., 2007a;Obleser and Kotz, 2010). Whilst Obleser et al. (2007a) showed intelligibility effects within the STS, they failed to show increased activity in the STS to conditions in which the effects of semantic facilitation were greatest. Rather they showed a network of regions which included the angular gyrus (AG), posterior cingulate and frontal regions including the superior and inferior frontal gyrus. Furthermore they showed an increase in the degree of correlation in activity between the left AG and prefrontal cortex when semantic facilitation was greatest. Davis et al. (2011) used a hybrid sparse-continuous data acquisition sequence, that allowed stimuli to be presented in silence whilst also acquiring multiple epi acquisitions per trial, to investigate the time course involved in the interaction between semantic information and signal degradation. They presented semantically coherent and incoherent sentences at a range of Signal to Noise Ratios (SNRs) and showed that for the timing of responses most indicative of compensation for distortion, the temporal lobe response preceded an inferior frontal response. This was interpreted as providing support for "bottom up" rather than "top down" accounts of compensation for distortion, and hence was not consistent with retuning of responses in lower level perceptual regions.

Eisner et al. (2010) taught participants to understand speech which had been spectrally degraded and shifted in frequency to simulate the effect of an incorrectly inserted cochlear implant and contrasted neural responses to equivalent stimuli which had been additionally rotated to remove intelligibility (an un-learnable condition). They showed increased activity in the left IFG and posterior STS to the learnable stimuli, whilst co-varying out differences in performance level between subjects. Activation to the learnable stimuli in the left IFG showed a significant positive correlation

with improvement in comprehension ability across the experiment, whilst activation in the left STS did not. As modulation in activity in left posterior STS was not correlated with increases in comprehension, this was interpreted as demonstrating that higher level language processes were unlikely to directly modulate lower level speech perceptual systems. They then examined correlations in activity between regions when subjects listened to the learnable and un-learnable conditions and showed significant correlation between the response in the STS and IFG, and between the AG and the supramarginal gyrus, for both the learnable and un-learnable conditions, and a selective correlation between the AG and IFG for the learnable stimuli.

The failure to find evidence for the retuning of lower level acoustic-phonetic processing by higher level linguistic information might be explained by a number of observations. Firstly the techniques used to date may not be sensitive to these kinds of effects. Often studies have either not looked directly at functional relationships in the activity between neural regions or they have used simple correlations to characterise relationships. Using correlations it is not possible to specify the direction of the relationship between regions, merely that there is an association in the activity between different regions. Secondly, efforts thus far to identify modulation of lower level processes by higher level information have tended to examine interactions between distant neural regions, such as the relationship between the STS, AG and IFG, whereas it is conceivable that integration of higher with lower level information might also occur within anatomical structures such as the temporal lobes, especially as regions of the temporal lobes have been implicated in both acoustic-phonetic processing (lower) and syntactic and semantic processing (higher level language processing).

DCM allows the direct testing of hypotheses concerning the causal relationships between neural regions, moving beyond correlations to ask if activity in one region causes changes in the dynamics of another. Here DCM is used as a method to test hypotheses concerning the role of bilateral anterior and posterior temporal cortex in the perception of degraded but intelligible speech.

5.3 METHOD

This study used the same data as was presented in Chapter 4. In the study 20 subjects listened passively to five stimuli conditions all of which were based on two formant sine wave speech in which the spectral shape of the synthetic formants had been additionally broadened with noise vocoding. Four unintelligible conditions were presented in which modulations of amplitude and spectrum was absent, applied singly or applied in combination (but with the modulations of spectrum and amplitude taken from different sentences to maintain an unintelligible percept). There was an additional intelligible condition in which spectral and amplitude modulations were taken from the same sentence and a silent condition in which no stimulus was played. Subjects were pre-trained on the intelligible condition. As a group they achieved around 62% key words correct pre-scanning (chance level = 0%) and showed a small but statistically significant increase in comprehension after scanning.

Data Analysis

Dynamic Causal Modelling (DCM), is a systems based neuroimaging analysis technique which investigates how brain regions interact with one another during different experimental contexts (Friston et al., 2003). DCM is causal in the sense that it attempts to specify how the dynamics in one neuronal population cause dynamics in another. Candidate models are specified by a set of endogenous connections that specify which regions are connected (A matrix), a set of exogenous inputs that perturb the system (C matrix), and a set of modulatory connections (B matrix) which specify how the endogenous connections are modulated by a subset of inputs. The use of a neurobiologically plausible model maps from synaptic activity to a measured BOLD response, with the likelihood of the neuronal model and its parameters estimated from the observed data via Bayesian inversion (Friston, 2007). Parameter values are specified as a measure of the rate of change induced

in one region by another, with the assumption that regions causing rapid changes in others are well connected.

Data were analysed using Statistical Parametric Mapping (SPM8; http://www.fil.ion.ucl.ac.uk/spm/) using DCM version 10. Scans were realigned, un-warped and spatially normalised using the parameters from the segmentation of each participant's T1-weighted image, and smoothed using an isotropic Gaussian kernel of 8 mm full-width at half maximum. At the first level movement parameters of no interest and two regressors of interest were modelled: all auditory events and intelligible speech, against an implicit silent baseline. Note that the design matrix was altered to optimise it for the purposes of DCM (see SPM8 manual). As such two orthogonal regressors were entered, the effect of all types of auditory stimuli and the effect of intelligibility; the BOLD response specific to intelligible contrasted with unintelligible stimuli. Responses were modelled with a canonical hemodynamic response function. At the group level two one sample t-tests were constructed using the con images (the effect of all auditory stimuli and intelligible speech) from the first level. Statistical parametric maps are presented at a threshold of $p<0.05$ FDR corrected with no cluster extent threshold. Peak level activations are reported in Tables with cluster extent greater than 40 contiguous voxels for the sake of brevity. Anatomical localisation was informed by the SPM anatomy toolbox (http://www.fz-juelich.de/inm/inm-1/spm_anatomy_toolbox)

5.4 RESULTS

Univariate Analysis

Activation to all auditory events resulted in wide spread activation. Peak level activations were found in the left STG, MTG, IFG (pars triangularis), cerebellum, IOG, fusifom gyrus and post central gyrus. In the right hemisphere peak level activations were found in the right STG, HG, MFG, IFG (triangularis), cerebellum, precentral gyrus and lingual gyrus (Figure 5.1A, Table 5.1). Clusters

of activation spread broadly across the superior and middle temporal gyrus and into HG bilaterally. Thus as expected activation was located within both primary and secondary auditory cortices bilaterally.

Peak level activations in response to intelligible stimuli were found in the left STG, MTG, precentral gyrus, supplementary motor area, posterior cingulate and thalamus. In the right hemisphere peak level activations were found in the right STG, temporal pole, caudate, inferior temporal gyrus and precentral gyrus (Figure 5.1B). Clusters of activation spread broadly across anterior and posterior superior and middle temporal gyri bilaterally. The activation for this contrast as expected was found predominantly in secondary auditory cortex. Note the similarity between this statistical map and the intelligibility effect shown in Chapter 3, despite the different design matrices and modelling with a canonical hemodynamic response function rather than using an FIR approach.

Figure 5.1 (A) All auditory stimulation and (B) [intelligible - unintelligible speech].

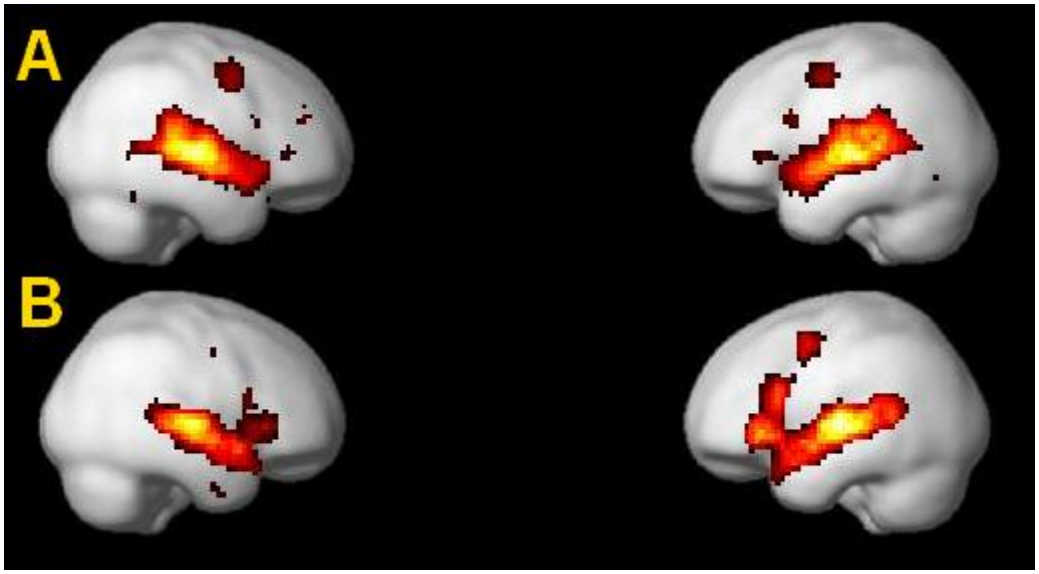


Table 5.1 Peak Level Activations, FDR p<0.05, cluster extent > 40.

Location	X	Y	Z	Extent	Z
Intelligible Speech					
Left mid-ant STS	-66	-16	1	2021	6.70
Left mid-ant STG	-57	-10	1		6.69
Left mid-post STS	-57	-22	4		6.68
Right Mid-Ant STG	63	-10	4	1256	6.41
Right Mid-Ant STS	57	-16	-5		6.02
Right Mid-Ant STS	63	-4	-5		6.01
Left Precentral	-51	-1	52	109	4.91
Left SMA	-3	11	58	81	3.92

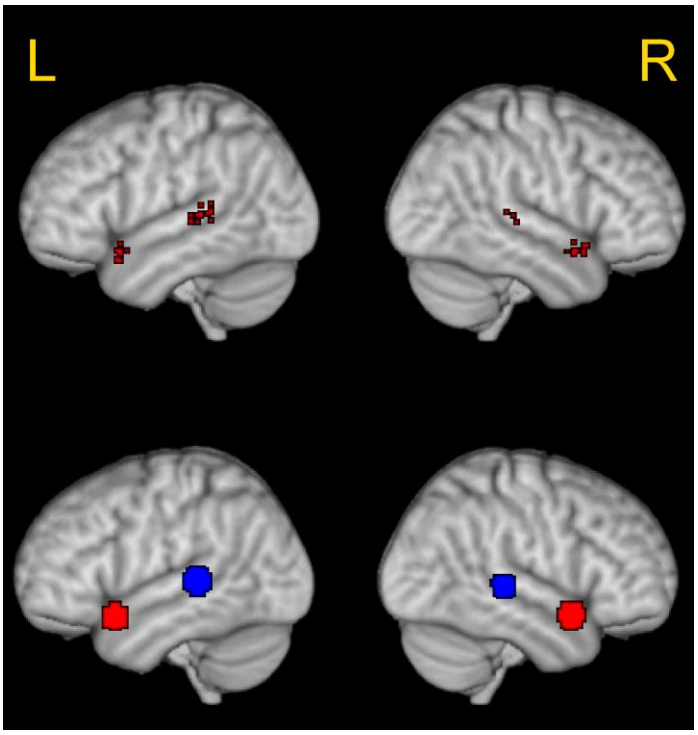
Dynamic Causal Modelling

Group "seed" coordinates for extracting Volumes of Interest (VOI) were established from regions that responded to intelligible speech at the group level. VOIs were chosen within bilateral anterior and posterior STS, this allowed hypotheses to be explored as to whether, and how these regions might be functionally connected. Group "seeds" were as follows: left anterior [-51 14 -14] and posterior STS [-60 -37 7], and right anterior [54 8 -14] and posterior STS [54 -31 4]. These coordinates were chosen to ensure that the regions were activated by intelligible speech, non-overlapping, roughly homotopic, and anatomically defined as anterior or posterior to Heschl's Gyrus (Penhune et al., 1996;Westbury et al., 1999).

Spherical VOIs were extracted with a radius of 8mm. They were allowed to vary in location by a small amount within each subject. Statistical maps for each subject were thresholded at a liberal $p<0.1$ uncorrected, taking the closest local maxima to the seed coordinates provided these lay no more than 8mm in any one direction from the group "seed" coordinate, or the closest suprathreshold voxel when the previous condition was not met. The mean locations across the subjects were as follows:

left posterior [-60 -37 7] and anterior [-52 12 -15] and right posterior [54 -31 4] and anterior [54 9 -14] temporal cortex. See Figure 5.2 for centre coordinates of all VOIs across subjects and the average location of each VOI. VOI data was extracted separately for each run.

Figure 5.2 Top – Locations of the centre coordinates for all subjects. Bottom – the mean centre coordinate across the group and a surrounding 8mm sphere for each VOI (red= anterior; blue=posterior).



Effective connectivity models were generated with full intrinsic connectivity between all regions (A matrix). The driving input of "all auditory stimulation" was fixed as entering the model at posterior temporal cortex in both the left and right hemisphere (C matrix), based on previous studies indicating the posterior STS as the best location for auditory stimuli to enter the system (Leff et al., 2008;Penny et al., 2010). The modulatory effect of intelligible speech (the B matrix) was permitted to vary across models as the manipulation of interest, with both feedforward and feedback connections permitted between anterior and posterior temporal cortex within each hemisphere and between

homotopic areas at the same level of the hierarchy. This precluded modulatory effects between the left anterior and right posterior regions and vice versa. This restriction was imposed to reduce the model space, which consisted of 256 different model structures ranging in complexity from a totally unconnected model to a fully connected model and all the models therein (see Figure 5.3 for an illustration). Data was extracted and modelled for both runs of data acquisition and included as a fixed effect in the analysis. In total there were 512 models per subject (256 x 2 runs).

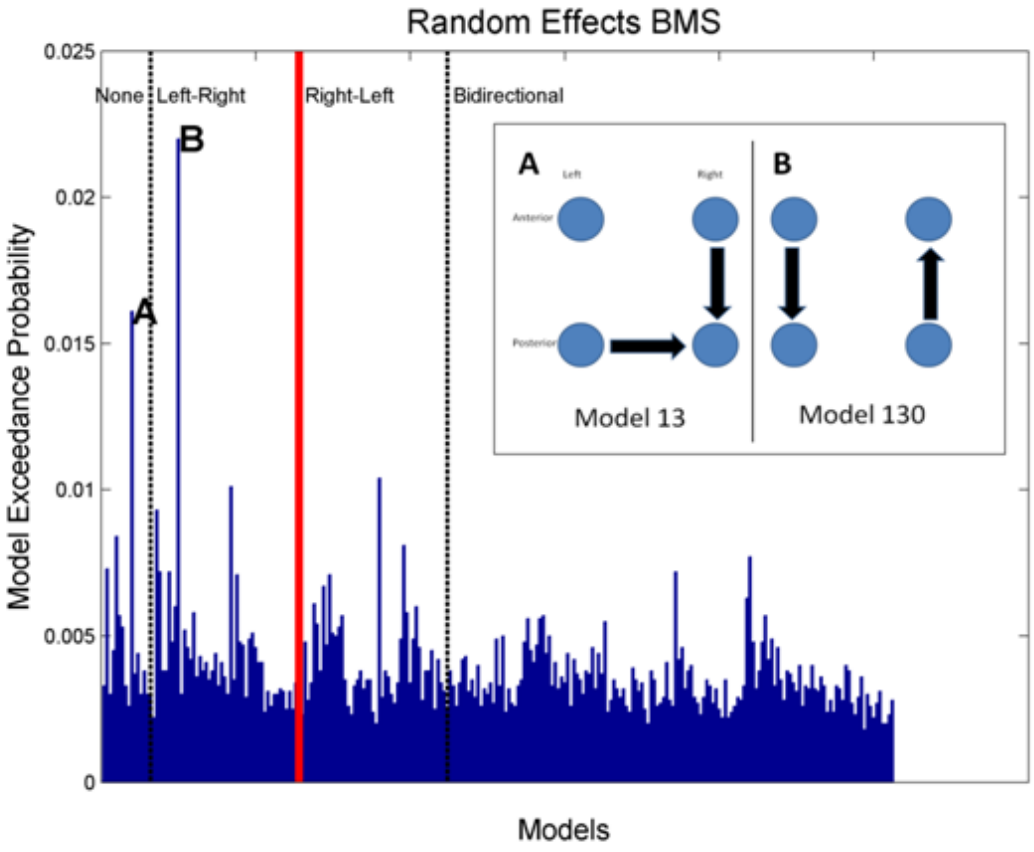
Figure 5.3 A selection of some possible models including a model with no connection between any region to a fully connected model. 256 unique models were generated in total.

In DCM random effects analyses are most appropriate when exploring hypotheses concerning higher level cognitive functions as was the case in this study (Stephan et al., 2010). In DCM analysis it has been common to select a winning model for which there is clearly the most evidence and then make inferences based on the parameters of that model, e.g. see Leff et al. (2008). When random effects Bayesian Model Selection (BMS) was carried out using all models in the model space, there was shown to be no clear winning model (see Figure 5.4), a clear winner would be expected to have an exceedance probability in excess of 0.95. Visual inspection demonstrated that there was relatively stronger evidence for models 13 and 130 compared to the rest of the models. These two models could be described as having 1.5 (model 13) to 2 times (model 130) greater probability than the third most probable model. Model 13 was characterised by the presence of a modulatory connection from left posterior to right posterior, and from right anterior to right posterior temporal cortex. Model 130 was characterised by a feedback connection from left anterior to left posterior and a feedforward

connection from right posterior to right anterior temporal cortex. See top right box in Figure 5.4 for a visual representation of these model structures.

Figure 5.4 Exceedance Probability for the full model space following BMS. The two models with relatively more evidence are marked as A and B.

Vertical partition lines represent partitioning into model families, everything to the left of the red line are models in the winning families (explained in later text).



It has recently been shown that BMS can become "brittle" when dealing with a large number of models, especially if different subjects use different models, as is more likely in random effects analysis (Penny et al., 2010). A new approach has been shown to be effective in the case of random effects analyses with many candidate models. This involves selecting the best "family" of models rather than an individual model, and making inferences on the parameters of those families using Bayesian Model Averaging (BMA) (Penny et al., 2010).

The models were first separated into mutually exclusive families for which there was either: (1) no interhemispheric modulatory effects (2) modulatory effects exclusively from the right to the left hemisphere (3) modulatory effects exclusively from the left to the right hemisphere (4) bi-directional modulatory effects between the hemispheres. BMS carried out on these families indicated that we could have a high level of confidence that intelligible speech either does not modulate connectivity across the hemispheres or that activity in the left hemisphere drives activity in the right, with a total exceedance probability of 0.90 (see Table 5.2). The two winning families included models 13 and 130. Note that in Figure 5.4 the exceedance probability for all the models (from the earlier BMS across the whole model space) is shown sorted into the above families, the dotted vertical partition line and text in the top half of the image indicate the families.

Table 5.2 Exceedance probabilities for the different model families following BMS.

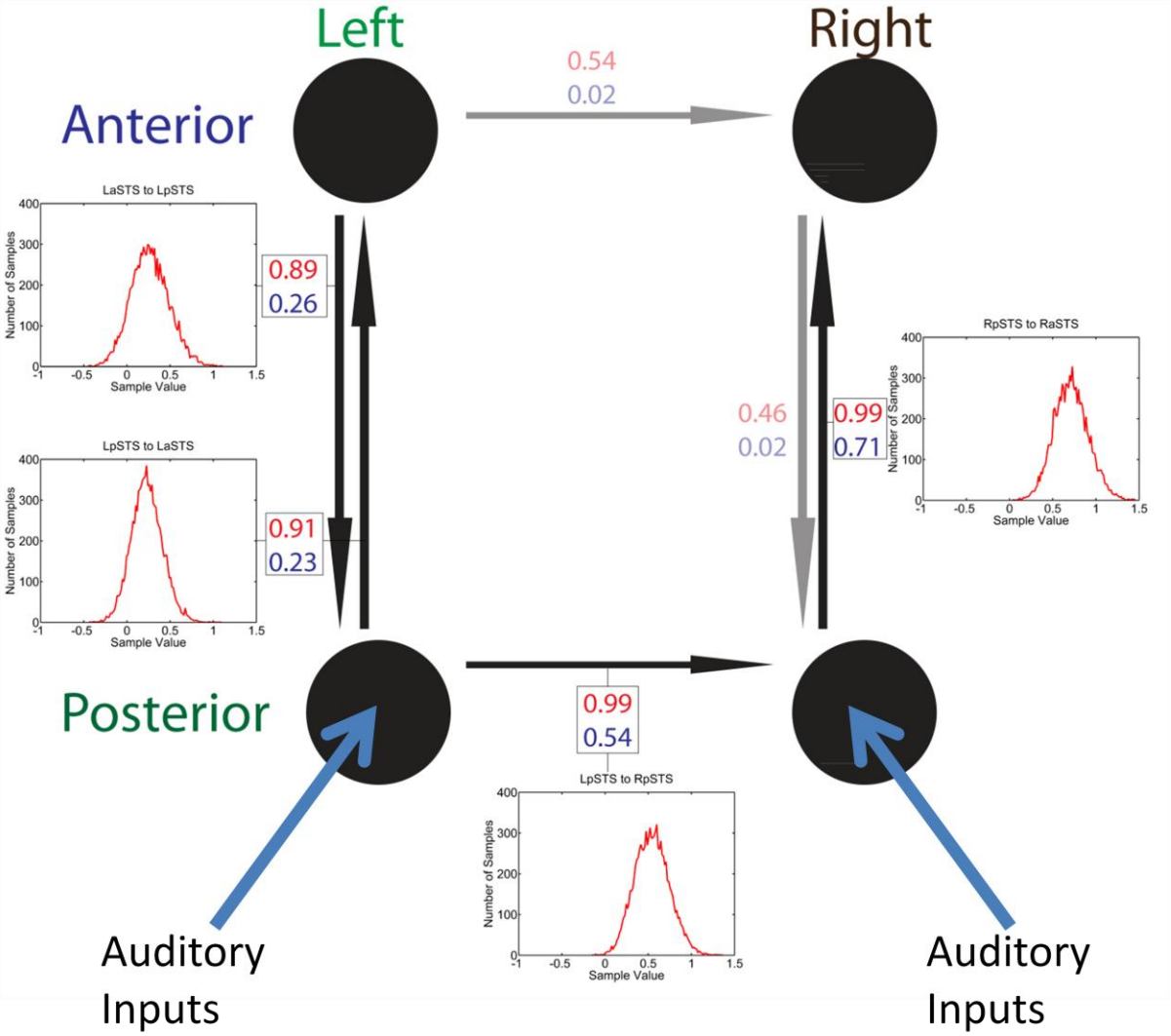
Model Family	Expected $<S_k\|Y>$	Exceedance Probability
None	0.47	0.67
Left - Right	0.27	0.23
Right - Left	0.17	0.08
Bi-directional	0.09	0.02

By considering only these two families of models the model space was reduced from 256 to just 64 unique models. Bayesian Model Averaging (BMA) was conducted within these two families to attain summary measures of likely parameter values to allow us to make inferences on specific connections (see David et al. (2011) for a similar approach). All of the averaging results were defined with an Occam's window defined using a minimal posterior odds ration of occ=1/20. In random effects model averaging each subject can have a different number of best models in Occam's window. There were an average of 32 models in the Occam's window of each subject, indicating again the fact that there was no clear winning model.

BMA, within the two designated families, showed highly probable forward (p=0.91) and backward (p=0.89) connections between the anterior and posterior VOIs in the left hemisphere, and a highly probable connection from left posterior to right posterior (p=0.99), and a further forward connection from right posterior to right anterior (p=0.99). Note that the threshold of p=0.90 is equivalent to a one tailed t-test; the backward connection between left anterior and posterior VOIs just misses out on significance by this criterion but it seems likely that this is a genuinely modulated connection, especially in light of the demonstrated low probability of the other connections. There was very little evidence that listening to intelligible speech increased the strength of connection between the left anterior and right anterior VOIs (p=0.54) or the backward connection from right anterior to right posterior (p=0.46). Figure 5.5 shows the "averaged" parameters of the most likely models. Note that features of both models 13 and 130 are represented in our "winning model": a driving connection from left to right posterior, a feedback connection from left anterior to left posterior and a feedforward connection from right posterior to right anterior temporal cortex.

Figure 5.5 The parameters of the most likely models derived from family level BMS and subsequent BMA.

Figure in red represent probability of each connection. Figure in blue indicates the median parameter value from the sampled distribution. Grey arrow represents low probability and black represents high probability connections. Distributions are also shown from the BMA sampling procedure.



5.5 DISCUSSION

Listening to degraded but mostly intelligible speech activates a network of regions, including bilateral temporal and frontal cortex. A sub network within that wider system was examined aimed at

understanding how bilateral anterior and posterior temporal cortices are functionally connected. Using DCM it was possible to model the direction of the relationship between these regions thereby moving beyond simple correlations to understand whether activity in a particular region causes activity changes in another. A number of interesting findings resulted from this analysis. Firstly, activity in the left temporal cortex was shown to drive activity within the right temporal cortex. Secondly the connection between the hemispheres was shown to be mediated via posterior rather than anterior temporal cortex. Thirdly, different functional connections were found within each hemisphere, with a reciprocal set of connections found in the left hemisphere but a purely feedforward system found in the right; the left anterior region was the only region shown to have a highly probable feedback connection.

In Chapter 4 it was demonstrated, using this same data set, that there was a subtle left hemisphere preference for processing intelligible but degraded speech. In that chapter it was shown that the learnable neural response in the left temporal cortex was more reliable in successfully separating intelligible from unintelligible sounds than the response in the right. Furthermore by examining the classifier weights it was possible to discern that the voxels contributing most to defining the classification boundary, which also showed an increase in signal to intelligible speech, were found more frequently in the left than the right hemisphere and vice versa for voxels coding for the unintelligible sounds. Using DCM these findings can be further embellished by demonstrating that the response in the left hemisphere drives the response in the right hemisphere, suggesting that the left hemisphere acts in some way to orchestrate responses in the right. Thus the results of this study demonstrate that the subtle left hemisphere preference observed for processing intelligible speech is functionally relevant.

Anterior and posterior temporal cortices were shown to be connected both within and across hemisphere. Anatomical evidence from the macaque monkey suggests that these connections are anatomically plausible (Pandya et al., 1969;Seltzer and Pandya, 1989). In the analysis, models were included in which no modulatory connections were specified between anterior and posterior regions. Therefore if it were the case that these regions were not connected it would have been possible to

determine this. The intelligibility contrast activated both anterior and posterior frontal cortex, suggestive that both the anterior "what" and the posterior "how" pathway were engaged. It is unclear whether the functional connectivity shown between anterior and posterior temporal cortex reflects the exclusive recruitment of the "what" stream or whether it reflects the interaction between the "what" and "how" streams. If it did exclusively represent the "what" pathway this might be interpreted as contradicting an account in which the stream runs anterio-laterally rather posteriorly from HG. DCM however can only provide relative evidence for models within the context of the model space assessed, and as a mid region was not included it is not possible to comment specifically on this. For the same reason these results do not corroborate the opposing view in which the "what" stream runs from HG to posterior inferior temporal cortex before running forward to anterior temporal cortex. The only way to evaluate these two hypotheses would be to directly contrast these models; as no activation was found in posterior inferior temporal regions there is no way that these two hypotheses could be directly tested. Including a mid region within the models was considered but it proved difficult to include a mid region without overlapping with anterior and posterior sites. In addition it would have been difficult to have just anterior and mid VOIs without sometimes forcing the mid region into a more posterior rather than mid location. The observation in the macaque that the connections from core, to rostral belt and rostral parabelt are largely parallel and separate to a similar set of connections running caudally (Romaski and Averbeck, 2009), and the identification of a similar set of pathways in humans connecting anterior HG to anterior STG and posterior HG to posterior STG (Upadhyay et al. 2008), suggests that communication between posterior and anterior temporal cortex was most likely via the STS and reflects the interaction between the pathways.

Fixing the VOIs at a specific coordinate would have allowed greater freedom in where VOIs could be placed allowing hypotheses to be explored with greater spatial sensitivity. VOIs in this study were defined at a group level in bilateral anterior and posterior STS, but then allowed to move slightly from these seed locations dependent on the functional activity of each subject. When the VOI centre coordinates for each subject are examined (Figure 5.2 top image) it is possible to see that the VOIs vary by small amounts in how much of the STS they include. By reference to Figure 5.2 (bottom

image) it is possible to see that the average group coordinate was located with its centre in posterior STS bilaterally, and slightly more orientated to the STG in the anterior sites, albeit with a lot of the VOI including the anterior STS. Allowing VOIs to move within each subject is a common practice, but reduces the precision of spatial localisation. It was decided to allow the VOIs to move slightly within each subject so as to allow greatest sensitivity to the experimental manipulation.

The pattern of connectivity within each hemisphere was shown to be different, with a reciprocal set of connections in the left and a feed forward system on the right. This suggests that the two hemispheres are likely to engage in different but complimentary computations. The frame work of predictive coding and more specifically the free energy principle can be invoked to understand these results (Rao and Ballard, 1999;Friston and Kiebel, 2009). In this framework the perceptual system seeks to minimise unexpected sensory experiences (referred to as free energy or "surprise"). To minimise "surprise", feedback connections send predictions of sensory experience back to lower levels, whilst lower levels communicate the amount of discrepancy between the predictions and the sensory experience forward to higher levels, with the system converging so that predictions and sensory experience align. A system can minimise unexpected sensory experience by changing the way that it samples the environment and/or by changing its expectations. This theoretical framework has been used previously to conceptualise the processes involved in learning (Friston and Stephan, 2007). The anterior temporal cortex was the only region associated with a feedback connection in our "winning model". As feedback connections are likely to facilitate the sending of predictions to lower cortical regions this would suggest a role for the anterior temporal cortex in linguistic representation. Left anterior STS has often been implicated in processing intelligible speech and as the location at which phonetic maps might be stored (Rauschecker and Scott, 2009). Equally however the anterior temporal lobes are associated with semantic (Pobric et al., 2010) and syntactic processing (Friederici et al., 2000). Thus it is not clear whether predictions fed back to posterior temporal cortex are likely to be of a phonemic, semantic or syntactic nature. It is argued here that the reciprocal connection between anterior and posterior temporal cortex represents a perceptual tuning process in which partially available higher level information is used to bootstrap the retuning of acoustic-phonetic

processes within posterior temporal cortex. It is acknowledged however that as there was no specific manipulation of the degree to which linguistic information could support speech perception, the conclusions that can be drawn are necessarily somewhat speculative.

Leff et al. (2008) also examined neural responses to intelligible speech using DCM. In a similar manner they explored the modulation of connectivity between anterior and posterior temporal cortex (as well as the inferior frontal gyrus) when subjects listen to intelligible as contrasted with unintelligible time reversed speech. In contrast to this study, they did not find reciprocal connections and instead demonstrated a purely feedfoward connection from left posterior to anterior temporal cortex. This discrepancy might arise from the fact that our intelligible stimuli were degraded necessitating perceptual re-tuning and thus a strong emphasis on the reciprocal connection between anterior and posterior regions. By way of contrast their stimuli were entirely intelligible and thus did not require the same perceptual retuning.

The connection between the hemispheres was mediated by posterior rather than anterior regions. This is surprising as in most instances when intelligible speech is contrasted with acoustically complex baselines and right hemisphere responses have been observed, activation is shown in anterior rather than posterior regions (Awad et al., 2007). rTMS of both the left and right anterior temporal lobes has been shown to impair performance on spoken semantic tasks (Pobric et al., 2010), one might then have expected that the DCM would show modulation of the connections in anterior rather than posterior regions. The fact that it did not, suggests that the connection between the hemispheres was unlikely to be modulated by semantics. Instead the connection between the hemispheres was modulated by left posterior temporal cortex, which we have already associated with a likely role in acoustic-phonetic processing. It might then be that the pattern of connectivity within the right hemisphere is reflective of acoustic-phonetic processes. Indeed Adank and Devlin (2010) found evidence to suggest differential perceptual learning processes in each hemisphere, with a more acoustically based adaptation response in the right compared to the left. One suggestion could be that the right hemisphere feed forward connection represents an attempt to integrate pitch or speaker identity information to aid comprehension of the degraded speech signal; the right hemisphere has

previously been associated with both pitch and voice processing (Patterson et al., 2002;von Kriegstein et al., 2003). Indeed the prosodic and voice quality of the intelligible stimuli are rather unusual thus within the framework of predictive coding the forward modulation may represent prediction error - the failure to use these acoustic features to support intelligibility.

5.6 CHAPTER CONCLUSION

In this chapter, using the same data as was used in Chapter 4, it was shown using DCM that activity in the right temporal cortex was likely to be driven by the response in the left, indicating that the left hemisphere preference for intelligible speech is also relevant from the perspective of functional connectivity. Further it was shown that the pattern of intra-hemispheric connections were different within the two hemispheres suggesting divergent roles in resolving intelligible speech; thus whilst the left hemisphere was characterized by both feedforward and feedback connections, the right was shown to be an exclusively feedforward system. In showing a set of reciprocal connections in the left hemisphere it was suggested that learning to understand degraded speech may involve the selective retuning of perceptual processes within left posterior STS by higher level linguistic knowledge held within left anterior STS.

Chapter 6 : EXPERIMENT 4

6.1 CHAPTER SUMMARY

In the previous two chapters neural responses were examined to a reduced representation of the speech signal. In the current chapter, intelligibility is degraded in a different way, by presenting natural speech in the presence of other concurrent sounds. A network of regions is identified that are engaged by listening to speech masked by other sounds. The effects of masking with speech from another talker and with other non-speech sounds are differentiated, and neural responses are identified that correlate with improved behavioural performance on masking tasks.

6.2 INTRODUCTION

We frequently encounter speech masked by other sounds in daily life. It is obvious to anyone who has tried to listen to a friend seated next to them at a noisy social event that listening to speech in the presence of background noise is cognitively demanding. It requires the separation and grouping of sounds from different sources and the selective attention to a particular auditory stream, alongside the concurrent need to decode the speech stream associated with the attended talker. The observation that it is easy to become distracted by a more entertaining story spoken by a friend in a different seat, suggests that listening to speech in noise is not just a problem of tuning into a desired speaker but also in tuning others out. Indeed there have been many experimental demonstrations that some features of unattended as well as attended speech are processed by listeners (Cherry, 1953). An elegant example of this is shown by Kouider and Dupoux (2005) who demonstrate that performance in a lexical decision task can be facilitated by a hidden subliminal prime.

Speech can be masked by the speech of other talkers or by non-speech sounds such as traffic and machinery noise. The masking processes in these two instances can be described as loading more

heavily on informational and energetic masking processes respectively. Energetic masking is defined as the masking that occurs when the target signal is obscured by a masker whose energy overlaps it in time and frequency; that is the representation of the target and the masker overlap in the cochlea. Informational masking is less well defined and has been described as any additional masking effects that are not accounted for by energetic masking alone (Shinn-Cunningham, 2008). Informational masking as such has been assigned a more central cognitive cause, in contrast to a sensory peripheral one. Note that informational masking is often associated with speech masking but can also occur with non-speech.

The effect of energetic masking from a steady state masker is well predicted by the speech intelligibility index, formerly the articulation index (Darwin, 2008). The intelligibility index is calculated by filtering speech and noise into frequency bands; within each band an audibility factor is derived from the Signal to Noise Ratio (SNR) within that band. The bands are then weighted by the band-importance function which indicates the degree to which each band contributes to intelligibility. The resulting index is determined by the accumulation of the audibility across the different frequency bands, weighted by the band-importance function. The success of the index in predicting the effects of stationary energetic noise indicates that the degree of spectral overlap between target and masker is important in understanding the processes involved in stationary energetic masking.

This index however fails to predict the intelligibility of speech masked by other speech. One reason for this is that it does not take into account the presence of dips in the spectrotemporal profile of maskers which allow the target signal to be momentarily glimpsed. Cooke (2006) used an automatic speech recognition system to identify consonants in noise; the proportion of the time-frequency plane glimpsed was shown to be a good predictor of intelligibility, with the resulting computational models showing a close fit to behavioural data. Thus any coherent account of informational and non-stationary energetic masking must take into account the effect of glimpsing. One way that researchers have attempted to differentiate energetic from informational masking has been to use noise maskers shaped to have a similar temporal and spectral profile to speech, in so doing the energetic component of informational masking and glimpsing effects are taken into account (Scott

et al., 2009;Brungart, 2001). This has allowed researchers to identify masking effects not accounted for solely by energetic masking.

Brungart (2001) showed that the intelligibility of speech in speech modulated noise, a signal that has the same long term average spectrum and a similar temporal profile to speech, evidenced a monotonic decrease with decreasing SNR (from 0dB); this was in contrast to speech maskers that showed a plateau in performance at SNRs below 0 dB and in some cases a slight increase in performance as SNR worsened. The speech maskers used in this study differed in their degree of similarity to the target, with same sex, different sex and the same talker used as a masker. Performance was shown to decrease with increasing similarity between the target and masker. At most SNRs performance was significantly better in the modulated noise as compared to the speech masking conditions. This difference was particularly pronounced when a same sex talker or the same talker was used as a masker, suggesting that the similarity between the target and masker increases the informational component of masking with speech. Rhebergen et al. (2005) masked speech with reversed speech and showed a relative release from masking when unintelligible reversed speech was used as compared to intelligible forward speech, suggesting a likely linguistic component to information masking. In support of this, Van Engen and Bradlow et al. (2007) also showed that sentence recognition was worse when the masker was from the same as compared to a different language. Note however earlier comments made in this thesis about the difficulty of using reversed and foreign speech as unintelligible control stimuli.

The exact processes contributing to informational masking are still relatively unknown. Mattys et al. (2010) suggested three component processes related both to language and cognitive factors: (1) competing attention required for stream segregation or selective attention, (2) increased cognitive load caused by the depletion of processing resources and (3) interference from a known language due to phonetic, lexical and semantic interference. To date there has been little work attempting to differentiate the relative contributions of cognitive and linguistic factors, and in identifying the specific levels of the linguistic hierarchy at which informational masking occurs.

Helfer et al. (2010) showed that older subjects (aged 60-69 yrs) performed significantly worse than younger subjects (20-38 yrs) when a masker was intelligible, as compared to when it had been reversed, suggesting a likely additional cognitive contribution to informational masking. Some of the older participants had a mild hearing loss, although as hearing loss did not correlate with recognition scores it seems unlikely that this result can be explained entirely by hearing loss. Boulenger et al. (2010) asked participants to make lexical decisions on target words presented in multi-talker babble with different numbers of simultaneous talkers speaking either high or low lexical frequency word lists. They showed a detrimental effect on reaction time when the two talker high lexical frequency babble was used, consistent with the suggestion that lexical competition plays a role in informational masking. Syntactic structure has also been shown to have an effect on masking, Kidd et al. (2008) showed that having a predictable syntactic structure to the speech of the target improved performance in an informational masking task in which target and masker words were presented in interleaved sequences so as to reduce energetic masking effects.

A number of studies have examined the neural correlates of speech masking. These studies have shown that listening to speech in the presence of masking sounds, as contrasted with speech without masking, recruits a large network of regions which include bilateral temporal, frontal, parietal and cingulate cortex (Wong et al., 2008;Zekveld et al., 2006;Scott et al., 2004;Wong et al., 2009). Wong et al. (2008) examined neural responses to multi-talker babble at two levels of masking, -5 and +20dB, and to speech in quiet. Behaviourally subjects performed equivalently listening in quiet and at +20dB (around 100% accuracy), but at around 85% accurate at -5dB. Relative to speech in quiet, both SNR conditions activated STG bilaterally (with seemingly greater relative activity in the left) as well as parietal, frontal and subcortical regions. When the two SNRs were directly compared, listening to the more difficult condition activated left posterior STG and left anterior insula, whereas the reverse contrast activated a number of regions including the left anterior temporal cortex. They also found a positive correlation between the response in the right STG (but not left STG) and behavioural performance on the -5dB condition. A further study using this same design but comparing responses between young and older adults showed that there was a relative deactivation in

the STG coupled with a relative increase in activity in prefrontal and precuneus in the older compared to the younger subjects, increasing activity in prefrontal and precuneus regions was correlated with improved behavioural performance, consistent with the suggestion of reduced sensory and increased compensatory processes in older subjects (Wong et al., 2009).

Other researchers have also shown differential responses associated with sensory as compared to decision processes. Binder et al. (2004) presented /ba/ /da/ syllables in a background of white noise at different SNRs. Participants were asked during the scanning session to listen out for a target syllable; the subjects' accuracy and reaction time were then used as regressors to identify activity correlated with either accuracy or the time taken to make a decision. Increases in activity within bilateral HG and lateral STG were associated with identification accuracy and activity in bilateral anterior insula and the medial frontal opercular cortex was associated with decision processes. Zekveld et al. (2006) also attempted to differentiate frontal from temporal lobe activations. They examined responses to a large range of SNRs using speech-spectrum-shaped noise. They found a network of regions that were activated when subjects listened to masked speech which was mostly intelligible (~70% correct) as contrasted with noise alone, implicating bilateral temporal, left middle and inferior frontal, and bilateral lingual gyri. They also demonstrated that the amplitude of response diverged at high SNRs in temporal lobe areas as compared to inferior frontal regions, with inferior frontal regions showing a relative increase in response relative to temporal lobe regions, in contrast to a similar amplitude of response in temporal and frontal regions at low SNRs. This they interpreted as reflecting bottom up stimulus driven responses in the temporal lobes and top down processes in inferior frontal regions.

Only two studies to date have attempted to differentiate the effects of energetic and informational masking at a neural level, both of which used PET rather than fMRI. Scott et al. (2004) presented speech against a background of speech spectrum noise and a competing speaker of different gender at a range of signal to noise ratios. They found greater activity in the left frontal pole, left dorsolateral prefrontal cortex, and the right posterior parietal cortex to noise masking as contrasted with speech masking. The reverse contrast, designed to isolate the informational component of

masking, activated bilateral STG anterior, lateral and posterior to HG. SNR dependent changes to the noise masker were identified in left inferior prefrontal cortex and left dorso-medial premotor area with responses in these regions increasing with reducing SNR. Using the average behavioural score across all conditions for each subject as a covariate revealed a region of anterior STG whose activity was positively correlated with increasing intelligibility.

By using stationary noise this study did not account for glimpsing of the masking signal, the activation may therefore have been driven by glimpses of the steady state masker when the two forms of masking were directly compared. This was directly addressed in a follow up study which included three masking conditions. A noise modulated by the envelope of speech was used to account for glimpsing effects, in addition to a rotated speech and a speech masker of different gender (Scott et al., 2009). The inclusion of a rotated speech masker was intended to allow differentiation of the effect of semantic/lexical as contrasted with acoustic-phonetic competition. A more constrained focus of activation was evidenced for the subtraction of speech masking from modulated noise masking than was observed in the previous study, with the contrast activating bilateral STG anterior to HG. The right STG was more greatly activated by masking with rotated speech than by modulated noise. There was no significant activation for the reverse contrast (which would have allowed the identification of regions responding more to lexical-semantic as contrasted with acoustic-phonetic competition). The results of the study were interpreted as suggesting differential hemispheric recruitment for masking by linguistic and non-linguistic stimuli.

The current study builds upon these two previous studies. In Scott et al. (2009) the masker and target speaker were of different genders. Informational masking was shown to exclusively activate bilateral STG and not engage additional non-auditory regions associated with increased demands on attention and cognitive control. One possible reason that these kinds of responses were not observed was because the two voices were very dissimilar making it easy to attend to the target speaker. In this study we use two sisters as masker and target respectively with exceptionally similar voices to increase the relative demands on the informational aspect of the masking task (Brungart, 2001). A clear speech baseline was also included in order to identify the whole masking network for speech and speech modulated noise masking. Further an additional level of informational masking, a

manipulation of narrative structure, is included to probe the extent to which subliminal higher level linguistic manipulations are processed at a neural level in unattended speech.

6.3 METHOD

A short behavioural experiment was conducted to select appropriate SNRs that would roughly equate the intelligibility of each masker.

Behavioural Experiment

Participants

Eight subjects were tested in initial piloting. An additional twenty-six subjects took part in the main behavioural experiment (14 male, mean age=25 years, range=18-40 years). Each subject contributed to a single experiment. All subjects had no known hearing, language or cognitive impairments, spoke English as their first language, and gave informed consent in accordance with UCL ethics committee approval.

Stimuli

Narratives were derived from the archives of a British national newspaper (The Daily Mirror, 1977-1989). Newspaper articles were chosen to equate the content and complexity of language used

across all the narratives. Specific narratives were chosen that were of a short duration (around 20 seconds when spoken) that would be attentionally engaging but of a relatively neutral emotional valence. The narratives did not reflect major news stories which the subjects might have remembered, for example a typical story was about a monkey escaping from its enclosure at a zoo. A single speaker was assigned as a target speaker and another as a masking speaker. The speakers were sisters raised in the same household and were of similar age (35 and 37 years of age) and accent (southern British English). Separate sets of narratives were recorded by the two speakers in an anechoic chamber (sampled at 44.1 kHz): 100 target narratives and 150 masker narratives. The speakers were instructed to speak in an engaging manner but with a relatively neutral prosody. Their voices were very similar making the masking of one voice on the other particularly effective.

Individual speech phrases were manually excised from each narrative at a zero crossing point. A phrase was judged to be the smallest standalone phrase that would make sense to a listener if presented to a listener in isolation, e.g. "the last wish of a widow finally came true | she has helped a young couple to buy the home of their dreams at a bargain price | the newlyweds are paying a fraction of the full cost for the kindly old ladies house". In order to maximise masking effects, an automated procedure was used to remove long silent periods, defined as sections of the waveform lasting in excess of 250ms which were less than the median value of the amplitude envelope, extracted via a Hilbert transform. This threshold was chosen by trial and error, and gave rise to natural sounding speech with few pauses. The automated procedure was used to process both the target and the masker narratives. 50ms of silence was added to the beginning and end of each phrase. This allowed the phrases to be reassembled in different orders whilst sounding relatively natural.

96 narratives were selected from one of the speakers as the target stimuli. Eight mutually exclusive randomised lists were created from these narratives, consisting of 12 target narratives per list. Each target stimulus was presented just once in the behavioural experiment, and was constructed by taking the first four consecutive speech phrases from each target narrative. Masking sentences were of the same duration as the target with an additional100 ms rise and fall time, thus ensuring that the masker always began before and finished fractionally after the target.

Two different types of speech maskers were constructed: a continuous (Con) and a discontinuous (Dis) narrative condition. Additionally two different non-speech maskers were constructed: speech modulated noise (SMN) and rotated speech (Rot), both of which were derived from the speech conditions. All stimuli were low-pass filtered at 3.8 kHz including the target stimuli. Rotated speech maskers were spectrally inverted, using a digital version of the simple modulation technique described by (Blesser, 1972), around 2kHz. As natural and spectrally inverted signals lead to sounds with different long-term spectra, the signal was equalized with a filter giving the inverted signal a similar long-term spectrum to the unrotated speech. This equalizing filter was constructed on the basis of the Long Term Average Speech Spectrum (LTASS) of the masking speaker; this is in contrast to previous methods which have used generic measurements (Byrne et al., 1994). This was found to increase the similarity in the LTASS of the rotated and original speech. The total RMS level of the inverted signal was set equal to that of the original signal.

SMN was created by modulating a speech-shaped noise with envelopes extracted from the original wide-band masker speech signal by full-wave rectification and second-order Butterworth low-pass filtering at 20 Hz. The speech-shaped modulated noise was based on a smoothed version of the LTASS of the speaker. Speech was subjected to a spectral analysis using a fast Fourier transform (FFT) of length 512 sample points (23.22 ms) with windows overlapping by 256 points, giving a value for the LTASS at multiples of 43.1 Hz. This 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–7 kHz. The smoothed spectrum was then used to construct an amplitude spectrum for an inverse FFT assuming a sampling rate of 11.025 kHz with component phases randomized with a uniform distribution over the range $0–2\pi$.

In order to create the continuous and discontinuous narrative maskers, a sub-selection of masker narratives were identified in which all the constituent phrases had a duration of 3.8s or less. Maskers with short duration phrases were chosen so as to increase the percept of discontinuity when they were randomly reassembled in the discontinuous condition. This constituted the 24 narrative stories from which both continuous and discontinuous narratives and all the non-speech maskers were derived. The continuous maskers were constructed online by concatenating consecutive masker

phrases from a specific narrative until it reached a duration which exceeded that of the target. The discontinuous masker was constructed online in each trial by concatenating randomly selected phrases from across all the speech narratives until this exceeded the duration of the target. The SMN and rotated speech were constructed by concatenating consecutive phrases which had been derived from the speech narratives. The target and maskers were then mixed at the appropriate signal to noise ratio. This signal to noise ratio was constructed by changing the intensity of the target while maintaining a constant level of the masker; hence the overall RMS level was allowed to vary for each trial.

Behavioural Testing

Each subject heard 12 trials from one of the four masking conditions (Con, Dis, Rot, SMN) at two different signal to noise ratios (from between -6 and +6dB at 3dB intervals). Stimuli were played out at 65 dBSPL over Senheiser 25HD headphones on a laptop in a quiet room using custom written MATLAB scripts. Participants were briefly trained, using an automated computerised procedure, to recognise the difference between the two speakers prior to the test. During testing participants were asked to report as much of the last phrase that they had heard from the target speaker. Participants were scored out of a possible four key words correct in their response, 48 key words per condition.

Initially pilot testing was carried out on 6 subjects using the SNRs established in Scott et al. (2009). Subjects performed rather poorly in this pilot testing, indicating that the SNR levels were inappropriate. This was not surprising considering the extra executive and working memory demands placed on subjects in recalling phrases from narrative (rather than reporting key words from single sentences as occurred in Scott et al. 2009), and the similarity of the target and masker voices. It became clear that a variety of SNRs would have to be sampled in order to fully characterise the response to these stimuli and task. To this end, a range of SNRs were tested dynamically in order to describe the effect of manipulating the SNR on the different masking conditions, with the purpose of

tracking around 80% key words correct across the group. Target stimuli list (1-8), masking condition (Con, Dis, Rot, SMN) and SNR (2 levels – an easier and harder level) were counterbalanced using two 8x8 Latin squares. The following number of measurements were made at each SNR; Con/Dis: -3 SNR (n=7), 0 SNR (n=15), +3 SNR (n=19), +6 SNR (n=11); SMN/Rot: -9 SNR (n=3), -6 SNR (n=6), -3 SNR (n=12), 0 SNR (n=9), +3 SNR (n=11), +6 SNR (n=11).

No subject noticed that there was a difference between the continuous and discontinuous masking conditions when asked. Logistic regression, analogous to analyses of variance and covariance, but appropriate for our task where the response variable is binomially distributed, was used to analyse the group data (Collett, 2003). SNR was assumed to be a continuous predictor and condition a categorical one. The analysis began with a maximal model (assuming both predictors and their interactions were significant) and then excised terms sequentially, using changes in deviance to assess statistical significance. F-tests were used as a suitable method to overcome the problems of overdispersion as described in Collett (2003). This showed that there was no significant difference between Con versus Dis, and SMN versus Rot respectively ($p=0.695$). A further analysis using a grouping term of intelligible and unintelligible maskers showed an interaction that was nearly significant ($p=0.067$), excising that term showed that both SNR and the new grouping term was highly significant ($p<0.001$). See Figure 6.1 for boxplots and Figure 6.2 for the logistic regression curves fitted separately to each condition. The following SNRs were chosen: Con=+3, Dis=+3, SMN=0, Rot=0 dB to obtain performance levels ~80% correct (range 79-82%). A relatively high level of performance was tracked to ensure that the neural correlates of effortful intelligibility, rather than responses to the absence of intelligibility were recorded.

Figure 6.1 Boxplots showing proportion key words correct as a function of SNR level for the four masking conditions

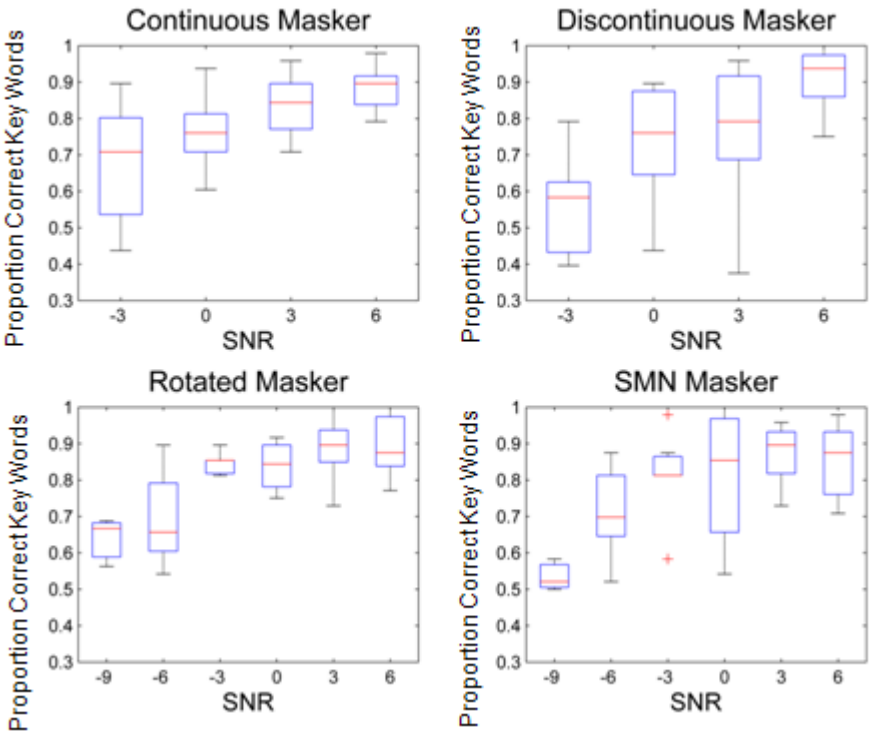
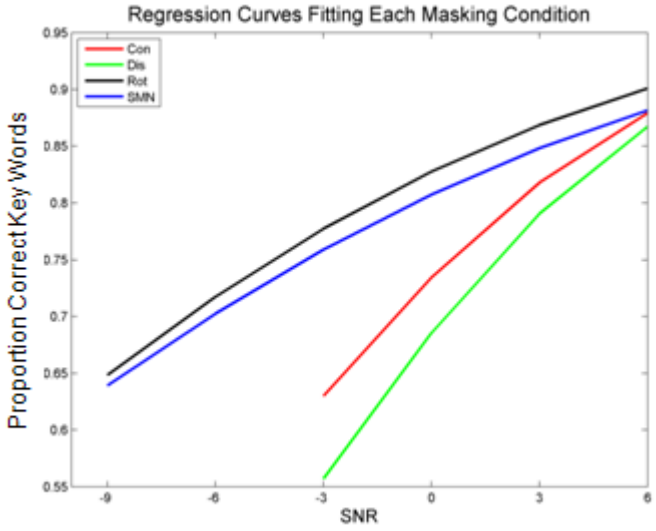


Figure 6.2 Fitted logistic regression curves for each condition



fMRI Experiment

Stimuli

The 90 target narratives were shortened where necessary (by removing individual speech phrases) to ensure that their length varied between 19 -23s. This was carried out with the proviso that the stimuli still sounded natural and made sense to the listener. Maskers were constructed to have a shorter duration than the target (17-19s). The maskers were also reduced in length where necessary, using the same criterion. A clear speech condition (target narrative in the absence of a masker) was also included. The inclusion of this condition was useful in orientating participants to the target speaker throughout the duration of the experiment to ensure that they remained on track in attending to the correct speaker.

The continuous masker was restricted to narratives in which all the constituent phrases had duration less than 3.8s. A subset of 18 narratives was chosen that met this criterion. All remaining individual masking phrases that had not been included within the "continuous set" that had duration less than 3.8s were collated from all remaining masker narratives. From this large selection of phrases a set of 18 discontinuous narrative stimuli were created by randomly concatenating phrases whilst ensuring that the mean length of the concatenated stimuli did not differ statistically from the continuous maskers ($t=1.160$, $p > 0.05$) and nor did the mean length of individual phrases that constituted them ($t=0.847$, $p > 0.05$). Note that the stimuli in the continuous and discontinuous conditions were mutually exclusive.

Rot and SMN were derived from the same (randomly selected) half of the continuous and discontinuous stimuli to create 18 rotated and 18 SMN masking stimuli. A randomisation was performed to create a single set of stimulus conditions in which targets and maskers were mixed (target+continuous masker, target+discontinuous masker, target+rotated speech, target+SMN, target

alone). The masker was exactly centred within the target so that the onset and offset of the masker were of the same duration. Note that the relative durations of masker and target were reversed compared to the behavioural task; thus in the scanner a longer target stimulus ensured that subjects would be able to orientate to the target before masking began. No target or masker stimulus was repeated between conditions. The target and maskers were mixed offline at the appropriate signal to noise ratios established by behavioural testing (Con=+3dB, Dis=+3dB, Rot=0dB, SMN=0dB). The output RMS was equalised for all conditions including the clear speech.

Scanning Procedure

Eighteen subjects underwent fMRI scanning (male=10, mean age=28 years, range=18-38 years). Scanning was performed at the Birkbeck-UCL Neuroimaging (BUCNI) Centre on a 1.5 T MR scanner (Siemens Avanto, Siemens Medical Systems, Erlangen, Germany). In the scanner auditory stimulation was delivered with the Cogent Toolbox (http://www.vislab.ucl.ac.uk/cogent.php) via electro-static headphones (MRConFon). Each subject was given the opportunity to hear the target talker speaking over the concurrent noise of the MR scanner to ensure that the sound level was appropriate. Stimuli were played out at the same fixed comfortable listening level for all subjects (with this level set following initial piloting), except in the case of two subjects who required a slight increase in volume. Two functional runs of data were acquired lasting around 15 minutes using a continuous acquisition sequence (TR=3s, TE=0.05,flip angle 90 degrees, 35 axial slices, 3x3x3 in-plane resolution). Continuous rather than sparse acquisition was used so as to increase statistical power, with the assumption that the noise arising from continuous acquisition would add to the ambient level of energetic masking. Note that a relatively quiet sequence (in MR terms) was used (~80dBSPL) along with sound attenuating headphones (~30 dB). A hi-resolution T1 structural image (HIRes MP-RAGE, 160 sagittal slices, voxel size=1 mm^3) was acquired following the functional runs. Nine narratives from each condition (and the target in clear) were played out during each run (45

narratives per run). Six, 18 second silent trials were inserted in the middle of each functional run. Five different pseudo-randomisations were used for presenting the stimuli in the scanner. Around 780 functional volumes were acquired across the duration of the experiment with the exact number of volumes variable, dependent on which stimulus randomisation was used. This pseudo-randomisation ensured that the clear speech "event" was fixed as occurring before each presentation of the four masking conditions with the four maskers randomised within this sequence. This ensured that participants always heard clear speech as the first stimulus of the experiment and heard the target talker in clear every four masking stimuli. This helped to ensure that the participants were orientated to the speaker that they needed to attend to throughout the experiment.

Subjects were briefed on the purpose of the study prior to scanning. They were trained to differentiate the two voices using the same familiarisation procedure as was conducted in the behavioural test. They were also played a sequence of stimuli exactly as they would be heard in the scanner (with different examples to those used in the scanning session). Before scanning commenced it was established that each participant knew which speaker they must attend to. Subjects were allowed as much familiarisation as was necessary to get them to this point. Participants were required to passively listen to the narratives in the scanner. Each participant was told to listen out for a "story about a bear" and were told that they would be questioned about the story when the scan had finished. This was to ensure that subjects maintained attention throughout the scanning; in reality there was no story about a bear for them to attend to. After scanning subjects were debriefed. Most subjects reported not hearing the story about the bear and were able to relate a number of the target stories. Subjects were given a behavioural test, identical to the testing used to previously establish the SNR levels, after the scanning session. This was to assess how well they were likely to have understood the sentences during the scanning session.

Imaging Analysis

The first five volumes from each run were discarded to allow longitudinal magnetization to reach equilibrium. Data were analyzed using SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/). Functional images were slice time corrected, realigned, unwarped and coregistered with the anatomical image. Normalisation was conducted using the parameters obtained from the segmentation of the anatomical image and data were smoothed using a Gaussian kernel of 8mm FWHM. Each stimulus was modelled as a canonical hemodynamic response function, with onsets modelled from the onset of the stimulus and with durations specified as the length of the narrative (with this length tailored to the length of each individual stimulus). At the first level each condition was modelled as a separate regressor in a GLM: Con, Dis, Rot, SMN and target in clear, against an implicit silent baseline. Six movement parameters were included as regressors of no interest. At the second level a succession of one sample t-tests were conducted. In each t-test, a covariate was modelled expressing each subject's performance on the post scanning speech in noise task relative to the group. This was derived by calculating a z-score for each subject relative to the group for each separate masking condition and averaging those scores.

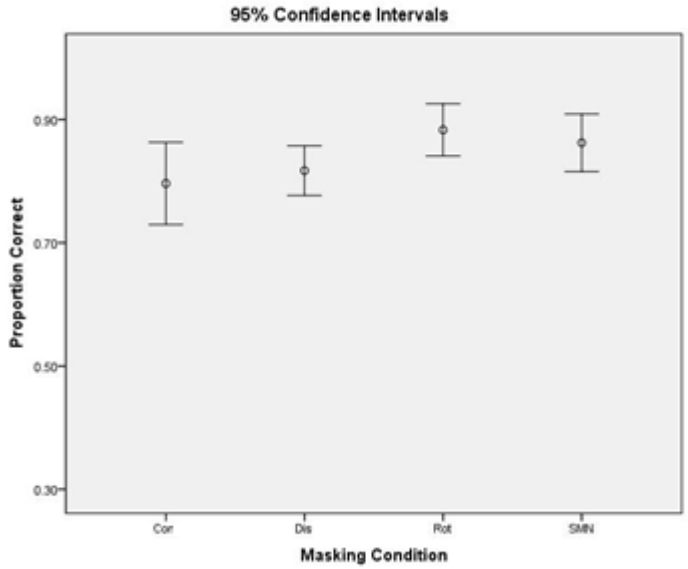
The cluster extent was corrected for multiple comparisons using Monte Carlo simulation (Slotnick et al., 2003). As clusters of activation are increasingly improbable as they become larger, it is possible to determine the probability of a given spatial extent of activity (or larger) and then enforce an extent threshold to yield the desired Type I error rate. After running 1,000 simulations of the null distribution it was determined that for an individual voxel threshold of $p < .001$, a cluster-extent threshold of 16 contiguous voxels was necessary to correct for multiple comparisons to achieve a corrected significance level of $p<.05$. Thus, only clusters of activation meeting or exceeding that size were considered significantly activated. Spatial localisation of responses was informed by SPM anatomy (http://www.fz-juelich.de/inm/inm-1/DE/Forschung/_docs/SPMAnantomyToolbox/SPMAnantomyToolbox_node.html).

6.4 RESULTS

Post-scanning behavioural testing

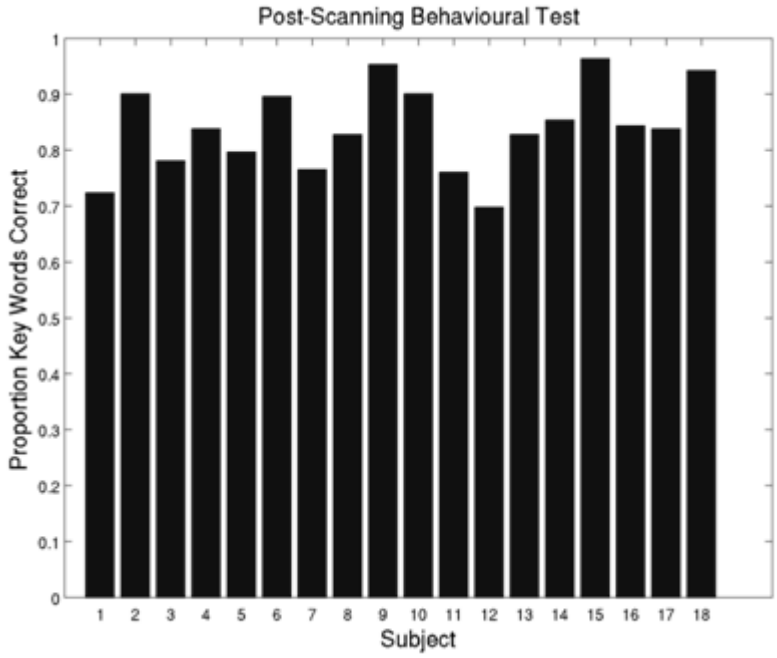
As would have been predicted, post scanning behavioural tests showed that the subjects performed at around ~80% key words correct across all the masking conditions, see Figure 6.3.

Figure 6.3 Post scanning behavioural test for the group showing 95% confidence intervals for each masking condition.



The mean accuracy of the subjects across all of the masking conditions was calculated (shown in Figure 6.4). This demonstrated that there was a moderate degree of variability in how well each subject was able to understand masked speech, this ranged from 70% to 96% correct with a mean performance across the group of 84%.

Figure 6.4 Mean accuracy across all masking conditions for each subject.



fMRI results

Responses to clear speech (above the noise of the scanner) were examined in the absence of the explicit masking. Clusters of activation spread broadly across bilateral primary auditory cortex, anterior and posterior superior and middle temporal gyri, and into left IFG. Peak level activations were found in the left and right PAC (TE 1.0), STG, STS and MTG and the left IFG (pars triangularis) (See Figure 6.5).

Figure 6.5 [Clear - silence (scanner noise)].

The masking network, undifferentiated by masking type: [all mask - Clear] was then examined. This gave rise to activation across a wide network of regions which included and extended beyond those regions identified as responding to clear speech alone. This bilateral symmetric network was concentrated predominantly within superior temporal, parietal (inferior and superior parietal lobule and precuneus), prefrontal (insular, inferior frontal and middle frontal) and cingulate cortex (see Figure 6.6).

Figure 6.6 [All Masking – Clear].

A cluster of activation focused predominantly within mid to posterior lateral STG (peak: [-63 -16 10]) was shown to correlate significantly with individuals' post scanning performance (averaged across conditions). 24.1% and 3.8% of the clusters fell in TE3 and TE1.2 respectively (Figure 6.7A). Plotting the averaged response from this cluster showed that subjects who performed better at the masking tasks tended to activate this region more when listening to masked speech (Figure 6.8).

Responses to increased intelligibility were explored by identifying regions which responded more to clear speech than to masked speech [clear - all mask] (see Figure 6.7B). The cluster of activation spread along most of the length of the STS in the left hemisphere with activation also identified in the Angular Gyrus (AG), and in the right hemisphere activation was focused exclusively within the anterior STS. The largest peak level activation was found in the left temporal pole; additional peaks were found in left mid-anterior and posterior STS and AG, and in the right anterior STS (see Table 6.1).

Figure 6.7 (A) Activity co-varying in [All Mask - Clear] with speech in noise abilities and (B) [Clear - All Mask].

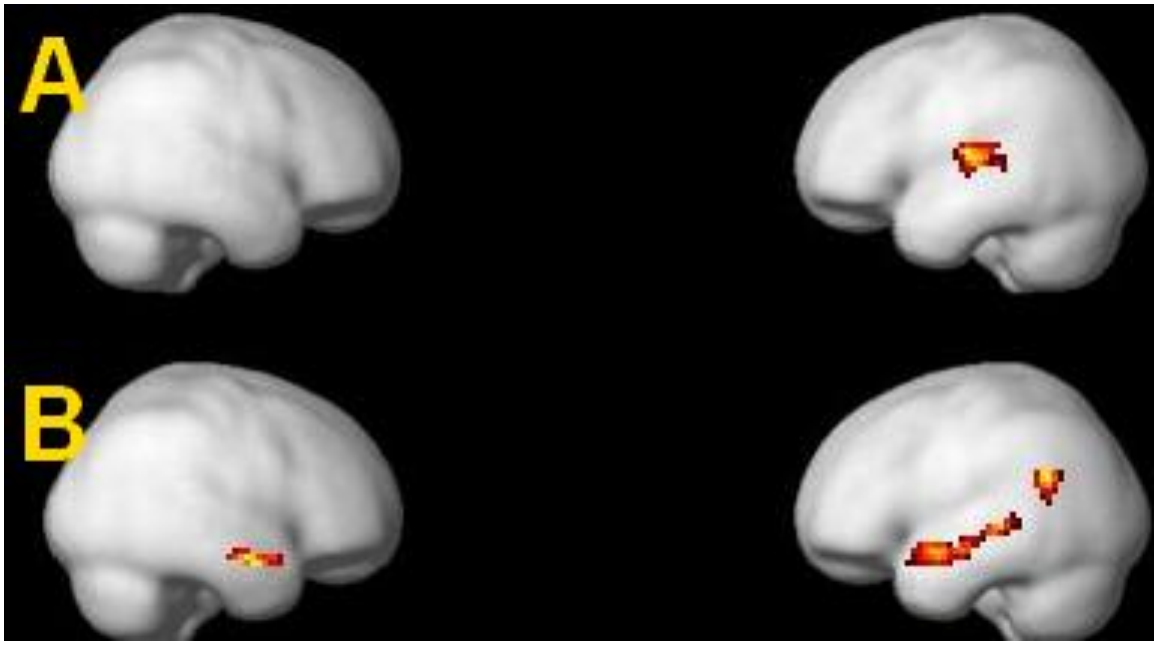
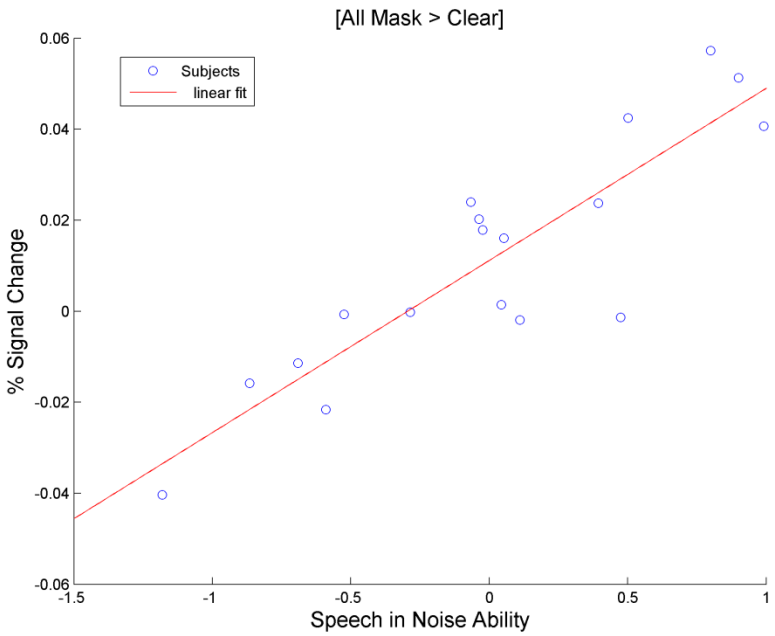
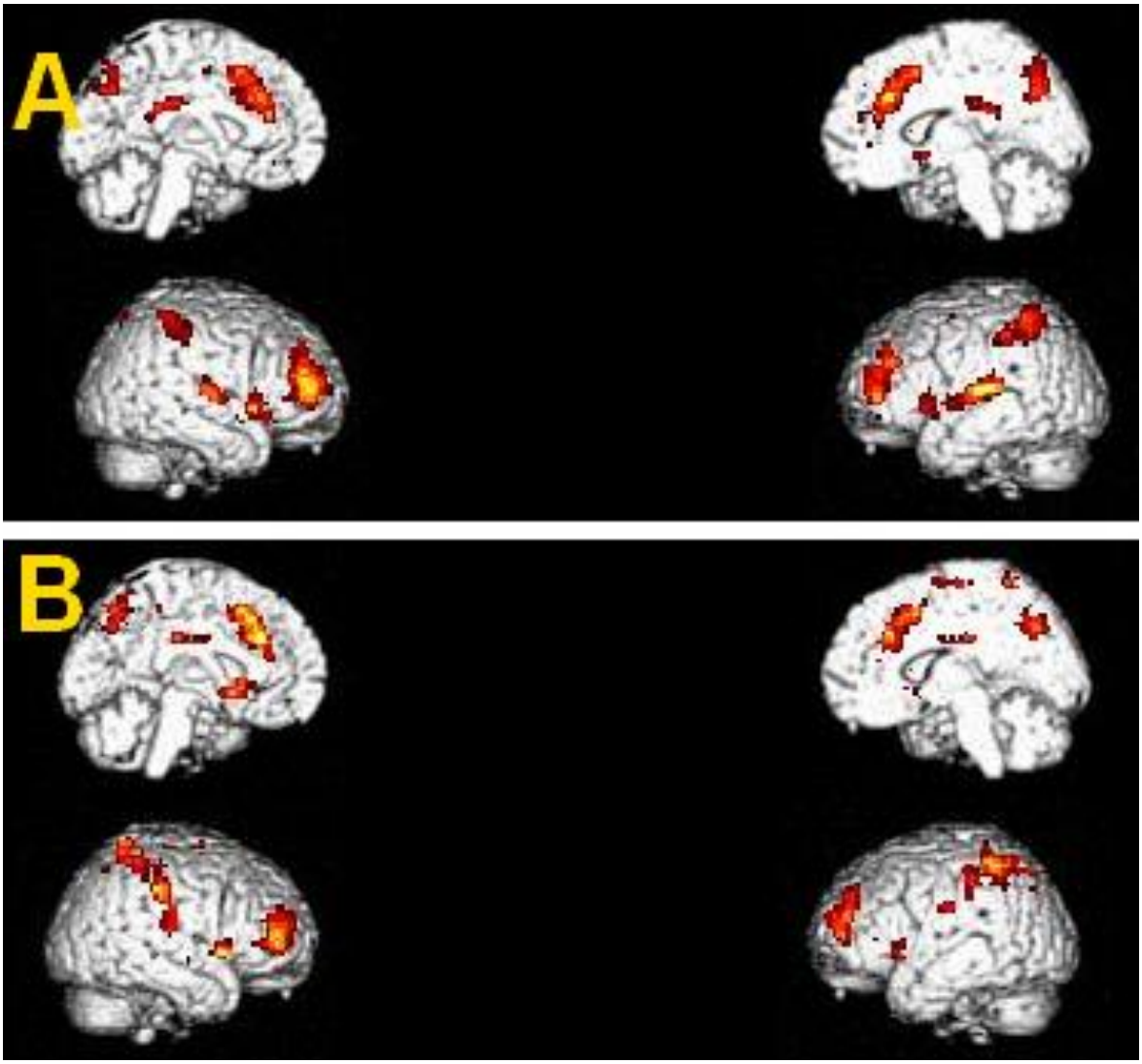


Figure 6.8 Percent signal change from [all mask - Clear] for the averaged response in the left mid-posterior STG cluster (Figure 6.7A above) plotted against speech in noise ability.



Masking responses which loaded more heavily on energetic and informational masking respectively were then examined. We began by examining the whole informational masking network, [Con+Dis - Clear]. Note that the Con and Dis conditions were collapsed as the separate [Con - Clear] and [Dis - Clear] gave rise to almost identical statistical maps. The [Con+Dis - Clear] contrast gave rise to a similar network of clusters as the all masking contrast; a bilateral network including temporal, parietal, prefrontal and cingulate cortex (see Figure 6.9A). When the energetic masking system was examined, [SMN - Clear], an almost identical pattern of activation was identified (see Figure 6.9B) albeit with an absence of significant activation in the bilateral temporal lobes. Note however that activation was identified in left STG when the cluster extent threshold was reduced to 12, and that [SMN - silence (scanner noise)] activated the bilateral temporal lobes strongly (not shown).

Figure 6.9 (A) [Con+Dis - Clear] (B) [SMN - Clear].



The contrast of [Con + Dis - SMN] was conducted to identify regions responding to informational masking when the contribution of energetic masking had been largely accounted for. This gave rise exclusively to clusters of activity in bilateral STG, with peak level activations in left anterior-mid STG and posterior STS/STG and right primary auditory cortex (TE3) (Figure 6.10 and Table 6.1), this pattern was not significantly altered by reducing the statistical threshold to $p<0.005$.

Subjects who performed well at masking tasks activated the left mid-posterior STG and left IFG more when presented with informational maskers, and just the left mid-posterior STG in the case

of energetic maskers (see Figure 6.11 for rendering on the same brain, and Figure 6.12A&B and 6.13 for plots from these clusters).

Table 6.1 Peak Level Activations, FDR p<0.05, cluster extent > 16.

Anatomical Label	MNI	Extent	Z
[Clear - All Mask]			
Left temporal pole	-54 8 -14	163	4.82
Left mid-ant STS	-51 -16 -11		4.62
Left post STS	-48 -31 -5		3.85
Right mid-ant STS	51 -1 -20	66	4.76
Right temporal pole	54 8 -20		
Left posterior STS	-54 -52 22	79	4.38
Left angular gyrus	-39 -52 22		3.32
[Con+Dis - SMN]			
Left mid-ant STG	-57 -16 7	183	5.46
Left posterior STG	-69 -31 10		4.47
Right PAC	66 -7 1	97	4.54

Figure 6.10 [Con+Dis - SMN]

Figure 6.11 Regions showing a correlation between activations revealed by [Con+Dis - clear] (blue) and [SMN - clear] (red) and individuals' speech in noise ability. Overlap between the two in purple.

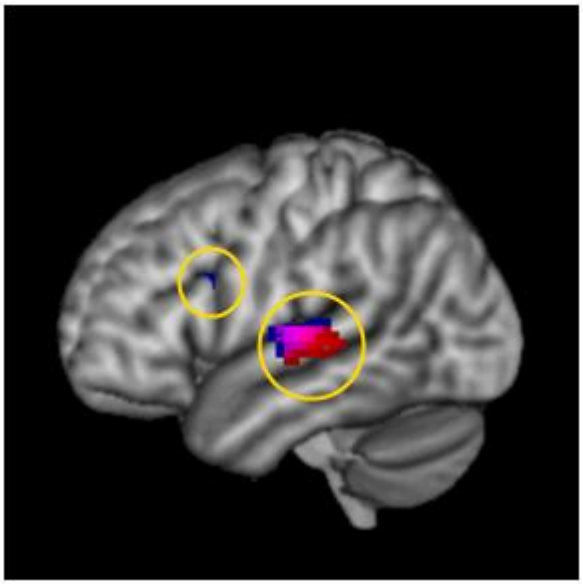


Figure 6.12 Plots of subjects' percent signal change in [Con+Dis - Clear] against speech in noise ability averaging the response in the cluster (A) in left mid-posterior STG (B) left IFG(blue on rendering in Figure 6.11).

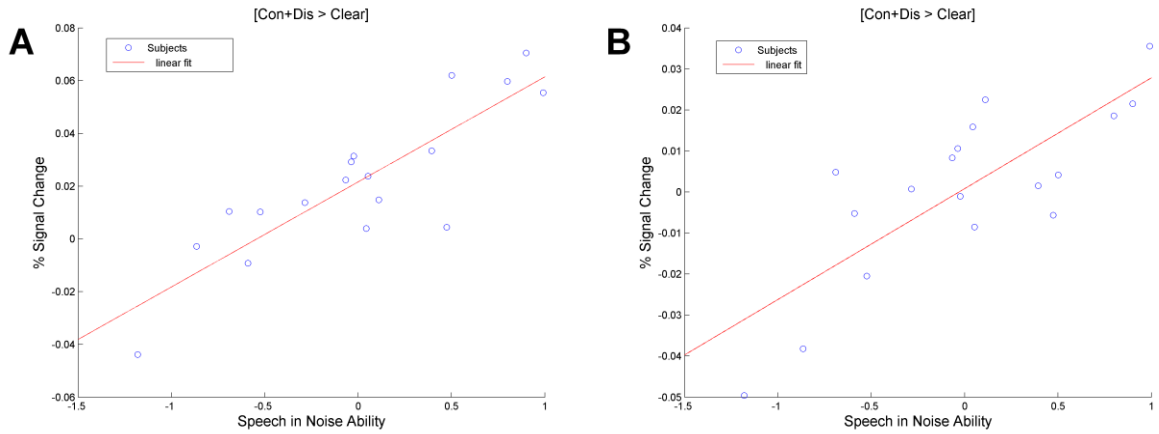
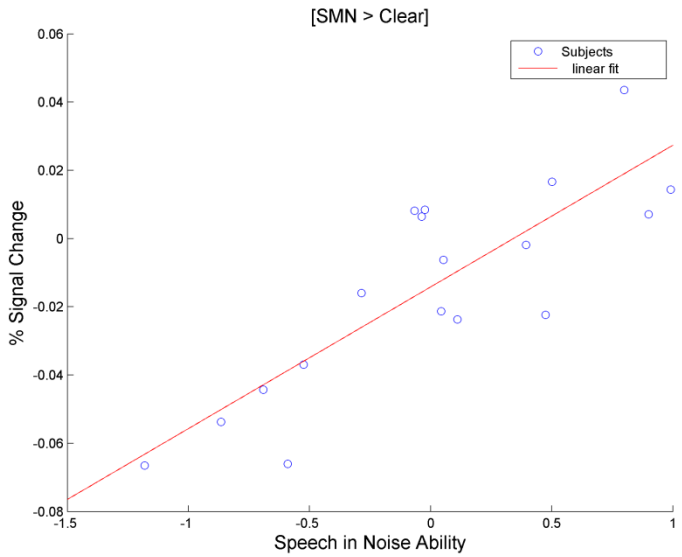


Figure 6.13 Plots of subjects' percent signal change in [SMN - Clear] against speech in noise ability averaging the response the cluster in left mid-posterior STG (red in rendering in Figure 6.11).



The effects of narrative continuity, [Con - Dis] yielded peak level activations in the inferior and superior parietal lobule, paracentral lobule and middle cingulate gyrus (see Figure 6.14), the reverse contrast, [Dis - Con], generated no significant activations regardless of threshold. Unfortunately as was also the case in Scott et al. (2009), the contrast of [Con+Dis - Rot] did not give rise to any significant activation, even when the threshold was reduced significantly. The contrast of [Rot - SMN] gave rise to bilateral activation of the STG, peak level activations were found within left STG, right STG and right primary auditory cortex (TE 1.0) (see Figure 6.15). This bilateral pattern was in contrast to the solely right lateralised activation identified for the same contrast in Scott et al. (2009).

Figure 6.14 [Con – Dis].

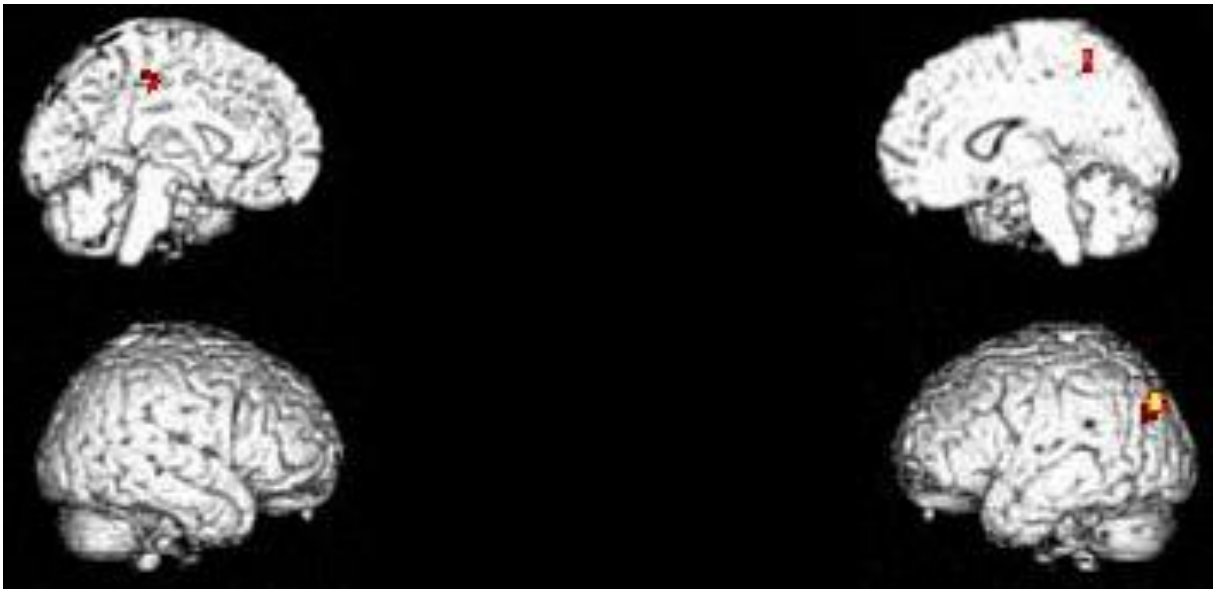


Figure 6.15 [Rot - SMN].

6.5 DISCUSSION

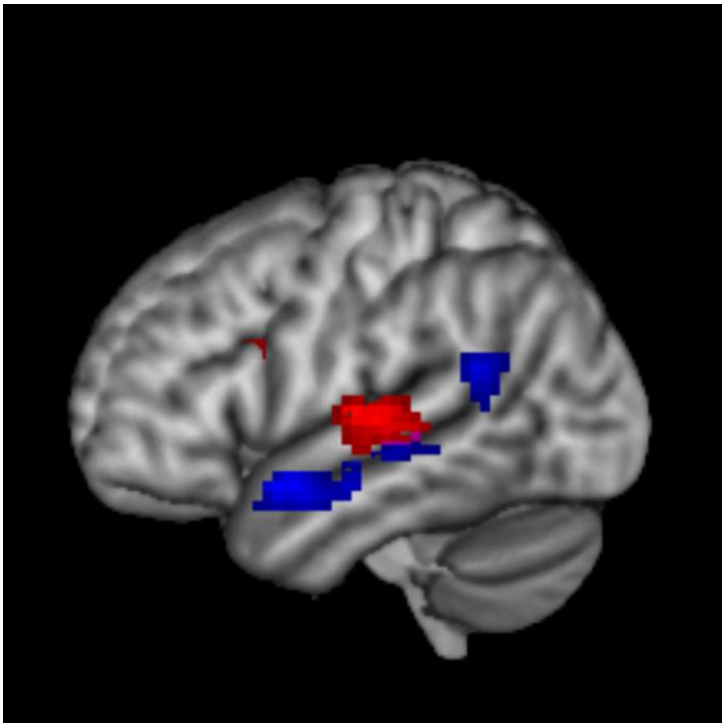
In this study a network of brain regions were identified that responded to the masking of speech by other sounds. This network included bilateral temporal, parietal, frontal and cingulate cortex. Subjects who performed well at masking tasks activated left mid-posterior lateral STG more than subjects who did not. Better performers activated the left IFG and left mid-posterior STG when

presented with informational maskers and the left mid-posterior STG in the context of energetic maskers. The majority of the left STS (in addition to the left AG) and the anterior portion in the right STS were recruited more strongly in response to fully intelligible speech as contrasted with masked speech, with the largest peaks located in anterior regions. The findings of Scott et al. (2009) were replicated in showing that bilateral STG (inclusive of primary auditory regions) were more activated by informational maskers than energetic maskers. Interestingly regions outside of auditory cortex were not shown to be more greatly activated by informational masking. Finally it was shown that the superior and inferior parietal cortex, paracentral lobule and middle cingulate cortex showed sensitivity to whether the masking speech constituted a coherent or an incoherent narrative, despite the fact that the subjects did not show an explicit awareness of this manipulation.

A region of left mid-posterior lateral STG was more greatly activated by subjects who performed well on an aggregate post scanning measure of speech in noise ability. Wong et al. (2008) found activity in right rather than left posterior STG that correlated with speech in noise ability. The difference in lateralisation might be explained by the maskers used in the respective studies. Wong et al. used a multi-talker masker. Right hemisphere temporal lobe responses (albeit anterior rather than posterior ones) have been associated with voice processing (von Kriegstein et al., 2003). One could imagine that masking with multiple talkers might place greater demands on voice processing and hence may explain the correlation in the right hemisphere found in that study.

The region more greatly activated by better performing individuals was spatially separate from a region which responded more to clear speech than to masked speech (see Figure 6.16). The fact that subjects who perform better at masking tasks activate the left mid-posterior STG more suggests that this region is likely to play an active role in acoustic-phonetic processing supporting intelligibility. And its spatial separation from a region within the STS which responds to increasing intelligibility suggests a degree of functional dissociation between regions engaged in the "process" and "product" of resolving intelligible speech.

Figure 6.16 Response to increasing intelligibility (blue) and the region activated more greatly by subjects who performed better in masking tasks (red).



The intelligibility response extended along most of the length of the STS in the left, with an additional cluster in the left AG, and the right anterior STS. This response was similar to responses observed in previous studies in which intelligibility responses have been found in right anterior and left posterior temporal cortex (Awad et al., 2007;Spitsyna et al., 2006), and are similar to the activation shown in Chapter 3 in which the main effect of intelligibility was examined whilst masking out the interaction. As was also the case in all previous studies in this thesis, the main peak of the intelligibility response was found in the left anterior STS.

Activation extending as far back as the AG has not often been noted in studies of speech intelligibility, except in the case where semantic information facilitates speech perception in the case of a degraded speech signal (Obleser et al., 2007a). The AG has often been associated with semantic processing (Binder et al., 2009), thus the activation in this contrast could reflect greater access to semantic information in the clear speech condition contrasted with the masking conditions. The AG

has been ascribed numerous other functions however including a role in the default network (Seghier et al., 2010). As part of the default network it has been demonstrated to show a relative deactivation during goal directed tasks. One could imagine in the context of this experiment that listening to speech in clear would be significantly "less active" compared to listening to speech in noise. Thus either a semantic processing or default network explanation might explain why increasing the level of intelligibility of the signal leads to the AG activation.

Speech in the absence of masking sounds activated the bilateral primary auditory cortex, superior and middle temporal gyri and left inferior frontal gyrus. When concurrent masking sounds were played to listeners, the same regions responding to a single speaker also responded to masking, in addition to a number of additional prefrontal, parietal and cingulate cortex regions. As all the speech stimuli were RMS equalised the activations in these additional regions cannot be explained solely by a greater overall level of auditory stimulation in the masking conditions. A number of other studies have shown a similar bilateral system responding more to masked than to clear speech (Wong et al., 2008;Wong et al., 2009). Previous studies examining responses to masked speech have indicated that the temporal lobe plays a sensory, and the inferior frontal cortex a decision based or "compensatory" role in masking (Zekveld et al., 2006;Binder et al., 2004). This is consistent with the suggestion that a network involving prefrontal and cingulate cortex is activated across a diverse range of tasks which involve response conflict, working memory and perceptual difficulty (Duncan and Owen, 2000). Indeed a prefrontal, parietal, cingulate network has been shown to be involved specifically in auditory based cognitive control (Falkenberg et al., 2011). Parietal cortex has been suggested to play a particularly important role in auditory scene analysis, as well as in the integration and structuring of sensory information across and within a range of modalities. The intraparietal sulcus in particular, in which activation falls in both masking conditions, has been suggested to play a particular role in stream segregation (Cusack, 2005) and auditory figure-background segregation (Teki et al., 2011), processes which are likely to be particularly important when listening to masked speech. Interestingly little activation extended into planum temporale, a region also often associated with scene analysis (Griffiths and Warren, 2002). Interpreted within the framework of streams of

processing, co-activation of temporal, parietal and prefrontal cortex during masked speech suggests that both the anterior "what" and the posterior "how" cortex are strongly engaged in processing speech in the presence of speech masking.

Correlations between neural activity when subjects listened to informational and energetic maskers and their speech in noise ability were examined. There was a common region in the left mid-posterior STG that correlated with behavioural performance. There was an additional region, the left IFG, which showed a correlation between activity during informational masking and masking ability, which was not present when subjects experienced energetic masking. The activation in the IFG cluster indicated that subjects who performed well at masking engaged the IFG more when listening in an informational masking context. Eisner et al. (2010) found activity in the left IFG to be associated with an individual's ability to learn to understand degraded speech, they suggested that the IFG's role in working memory processes may have facilitated this improvement. It is difficult to understand why working memory processes would confer benefit in an informational but not an energetic masking context. An alternative explanation might be that activating the IFG more facilitates more efficient processing of semantic competition that arises from masking with meaningful speech (Bedny et al., 2008).

Activity in bilateral STG, which spread both anteriorly and posteriorly, was associated with informational masking when non-speech energetic masking effects were mostly accounted for. This result replicates the findings of Scott et al. (2009) using fMRI rather than PET, and helps to corroborate the authenticity of both sets of results. This suggests that non-attended speech, once the energetic masking effects have been accounted for, engages a similar processing stream as attended speech. This should however be tempered by observations concerning the limitations of the energetic masking stimulus; speech modulated noise accounts for the averaged spectrum of speech but does not contain spectral dynamics or a pitch percept. Further work addressing these limitations is suggested for the future. Interestingly the informational aspect of masking did not engage areas outside regions traditionally associated with sensory/linguistic functions. This was despite the fact that two sisters were used as masker and target with the aim of increasing the informational aspect of the masking

effect. By including a clear speech baseline it was possible to demonstrate that the energetic component of informational masking was likely to engage a wider prefrontal, parietal and cingulate network, something which could not be established in the Scott et al. (2009) which was restricted to having fewer stimulus conditions because of the imaging methodology used. One explanation for our results is that similar levels of attention, cognitive control and stream segregation are required when processing speech and non-speech maskers, and that the informational aspect of speech masking places greatest emphasis on extracting the phonemes of the target speech stream. This does concur with behavioural findings that show that subjects sometimes report the masker rather than the target during speech masking (Brungart, 2001) which was also noted in the behavioural experiment of this study. It is acknowledged however that it is impossible to rule out the possibility that a lack of power may explain the failure to see activation in the wider network, however the fact that reducing the statistical threshold did not change the statistical map argues against this.

Whilst this study succeeded in replicating the same effect for [Con+Dis - SMN] in Scott et al. (2009), the same right lateralisation found in Scott et al. (2009) was not found when [Rot - SMN] was examined. Instead strong bilateral activation was found with a slightly larger cluster extent and Z values in the left. It is difficult to explain this discrepancy between the studies. One factor that may be important is the fact that the rotated speech was filtered to have the same long term average spectrum as the original sentences derived from the masking speaker. This was in contrast to the filtering conducted in the original study which used generic measurements made by Byrne et al. (1994). It is possible that small differences in the spectrum of the rotated speech compared to the spectrum of the SMN (which was derived from the masked speaker in the original study) drove the right hemisphere lateralisation in that study, although this seems unlikely. Perhaps more likely it reflects the greater statistical power in this study afforded by a larger number of subjects. The fact that [Con+Dis - SMN] and [Rot - SMN] gave rise to similar patterns of activation might suggest that the informational component of masking arises at the level of phonetic competition irrespective of intelligibility.

When the continuous and discontinuous maskers were contrasted we found activations in the superior and inferior parietal and paracentral lobule and the middle cingulate. The fact that the activated regions were located beyond primary and secondary auditory cortices suggest that significant acoustic confounds were not introduced when these two stimulus conditions were constructed. One speculative suggestion is that the neural activation identified in this contrast reflects modulation of attention; specifically the continuous narrative might cause greater attentional capture due to its coherence. In a visual search task, a unique stimulus that differs along a non-relevant dimension which is not the target item can be shown to "capture the attention" of searcher increasing the search time. Watkins et al. (2007) recently demonstrated this effect using simple auditory stimuli and showed that the parietal cortex was implicated in this "attentional capture".

6.6 CHAPTER CONCLUSION

In this chapter the intelligibility of speech was degraded by presenting natural speech concurrently with other competing speech and non-speech sounds. Speech masking was shown to engage a network including bilateral temporal, parietal, prefrontal and cingulate cortex. Subjects who performed well on post scanning speech in noise tasks, activated the left mid-posterior STG more during energetic masking, and this same region in addition to the left IFG during informational masking. The region of left mid-posterior STG was spatially separate from a region that responded more to clear speech than to masked speech, which was found in bilateral anterior and left posterior STS. This suggests partially dissociable neural systems involved in the "process" as contrasted with the "product" of resolving speech. Masking speech with other speech (and also with non-speech signals with phonetic content, i.e. rotated speech) activated the bilateral mid-anterior temporal lobes more than masking with a non-speech stimuli without the same phonetic content. Finally masking speech with other speech (and also with non-speech signals with phonetic content, i.e. rotated speech) activated the bilateral STG more than masking with a non-speech stimuli without the same phonetic

content, suggestive that unattended speech is processed in the same processing stream as attended speech.

Chapter 7 : CONCLUSIONS

7.1 Summary of aims

The central aim of this thesis was the characterisation of neural responses to speech intelligibility, with a particular emphasis on differentiating the roles of bilateral anterior and posterior temporal cortex. The following two questions have been addressed:

1) Where are neural responses to intelligible, and intelligible but degraded speech, separated from responses to acoustic complexity?
2) Are the resulting patterns of lateralisation driven by the acoustic or linguistic properties of speech?

The following sections review the novel experimental findings of this thesis.

7.2 Which regions of the temporal lobes respond most selectively to intelligible speech?

There is an ongoing debate as to whether intelligible speech is first resolved in bilateral posterior or left anterior STS (Rauschecker and Scott, 2009;Hickok and Poeppel, 2007). The results of this thesis provide partial support for both models and thus may help to reconcile the opposing views.

A recurrent finding across the studies within this thesis was that intelligible speech, whether fully or partially intelligible, preferentially recruited the left anterior STS. The left anterior STS/STG

showed the largest peak level responses across all studies (Ch 3-6) and was most predictive in coding for intelligible speech (Ch 3). This is in keeping with previous human functional imaging studies which have emphasised the importance of left anterior STS in responding to intelligible speech (Scott et al., 2000;Narain et al., 2003;Scott et al., 2006). It is also consistent with the neurophysiology of the macaque, in which con-specific vocalisations have been found to engage an anterior stream with a left hemisphere bias that begins in core regions and selectively engages rostral belt, parabelt and the temporal pole (Tian et al., 2001;Poremba et al., 2004).

These findings are also in accord with the observation that the anterior temporal lobe is integral to semantic memory. Patterson et al. (2007) have argued that the anterior temporal lobe operates as a semantic hub integrating conceptual knowledge across modalities, which is consistent with the observation that relatively selective degeneration of the anterior temporal lobes is a defining feature of semantic dementia. If the anterior temporal lobe does operate as a semantic hub, then responses found to intelligible speech in the anterior STS would be ideally situated to interface with this hub. It may be the case that activation would have been found further anteriorly if another imaging modality had been used, as the signal to noise ratio of the BOLD signal reduces with proximity to the air filled sinuses of the anterior temporal lobe (Devlin et al., 2000). Future imaging studies using techniques less affected by susceptibility artefact may help to fully map the responses to intelligibility in the temporal pole (Wang et al., 2004).

Additional responses to intelligibility were also found in the left posterior and the right anterior STS, but these responses were not as consistent. That is in Chapter 3, the right anterior and left posterior STS responses were only identified with uncorrected thresholds or when examining the main effect of intelligibility whilst masking out the interaction. In all cases when responses were found outside the left anterior STS, the peaks were largest in the left anterior STS (Ch 3-6). Responses were found in additional areas in the instance when speech was intelligible but degraded (Ch 4) and when responses to clear as contrasted with less intelligible masked speech were examined (Ch 6).

A degree of functional disassociation was shown between the mid-posterior lateral STG and the STS. In Chapter 6 it was shown that subjects who performed well at masking tasks activated the left mid-posterior STG more when listening to masked speech, whereas the adjacent STS responded more to clear speech than to less intelligible speech. This suggests that there might be a degree of disassociation between regions involved in the acoustic-phonetic processing involved in resolving intelligible speech (the process) and regions responding to the resolved intelligible percept (the product). In this thesis the left posterior STS was associated with both those roles. In Chapter 3 a region of the left posterior STS: [-63 -35 2] showed no differentiation in its response to clear, rot and NV speech. This was very close to a peak showing a similar response in Scott et al. (2000). As left posterior STS responded alike to intelligible and unintelligible speech that contained phonetic cues, it may implicate this region as playing a greater acoustic-phonetic processing role relative to anterior regions. This was also suggested by the DCM analysis which implicated reciprocal connections between left anterior and left posterior STS, suggestive that left anterior STS was involved in retuning acoustic-phonetic processes within posterior STS (Ch 5). Note that the centre coordinate for the group averaged VOI in the DCM: [-60 -37 7], was within 5 mm in any one direction from the peak identified as responding to rotated speech in Chapter 3. These findings therefore posit a specific role for the left posterior STS in the acoustic-phonetic processing required to resolve intelligible speech.

It is unclear whether the reciprocal connections identified using DCM are characteristic of a response to clear speech or whether this pattern of connections was only observed as a function of the degradation of the speech signal. Indeed Leff et al. (2008) did not observe backward connections between anterior and posterior STS when they contrasted clear speech with reversed speech. By way of counter point, predictive coding frameworks would argue that backward connections are an essential part of veridical perception. Further experiments addressing these issues would be pertinent in the light of ongoing debates in the behavioural literature concerning whether higher level information can directly influence prelexical processing (Norris et al., 2000;McClelland et al., 2006). Indeed it would be interesting to return to the data in Chapter 3 and carry out a DCM analysis to understand whether equivalent results to Leff et al. would be found using rotated rather than reversed

speech. Further, using the data in Chapter 5 it would be interesting to understand how the connectivity parameters between anterior and posterior STS related to each individuals' ability to understand degraded speech. A future experiment, in which access to higher level linguistic information and the amount of signal degradation were co-varied within the context of a DCM analysis, could provide more direct supportive evidence for the selective retuning of acoustic-phonetic processes by anterior STS (cf. Davis et al. 2011; Obleser et al. 2007b).

The finding of acoustic-phonetic sensitivity in the left posterior STS does not necessarily suggest that this region has no involvement in responding to fully resolved intelligible speech. Indeed posterior STS was shown to respond to intelligible speech when the statistical criterion was reduced in Chapter 3 and when clear speech was contrasted with masked speech in Chapter 6. It may be that either, different sub regions of the posterior STS are involved in the process versus the product of resolving intelligibility, or that there are gradients of response across the temporal lobes (cf. Davis and Johnsrude et al. (2003)) with the peak of the gradient for responding to resolved intelligible speech found in left anterior STS and the peak of the gradient for acoustic-phonetic processing found in mid-posterior lateral STG. In support of this we show in Chapter 3 that the most selective classifier weights coding for unintelligible sounds were located closest to mid-posterior lateral STG/PT, and those for intelligible speech were found closest to the left anterior STS, with these weights becoming less well spatially separated as the selectivity of the weight banding was reduced (Figure 3.16). This might place left posterior STS as falling somewhere in the middle of these two gradients explaining why this region is capable to some extent of performing both functions and why the effect size in posterior regions is weaker than the anterior.

It remains to be seen how the intelligibility/acoustic-phonetic sensitivities of anterior and posterior STS relate to the anterior and posterior processing pathways. Evidence from the macaque suggest that the anterior and posterior pathways are largely separate and parallel (Romanski and Averbeck, 2009), although note that Tian et al. (2001) found relatively greater, rather than absolute, selectivity in rostral versus caudal belt for type of call versus its location. Upadhyay et al. (2008) showed using structural connectivity similar separate anterior and posterior pathways in human

subjects, with one pathway projecting from rostral HG to anterior STG, and the other from caudal HG to posterior STG. The connectivity analysis in Chapter 5 suggests that anterior and posterior STS directly communicated with one another. It seems likely given the evidence of largely separate processing streams that communication between these regions was predominantly facilitated by the feedforward and feedback connections within the STS that have been identified in the macaque (Seltzer and Pandya, 1989). The degree of communication between these regions may be relatively more enhanced in circumstances when speech is degraded, when one might imagine that the demands placed on working memory and sensori-motor are increased. This may explain why we identified strong activation in posterior temporal cortex when speech was degraded but not so consistently when speech was fully intelligible.

The results of this thesis may shed some light on why damage to left posterior temporal cortex causes such pervasive impairments in speech perception in individuals with aphasia. In this thesis posterior STS has been shown to evidence a degree of sensitivity to intelligibility and has also been shown to play a likely role in the acoustic-phonetic processes supporting speech perception, it is thus likely that damage to this region could directly impair speech perception. In addition it may exert an indirect influence by reducing communication with left anterior STS, where the intelligibility response is strongest, and by damaging connections to the right hemisphere (note that the connection between the hemispheres was shown to be mediated by the posterior in Chapter 5). The fact that we observed an intelligibility response in the right anterior STS may explain why bilateral lesions have the most profound effect on speech perception, and why some albeit limited speech perception abilities remain after the left hemisphere is incapacitated. Note also that the intelligibility response was strongest in the left hemisphere explaining why left unilateral lesions result in more profound impairments than right sided ones.

Contrary to the theoretical position of Hickok and Poeppel, no evidence was found in any of the studies within this thesis supporting a role for right posterior STS in responding to intelligible speech. Indeed this seems in keeping with most studies which have contrasted intelligible speech with unintelligible sounds (Awad et al., 2007;Spitsyna et al., 2006;Scott et al., 2000;Narain et al.,

2003;Friederici et al., 2010;Obleser et al., 2007b;Scott et al., 2006). Indeed Okada et al. (2010) were equally unable to find convincing evidence for it playing a role in responding to intelligible speech.

7.3 Is lateralisation to speech driven by its acoustic or linguistic properties?

It has been argued that bilateral responses to intelligible speech are driven by the acoustics of the speech signal, with the left hemisphere showing a preferential response to information varying over short time intervals and the right to information evolving over longer intervals or to spectral information more generally (Poeppel, 2003;Zatorre and Gandour, 2008).

In Chapter 4, four unintelligible stimuli were generated in which speech derived modulations of spectrum and amplitude were absent, applied singly or in combination (with modulations taken from different sentences to maintain an unintelligible percept), in addition a fifth intelligible condition was created in which the modulations were taken from the same sentence. Previous demonstrations of hemispheric lateralisation to spectral and temporal processing have used simple non-speech stimuli and have often used non-independently defined regions of interest of arbitrarily size and shape. In this thesis a pattern classification approach was conducted using a SVM that allowed the use of large anatomical ROIs with a high degree of sensitivity to experimental effects.

It was shown using this approach that the response of the left as compared to the right hemisphere was more successful in classifying intelligible speech from unintelligible sounds. Furthermore the most discriminant voxels coding for a response to intelligible speech were found in the left rather than the right hemisphere, and vice versa for voxels coding for unintelligible speech. The same evidence of lateralisation could not be found for the acoustic manipulations of spectrum and amplitude, from which it was concluded that it was most likely that the left lateralisation evidenced for speech was driven by access to linguistic representations rather than acoustic features. These results may have implications for theories of dyslexia and specific language impairment which posit a

selective impairment in left hemisphere structures associated with temporal auditory processing (Vandermosten et al., 2010).

A DCM analysis was conducted examining the connectivity between bilateral anterior and posterior temporal cortex in order to understand whether the observed lateralisation for intelligible speech was meaningful (Ch 5). This analysis showed that the left hemisphere preference for intelligible speech identified in the previous chapter was also evident in the effective connectivity between the hemispheres, with responses in the left hemisphere shown to drive the response in the right.

7.4 Are responses to degraded speech different to responses to fully intelligible speech, and does the type of degradation affect the response seen?

In Chapter 3 clear speech was contrasted with unintelligible sounds, whilst in Chapter 4 degraded but intelligible speech was contrasted with unintelligible sounds. Degraded speech activated more extensive regions of the bilateral temporal lobes with activation spreading across posterior and anterior STG and STS. This was in contrast to the more constrained region of the STS shown to be activated by clear speech in Chapter 3. This is likely to reflect the additional acoustic-phonetic and working memory demands associated with degraded signals. However, it wasn't just in the temporal lobes where differences were observed. In addition, signal degradation activated much more extensive regions of frontal cortex in keeping with findings showing prefrontal activity associated with increased listening effort (Davis and Johnsrude, 2003). Further work would be required to differentiate the response within these regions; possible cognitive correlates include activation related to sensori-motor integration, sub vocal articulation and working memory processes supporting comprehension.

In Chapter 4 speech was degraded by creating a reduced representation of the speech signal, whilst in Chapter 6 natural speech was degraded by presenting concurrent competing sounds.

Comparing these two studies, a number of shared areas of activation were observed in temporal and prefrontal cortex. A striking finding however was the additional recruitment of bilateral parietal cortex in responding selectively to masked speech. This seems likely to be explained by the requirement to attend to a target speaker within a mixture of sounds. It seems that both types of signal degradation engaged both the anterior "what" and posterior "how" streams strongly, evidenced by the parietal activity during speech masking (Ch 4) and the posterior frontal activation in the case of the two formant speech (Ch 6). This was in contrast to clear speech which was only shown to strongly engage the anterior stream, evidenced by activation solely within anterior temporal regions (Ch 3).

In Chapter 6 it was demonstrated that masking with different types of sound affected the activation patterns observed. Masking speech with other speech (and also with non-speech signals with phonetic content, i.e. rotated speech) activated the bilateral STG more than masking with a non-speech stimuli without the same phonetic content. This suggested that unattended speech is processed within the same neural stream as attended speech, likely placing greater demands on acoustic-phonetic processing. Interestingly masking with speech did not engage areas supplementary to the auditory cortices more than masking with non-speech sounds. Although subjects who performed well at masking tasks did tend to activate the IFG (as well as mid-posterior STG) during masking with speech, compared to only the left mid-posterior STG during masking with non-speech. One explanation for this finding is that activating the IFG more increases the efficiency with which individuals can deal with the semantic competition between the target and masker. A future study contrasting neural responses in the instance when the target and maskers vary in their degree of semantic competition may help to confirm this interpretation.

The results in this thesis, in associating increased activation during speech masking within mid-posterior STG and the IFG with individuals who perform better in speech in noise tasks is in accord with the findings of Wong et al. (2009) who showed that older subjects without hearing loss (who performed worse than younger controls in masking tasks) activated regions of prefrontal cortex more and the temporal cortex less during masking tasks. Further to this, the increases in activation in prefrontal cortex in the older group were associated with better behavioural performance suggesting

some kind of compensatory role. Older subjects have been shown to evidence greater difficulty in listening to speech in noise than might be expected given their pure tone thresholds (Helfer et al., 2010), as do individuals with hearing loss (Darwin, 2008). Having a specific difficulty in listening to speech in noisy environments can be socially isolating, further work is required in differentiating the factors which contribute to this difficulty in hearing impaired and elderly groups. A framework which identifies the informational and energetic components of masking is well suited to identifying the neural correlates associated with better comprehension of speech in noise. In the long term it may be possible to identify neural interventions aimed at improving speech in noise abilities in these groups; by separating the energetic effects, which may prove more resistant to change as they arise from the auditory periphery, from the informational effects, might allow more effective targeting of interventions.

7.5 Summary of key findings

- The left anterior STS was shown to be the key region involved in responding to intelligible and degraded speech.
- The right anterior and left posterior STS showed a less consistent and reduced level of response to intelligible speech relative to the left anterior STS.
- The right posterior STS did not show any sensitivity to intelligible speech.
- Unlike bilateral anterior areas, left posterior STS showed an additional acoustic-phonetic sensitivity irrespective of intelligibility.
- The relative left lateralisation for intelligible speech was shown to be more likely to be driven by its linguistic rather than acoustic properties.
- In the case of degraded speech, left anterior STS was shown to be reciprocally connected to the left posterior STS suggestive of a retuning of lower level acoustic-phonetic processes by higher level linguistic information.

- Degraded speech was shown to activate a bilateral fronto-temporal network; masking speech with concurrent sounds additionally recruited the parietal cortices.
- Individuals who performed well at making tasks, showed increased activation of the left mid-posterior lateral STG when engaged in non-speech masking, and both mid-posterior lateral STG and the IFG when engaged in speech masking.

Reference List

Adank P, Devlin JT (2010) On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. Neuroimage 49:1124-1132.

Ahveninen J, Jaaskelainen IP, Raij T, Bonmassar G, Devore S, Hamalainen M, Levanen S, Lin FH, Sams M, Shinn-Cunningham BG, Witzel T, Belliveau JW (2006) Task-modulated "what" and "where" pathways in human auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 103:14608-14613.

Awad M, Warren JE, Scott SK, Turkheimer FE, Wise RJS (2007) A common system for the comprehension and production of narrative speech. Journal of Neuroscience 27:11455-11464.

Bedny M, McGill M, Thompson-Schill SL (2008) Semantic Adaptation and Competition during Word Comprehension. Cereb Cortex 18:2574-2585.

Belin P, Zilbovicius M, Crozier S, Thivard L, Fontaine A (1998) Lateralization of speech and auditory temporal processing. Journal of Cognitive Neuroscience 10:536-540.

Bench J, Kowal A, Bamford J (1979) The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. British Journal of Audiology 13:108-112.

Benson RR, Whalen DH, Richardson M, Swainson B, Clark VP, Lai S, Liberman AM (2001) Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain and Language 78:364-396.

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cereb Cortex 19:2767-2796.

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.

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD (2004) Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience 7:295-301.

Blesser B (1972) Speech Perception Under Conditions of Spectral Transformation .1. Phonetic Characteristics. Journal of Speech and Hearing Research 15:5-&.

Blumstein SE, Baker E, Goodglass H (1977) Phonological Factors in Auditory Comprehension in Aphasia. Neuropsychologia 15:19-30.

Blumstein SE, Myers EB, Rissman J (2005) The perception of voice onset time: An fMRI investigation of phonetic category structure. Journal of Cognitive Neuroscience 17:1353-1366.

Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nature Neuroscience 8:389-395.

Boulenger V, Hoen M, Ferragne E, Pellegrino F, Meunier F (2010) Real-time lexical competitions during speech-in-speech comprehension. Speech Communication 52:246-253.

Brett M, Anton JL, Valabregue R, Poline JB (2002) Region of interest analysis using an SPM toolbox. In: 8th International Conference on Functional Mapping of the Human Brain.

Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. Journal of the Acoustical Society of America 109:1101-1109.

Buchsbaum BR, D'Esposito M (2008) The search for the phonological store: From loop to convolution. Journal of Cognitive Neuroscience 20:762-778.

Byrne D, et al. (1994) An International Comparison of Long-Term Average Speech Spectra. Journal of the Acoustical Society of America 96:2108-2120.

Carlson TA, Schrater P, He S (2003) Patterns of activity in the categorical representations of objects. Journal of Cognitive Neuroscience 15:704-717.

Cherry EC (1953) Some Experiments on the Recognition of Speech, with One and with 2 Ears. Journal of the Acoustical Society of America 25:975-979.

Collett (2003) Modelling Binary Data. Florida: Chapman & Hall/CRC.

Cooke M (2006) A glimpsing model of speech perception in noise. Journal of the Acoustical Society of America 119:1562-1573.

Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261-270.

Crinion JT, Warburton EA, Lambon-Ralph MA, Howard D, Wise RJS (2006) Listening to narrative speech after aphasic stroke: The role of the left anterior temporal lobe. Cereb Cortex 16:1116-1125.

Cusack R (2005) The intraparietal sulcus and perceptual organization. Journal of Cognitive Neuroscience 17:641-651.

Darwin CJ (2008) Listening to speech in the presence of other sounds. Philosophical Transactions of the Royal Society B-Biological Sciences 363:1011-1021.

David O, Maess B, Eckstein K, Friederici AD (2011) Dynamic Causal Modeling of Subcortical Connectivity of Language. Journal of Neuroscience 31:2712-2717.

Davis M, Ford M, Kherif F, Johnsrude I (2011) Does Semantic Context Benefit Speech Understanding through "Top Down" Processes? Evidence from Time-Resolved Sparse fMRI. Journal of Cognitive Neuroscience.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. Journal of Neuroscience 23:3423-3431.

Davis MH, Johnsrude IS (2007) Hearing speech sounds: Top-down influences on the interface between audition and speech perception. Hearing Research 229:132-147.

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C (2005) Lexical information drives; Perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology-General 134:222-241.

de Zubicaray G, McMahon K, Eastburn M, Pringle AJ, Lorenz L, Humphreys MS (2007) Support for an auto-associative model of spoken cued recall: Evidence from fMRI. Neuropsychologia 45:824-835.

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

Della-Maggiore V, Chan W, Peres-Neto PR, McIntosh AR (2002) An empirical comparison of SPM preprocessing parameters to the analysis of fMRI data. Neuroimage 17:19-28.

Demonet JF, Chollet F, Ramsay S, Cardebat D, Nespoulous JL, Wise R, Rascol A, Frackowiak R (1992) The Anatomy of Phonological and Semantic Processing in Normal Subjects. Brain 115:1753-1768.

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1-30.

Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53:1-15.

Devlin JT, Russell RP, Davis MH, Price CJ, Wilson J, Moss HE, Matthews PM, Tyler LK (2000) Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task. Neuroimage 11:589-600.

DeWitt I, Rauschecker JP (2010) Meta-analysis of acoustic-phonetic processing: Evidence of a ventral stream hierarchy. In: Organization for Human Brain Mapping.

Downar J, Crawley AP, Mikulis DJ, Davis KD (2002) A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. Journal of Neurophysiology 87:615-620.

Dronkers NF (1996) A new brain region for coordinating speech articulation. Nature 384:159-161.

Dronkers NF, Wilkins DP, Van Valin RD, Redfern BB, Jaeger JJ (2004) Lesion analysis of the brain areas involved in language comprehension. Cognition 92:145-177.

Duncan J, Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends in Neurosciences 23:475-483.

Dupoux E, Green K (1997) Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. Journal of Experimental Psychology-Human Perception and Performance 23:914-927.

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

Efron R (1963) Temporal Perception, Aphasia and Deja Vu. Brain 86:403-&.

Eisner F, McGettigan C, Faulkner A, Rosen S, Scott SK (2010) Inferior Frontal Gyrus Activation Predicts Individual Differences in Perceptual Learning of Cochlear-Implant Simulations. Journal of Neuroscience 30:7179-7186.

Etzel JA, Gazzola V, Keysers C (2009) An introduction to anatomical ROI-based fMRI classification analysis. Brain Research 1282:114-125.

Etzel JA, Valchev N, Keysers C (2011) The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. Neuroimage 54:1159-1167.

Falkenberg LE, Specht K, Westerhausen R (2011) Attention and cognitive control networks assessed in a dichotic listening fMRI study. Brain and Cognition 76:276-285.

Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40:859-869.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science 322:970-973.

Friederici AD, Kotz SA, Scott SK, Obleser J (2010) Disentangling Syntax and Intelligibility in Auditory Language Comprehension. Human Brain Mapping 31:448-457.

Friederici AD, Meyer M, von Cramon DY (2000) Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. Brain and Language 74:289-300.

Friston K (2007) Dynamic Causal Models for fMRI. In: Statistical Parametric Mapping: The Analysis of Functional Brain Images (Friston K, Ashburner J, Kiebel SJ, Nichols T, Penny W, eds), London: Elsevier.

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B-Biological Sciences 364:1211-1221.

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. Neuroimage 19:1273-1302.

Friston KJ, Penny WD, Glaser DE (2005) Conjunction revisited. Neuroimage 25:661-667.

Friston KJ, Stephan KE (2007) Free-energy and the brain. Synthese 159:417-458.

Ganong WF (1980) Phonetic Categorization in Auditory Word Perception. Journal of Experimental Psychology-Human Perception and Performance 6:110-125.

Golland P, Fischl B (2003) Permutation tests for classification: Towards statistical significance in image-based studies.

Greenwood DD (1990) A Cochlear Frequency-Position Function for Several Species - 29 Years Later. Journal of the Acoustical Society of America 87:2592-2605.

Griffiths TD, Warren JD (2002) The planum temporale as a computational hub. Trends in Neurosciences 25:348-353.

Grill-Spector K, Malach R (2001) fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychologica 107:293-321.

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46:389-422.

Hackett TA (2011) Information flow in the auditory cortical network. Hearing Research In Press, Corrected Proof.

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

Hall DA, Johnsrude IS, Haggard MP, Palmer AR, Akeroyd MA, Summerfield AQ (2002) Spectral and Temporal Processing in Human Auditory Cortex. Cereb Cortex 12:140-149.

Hart HC, Palmer AR, Hall DA (2003) Amplitude and frequency-modulated stimuli activate common regions of human auditory cortex. Cereb Cortex 13:773-781.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7:523-534.

Helfer KS, Chevalier J, Freyman RL (2010) Aging, spatial cues, and single- versus dual-task performance in competing speech perception. Journal of the Acoustical Society of America 128:3625-3633.

Hickok G (2009) The functional neuroanatomy of language. Physics of Life Reviews 6:121-143.

Hickok G, Poeppel D (2004) Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92:67-99.

Hickok G, Poeppel D (2007) Opinion - The cortical organization of speech processing. Nature Reviews Neuroscience 8:393-402.

Jacquemot C, Scott SK (2006) What is the relationship between phonological short-term memory and speech processing? Trends Cogn Sci 10:480-486.

Jamison HL, Watkins KE, Bishop DVM, Matthews PM (2006) Hemispheric specialization for processing auditory nonspeech stimuli. Cereb Cortex 16:1266-1275.

Joanisse MF, Zevin JD, McCandliss BD (2007) Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. Cereb Cortex 17:2084-2093.

Josse G, Seghier ML, Kherif F, Price CJ (2008) Explaining Function with Anatomy: Language Lateralization and Corpus Callosum Size. Journal of Neuroscience 28:14132-14139.

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proceedings of the National Academy of Sciences of the United States of America 97:11793-11799.

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8:679-685.

Kass RE, Raftery AE (1995) Bayes Factors. Journal of the American Statistical Association 90:773-795.

Kidd G, Best V, Mason CR (2008) Listening to every other word: Examining the strength of linkage variables in forming streams of speech. Journal of the Acoustical Society of America 124:3793-3802.

King AJ, Nelken I (2009) Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat Neurosci 12:698-701.

Kluender KR, Alexander JM (2010) Perception of Speech Sounds. In: The Senses: A Comprehensive Reference (Basbaum AI, Kaneko A, Shephard GM, Westheimer G, Albright TD, Masland RH, Dallos P, Oertel D, Firestein S, Beauchamp GK, Bushnell C, Kass JH, Gardner E, eds), Elsevier.

Kluender KR, Diehl RL, Killeen PR (1987) Japanese-Quail Can Learn Phonetic Categories. Science 237:1195-1197.

Kouider S, Dupoux E (2005) Subliminal speech priming. Psychological Science 16:617-625.

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103:3863-3868.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 12:535-540.

Ku SP, Gretton A, Macke J, Logothetis NK (2008) Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys. Magnetic Resonance Imaging 26:1007-1014.

Kumar S, Stephan KE, Warren JD, Friston KJ, Griffiths TD (2007) Hierarchical processing of auditory objects in humans. Plos Computational Biology 3:977-985.

Leech R, Holt LL, Devlin JT, Dick F (2009) Expertise with Artificial Nonspeech Sounds Recruits Speech-Sensitive Cortical Regions. Journal of Neuroscience 29:5234-5239.

Leff AP, Schofield TM, Stephan KE, Crinion JT, Friston KJ, Price CJ (2008) The Cortical Dynamics of Intelligible Speech. Journal of Neuroscience 28:13209-13215.

Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The Discrimination of Speech Sounds Within and Across Phoneme Boundaries. Journal of Experimental Psychology 54:358-368.

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. Cereb Cortex 15:1621-1631.

Lisker L (1977) Rapid Versus Rabid - Catalog of Acoustic Features That May Cue Distinction. Journal of the Acoustical Society of America 62:S77-S78.

Mattys SL, Carroll LM, Li CKW, Chan SLY (2010) Effects of energetic and informational masking on speech segmentation by native and non-native speakers. Speech Communication 52:887-899.

McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? Trends Cogn Sci 10:363-369.

Miller GA, Heise GA, Lichten W (1951) The Intelligibility of Speech As A Function of the Context of the Test Materials. Journal of Experimental Psychology 41:329-335.

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53:103-118.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang XR, Just M, Newman S (2004) Learning to decode cognitive states from brain images. Machine Learning 57:145-175.

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.

Mourao-Miranda J, Bokde ALW, Born C, Hampel H, Stetter M (2005) Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage 28:980-995.

Mourao-Miranda J, Reynaud E, McGlone F, Calvert G, Brammer M (2006) The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMFI data. Neuroimage 33:1055-1065.

Mummery CJ, Ashburner J, Scott SK, Wise RJS (1999) Functional neuroimaging of speech perception in six normal and two aphasic subjects. Journal of the Acoustical Society of America 106:449-457.

Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRIan introductory guide. Social Cognitive and Affective Neuroscience 4:101-109.

Myers EB, Blumstein SE, Walsh E, Eliassen J (2009) Inferior Frontal Regions Underlie the Perception of Phonetic Category Invariance. Psychological Science 20:895-903.

Narain C, Scott SK, Wise RJS, Rosen S, Leff A, Iversen SD, Matthews PM (2003) Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb Cortex 13:1362-1368.

Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653-660.

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences 23:299-+.

Norris D, McQueen JM, Cutler A (2003) Perceptual learning in speech. Cognitive Psychology 47:204-238.

O'Toole AJ, Jiang F, Abdi H, Penard N, Dunlop JP, Parent MA (2007) Theoretical, statistical, and practical perspectives on pattern-based classification amoroaches to the analysis of functional neuroimaging data. Journal of Cognitive Neuroscience 19:1735-1752.

Obleser J, Eisner F, Kotz SA (2008) Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. Journal of Neuroscience 28:8116-8123.

Obleser J, Kotz SA (2010) Expectancy Constraints in Degraded Speech Modulate the Language Comprehension Network. Cereb Cortex 20:633-640.

Obleser J, Wise RJS, Dresner MA, Scott SK (2007a) Functional integration across brain regions improves speech perception under adverse listening conditions. Journal of Neuroscience 27:2283-2289.

Obleser J, Zimmermann J, Van Meter J, Rauschecker JP (2007b) Multiple stages of auditory speech perception reflected in event-related fMRI. Cereb Cortex 17:2251-2257.

Obleser JLAVJRJ (2010) Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Frontiers in Psychology 1:232.

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G (2010) Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. Cereb Cortexbhp318.

Pandya DN, Hallett M, Mukherje SK (1969) Intra-Hemispheric and Interhemispheric Connections of Neocortical Auditory System in Rhesus Monkey. Brain Research 14:49-&.

Patterson K, Nestor PJ, Rogers TT (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8:976-987.

Patterson RD, Johnsrude IS (2008) Functional imaging of the auditory processing applied to speech sounds. Philosophical Transactions of the Royal Society B-Biological Sciences 363:1023-1035.

Patterson RD, Uppenkamp S, Johnsrude IS, Griffiths TD (2002) The processing of temporal pitch and melody information in auditory cortex. Neuron 36:767-776.

Penhune VB, Zatorre RJ, MacDonald JD, Evans AC (1996) Interhemispheric anatomical differences in human primary auditory cortex: Probabilistic mapping and volume measurement from magnetic resonance scans. Cereb Cortex 6:661-672.

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP (2010) Comparing Families of Dynamic Causal Models. Plos Computational Biology 6.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45:S199-S209.

Pisoni DB (1977) Identification and Discrimination of Relative Onset Time of 2 Component Tones - Implications for Voicing Perception in Stops. Journal of the Acoustical Society of America 61:1352-1361.

Pobric G, Jefferies E, Ralph MAL (2010) Amodal semantic representations depend on both anterior temporal lobes: Evidence from repetitive transcranial magnetic stimulation. Neuropsychologia 48:1336-1342.

Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Communication 41:245-255.

Poeppel D, Hickok G (2004) Towards a new functional anatomy of language. Cognition 92:1-12.

Poldrack RA (2007) Region of interest analysis for fMRI. Social Cognitive and Affective Neuroscience 2:67-70.

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. Journal of Cognitive Neuroscience 13:687-697.

Poremba A, Malloy M, Saunders RC, Carson RE, Herscovitch P, Mishkin M (2004) Species-specific calls evoke asymmetric activity in the monkey's temporal poles. Nature 427:448-451.

Price CJ, Friston KJ (1997) Cognitive conjunction: A new approach to brain activation experiments. Neuroimage 5:261-270.

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.

Ramus F, Rosen S, Dakin SC, Day BL, Castellote JM, White S, Frith U (2003) Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. Brain 126:841-865.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2:79-87.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience 12:718-724.

Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981) Speech-Perception Without Traditional Speech Cues. Science 212:947-950.

Rhebergen KS, Versfeld NJ, Dreschler WA (2005) Release from informational masking by time reversal of native and non-native interfering speech (L). Journal of the Acoustical Society of America 118:1274-1277.

Rivier F, Clarke S (1997) Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: Evidence for multiple auditory areas. Neuroimage 6:288-304.

Romanski LM, Averbeck BB (2009) The Primate Cortical Auditory System and Neural Representation of Conspecific Vocalizations. Annual Review of Neuroscience 32:315-346.

Rosen S (2003) Auditory processing in dyslexia and specific language impairment: is there a deficit? What is its nature? Does it explain anything? Journal of Phonetics 31:509-527.

Sato JR, Fujita A, Thomaz CE, Martin MDM, Mourao-Miranda J, Brammer MJ, Amaro E (2009) Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction. Neuroimage 46:105-114.

Schmidt CF, Zaehle T, Meyer M, Geiser E, Boesiger P, Jancke L (2008) Silent and continuous fMRI scanning differentially modulate activation in an auditory language comprehension task. Human Brain Mapping 29:46-56.

Schonwiesner MRRvC (2005) Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. [References]. European journal of neuroscience 22:1521-1528.

Schouten B, Gerrits E, van Hessen A (2003) The end of categorical perception as we know it. Speech Communication 41:71-80.

Scott SK, Evans S (2010) Categorizing speech. Nature Neuroscience 13:1304-1306.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 26:100-107.

Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJS (2009) The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. Journal of the Acoustical Society of America 125:1737-1743.

Scott SK, Rosen S, Lang H, Wise RJS (2006) Neural correlates of intelligibility in speech investigated with noise vocoded speech- A positron emission tomography study. Journal of the Acoustical Society of America 120:1075-1083.

Scott SK, Rosen S, Wickham L, Wise RJS (2004) A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. Journal of the Acoustical Society of America 115:813-821.

Scott SK, Wise RJS (2004) The functional neuroanatomy of prelexical processing in speech perception. Cognition 92:13-45.

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.

Seghier ML, Fagan E, Price CJ (2010) Functional Subdivisions in the Left Angular Gyrus Where the Semantic System Meets and Diverges from the Default Network. Journal of Neuroscience 30:16809-16817.

Seltzer B, Pandya DN (1989) Intrinsic Connections and Architectonics of the Superior Temporal Sulcus in the Rhesus-Monkey. Journal of Comparative Neurology 290:451-471.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech Recognition with Primarily Temporal Cues. Science 270:303-304.

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends Cogn Sci 12:182-186.

Slotnick SD, Moo LR, Segal JB, Hart J (2003) Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. Cognitive Brain Research 17:75-82.

Spitsyna G, Warren JE, Scott SK, Turkheimer FE, Wise RJS (2006) Converging language streams in the human temporal lobe. Journal of Neuroscience 26:7328-7336.

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E (2009) Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. Current Biology 19:498-502.

Stephan KE, Penny WD, Moran RJ, den Ouden HEM, Daunizeau J, Friston KJ (2010) Ten simple rules for dynamic causal modeling. Neuroimage 49:3099-3109.

Stephan KE, Weiskopf N, Drysdale PM, Robinson PA, Friston KJ (2007) Comparing hemodynamic models with DCM. Neuroimage 38:387-401.

Tallal P (1980) Auditory Temporal Perception, Phonics, and Reading Disabilities in Children. Brain and Language 9:182-198.

Tallal P, Piercy M (1973) Defects of Nonverbal Auditory-Perception in Children with Developmental Aphasia. Nature 241:468-469.

Teki S, Chait M, Kumar S, von Kriegstein K, Griffiths TD (2011) Brain Bases for Auditory Stimulus-Driven Figure-Ground Segregation. Journal of Neuroscience 31:164-171.

Theintun U (1987) Phonetic Cue Trading, Categorical Perception and the Order of Speech Processing. Speech Communication 6:353-362.

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

Tomasi D, Caparelli EC, Chang L, Ernst T (2005) fMRI-acoustic noise alters brain activation during working memory tasks. Neuroimage 27:377-386.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. Neuroimage 15:273-289.

Upadhyay J, Silver A, Knaus TA, Lindgren KA, Ducros M, Kim DS, Tager-Flusberg H (2008) Effective and structural connectivity in the human auditory cortex. Journal of Neuroscience 28:3341-3349.

Uppenkamp S, Johnsrude IS, Norris D, Marslen-Wilson W, Patterson RD (2006) Locating the initial stages of speech-sound processing in human temporal cortex. Neuroimage 31:1284-1296.

Vadlamudi L, Hatton R, Byth K, Harasty J, Vogrin S, Cook MJ, Bleasel AF (2006) Volumetric analysis of a specific language region - the planum temporale. Journal of Clinical Neuroscience 13:206-213.

van der Zwaag W, Gentile G, Gruetter R, Spierer L, Clarke S (2011) Where sound position influences sound object representations: A 7-T fMRI study. Neuroimage 54:1803-1811.

Van Engen KJ, Bradlow AR (2007) Sentence recognition in native- and foreign-language multi-talker background noise. Journal of the Acoustical Society of America 121:519-526.

Vandermosten M, Boets B, Luts H, Poelmans H, Golestani N, Wouters J, Ghesquiere P (2010) Adults with dyslexia are impaired in categorizing speech and nonspeech sounds on the basis of temporal cues. Proceedings of the National Academy of Sciences of the United States of America 107:10389-10394.

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Cognitive Brain Research 17:48-55.

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. Journal of Cognitive Neuroscience 13:994-1005.

Wang JJ, Li L, Roc AC, Alsop DC, Tang K, Butler NS, Schnall MD, Detre JA (2004) Reduced susceptibility effects in perfusion fMRI with single-shot spin-echo EPI acquisitions at 1.5 Tesla. Magnetic Resonance Imaging 22:1-7.

Warren RM (1970) Perceptual Restoration of Missing Speech Sounds. Science 167:392-&.

Warrier C, Wong P, Penhune V, Zatorre R, Parrish T, Abrams D, Kraus N (2009) Relating Structure to Function: Heschl's Gyrus and Acoustic Processing. Journal of Neuroscience 29:61-69.

Watkins S, Dalton P, Lavie N, Rees G (2007) Brain mechanisms mediating auditory attentional capture in humans. Cereb Cortex 17:1694-1700.

Wessinger CM, VanMeter J, Tian B, Van Lare J, Pekar J, Rauschecker JP (2001) Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. Journal of Cognitive Neuroscience 13:1-7.

Westbury CF, Zatorre RJ, Evans AC (1999) Quantifying variability in the planum temporale: A probability map. Cereb Cortex 9:392-405.

Wise RJS, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA (2001) Separate neural subsystems within 'Wernicke's area'. Brain 124:83-95.

Wolmetz M, Poeppel D, Rapp B (2011) What Does the Right Hemisphere Know about Phoneme Categories? Journal of Cognitive Neuroscience 23:552-569.

Wong PCM, Jin JXM, Gunasekera GM, Abel R, Lee ER, Dhar S (2009) Aging and cortical mechanisms of speech perception in noise. Neuropsychologia 47:693-703.

Wong PCM, Uppunda AK, Parrish TB, Dhar S (2008) Cortical mechanisms of speech perception in noise. Journal of Speech Language and Hearing Research 51:1026-1041.

Yamashita O, Sato M, Yoshioka T, Tong F, Kamitani Y (2008) Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. Neuroimage 42:1414-1429.

Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. Cereb Cortex 11:946-953.

Zatorre RJ, Evans AC, Meyer E, Gjedde A (1992) Lateralization of Phonetic and Pitch Discrimination in Speech Processing. Science 256:846-849.

Zatorre RJ, Gandour JT (2008) Neural specializations for speech and pitch: moving beyond the dichotomies. Philosophical Transactions of the Royal Society B-Biological Sciences 363:1087-1104.

Zekveld AA, Heslenfeld DJ, Festen JM, Schoonhoven R (2006) Top-down and bottom-up processes in speech comprehension. Neuroimage 32:1826-1836.

Zevin J, McCandliss B (2004) Dishabituation to phonetic stimuli in a 'silent' event-related fMRI design. International Journal of Psychology 39:102.

Zevin JD, Yang JF, Skipper JI, McCandliss BD (2010) Domain General Change Detection Accounts for "Dishabituation" Effects in Temporal-Parietal Regions in Functional Magnetic Resonance Imaging Studies of Speech Perception. Journal of Neuroscience 30:1110-1117.