Published in final edited form as:
Hear Res. 2013 September ; 303: 58–66. doi:10.1016/j.heares.2013.05.001.

The neural processing of masked speech

Sophie K Scott[1] and **Carolyn McGettigan**[2]

[1]Institute of Cognitive Neuroscience, UCL, 17 Queen Square, London WC1N 3AR

[2]Department of Psychology, Royal Holloway University of London, Egham, Surrey, TW20 0EX

Abstract

Spoken language is rarely heard in silence, and a great deal of interest in psychoacoustics has focused on the ways that the perception of speech is affected by properties of masking noise. In this review we first briefly outline the neuroanatomy of speech perception. We then summarise the neurobiological aspects of the perception of masked speech, and investigate this as a function of masker type, masker level and task.

1. Introduction

1.1 The Neural basis of speech perception

The perceptual processing of heard speech is associated with considerable subcortical and cortical processing, and in humans the auditory cortical fields lie in the left and right dorsolateral temporal lobes. Primary auditory cortex (PAC), which receives all its input from the auditory thalamus, lies on Heschl's gyrus, which is on the supratemporal plane - that is, the cortical surfaces of the dorsolateral temporal lobes that extend into the Sylvian fissure (Figure 1). Auditory association cortex, to which PAC projects, and which also receives inputs from the auditory thalamus, surrounds PAC on the supra temporal plane and extends along the superior temporal gyrus (STG), to the superior temporal sulcus (STS).

Historically, our understanding of the kinds of perceptual mechanisms found in auditory cortex has been driven by studies of neuropsychological patients, for example Wernicke's pioneering work outlining the effect that lesions of left STG played in deficits of the reception of speech – that is, sensory aphasia. This neuropsychological work with patients who have acquired brain damage has continued. For example, Johnsrude and colleagues (Johnsrude et al., 2000) demonstrated that patients with temporal lobe resections, both left and right, had preserved pitch perception, but patients with right temporal lobe resections had specific problems identifying the *direction* of pitch change. This work suggested strongly that right temporal lobe mechanisms were not associated with determining pitch *per se*, but were critical to aspects of the detection of structure in pitch variation. However recent developments in functional imaging have allowed us both to better characterize the extent and locations of lesions in neuropsychological investigations, and to determine the relationship(s) between structure and function in the brains of healthy adults. This has been a

Address for correspondence: sophie.scott@ucl.ac.uk.

tremendous development in the field of cognitive neuroscience generally, and for the fields of speech and hearing in particular, as it has permitted us to move beyond constructs such as ‘Broca’s area’ and ‘Wernicke’s area’ when discussing the neurobiology of speech and sound processing (Rauschecker et al., 2009; Scott et al., 2003a; Wise et al., 2001b). Though these concepts have been critical to our understanding of different profiles of aphasia, they tend to be less helpful as anatomical frameworks, not least because the boundaries of the cortical fields people would be prepared to call Wernicke’s area have widened considerably over the intervening 150 years (Wise et al., 2001b). Such variability weakens the specificity of “Wernicke’s area” as an anatomical construct.

1.2 Functional Neuroimaging

Functional neuroimaging techniques, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) allow us to try and map between structure and function in the human brain. In this review we will address the neural systems recruited during both energetic and informational masking, and also discuss the ways that these have been used to modulate intelligibility in neuroimaging studies. Energetic masking has been largely associated with the masking effects of ‘steady’ state noise, although it has recently been demonstrated that there are important contributions from random amplitude fluctuations in the noise (Stone et al., 2012; Stone et al., 2011). As these contributions have not been investigated with functional imaging, this manuscript we will refer to speech masked with noise as energetic/modulation masking. Informational masking has been associated with elements of masking which extend beyond the influence of energetic/modulation masking, for example, associated with masking speech or other complex sounds (Brungardt, 2001).

We confine the present discussion to fMRI and PET. The spatial and temporal resolution of these techniques is relatively poor, but they permit the resolution of peaks of neural activity within millimeters, which allows for localization of activity within gyral anatomy. Furthermore, PET and fMRI both permit the detection of neural activity across the whole brain, including the cerebral cortex and subcortical structures, which means that they are ideal techniques for looking at networks of activity. We note that several electrophysiological studies have investigated neural responses to speech in masking sounds (Dimitrijevic et al., in press; Parbery-Clark et al., 2011; Riecke et al., 2012; Xiang et al., 2010). These techniques have the advantage of very fine temporal analyses of neural responses, and the increasing use of combined methods will be able to exploit the benefits of both approaches.

Both PET and fMRI utilize local changes in blood flow as an index of neural activity, since where there are regional increases in neural activity, there are immediate increases in the blood supply to that region. PET measures changes in the blood flow by detecting increases in the number of gamma rays emitted following decay of a radiolabelled tracer in the blood. FMRI detects differences in the proportion of oxygenated to deoxygenated blood - the BOLD (blood oxygenation level dependent) response (Ogawa et al., 1990). Though equally sensitive across the whole brain, and silent, PET scans are limited by the half life of the tracers used, which typically take minutes to decay, and regulations regarding the total amount of radioactivity that can be delivered to any one participant. This has the effect of reducing the statistical power of PET, which is nowadays used less often for non-clinical functional scans. FMRI, in contrast, allows many more scans per condition, and thus has greater potential power. However, there can be serious signal loss in fMRI due to artefacts caused by air-bone interfaces, such as those found at the anterior temporal lobes. Signal lost in this way cannot be recovered (but ongoing work in MRI physics aims to reduce these artefactual limitations; Posse, 2012). Finally, the echo-planar imaging sequences used to acquire fMRI data generate a great deal of acoustic noise. While there is continued interest in developing silent sequences (e.g. Peelle et al., 2010), the vast majority of fMRI studies of audition either use continuous scanning (where scanner noise is always present), or employ a sparse scanning technique (Edmister et al., 1999; Hall et al., 1999). The latter takes advantage of the fact that the BOLD response builds to a peak over a lag of several seconds (4-6 seconds) after the onset of a perceptual/cognitive event. This means that it is possible to deliver acoustic stimuli during a quiet period and to time the noisy echo-planar image acquisition to occur later and coincident with the estimated peak of the bloodflow response. Importantly, the maps of activation seen in standard neuroimaging studies are made by contrasting the condition of interest with a baseline condition, meaning that we are looking at maps of activation relative to some other state of whole-brain activation. This subtractive approach means that the selection of a baseline condition can have profound implications for the pattern of observed results.

1.3 Functional neuroimaging studies of speech perception

Functional imaging has revealed considerable complexity in the ways that the human brain engages with speech and sound (Scott et al., 2003a). The perception of spoken language was addressed in some very early functional imaging studies, and there is a largely uniform finding of strong bilateral activation of the superior temporal gyri in response to human speech, relative to silence (Wise et al., 1991). Within this, primary auditory cortex has been shown to have no specific role in speech perception (e.g. Mummery et al., 1999), and the planum temporale is similarly not selective in its response to speech (Binder et al., 1996). Speech-selective responses (defined as neural activity which is significantly greater than that seen in response to a complex auditory baseline) have been described in left anterior temporal fields, lying in the STG/STS, lateral and anterior to PAC. These speech selective responses have been seen for a number of different kinds of linguistic information, from intelligible speech (Narain et al., 2003; Okada et al., 2010; Scott et al., 2000), including sentences and single words (Cohen et al., 2004; Scott et al., 2004a), to phonotactic structure (Jacquemot et al, 2003) and single phonemes (Agnew et al., 2011). This profile has been linked to the rostrally directed ‘what’ pathways described in primate neurophysiology and neuroanatomy, where cells in the anterior STS have been linked to the processing of conspecific vocalizations (Rauschecker et al., 2009; Scott et al., 2003a). It seems that the early perceptual processing of intelligible speech is associated with this same rostral ‘what’ pathway, and this may well be linked to amodal semantic representations in the anterior and ventral temporal lobe(s) (Patterson et al., 2007). In contrast, auditory fields lying posterior and medial have been associated with sensori-motor links in speech perception and production tasks: the left medial planum has been shown to respond during speech production, even if the speech is silently mouthed (i.e. there is no acoustic consequence of the articulation; see Wise et al., 2001a). This ‘how’ pathway has been linked to auditory-somatosensory responses in non-human primates, and has been suggested to form a ‘sounds do-able’ pathway whereby sounds are analyzed in terms of the actions and events that have caused them (Warren et al., 2005), perhaps with specific reference to the production of sound via the human vocal tract (Hickok et al., 2009; Hickok et al., 2003; Pa et al., 2008).

In terms of hemispheric asymmetries, there are clear differences between the left and the right temporal lobes in terms of their responses to speech and sound. The left auditory areas show a relatively specific response to linguistic information at several different levels: from phonotactic structure (Jacquemot et al., 2003) and categorical perception of syllables (Liebenthal et al., 2005), to word forms (Cohen et al., 2004) and semantics/syntax (Friederici et al., 2003). This left hemisphere dominance for spoken language may reflect purely linguistic mechanisms (McGettigan et al., 2012a) or more domain general mechanisms. For example, musicians with perfect pitch show greater activation of the left anterior STS than non-perfect pitch musicians (Schulze et al., 2009), which may suggest that left hemisphere mechanisms could be associated with categorization and expertise in auditory processing. However, whatever is driving the left dominance for language, this appears to reflect a specialization in the left temporal lobe that is not purely acoustic (McGettigan et al., 2012a). In contrast, it is relatively easy to modulate right temporal lobe activity with acoustic factors - in addition to an enhanced response to voice-like (Rosen et al., 2011) and speaker-related information (Belin et al., 2003), right auditory areas respond more strongly than the left to longer sounds (Boemio et al., 2005), and to sounds that have pitch variation (Zatorre et al., 2001). Future developments will be able to outline how these hemispheric asymmetries interact with the rostral/caudal auditory processing pathways discussed above.

2. Masking speech

Speech is often hard to understand in the presence of other sounds, and psychophysical approaches to understanding this have typically distinguished between informational and energetic/modulation components to the masking signal (e.g. Brungardt, 2001). Energetic/modulation masking effects are those associated with masking at the auditory periphery, where the masking sound competes with the target speech at the level of the cochlea. Signals that lead to the highest amounts of energetic/modulation masking are those with a wide spectrum (e.g. noise, multi-talker babble) presented at a low signal to noise (SNR) level. There is typically an ogive relationship between the intelligibility of the target speech and the SNR level. Signals leading to informational masking, in addition to their energetic/modulation masking effects, are those which cause competition for resources as a result of some higher-order property of the acoustic signal – for example, linguistic information (Dirks et al., 1969; Festen et al., 1990). An example of this would be trying to comprehend one talker against the concurrent speech of another talker. This is generally held to represent some *central* auditory processing of the masking signal, which is competing for resources with the target stimulus. This competition need not be lexical – reversed speech is an effective informational masker (Brungart et al., 2002). One feature of informational masking is the possibility of intrusions of items from the masking signal into the participants’ spoken repetitions of the target signal. These intrusions can be taken as indicating some central processing of the informational masker (Brungardt, 2001).

3. Functional imaging studies using speech in the presence of maskers

There are two distinct differences in the approaches taken to the neurobiological study of masked speech. First, there are studies that address the cortical and subcortical systems associated with the perception of speech in a masker, and the way that pattern of activation can be affected by different masker types and characteristics. Several of these studies are motivated by the findings that the perception of speech in noise is something which worsens with age, due to presbycusis as well as to central changes in auditory processing capacity (e.g. Helfer et al., 2010). Other studies are motivated by more specific questions about the mechanisms underlying different aspects of masking phenomena. These motivations tend to influence the design of the experiment. For example, studies overtly comparing energetic/modulation masking with informational masking do not typically include a speech-in-quiet condition, as they have matched instead for intelligibility scores across the two masking conditions (Scott et al., 2004b; Scott et al., 2009b). In a largely separate set of studies, target speech items are presented with a masking noise as a way of expressly manipulating the intelligibility of speech and investigating the neural systems supporting comprehension. In such studies, the relationship between speech comprehension and the SNR of energetic/modulation or informational maskers, which compete with the target speech for resources at the auditory periphery, is used as one of a variety of ways in which speech intelligibility can be affected.

3.1 Investigations of masking speech: energetic/modulation masking studies

Salvi and colleagues (2002) conducted a study of the perception of speech in noise, imaging neural activity with PET, and using the following conditions: quiet, noise, speech-in-quiet, speech-in-noise (where the noise was composed of 12-talker babble - SNR not reported). Energetic/modulation masking effects might therefore be expected to dominate in this experiment. There was an overt task: during each PET scan, the subjects were required to repeat the last word of each sentence aloud (or say ‘nope’ if they could not hear the word, or when they thought a sentence ended in the Noise condition). The contrast of speech-in-noise (SPIN) over speech-in-quiet revealed widespread midline cortical and cerebellar activation. The speech-in-noise level was selected such that the participants would be expected to get approximately 50% of the sentences correct, and it is possible that the activations seen in the SPIN condition reflect cognitive resources (attentional and/or semantic) which support this task, as the energetic/modulation masking makes accurate repetition more difficult (Obleser et al., 2007). The opposite contrast, of speech-in-quiet over SPIN, was not reported.

Hwang and colleagues (2006) used fMRI to measure neural activation while participants listened to stories in Chinese. They used a block design, where the participants heard the stories either in quiet or with a continuous white noise masker at +5 dB SNR. The use of a continuous fMRI paradigm means that the scanner noise was a further source of acoustic stimulation (i.e. speech-in-quiet should be thought of as speech-in-scanner-noise). However as the scanner noise was present in every scan, it should have been effectively subtracted from the contrasts of interest (though it may still have influenced overall levels of activation). Hwang and colleagues (2006) found reduced activation in left STS (and elsewhere) for the SPIN condition relative to the speech-in-quiet condition. It is possible that this reflects reduced intelligibility of the speech-in-noise relative to the speech-in-quiet condition, since the left STS has a strong response to intelligibility in speech (Eisner et al., 2010; McGettigan et al., 2012b; Narain et al., 2003; Rosen et al., 2011; Scott et al., 2000). The opposite contrast, of SPIN over speech-in-quiet, was not reported.

Hwang and colleagues (2007) performed the same experiment with older adults and found comparable effects, with some evidence for a greater signal decrease for the speech in noise condition in the left temporal lobe for the older adults than for the younger group. They specifically identified posterior left STG as a site for central effects of presbycusis.

Wong and colleagues (2008) used fMRI to study the perception of speech in masking noise, and utilized a sparse imaging design to reduce the contamination of the acoustic stimuli with scanner noise. Their task involved the presentation of single words in a picture-matching task, with one word (out of a possible 20) presented per trial. The noise was multi-talker babble, so one would again expect energetic/modulation masking effects to dominate, and the noise was presented so that it coincided with the onsets and offsets of the presented words (i.e. the noise had the same duration as the word). The SNRs chosen were −5 dB and +20 dB: these were selected as the +20 dB SNR masking levels elicited ‘normal’ performance, while the −5 dB SNR led to a significant decrement in performance. A speech-in-quiet condition was also included. The results were complex: the contrast of (−5 dB SNR SPIN) > (+20 dB SNR SPIN) showed activation in posterior STG and left anterior insula, while the opposite contrast revealed extensive MTG and superior occipital gyrus activation, plus fusiform, inferior temporal gyrus and posterior cingulate gyrus. This is somewhat surprising as the (+20 dB SNR SPIN) > (−5 dB SNR SPIN) contrast might be expected to reveal the more extensive auditory cortical fields associated with the better comprehension of speech at higher SNRs. This pattern of activations may well reflect the task used, in which the activation to the single word in the context of a visual matching task may not entail extensive cortical activations compared to that seen when listeners perceive connected speech, such as sentences or stories (Hwang et al., 2006; Hwang et al., 2007; Salvi et al., 2002). The repetition of the same 20 words in each condition may also have been detrimental to responses in auditory cortical fields, which can show habituation to repeated spoken items (Zevin et al., 2004). However, using a region-of-interest (ROI) analysis, the authors identified sensitivity in the left STG to the level of the masking noise, where there was greater activation associated with *increasing* noise levels. This result is directly at odds with the finding of Hwang and colleagues of less activation in the STG for the SPIN listening condition than for the speech-in-quiet listening condition (Hwang et al., 2006; Hwang et al., 2007; see also Scott et al., 2004b).

In a further study (Wong et al., 2009), the design was repeated with older adult participants, with the difference that all three sets of stimuli, regardless of SNR were adjusted to have a level of 70 dB SPL. There was no behavioural difference between the older and the younger groups for the speech-in-quiet and +20 dB SNR conditions, but at −5 dB SNR the older adults had lower scores. The results indicated that the older adults, especially at −5 dB SNR, showed reduced activation in auditory areas, and recruited prefrontal and parietal areas associated with attentional and/or working memory processes to support the difficult listening task. The authors specifically linked this to compensatory listening strategies.

There may be some specific reasons for the differences between the studies reported by Hwang and colleagues and Wong and colleagues. Wong et al. found greater activation in STG regions for speech in noise at a lower SNR, while Hwang et al. found higher levels of activation for the speech-in-quiet than the speech in noise. These differences may be a consequence of the precise implementation of the study designs: Hwang et al used stories as the target speech, and a continuous masking noise, while Wong et al used single words presented in noise of matched duration. Hwang et al asked people to respond when they had understood a sentence, and Wong et al asked participants to match the heard word to pictures, which may emphasize semantic and executive processes over and above normal speech perception, while simultaneously reducing the total amount of speech heard. We discuss such factors in detail later in the paper, along with consideration of other aspects of the study design (such as sparse vs continuous scanning, the SNRs chosen, and how these might affect the patterns of results observed across different studies).

These studies show that there are differences in the cortical processing of speech-in-quiet and speech in a masking noise: however the methodological dissimilarities (and variation in how the results are analysed) mean that the precise nature of these differences remains elusive.

3.2 Investigations of masking speech: Informational masking

In an fMRI study of perception of speech-in-speech (SPIS) Nakai et al (2005) presented listeners with a story to follow, read aloud by a female talker. Periodically the story was masked with speech from a different (male) talker (DV) or speech from the same talker (SV) (these different maskers were presented in different blocks – i.e. the type of masking speech was not randomly varied). In addition to any energetic/modulation masking effects, one would expect masking speech to lead to considerable information masking, and that this would be more severe for the SV condition than for the different talker (DV) condition (Brungardt, 2001). Continuous fMRI scanning was carried out with a blocked presentation of baseline speech with no masker, alternated with the DV or SV conditions. There was bilateral STG activation for the contrast of the DV condition with the baseline speech condition, indicating that the masking speech was indeed being processed centrally to some extent, i.e. that central, cortical auditory areas were processing the unattended masking speech as well as the attended target speech. The SV condition led to significantly greater activation in the bilateral temporal lobes, and also in prefrontal and parietal fields. A direct comparison of the SV and DV conditions revealed greater activation for the SV masker condition in pre-supplementary speech area (pre-SMA), right parietal and bilateral prefrontal areas, suggesting that these are recruited to compensate for the strong competition for central resources produced by the SV masker. This study indicates that as informational masking effects increase, there is greater activity in non-auditory areas, potentially reflecting the recruitment of brain areas to cope with the increased attentional demands of a difficult listening condition.

Hill and Miller (2010) ran an fMRI study involving auditory presentation of 3 simultaneous talkers, all continuously reciting IEEE sentences (IEEE Subcommittee on Subjective Measurements, 1969). All the stimuli were generated from sentences produced by the same talker, with additional ‘talkers’ produced by shifting the pitch of the original speaker up or down. The ‘three talkers’ were presented at three virtual spatial locations, using headphones and a head-related transfer function. The sequences were constructed such that sentence onsets were not signaled by longer gaps, nor were there synchronous sentence onsets across the different voices. Prior to each trial, participants were cued to attend to either the pitch or location of the target talker, or to rest. In the attending trials, the participants were required to identify when the target talker began a new sentence. Continuous scanning was employed. As there were no conditions where the number of masking speakers or the type of maskers was varied, this study can be considered to be an investigation of attentional control and strategy within an informational masking context, rather than a study of masking *per se*. During stimulus presentation, they reported activation in bilateral STG/STS and beyond (but not in PAC) and the same system was recruited whether listening was based on pitch or location. This activation therefore probably reflects the objects of perception, rather than the strategies used to attend to them.

3.3 Investigations of masking speech: Contrasting informational and energetic/modulation masking

Informational and energetic/modulation masking have been identified as involving somewhat different mechanisms, although a complex masking sound (like masking speech) will produce both informational and energetic/modulation masking factors. One interesting distinction is the effects of level – while there is a relationship between SNR and intelligibility for energetic/modulation maskers, especially for SNRs below 0 dB, masking speech items can intrude into responses during informational masking even at high SNRs (Brungardt, 2001). An early PET paper from the first author’s lab aimed to investigate this by requiring participants to listen to short (BKB; Bench et al., 1979) sentences produced by a female talker, and contrasting two masker types - continuous speech spectrum noise as the energetic/modulation masker, and a male talker producing similar (though not identical) sentences as the informational masker. Passive presentation was used, to avoid neural activation associated with an overt task: such a design is arguably preferable when the processes under consideration are held to be both automatic and obligatory (Scott et al., 2004a). Pre-testing for intelligibility was used to establish that the SNRs used for each masking type did not result in overall differences of intelligibility across maskers. This resulted in use of the following SNRs: −3, 0 +3 and +6 dB for the speech in noise, and −6, −3, 0 and +3 for the speech in speech.

The study of Scott et al. (2004) has been criticized for not including a speech-in-quiet condition (Wong et al, 2008). However it would have been impossible to avoid intelligibility differences between the speech-in-quiet and masked speech, which would have led to differences in neural activation which reflected this modulation of intelligibility, rather than being an index of masking mechanisms. The logic of the chosen design was that contrasting speech in speech with speech in noise will remove perceptual/linguistic activations associated with the comprehension of the target talker’s speech, and reveal activation more specifically associated with the processing of the masking noise/talker. Also, a goal was to reveal neural activation that varied with the SNR, separately for each masking condition.

The study revealed greater overall activity for energetic/modulation masking than informational masking (independent of SNR) in rostral and dorsal prefrontal cortex and posterior parietal cortex (Figure 2), but this contrast revealed no activation in the temporal lobes. Within the energetic/modulation masking conditions, there was evidence for level-dependent activations in left ventral prefrontal cortex and the supplementary motor area (SMA) (Figure 2). This was interpreted in terms of the increased recruitment of semantic access and articulatory processes as the level of the masking noise increased. In contrast, there was extensive, level-independent activation in the dorsolateral temporal lobes associated with the contrast of speech-in-speech over speech-in-noise (Figure 2, see also Figure 4). As predicted from the behavioural data, there were no level-dependent effects within the speech-in-speech condition. The lack of activity beyond the temporal lobes may have been because the masking talker was male, and the target talker was female, so the competition for central resources was potentially reduced relative to more extreme information masking conditions (e.g. masking a talker with their own voice). The results were interpreted as showing a distributed, non-linguistic-specific network to support the perception of speech in noise, with both semantic and articulatory processes and representations recruited more strongly as the level of the noise masker increases. The strong bilateral activation of the dorsolateral temporal lobes in response to the masking speech was attributed to the obligatory and automatic central processing of the unattended speech: as has been shown by early investigations into selective attention, irrelevant or unattended speech is not discarded at the auditory periphery but is processed to some degree. However we cannot distinguish from this design which aspects of the unattended speech masker drive this central processing – is it the linguistic nature of speech, or some aspect of its acoustic structure? Considering the latter, it could be that because masking speech is highly amplitude modulated, it allows for brief ‘glimpses’ of the target speech (Festen et al., 1990).

A follow-up study (Scott et al., 2009b) addressed these issues by presenting participants with speech in signal-correlated noise (Schroeder, 1968), spectrally rotated speech, and speech. All three maskers allowed ‘glimpses’ of the target speech, and the three maskers varied in the degree to which they are dominated by energetic/modulation masking, and informational masking. The target speech again comprised simple sentences read by a female talker. Signal correlated noise (Schroeder, 1968) was created by modulating noise (in this experiment, noise with a speech-shaped spectrum) with the amplitude envelope of sentences spoken by the male talker from the previous study (Scott et al., 2004b). The speech in signal-correlated noise was considered to be dominated by energetic/modulation masking, but because of the amplitude modulated profile of the signal correlated noise, ‘glimpses’ of the target speech were possible. The spectrally rotated speech was again generated from the original male masking talker. Spectral rotation involves inverting a low-passed speech signal around a frequency point (e.g. 4 kHz low-passed filtered speech, inverted around 2 kHz), and has been used extensively as a baseline condition in speech perception studies as it preserves many aspects of the original speech signal, such as pitch, spectral structure (though inverted) and amplitude modulation (Blesser, 1972). This results in a signal highly analogous to speech in terms of complexity, but which cannot be understood. This was included as a masking condition as it was expected to contribute to informational masking phenomena, but without contributions from linguistic competition. Finally, there was a speech-in-speech condition, which was identical to the speech-in-speech informational masking condition in the previous study (Scott et al., 2004b). Pretesting was used to select SNRs leading to comparable levels of around 80% intelligibility. These were −3 dB for the signal correlated noise masker, and −6 dB for the speech and rotated speech maskers.

The contrast of speech in speech with speech in signal-correlated noise revealed much more restricted activation of the lateral STG by the masking speech than was seen in the earlier study (Figure 3, Figure 4), suggesting that much of this activation was indeed associated with unmasked ‘glimpses’ of the target speech in the speech-in-speech condition (Festen et al., 1990). Once this was controlled for (by having glimpses possible in both speech in signal-correlated noise and speech in speech) the activation no longer extended into primary auditory fields. Notably, there was also activation in the STG seen in the contrast of speech in rotated speech over speech in signal-correlated noise, though it was restricted to the right STG (Figure 3). The non-intelligibility of the spectrally rotated speech thus results in some informational masking effects through competition for central auditory processing resources, but not via linguistic mechanisms (which we would expect to yield modulations of left-dominant auditory regions).

Figure 4 shows the speech-in-speech>speech-in-noise contrasts for the Scott et al (2004) study, where continuous noise was used for the speech in noise (A), and the Scott et al (2009) study, in which signal-correlated noise was used to construct the speech-in-noise condition (B). The activation from the Scott et al (2004) study is more extensive than that seen from the Scott et al, (2009) study. The results of Scott et al. (2009) therefore indicated that some (but not all) of the greater STG activation seen for informational masking relative to energetic/modulation masking is a consequence of the glimpses of the target speech afforded by (naturally amplitude-modulated) speech maskers, but not by continuous noise maskers, which are commonly used. Furthermore, the study suggested that a masker need not be formed of intelligible speech to lead to central auditory processing: there was significant activation of the right STG by the masking spectrally rotated speech, and this is consistent with theories of informational masking which suggest that these may involve linguistic effects, but that informational masking can also arise due to other, paralinguistic properties of the stimuli.

4. Cognitive studies using speech in noise

Several studies have employed masking signals to explore the neural systems supporting the comprehension of speech, and cognitive processes within this. Here, the primary aim is often not primarily to explore the specific perceptual effects of presenting competing sounds alongside speech, but rather to use the presentation of speech against masking noise as a way of varying the comprehension of speech, and hence as a way of exploring linguistic processes relevant to comprehension. Binder et al (Binder et al., 2004) carried out an fMRI study of speech perception using a parametric modulation of SNR (−6 dB, −1 dB, +4 dB,

+14 dB and +24 dB) with syllables (‘ba’ and ‘da’) masked by white noise. Increasing SNR led to increases in the intelligibility of the speech signal, while reaction times (RTs) showed an inverted U-shape function with slowest responses occurring around +4 dB SNR. The authors related the RT profile to decision making processes in the task, suggesting that intermediate SNRs are those at which these processes are most strongly engaged (i.e. where the speech percept is likely to be most ambiguous). In a clever analysis approach, the authors identified brain regions whose activation profiles across the conditions matched the behavioural pattern of accuracy and RT change with SNR. They found that activation in bilateral superior temporal cortex showed a positive correlation with accuracy, while ventral areas of the pars opercularis in both hemispheres, as well as the left inferior frontal sulcus, showed the opposite pattern. In contrast, activation in bilateral anterior insula and portions of the frontal operculum showed a profile similar to the pattern of RTs across SNR. This study, as in some of the energetic/modulation masking studies, showed a sensitivity to speech intelligibility in temporal lobe fields, and an involvement of left prefrontal areas as the masked speech became harder to understand. Activity in other prefrontal areas (bilateral anterior insula and frontal operculum) was associated with the speed of response, suggesting an involvement in task related aspects of performance. This study is an elegant example of how fMRI data can be used to dissociate different aspects of the perception of speech in noise.

A more recent study (Davis et al., 2011) took a similar approach, in using behavioural profiles of speech comprehension in noise to explore ‘top-down’ versus ‘bottom-up’ aspects of speech processing. Similar to Binder and colleagues (Binder et al., 2004), they presented speech in noise (this time signal-correlated noise with English sentences) across 6 different SNRs equally spaced from −5 dB to 0 dB. However, in an additional manipulation, half of the sentences were semantically anomalous (e.g. ‘Her good slope was done in carrot’). The authors used a cluster analysis approach to fMRI responses across these conditions in order to identify neural regions showing significantly different profiles of activation. Similar to Binder et al. (2004), the left posterior STG showed a profile of increasing signal as SNR increased. However, this profile did not distinguish between semantically anomalous and semantically coherent sentence types. The left inferior frontal gyrus (IFG), as well as anterior STG, in contrast, showed a profile in which the response to semantically coherent items increased with increasing SNR, but decreased at the highest SNRs, while activation to the anomalous items continued to increase for these conditions. The authors interpreted this as a reflection of ‘effortful semantic integration’ (p. 3928) taking place in the IFG. That is, this region is engaged strongly at intermediate SNRs where there is sufficient low-level perception of the signal to allow higher-order linguistic computations to be engaged, but this is no longer necessary at high SNRs for semantically coherent (and to some extent, predictable) sentences. On the other hand, semantically anomalous sentences demand additional processing to facilitate the extraction of a sensible sentence percept, even at high SNRs. Although the posterior STG showed no distinction between the two sentence types, and therefore may be more concerned with the lower-level acoustic-phonetic aspects of perception, the authors suggest that this region is likely to be functionally connected with inferior frontal regions as part of a network of regions supporting comprehension.

Similar findings related to the IFG have recently been reported (Adank et al., 2012) in a fMRI experiment exploring the differences in the neural responses between two challenging listening conditions – SPIN and accented speech – and undistorted speech. Their SPIN stimuli were simple sentences in Dutch that were presented in speech spectrum noise at an SNR of +2 dB. In a sentence verification task, in which participants responded with a yes/no evaluation of the semantic plausibility of the sentence content, the SPIN condition elicited 18.1% errors (compared with 8% errors for undistorted, speech-in-quiet sentences). When comparing the BOLD responses to SPIN compared with speech-in-quiet, the authors found greater signal bilaterally in the opercular part of the IFG, as well as cingulate cortex, parahippocampal gyrus, caudate and frontal pole. They relate this to the engagement of higher-order linguistic representations in the attempt to understand speech that has been masked by an external signal. As a fully factorial design was not employed (that is, there was no SPIN+Accented speech condition), however, it remains unknown how the effects of masking noise and accented speech interact.

Scott et al. (Scott et al., 2004b) conducted a study described in a previous section, in which the neural consequences of energetic/modulation and informational masking were explored. However, they also included one analysis that addressed the correlates of speech intelligibility across all conditions (SPIN, and speech-in-speech), which they identified using the behavioural speech repetition scores from every participant (which were collected prior to scanning). This revealed a single cluster of activation in left anterior STG that showed a positive correlation with stimulus intelligibility. This is in line with their previous work on factorial comparisons of intelligible and unintelligible sentence stimuli (Scott et al., 2000), while the results of Davis et al. (2011) also point toward higher-order processes in this region supporting the extraction of a meaningful perception from a noisy stimulus.

Zekveld et al (Zekveld et al., 2006) masked spoken sentences with speech spectrum noise at 144 SNRs from −35 dB to 0 dB (−30 dB to 5 dB in some participants). The chosen range yielded speech comprehension accuracy scores of around 50% on average. The authors explored the interaction between masking and perception, dividing the stimuli into a factorial array describing SNR (low, medium and high), and intelligible versus unintelligible. Analyses exploring responses of the speech perception network across these conditions showed that unintelligible stimuli with low SNR yielded significant activation (compared with noise alone) in left Brodmann Area (BA) 44 in the opercular part of the IFG. They link this back to the findings of Binder et al. (2004) and suggest that BA 44 is recruited to support articulatory strategies, which affect speech perception in a top-down fashion and help the comprehension of severely distorted speech stimuli. In a further contrast in which SNR was controlled (at ‘medium’) but intelligibility differed, the authors identified increased signal in bilateral superior temporal cortex and a more anterior portion of the left IFG, in BA 45. There was some differentiation between the neural responses in temporal and frontal areas with increasing SNR, with temporal regions showing a (non-significant) ‘earlier’ onset response at around −13 dB, and a significantly larger asymptotic response than frontal areas.

There are some clear themes emerging from this work. Most consistently, there appears to be a distinction between the role of mainly left-lateralized regions of the opercular part of the IFG and sites in superior temporal cortex, where the frontal regions are associated with strategic and top-down effects, while temporal cortex more faithfully tracks the changes in speech audibility with varying SNR. At the same time, two studies have indicated that temporal responses in the comprehension of noisy speech precede those of the frontal cortex, both in the timing of the response, and in terms of the SNRs at which they ‘emerge’ (Davis et al., 2011; Zekveld et al., 2006). What are the roles of these higher-order influences of frontal cortex? Is the role evaluative, for example in the decision-making aspects of the Binder et al. (2004) task? Or is the involvement more dynamic, with, for example, an articulatory strategy guiding the extraction of phonemic content from the incoming stimulus? A parallel pattern in other studies is that direct contrasts of SPIN with speech-in-quiet or undistorted speech reveal activations in left inferior frontal cortex. Some authors have linked motoric simulation strategies to difficult listening situations (Adank, 2012; Hervais-Adelman et al., 2012). However, as we have previously argued (McGettigan et al., 2010; Scott et al., 2009a), the choice of stimulus and task can also have effects on the extent to which certain strategies come into play, for example, the engagement of articulatory representations to perform phonetic decomposition of a heard syllable (Sato et al., 2009). Across the studies described above (and for the moment not considering the type of noise/ masker used), authors have presented participants with stimuli ranging from a closed set of syllables, to words, to sentences, with tasks including repetition, semantic evaluation and passive listening (sometimes subsequent to an active behavioural task on the same stimuli). These different tasks would be expected to result in different patterns of neural activation. For example, the cognitive processes involved in evaluating whether a sentence ‘makes sense’ or not may more consistently recruit more prefrontal cortical fields than when deciding whether a heard syllable is a ‘da’ or a ‘ba’ (Scott et al., 2003b). This may partly explain why some studies in the literature on SPIN implicate frontal and auditory/auditory association cortices to varying degrees. In terms of task difficulty, the choice of SNR(s) used, in combination with task, is likely to have a major role in which brain regions are uncovered, and the profile exhibited by these regions. In factorial designs, the precise SNR used, and the degree of its effects on intelligibility, presents the authors with a certain ‘snapshot’ view of the activation of brain networks, and the relative involvement of components within them. A good example is Adank et al. (2012), who used two noise conditions – one intrinsic (sentences spoken with a novel accent) and one extrinsic (sentences in speech-shaped noise). When each of these was contrasted with rest, different profiles of activation were found. Additional activations were observed in the direct comparison of the two noise conditions. However, it is also the case that the two ‘noise’ conditions, aside from their likely differing effects at the auditory periphery, were significantly different in the proportion of errors they generated in the behavioural task. Thus, the direct comparison of the two conditions reflects an unknown ratio of contributions of basic acoustic and higher-order task related effects. Consideration of task difficulty effects may help us to understand the some of the inconsistencies across studies. For example IFG has been shown to be more engaged by SPIN than by speech-in-quiet (Adank et al., 2012), yet IFG activation also shows positive correlations with SNR (and therefore increasing intelligibility) (Davis et al., 2011; Zekveld et al., 2006). In a study by Zekveld et al. (2012), where performance in the scanner was on average less than 29% correct, a comparison of SPIN with speech-in-quiet gave an array of fronto-parietal activations more commonly associated with the default mode network (networks of activation which have been consistently associated with the brain ‘at rest’ and which typically indicate that participants are focussing on their own thoughts rather than on the experimental stimuli (Raichle et al., 2001)). In contrast, the IFG appeared in the converse comparison of speech-in-quiet > SPIN (Zekveld et al, 2012). We suggest that, where factorial designs have been employed to compare different types of masker, authors should endeavour to control for intelligibility across conditions (see Scott et al., 2004) and consider the overall *level of difficulty* at which their task is pitched. Alternatively, the parametric designs used by Davis et al. (2011) and Zekveld et al. (2006) offer the chance for authors to characterize an ‘intelligibility landscape’ across the brain, showing the interplay of different key sites in the comprehension network across different levels and types of masker. These designs could be developed to interrogate, for example, the effects of masking on the neural systems supporting comprehension at different levels of the linguistic hierarchy.

5. Conclusions

From a functional neuroimaging perspective, the understanding of speech presented against different maskers is still developing, not least because of the considerable technical problems of dealing with the acoustic noise generated by the scanning sequences used in most fMRI paradigms. However, a pattern is starting to emerge, where there is now considerable evidence for widespread prefrontal and parietal activation associated with dealing with speech in masking noise (Adank et al., 2012; Hwang et al., 2006; Hwang et al., 2007; Scott et al., 2004b; Scott et al., 2009b; Wong et al., 2008; Wong et al., 2009) and some evidence for level specific recruitment of prefrontal and premotor fields (Davis et al., 2011; Zekveld et al., 2012). In contrast, the claim that informational masking arises due to competition for resources at central *auditory* cortical processing levels has been largely supported by the findings of considerable processing of the unattended masking speech in bilateral superior temporal lobes, which is seen in addition to the activation associated with the target speech. As might be expected, these central informational masking effects are not restricted to intelligible speech, and can also be seen for spectrally rotated speech (Scott et al., 2009b). Greater activation (outside auditory cortex) is also seen when the masking speech is highly similar to the target speech (Nakai et al., 2005), as would be predicted from the behavioural literature (Brungardt, 2001).

There are outstanding questions about what this increased STG activation reflects: is it showing parallel processing of the different auditory streams associated with the target speech and with the informational masker? Is it a result of increased competition for resources? Does this processing profile vary between the two hemispheres, and does this vary with functional differences? Further studies will be needed to refine our understanding of these responses. However, if the objects of auditory perception are processed along neural pathways that are both plastic and capable of processing multiple streams of information in parallel (Scott, 2005), all of these possibilities may be found to play a role.

We have outlined several factors which affect the patterns of activation seen above and beyond those specifically associated with masking: the task used, the speech stimuli selected and SNR all affect the kinds of activation seen. Further outstanding challenges will be to identify cortical signatures that are masker-specific, and that might be recruited for both energetic/modulation masking and informational masking (not possible from Scott et al.'s 2004 or 2009 studies), and address the ways that ageing affects the perception of masked speech, while controlling for intelligibility (or performance differences). Finally, another important dimension to address will be individual variation in performance with different maskers. The neural activations associated with variation in adaptation to noise-vocoded speech have been reported (Eisner et al, 2010). In this study better learning was associated with greater recruitment of left IFG, rather than greater recruitment of auditory cortical fields. Is this similar or different for the perception of speech in masking sounds? How are the neural mechanisms altered by the kinds of experience-based changes (e.g. musical training) which have been argued to lead to enhanced perception of speech in noise (Parbery-Clark et al., 2012; Strait et al., 2012)? We anticipate the next decade of research to build upon the strong foundations reported here, with the real possibility of important changes in our understanding of how our brains cope with listening in a noisy world.

References

Adank P. The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. Brain Lang. 2012; 122:42–54. [PubMed: 22633697]

Adank P, Davis MH, Hagoort P. Neural dissociation in processing noise and accent in spoken language comprehension. Neuropsychologia. 2012; 50:77–84. [PubMed: 22085863]

Agnew ZK, McGettigan C, Scott SK. Discriminating between auditory and motor cortical responses to speech and nonspeech mouth sounds. J Cogn Neurosci. 2011; 23:4038–47. [PubMed: 21812557]

Belin P, Zatorre RJ. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 2003; 14:2105–2109. [PubMed: 14600506]

Bench J, Kowal A, Bamford J. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. Br J Audiol. 1979; 13:108–12. [PubMed: 486816]

Binder JR, Frost JA, Hammeke TA, Rao SM, Cox RW. Function of the left planum temporale in auditory and linguistic processing. Brain. 1996; 119:1239–47. [PubMed: 8813286]

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD. Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci. 2004; 7:295–301. [PubMed: 14966525]

Blesser B. Speech perception under conditions of spectral transformation. I. Phonetic characteristics. J Speech Hear Res. 1972; 15:5–41. [PubMed: 5012812]

Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci. 2005; 8:389–395. [PubMed: 15723061]

Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am. 2001; 109:1101–1109. [PubMed: 11303924]

Brungart DS, Simpson BD. Within-ear and across-ear interference in a cocktail-party listening task. J Acoust Soc Am. 2002; 112:2985–95. [PubMed: 12509020]

Cohen L, Jobert A, Le Bihan D, Dehaene S. Distinct unimodal and multimodal regions for word processing in the left temporal cortex. Neuroimage. 2004; 23:1256–70. [PubMed: 15589091]

Davis MH, Ford MA, Kherif F, Johnsrude IS. Does semantic context benefit speech understanding through "top-down" processes? Evidence from time-resolved sparse fMRI. J Cogn Neurosci. 2011; 23:3914–3932. [PubMed: 21745006]

Dimitrijevic A, Pratt H, Starr A. Auditory cortical activity in normal hearing subjects to consonants and vowels presented in quiet and in noise. Clin Neurophysiol. 2013; 124:1204–1215. [PubMed: 23276491]

Dirks DD, Bower DR. Masking effects of speech competing messages. J Speech Hear Res. 1969; 12:229–&. [PubMed: 5808851]

Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM. Improved auditory cortex imaging using clustered volume acquisitions. Hum Brain Mapp. 1999; 7:89–97. [PubMed: 9950066]

Eisner F, McGettigan C, Faulkner A, Rosen S, Scott SK. Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. J Neurosci. 2010; 30:7179–86. [PubMed: 20505085]

Festen JM, Plomp R. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J Acoust Soc Am. 1990; 88:1725–1736. [PubMed: 2262629]

Friederici AD, Ruschemeyer SA, Hahne A, Fiebach CJ. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cereb Cortex. 2003; 13:170–7. [PubMed: 12507948]

Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW. “Sparse” temporal sampling in auditory fMRI. Hum Brain Mapp. 1999; 7:213–223. [PubMed: 10194620]

Helfer KS, Chevalier J, Freyman RL. Aging, spatial cues, and single-versus dual-task performance in competing speech perception. J Acoust Soc Am. 2010; 128:3625–33. [PubMed: 21218894]

Hervais-Adelman AG, Carlyon RP, Johnsrude IS, Davis MH. Brain regions recruited for the effortful comprehension of noise-vocoded words. Lang Cogn Process. 2012; 27:1145–1166.

Hickok G, Okada K, Serences JT. Area Spt in the human planum temporale supports sensory-motor integration for speech processing. J Neurophysiol. 2009; 101:2725–2732. [PubMed: 19225172]

Hickok G, Buchsbaum B, Humphries C, Muftuler T. Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J Cogn Neurosci. 2003; 15:673–82. [PubMed: 12965041]

Hill KT, Miller LM. Auditory attentional control and selection during cocktail party listening. Cereb Cortex. 2010; 20:583–90. [PubMed: 19574393]

Hwang JH, Wu CW, Chen JH, Liu TC. The effects of masking on the activation of auditory-associated cortex during speech listening in white noise. Acta Otolaryngol. 2006; 126:916–20. [PubMed: 16864487]

Hwang JH, Li CW, Wu CW, Chen JH, Liu TC. Aging effects on the activation of the auditory cortex during binaural speech listening in white noise: an fMRI study. Audiol Neurootol. 2007; 12:285–94. [PubMed: 17536197]

Jacquemot C, Pallier C, LeBihan D, Dehaene S, Dupoux E. Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. J Neurosci. 2003; 23:9541–6. [PubMed: 14573533]

Johnsrude IS, Penhune VB, Zatorre RJ. Functional specificity in the right human auditory cortex for perceiving pitch direction. Brain. 2000; 123:155–163. [PubMed: 10611129]

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. Neural substrates of phonemic perception. Cereb Cortex. 2005; 15:1621–31. [PubMed: 15703256]

McGettigan C, Scott SK. Cortical asymmetries in speech perception: what’s wrong, what’s right and what’s left? Trends Cogn Sci. 2012a; 16:269–76. [PubMed: 22521208]

McGettigan C, Agnew ZK, Scott SK. Are articulatory commands automatically and involuntarily activated during speech perception? Proc Natl Acad Sci U S A. 2010; 107:E42–E42. [PubMed: 20304788]

McGettigan C, Evans S, Rosen S, Agnew ZK, Shah P, Scott SK. An application of univariate and multivariate approaches in FMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. J Cogn Neurosci. 2012b; 24:636–52. [PubMed: 22066589]

Mummery CJ, Ashburner J, Scott SK, Wise RJS. Functional neuroimaging of speech perception in six normal and two aphasic subjects. J Acoust Soc Am. 1999; 106:449–457. [PubMed: 10420635]

Nakai T, Kato C, Matsuo K. An fMRI study to investigate auditory attention: A model of the Cocktail Party phenomenon. Magn Reson Med Sci. 2005; 4:75–82. [PubMed: 16340161]

Narain C, Scott SK, Wise RJS, Rosen S, Leff A, Iversen SD, Matthews PM. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb Cortex. 2003; 13:1362–1368. [PubMed: 14615301]

Obleser J, Wise RJS, Dresner MA, Scott SK. Functional integration across brain regions improves speech perception under adverse listening conditions. J Neurosci. 2007; 27:2283–2289. [PubMed: 17329425]

Ogawa S, Lee TM, Kay AR, Tank DW. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A. 1990; 87:9868–72. [PubMed: 2124706]

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G. Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex. 2010; 20:2486–2495. [PubMed: 20100898]

Pa J, Hickok G. A parietal-temporal sensory-motor integration area for the human vocal tract: Evidence from an fMRI study of skilled musicians. Neuropsychologia. 2008; 46:362–368. [PubMed: 17709121]

Parbery-Clark A, Marmel F, Bair J, Kraus N. What subcortical-cortical relationships tell us about processing speech in noise. Eur J Neurosci. 2011; 33:549–557. [PubMed: 21255123]

Parbery-Clark A, Anderson S, Hittner E, Kraus N. Musical experience strengthens the neural representation of sounds important for communication in middle-aged adults. Front Aging Neurosci. 2012; 4:30. [PubMed: 23189051]

Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 2007; 8:976–87. [PubMed: 18026167]

Peelle JE, Eason RJ, Schmitter S, Schwarzbauer C, Davis MH. Evaluating an acoustically quiet EPI sequence for use in fMRI studies of speech and auditory processing. Neuroimage. 2010; 52:1410–1419. [PubMed: 20483377]

Posse S. Multi-echo acquisition. Neuroimage. 2012; 62:665–671. [PubMed: 22056458]

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. A default mode of brain function. Proc Natl Acad Sci U S A. 2001; 98:676–82. [PubMed: 11209064]

Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci. 2009; 12:718–24. [PubMed: 19471271]

Riecke L, Vanbussel M, Hausfeld L, Baskent D, Formisano E, Esposito F. Hearing an illusory vowel in noise: Suppression of auditory cortical activity. J Neurosci. 2012; 32:8024–8034. [PubMed: 22674277]

Rosen S, Wise RJS, Chadha S, Conway E-J, Scott SK. Hemispheric asymmetries in speech perception: Sense, nonsense and modulations. PloS One. 2011:6. 10.1371/journal.pone.0024672.

Salvi RJ, Lockwood AH, Frisina RD, Coad ML, Wack DS, Frisina DR. PET imaging of the normal human auditory system: responses to speech in quiet and in background noise. Hear Res. 2002; 170:96–106. [PubMed: 12208544]

Sato M, Tremblay P, Gracco VL. A mediating role of the premotor cortex in phoneme segmentation. Brain Lang. 2009; 111:1–7. [PubMed: 19362734]

Schroeder MR. Reference signal for signal quality studies. J Acoust Soc Am. 1968:44.

Schulze K, Gaab N, Schlaug G. Perceiving pitch absolutely: Comparing absolute and relative pitch possessors in a pitch memory task. BMC Neurosci. 2009; 10:106. [PubMed: 19712445]

Scott SK. Auditory processing - speech, space and auditory objects. Curr Opin Neurobiol. 2005; 15:197–201. [PubMed: 15831402]

Scott SK, Johnsrude IS. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 2003a; 26:100–107. [PubMed: 12536133]

Scott SK, Wise RJS. The functional neuroanatomy of prelexical processing in speech perception. Cognition. 2004a; 92:13–45. [PubMed: 15037125]

Scott SK, Leff AP, Wise RJS. Going beyond the information given: a neural system supporting semantic interpretation. Neuroimage. 2003b; 19:870–876. [PubMed: 12880815]

Scott SK, McGettigan C, Eisner F. A little more conversation, a little less action--candidate roles for the motor cortex in speech perception. Nat Rev Neurosci. 2009a; 10:295–302. [PubMed: 19277052]

Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 2000; 123:2400–6. [PubMed: 11099443]

Scott SK, Rosen S, Wickham L, Wise RJ. A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. J Acoust Soc Am. 2004b; 115:813–21. [PubMed: 15000192]

Scott SK, Rosen S, Beaman CP, Davis JP, Wise RJS. The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes. J Acoust Soc Am. 2009b; 125:1737–1743. [PubMed: 19275330]

Stone MA, Füllgrabe C, Moore BCJ. Notionally steady background noise acts primarily as a modulation masker of speech. J Acoust Soc Am. 2012; 132:317–326. [PubMed: 22779480]

Stone MA, Füllgrabe C, Mackinnon RC, Moore BCJ. The importance for speech intelligibility of random fluctuations in “steady” background noise. J Acoust Soc Am. 2011; 130:2874–2881. [PubMed: 22087916]

Strait DL, Parbery-Clark A, Hittner E, Kraus N. Musical training during early childhood enhances the neural encoding of speech in noise. Brain Lang. 2012; 123:191–201. [PubMed: 23102977]

Warren JE, Wise RJ, Warren JD. Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends Neurosci. 2005; 28:636–43. [PubMed: 16216346]

Wise R, Chollet F, Hadar U, Friston K, Hoffner E, Frackowiak R. Distribution of cortical neural networks involved in word comprehension and word retrieval. Brain. 1991; 114:1803–17. [PubMed: 1884179]

Wise RJ, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA. Separate neural subsystems within ‘Wernicke’s area’. Brain. 2001a; 124:83–95. [PubMed: 11133789]

Wise RJS, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA. Separate neural subsystems within ‘Wernicke’s area’. Brain. 2001b; 124:83–95. [PubMed: 11133789]

Wong PC, Uppunda AK, Parrish TB, Dhar S. Cortical mechanisms of speech perception in noise. J Speech Lang Hear Res. 2008; 51:1026–41. [PubMed: 18658069]

Wong PC, Jin JX, Gunasekera GM, Abel R, Lee ER, Dhar S. Aging and cortical mechanisms of speech perception in noise. Neuropsychologia. 2009; 47:693–703. [PubMed: 19124032]

Xiang J, Simon J, Elhilali M. Competing Streams at the Cocktail Party: Exploring the Mechanisms of Attention and Temporal Integration. J Neurosci. 2010; 30:12084–12093. [PubMed: 20826671]

Zatorre RJ, Belin P. Spectral and temporal processing in human auditory cortex. Cereb Cortex. 2001; 11:946–53. [PubMed: 11549617]

Zekveld AA, Heslenfeld DJ, Festen JM, Schoonhoven R. Top-down and bottom-up processes in speech comprehension. Neuroimage. 2006; 32:1826–36. [PubMed: 16781167]

Zekveld AA, Rudner M, Johnsrude IS, Heslenfeld DJ, Ronnberg J. Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. Brain Lang. 2012; 122:103–13. [PubMed: 22728131]

Zevin J, McCandliss B. Dishabituation to phonetic stimuli in a “silent” event-related fMRI design. Int J Psychol. 2004; 39:102–102.

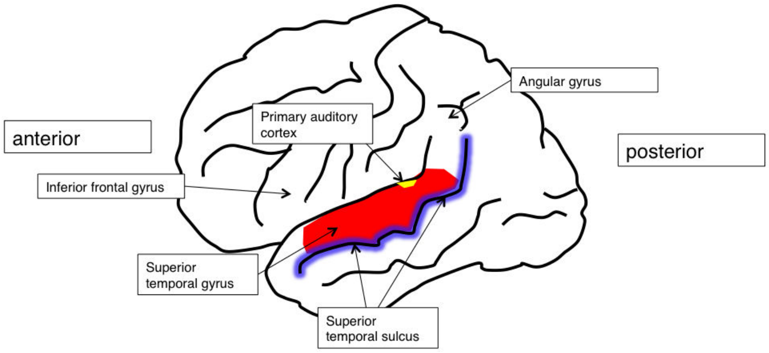

Figure 1. Lateral view of the left cortical hemisphere, showing cortical fields important in speech perception and comprehension.

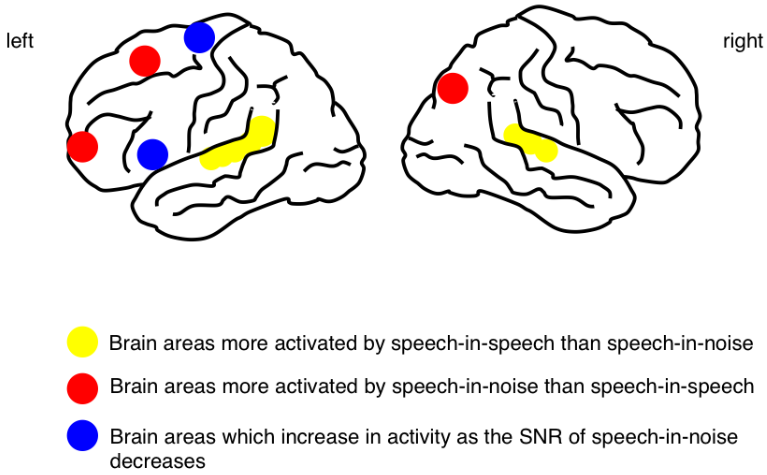
left
right
● Brain areas more activated by speech-in-speech than speech-in-noise
● Brain areas more activated by speech-in-noise than speech-in-speech
● Brain areas which increase in activity as the SNR of speech-in-noise decreases

Figure 2. Lateral view of the left and right cortical hemispheres, showing cortical areas associated with different aspects of the perception of speech-in-speech and speech-in-noise (from Scott et al, 2004).

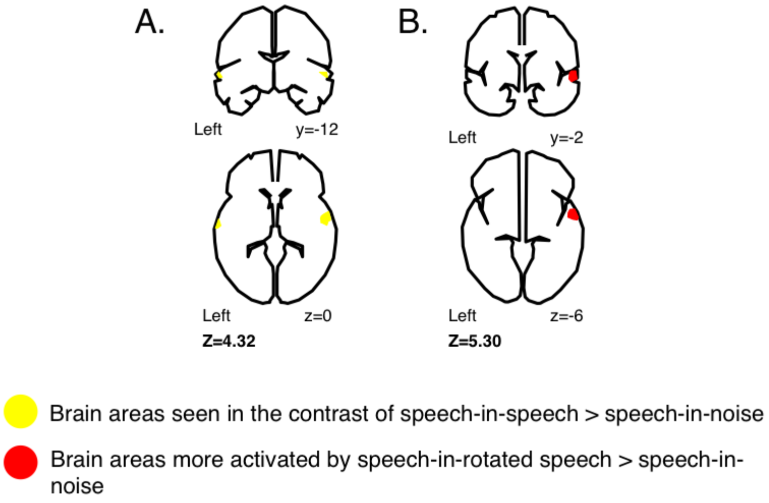
A.
Left
y=-12
Left
z=0
Z=4.32
B.
Left
y=-2
Left
z=-6
Z=5.30
● Brain areas seen in the contrast of speech-in-speech > speech-in-noise
● Brain areas more activated by speech-in-rotated speech > speech-in-noise

Figure 3. Cortical areas associated with (a) the contrast of speech-in-speech> speech-in-noise and (b) the contrast of speech-in-rotated speech> speech-in-noise (from Scott et al, 2009).

Upper panels show activity on coronal brain sections, and lower panels show activity on transverse brain sections. Figures have been redrawn from images thresholded at p<0.0001, cluster size>40 voxels.

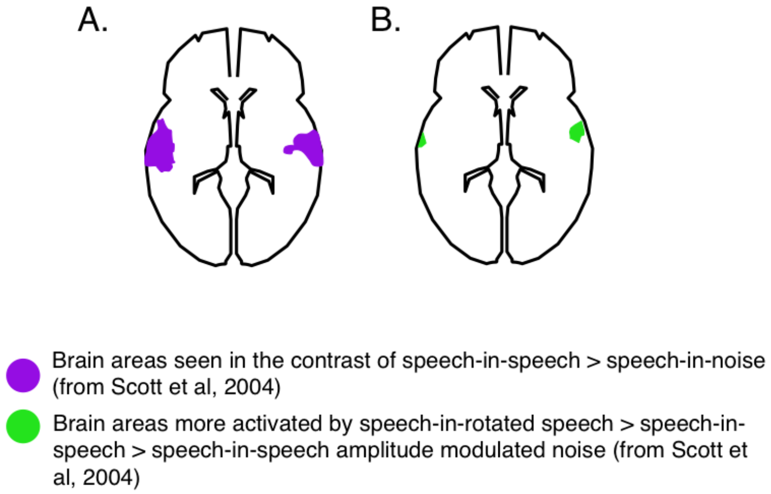
A.
B.
🟣 Brain areas seen in the contrast of speech-in-speech > speech-in-noise (from Scott et al, 2004)
🟢 Brain areas more activated by speech-in-rotated speech > speech-in-speech > speech-in-speech amplitude modulated noise (from Scott et al, 2004)

Figure 4. Cortical areas associated with the contrast of speech-in-speech>speech-in-noise from (a) Scott et al, 2004, where the noise in speech-in-noise is continuous and (b) Scott et al, 2009, where the noise in speech-in-noise is speech amplitude modulated, that is, glimpses of the target speech are possible in both speech-in-speech and speech-in-noise.

Upper panels show activity on coronal brain sections, and lower panels show activity on transverse brain sections. Figures have been redrawn from images thresholded at p<0.0001, cluster size>40 voxels.