Getting the Cocktail Party Started: Masking Effects in Speech Perception

Samuel Evans[1], Carolyn McGettigan[1,2], Zarinah K. Agnew[1,3], Stuart Rosen[1], and Sophie K. Scott[1]

Abstract

■ Spoken conversations typically take place in noisy environments, and different kinds of masking sounds place differing demands on cognitive resources. Previous studies, examining the modulation of neural activity associated with the properties of competing sounds, have shown that additional speech streams engage the superior temporal gyrus. However, the absence of a condition in which target speech was heard without additional masking made it difficult to identify brain networks specific to masking and to ascertain the extent to which competing speech was processed equivalently to target speech. In this study, we scanned young healthy adults with continuous fMRI, while they listened to stories masked by sounds that differed in their similarity to speech. We show that auditory attention and control networks are activated during attentive listening to masked speech in the absence of an overt behavioral task. We demonstrate that competing speech is processed predominantly in the left hemisphere within the same pathway as target speech but is not treated equivalently within that stream and that individuals who perform better in speech in noise tasks activate the left mid-posterior superior temporal gyrus more. Finally, we identify neural responses associated with the onset of sounds in the auditory environment; activity was found within right lateralized frontal regions consistent with a phasic alerting response. Taken together, these results provide a comprehensive account of the neural processes involved in listening in noise. ■

INTRODUCTION

Spoken conversations typically take place in noisy acoustic environments, unlike the quiet conditions of the laboratory. Conversing in noise is cognitively demanding, requiring the segregation and grouping of sounds from different sources and the selective attention to and decoding of the target auditory stream (Shinn-Cunningham, 2008). In terms of neural systems, we know much more about how the brain processes speech in quiet than we do about how it processes speech in background noise. This is problematic because many individuals find listening to speech in noise particularly difficult, for example, those with specific language impairment (Ziegler, Pech-Georgel, George, Alario, & Lorenzi, 2005) and dyslexia (Dole, Hoen, & Meunier, 2012), older adults (Wong et al., 2009), and people with hearing impairment (Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006). Identifying the cortical systems supporting perception in noise and relating individual differences in neural activity to perceptual ability in healthy adults is an important step toward understanding how these mechanisms are impaired in clinical groups and may help to guide future interventions.

In everyday life, speech is obscured by different kinds of sounds, for example, traffic and machinery noise, music, and the speech of others. The precise acoustic structure of masking sounds place differing demands on cognitive resources. Current frameworks for understanding how perception is affected by noise suggest a broad distinction between energetic, including modulation masking (Stone, Füllgrabe, Mackinnon, & Moore, 2011), and informational masking effects (Brungart, 2001). Energetic/modulation masking involves a direct interaction of the target signal and the masker within the auditory periphery (e.g., at the cochlea), resulting in a direct disruption of the target either by the energy in the masker (EM) or by the modulations in the masker interfering with those in the target (MM). By contrast, informational masking (IM) refers to additional effects not accounted for by energetic/modulation masking and is associated with "central" cognitive processes such as object formation and selection, and linguistic processing (Scott & McGettigan, 2013; Boulenger, Hoen, Ferragne, Pellegrino, & Meunier, 2010; Shinn-Cunningham, 2008; Brungart, 2001). Sounds can be described by whether they dominate in EM/MM or IM effects. For example, continuous white noise is an archetypal EM/MM as it obscures target speech at the auditory periphery: White noise is low in informational masking, as it is not perceptually confusable with speech. By contrast, the sound of

[1]University College London, [2]University of London, Royal Holloway, [3]University of California, San Francisco

Journal of Cognitive Neuroscience 28:3, pp. 483–500
doi:10.1162/jocn_a_00913

another talker, speaking the same language, will have some energetic/modulation masking properties, but speech is a stronger informational masker for target speech as both sources of speech are perceptually similar and include semantic and syntactic information. This distinction between EM/MM and IM is also a useful framework for interrogating the level of processing at which speech comprehension breaks down (Huang, Xu, Wu, & Li, 2010) or is enhanced in different listener groups (Boebinger et al., 2015; Oxenham, Fligor, Mason, & Kidd, 2003).

EM/MM and IM affect speech comprehension differently. For example, the comprehension of target speech is affected linearly by the signal-to-noise ratio (SNR) of EM/MM, but not IM (Brungart, 2001). Similarly, there is a greater benefit to comprehension when a target and informational masker are spatially separated compared with an equivalent energetic masker (Freyman, Helfer, Mccall, & Clifton, 1999). In a recent study by Ezzatian, Li, Pichora-Fuller, and Schneider (2012), participants listened to semantically anomalous but syntactically correct sentences masked by speech or a steady-state noise. They showed that, within a trial, performance improved over time when masking with speech but remained stable for steady-state noise, suggesting that speech maskers are particularly disruptive to perception in the early time window. This likely reflects the build-up of separate auditory streams, which takes longer for speech on speech masking because of its greater similarity to target speech. Indeed, stream segregation is an important component of perception in noise, and differences in stream segregation abilities may underlie some of the difficulties experienced by individuals who find listening in noise difficult (Ben-David, Tse, & Schneider, 2012). An additional factor known to affect comprehension in noise is the extent to which a masking stimulus affords "glimpses" of target speech (Cooke, 2006). Some maskers, for example, speech, are inherently modulated offering spectrotemporal dips in which target speech can be heard more easily, improving comprehension in noise (Brungart, 2001). However, in a similar manner to stream segregation abilities, different groups of individuals, for example, those with hearing impairment, differ in the extent to which they are able to take advantage of "glimpsing" (Peters, Hill, Carolina, Moore, & Baer, 1998).

At a neural level, sound engages multiple streams of processing that radiate from primary auditory cortex (Peelle, Johnsrude, & Davis, 2010; Rauschecker & Scott, 2009; Davis & Johnsrude, 2007; Hickok & Poeppel, 2007). These streams include a ventral pathway associated with extracting meaning from speech and a dorsal pathway associated with integrating perception and production. The ventral stream is hierarchically organized such that primary auditory cortex responds strongly to simple stimuli like pure tones, whereas surrounding regions respond to more complex sounds like band pass noise (Wessinger et al., 2001). Identifying regions that respond specifically to speech has proved difficult because of the inherent acoustic complexity of the signal; low-level auditory baselines such as tones and noise bursts make it difficult to distinguish between neural responses that are specific to speech and those that are a consequence of the perception of a complex sound. When speech is compared with complex nonspeech baseline sounds like rotated speech, selective responses extend anteriorly along the STS (Evans et al., 2014; Friederici, Kotz, Scott, & Obleser, 2010; Awad, Warren, Scott, Turkheimer, & Wise, 2007; Spitsyna, Warren, Scott, Turkheimer, & Wise, 2006; Narain et al., 2003; Scott, Blank, Rosen, & Wise, 2000). In terms of laterality, these responses are of higher amplitude and are more reliably encoded in the left hemisphere (Evans et al., 2014; McGettigan et al., 2012).

How is the perceptual system affected when speech is processed in background noise? Listening to speech in noise generates additional activity within prefrontal, parietal, and cingulate cortex (Golestani, Hervais-Adelman, Obleser, & Scott, 2013; Adank, 2012; Wong et al., 2009; Wong, Uppunda, Parrish, & Dhar, 2008). This is consistent with the notion that perception in noise recruits additional domain general cognitive control networks (Vaden et al., 2013; Petersen & Posner, 2012; Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008; Duncan & Owen, 2000). In support of this, a number of studies have shown a functional dissociation between periauditory regions and pFC, with auditory regions shown to respond in a "bottom–up" stimulus-driven manner and prefrontal regions evidencing "top–down" decision based or supplementary processes (Davis & Johnsrude, 2003, 2007; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006; Binder, Liebenthal, Possing, Medler, & Ward, 2004). However, as these studies included an "active" perceptual task, it is unclear whether these same regions would be equivalently activated during passive perception. Indeed, neural activity in frontal cortex during speech perception has been argued to be driven by metacognitive and task-based perceptual processes (McGettigan, Agnew, & Scott, 2010; Scott, McGettigan, & Eisner, 2009).

Recent studies using electrocorticography have further specified the nature of the information represented in the temporal lobes during perception in noise. Mesgarani and Chang (2012) recorded neural responses within the mid-posterior superior temporal gyrus (STG) to hearing two speakers presented in quiet and when the speakers were mixed together. Participants were instructed to attend to one or the other speaker. Neural reconstructed spectrograms demonstrated that, although the speech of both speakers was represented in cortical recordings, the response to the unattended speaker was suppressed relative to the attended one. Extending the number of recording sites, Golumbic et al. (2013) showed that regions close to primary auditory cortex tracked both target and masking sounds, whereas downstream regions tracked the target speaker alone. This is suggestive of a

gradient of enhancement in the representation of the target, as compared with the masker, with distance from primary auditory cortex. The ability to formulate and maintain separate sound streams is critical to the successful comprehension of speech in masking sounds. Functional imaging studies examining the neural basis of stream segregation have typically investigated responses to auditory stimuli and tasks that are not speech based, for example, in the context of auditory figure-ground perception or tone sequences that vacillate in percept between single and multiple streams over time. These studies have identified neural activity associated with the processing of multiple auditory streams in primary auditory cortex and the planum temporale (Gutschalk, Oxenham, Micheyl, Wilson, & Melcher, 2007; Wilson, Melcher, Micheyl, Gutschalk, & Oxenham, 2007; Deike, Gaschler-Markefski, Brechmann, & Scheich, 2004) and the inferior parietal sulcus (Teki, Chait, Kumar, von Kriegstein, & Griffiths, 2011; Cusack, 2005).

Two previous PET studies have examined the effect of manipulating the acoustic and linguistic properties of masking sounds to identify the neural basis of different types of perceptual competition. Scott, Rosen, Wickman, and Wise (2004) examined neural responses to speech masked by another talker and speech masked by a speech-shaped steady-state noise at a range of SNRs. The neural response-to-target speech presented in the context of these masking sounds was directly compared to isolate the effect of the masker itself. Increased responses to speech on speech masking, relative to noise masking, were found in the bilateral STG. The reverse contrast identified greater activity in the right posterior parietal and left prefrontal cortices in response to the steady-state noise. The fact that speech masking activated the STG is suggestive that competing speech was processed within the same processing stream as target speech. However, the absence of an unmasked single talker condition made it difficult to ascertain whether competing speech was treated equivalently within this pathway and did not allow identification of the broader masking network. A follow-up study by Scott, Rosen, Beaman, Davis, and Wise (2009) included a modulated rather than steady-state noise to equate "glimpsing" opportunities between masking conditions, which may have accounted for some of the energetic effects described in the previous study. A rotated speech masker was also included to isolate neural responses associated with the linguistic properties of competing speech. In the study, speech on speech masking relative to modulated noise masking activated the bilateral STG, and rotated speech masking relative to modulated noise masking activated the right STG. These results suggest hemispheric asymmetries in the processing of speech as compared with nonspeech maskers; however, the degree of asymmetry was not directly quantified, and therefore, the degree of lateralization remains equivocal. Also, no activation was identified for an increased response to speech masking as compared with rotated speech masking, the strongest test for sensitivity to the intelligibility of competing sounds. It is difficult to ascertain, given the small number of participants and the imaging modality used, whether the absence of this effect reflected a lack of statistical power or an absence of neural sensitivity.

In the current study, we used fMRI to address how different masking sounds are processed in the human brain. Here, with greater statistical power, and now crucially including a condition in which participants listened to target speech without additional masking sounds, we asked whether masking engages domain general attentional control systems in the context of an attentive listening task in which overt behavioral responses were not required. We hypothesized that we would find evidence that the informational properties of masking sounds modulate neural activity in the STG and that competing speech would be associated with left lateralized activity but would not be processed equivalently to target speech. Finally, we addressed how the onset of these different kinds of sounds modulated neural responses.

METHODS

Participants

Twenty right-handed native British English speakers (mean age = 25 years, age range = 19–36 years, 10 men) took part in the study. All participants reported having no known hearing, language, or cognitive impairments and gave informed consent in accordance with the University College London ethics committee.

Stimuli

All recorded stimuli were based on tabloid newspaper articles published from 1977 to 1979 in the *Daily Mirror*, a British national newspaper. These newspaper stories were of a short duration when read and were consistent in style, with simple syntax and vocabulary. The narratives used for the masking and target stimuli were mutually exclusive, and there was no repetition of target or masker stories within the behavioral or fMRI testing. Two female Southern British English speakers read aloud the narratives in an anechoic chamber, recorded at a sampling rate of 44.1 kHz and 16-bit quantization. One speaker was assigned to be the target speaker and the other the masking speaker. The two speakers were chosen to maximize the masking potential of the two voices; the speakers were sisters aged 35 and 37 years old, and both had lived in the South East of England for the majority of their lives. An automated procedure was used to remove long silent periods in the recordings of both speakers, defined as sections lasting in excess of 250 msec that were less than the median value of the amplitude envelope (extracted via a Hilbert transform).

This gave rise to natural-sounding speech with very few pauses.

Each target narrative was presented as clear (CL; i.e., without the presence of a masker) or in the presence of competing speech (SP), rotated speech (ROT), or speech modulated noise (SMN). See Figure 1A for spectrograms and oscillograms of example stimuli.

As rotated speech can only contain energy up to twice the rotation frequency, all stimuli were low-pass filtered at 3.8 kHz, including the target speech, to ensure a similar distribution of spectral energy across all the conditions. Two speech maskers were constructed: a continuous and a discontinuous narrative masker. In the continuous narrative masking condition, the masking speech was a single coherent narrative. In the discontinuous narrative condition, speech phrases from random stories were reassembled to construct a disjointed narrative, where each randomly selected phrase was syntactically complete. As there was no evidence of any behavioral or neural differences between these conditions, they were collapsed into a single condition using contrast weights in the fMRI analysis (e.g., SP = 0.5 × continuous + 0.5 × discontinuous). Two nonspeech maskers, SMN and ROT, were constructed from a random half split of the continuous and discontinuous speech conditions. 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 SMN was given the same long-term average spectrum as the original speech. This was achieved by subjecting the speech signal to a spectral analysis using a fast Fourier transform of length 512 sample points (23.22 msec) 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 50–7000 Hz. The smoothed spectrum was then used to construct an amplitude spectrum for an inverse fast Fourier transform with component phases randomized with a uniform distribution over the range 0–2π. Rotated speech was constructed by spectrally inverting speech around a 2-kHz axis using a digital version of the simple modulation technique described by Blesser (1972). As natural and spectrally inverted signals have different long-term spectra, the signal was equalized with a filter giving the inverted signal approximately the same long-term spectrum as the original speech. Rotated speech preserves some features of the original speech. It has a largely unchanged pitch profile, where some vowels remain relatively unchanged and some voice and manner cues are preserved. However, it is still unintelligible without significant training (Green, Rosen, Faulkner, & Paterson, 2013; Azadpour & Balaban, 2008; Blesser, 1972).

The stimuli were chosen to represent a broad parametric manipulation in similarity to speech. For example,

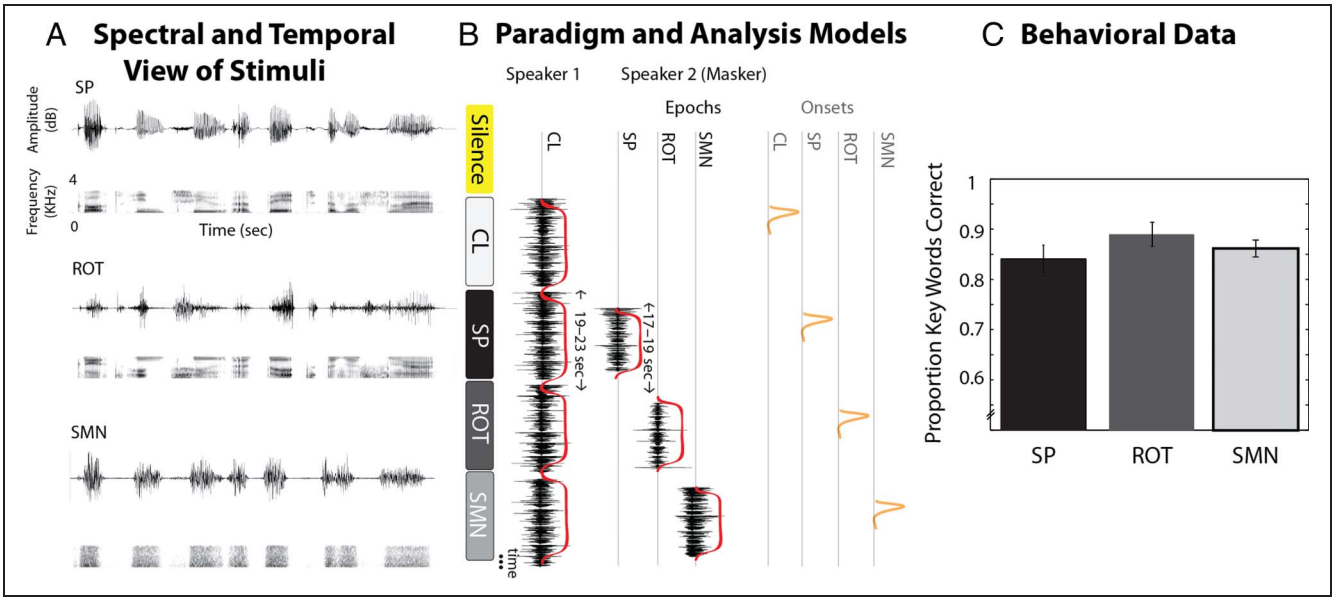

Figure 1. (A) Oscillograms and spectrograms of masking stimuli. SP = speech; ROT = rotated speech; SMN = Speech modulated noise. (B) The organization of a set of trials (left, rounded boxes) and the statistical models (right). Epoch Model (red), the first column in the design matrix models the presence of the voice of Speaker 1 and thus models all auditory trials for their full duration, excluding "silent" implicit baseline trials. Additional columns model the presence of competing masking sounds derived from Speaker 2, with each column representing a different masking sound. These events partially overlap with events specified in the first column. This design identifies unique variance associated with hearing Speaker 1 in clear and the additional effect of masking with competing sounds. Onset Model (red + orange), the design matrix models events in the same way as the Epoch Model (red) with additional events modeling the onset of clear speech and masking (orange). This allows the identification of unique variance associated with the onset of masking. (C) Behavioral postscanning accuracy for each condition. Bar graphs of the mean beta value for each condition with within-subject error bars representing one standard error (Loftus & Masson, 1994).

SMN is the least like speech: Although it has the temporal modulations and the same long-term average spectrum of speech, it does not have spectrotemporal dynamics such as formant or harmonic structure. ROT has spectrotemporal dynamics—for example, it has evenly spaced spectral components and formants—but is largely unintelligible to naive listeners. SP is obviously intelligible and includes semantic and syntactic information. In terms of different kinds of masking, we would expect EM/MM effects to be associated with all the masking stimuli, but we expect an increasing informational component as the masking stimuli show greater similarity to speech, for example, from SMN, ROT to SP. The conditions are also likely to differ in a small amount in their energetic properties, although as SMN and ROT have amplitude modulations and long-term spectra derived from the speech stimulus they are well matched in energetic masking to the SP condition. Note that we do not claim that SMN, ROT, and SP increase in informational content in necessarily equal steps, rather that the manipulation reflects an ordinal increase in information.

The target and masker narratives were mixed offline at SNRs, giving rise to ~85% key words correct (as established in pilot behavioral testing): SP (+3 dB), SMN (0 dB), and ROT (0 dB). Note that, in the case of positive SNRs, the target signal was presented at a more intense level than the masker. Therefore, higher SNRs reflect the fact that perception was harder in that condition during pilot testing. These SNRs were attained in behavioral testing by fixing the level of the masker at 68 dB SPL (measured by a Bruel & Kjaer 4153 artificial ear) and changing the level of the signal. Favorable SNRs were chosen to ensure a relatively high level of accuracy. This ensured that we recorded neural responses to effortful intelligibility, rather than an absence of intelligibility. Note that the SNRs were higher than those used in Scott, Rosen, et al. (2009), which included the same masking conditions—that is, in the current study the relative intensity of the target speech needed to be higher to achieve comparable comprehension performance. This likely reflects the use of narratives and the greater similarity between the voices; in the previous study, simple sentences recorded by a male and female speaker were used. Following adjustment of the relative SNR (by changing the level of the signal/target), all stimuli, including clear speech, were equated to the same output RMS level for presentation in the scanner (cf. Scott, Rosen, et al., 2009).

Scanning Procedure

Before scanning, participants were trained to differentiate the two speakers using a computerized task. It was established that every participant was able to discriminate between the speakers and understood which speaker they were tasked with attending to. Before the main experiment, each participant heard example stimuli inside the scanner while the scanner was acquiring data. This served to familiarize the participants with the stimuli in the presence of scanner noise. During the experiment, the participants were told to listen carefully to the stories spoken by the target speaker, as they would be asked questions about them after scanning. In particular, they were asked to listen for a story about a bear (which was in fact not presented) to encourage them to listen carefully throughout both runs of data collection. After scanning, all participants correctly reported that they had not heard a story about a bear. They were also informally asked whether they could recall any narratives. All participants were able to recall at least one target, and most participants recalled multiple stories. As in previous studies, we did not ask participants to engage in an explicit behavioral task in the scanner (Scott, Rosen, et al., 2009), except to remain attentive to the target speaker, as we wished to understand the neural mechanisms involved in listening to speech in noise in a more ecologically valid context. The absence of an explicit behavioral task allows us to be confident that observed activation was the consequence of attentive speech recognition, rather than reflecting the requirement to provide an overt response on each trial.

In the scanner, the spoken target narratives varied in duration from 20–23 sec, and the masking sounds lasted 17–19 sec. In each masking trial, the masking sounds were temporally aligned to the center point of the target narrative so that the participant heard the target speaker at the beginning and end of each trial without competing sounds. This helped participants to remain orientated to the target speaker. It also introduced a natural jitter (0.5–3.5 sec) to the onset of the masking stimulus relative to the onset of the target speaker. Further jitter was ensured consequent to the differing durations of the target narratives. We used masking trials of a long duration, in which masking sounds began after a short delay, so as to more closely mimic the experience of masking as it occurs in everyday situations. This also allowed us to examine neural responses to the onset of sounds (described in more detail below). Nine unique narratives from each condition were played during each run, including the target speaker presented in clear, making a total of 45 narratives. There were an additional six "silent" trials in each run, lasting 18 sec, in which the scanner was heard in the absence of additional auditory stimulation. The order of the conditions was pseudorandomized with the constraint that, within a subblock of five trials, the target speaker in clear preceded a single repetition from each masking condition, with the order of those masking conditions randomly permuted (see Figure 1B). This ensured that participants always heard the target speaker in the absence of masking as the first stimulus of the experiment and then regularly thereafter (after every four trials) to help them remain orientated to the target speaker.

Scanning was performed at the Birkbeck-UCL Neuroimaging Centre on a 1.5-T MR scanner (Siemens Avanto,

Siemens Medical Systems, Erlangen, Germany). In the scanner, auditory stimuli were presented using the Cogent Toolbox (www.vislab.ucl.ac.uk/cogent.php) via electrostatic headphones (MRCONFON, Magdeburg, Germany). The stimuli were played out at the same comfortable listening level for all participants. Unfortunately, it was not possible to measure the intensity level at the electrostatic headphones as the sound intensity changes in relation to the magnetic field when using these headphones. As participants' heads were placed at the same location within the scanner bore, each participant would have heard the stimuli at the same overall intensity level. It was not possible to measure this overall level as MR safe calibration equipment was not available to us. However, as the intensity remained fixed and the stimuli were all RMS equalized (including the clear unmasked speech) after adjusting to the appropriate SNR and the difference in intensity between target and masker was small (between 0 and 3 dB dependent on masking condition), we can be confident that the effects identified reflect masking effects rather than the audibility of the target signal per se. Hence, the absence of dB SPL measurement of the sound mixtures does not affect the interpretation of our findings.

Two functional runs of data lasting around 20 min were acquired using a continuous acquisition sequence (repetition time = 3 sec, echo time = 50 msec, flip angle = 90°, 35 axial slices, matrix size = 64 × 64 × 35, 3 × 3 × 3 mm in-plane resolution). Slices were angled away from the eyeballs to avoid ghosting artifacts from eye movements. The field of view included the frontal and parietal cortex at the expense of the inferior-most part of the temporal lobes and the cerebellum. Data were acquired with continuous rather than sparse acquisition to allow the differentiation of neural responses associated with the onset of sounds: sparse scanning entails more prominent onsets and offsets of scanner noise, which would have interfered with the analysis of the onsets of masking noises. A relatively quiet MRI sequence was used (~80 dB SPL) along with sound attenuating headphones (~30 dB attenuation). A high-resolution T1 structural image (HIRes MP-RAGE, 160 sagittal slices, matrix size: 224 × 256 × 160, voxel size = 1 mm^3) was acquired following the functional runs.

The first five volumes from each run were discarded to allow longitudinal magnetization to reach equilibrium. Data were analyzed with SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK; www.fil.ion.ucl.ac.uk/spm/). Functional images were slice time corrected to the middle slice and realigned to the mean functional image. The anatomical image was coregistered to the mean functional image. Normalization was conducted using the parameters obtained from the unified segmentation of the structural image (SPM8, segment function) using the International Consortium for Brain Mapping tissue probability maps, with voxels resampled to 2 mm^3 and the data smoothed with a Gaussian kernel of 8 mm FWHM. Two design matrices were constructed at the first level: (1) a design matrix specifying the effect of each condition and (2) a design matrix differentiating the effect of the onset of masking. For both designs, each stimulus was modeled by a canonical hemodynamic response function with condition effects modeled alongside six movement parameters of no interest. A high-pass filter of 250 sec and AR(1) correction were applied.

In the first design (Epoch Model), a regressor modeled the target speaker as present in all trials except the rest trials, with additional overlapping regressors modeling the masking epochs for each condition, against the implicit "silent" rest baseline (see Figure 1B, red). This design, with overlapping events, allowed us to identify variance explained by masking beyond that explained by listening to a single speaker and best reflected the experimental paradigm as experienced by the participants. Note that the regressor coding for each masking condition implicitly represents the subtraction of the clear speech from the masked speech condition. In the second design (Onset Model), the conditions were modeled as above with additional regressors modeling the onset of clear speech and each masking condition (see Figure 1B, red + orange). These additional events were modeled with duration of 0 sec indicating transient events. This allowed the identification of additional variance associated with the onset of masking (e.g., variance not explained by the trial length regressors). Note that these regressors implicitly represent the subtraction of onsets from sustained masking epochs.

At the second level, one sample t tests and within-subject one-way ANOVAs were conducted by entering contrast images from each participant into random effects models using the summary statistic approach. All statistical analyses are presented at $p < .001$ uncorrected at the voxel level, $q < 0.05$ corrected at the cluster level for the whole-brain volume using a nonstationary correction (Hayasaka, Phan, Liberzon, Worsley, & Nichols, 2004). Spatial localization of significant activations was carried out using the SPM Anatomy Toolbox (Eickhoff et al., 2005). ROI analyses were conducted using the Marsbar Toolbox (Brett, Anton, Valabregue, & Poline, 2002). Lateralization analyses were conducted using an iterative bootstrap approach implemented within the LI toolbox (Wilke & Lidzba, 2007; Wilke & Schmithorst, 2006). This is a well-established method for quantifying the relative lateralization of neural activity across different statistical thresholds. The LI tool box calculates laterality using the following formula:

$$\mathrm{LI} = \frac{\sum \text{activation left} - \sum \text{activation right}}{\sum \text{activation left} + \sum \text{activation right}}$$

Voxel activation values rather than voxel counts were used in the calculation of the index. Laterality analyses were conducted on second-level rather than first-level images to extend inferences concerning laterality to the population as a whole (i.e., a random effects inference).

The default anatomical regions within the toolbox for parietal, temporal, and frontal cortex (Tzourio-Mazoyer et al., 2002) were used masking out midline structures (±5 mm). Laterality curves were calculated by sampling, with 25% replacement, above threshold voxels in each hemisphere to generate 100 example vectors from which all possible lateralization index combinations are then calculated (10,000 combinations) across a range of statistical thresholds. The resulting mean lateralization curves were plotted. Analyses were conducted without clustering or variance weighting. We report the weighted average, which gives greater influence to LI values at higher statistical thresholds. For the sake of completeness, we also report the trimmed mean, which excludes the upper and lower quartile of the resampled laterality values. Laterality values are expressed in the radiological convention. Values can vary from +1 (total left lateralization) to −1 (total right lateralization). Weighted laterality values ≥+0.2 or ≤−0.2 indicate significant lateralization (Norrelgen, Lilja, Ingvar, Åmark, & Fransson, 2015; Gelinas, Fitzpatrick, Kim, & Bjornson, 2014; Nagel, Herting, Maxwell, Bruno, & Fair, 2013; Pahs et al., 2013; Badcock, Bishop, Hardiman, Barry, & Watkins, 2012; Lidzba, Schwilling, Grodd, Krägeloh-Mann, & Wilke, 2011; Lebel & Beaulieu, 2009; Wilke & Schmithorst, 2006; Wilke et al., 2006).

Behavioral Testing

All participants who took part in the main fMRI experiment were tested after scanning to assess their comprehension of speech in noise in a behavioral test completed outside the scanner. Stimuli were played out over Sennheisser 25HD headphones on a laptop in a quiet room. Each participant listened to speech presented in the same noise conditions and at the same SNRs as were used in the scanner. Each participant heard 12 trials. During a trial, the participant heard the target and masker presented for variable durations (ranging from 3 to 15 sec). They were required to report back as much of the last phrase spoken by the target speaker as they could. Each target phrase contained four key words, on which report accuracy was scored. The masking conditions were counter balanced with a randomized latin square.

To address concerns that the scanner noise may have unduly affected the perception of the auditory stimuli, we also ran three participants on a modified 15-min version of the experimental paradigm used in the pilot testing, within the scanner, while it ran. These were not functional scans: The aim was to determine whether the behavioral effects of masking sounds were affected by the noise of the scanner running during continuous acquisition. In this task, as in the pilot experiment which was used to determine the levels for each masking condition, each participant attended to 10 narratives from each condition (CL, SP, ROT, and SMN) and repeated back the last phrase of each narrative. Their responses were recorded with a noise attenuating optical microphone (Optoacoustics fOMRI-III, Moshav Mazor, Israel), and performance was scored offline for the number of correct key words with a maximum of 40 possible key words in each condition (160 key words in total).

RESULTS

Behavioral Testing

In behavioral testing after scanning, three participants scored an average of less than 0.65 key words correct (or <31/48 words correct overall) in one or more of the masking conditions. The behavioral and fMRI data from these three participants were removed from the analysis to ensure high levels of intelligibility and greater homogeneity across the group of participants. Scores were converted to rationalized arc sine units (RAU; Studebaker, 1985) and submitted to a one-way repeated-measures ANOVA. This showed that there was no evidence of a difference in intelligibility between the different masking conditions, $F(3, 48) = 2.019$, $p = .124$, $\eta^2 = 0.112$. The mean proportion of correct key words across participants for each masking condition was SP = 0.84, ROT = 0.89, SMN = 0.86 (see Figure 1C). These data were used as a regressor to identify neural activity associated with masking performance—described in the Imaging Data section.

When we tested the three additional participants on a modified version of the scanner task to ascertain the effect of the scanner noise on performance, we found that accuracy was slightly reduced compared with when assessed in quiet outside the scanner, scores were as follows (proportion correct): CL = 0.95, SP = 0.78, ROT = 0.77, SMN = 0.77. These results suggest that the target speaker presented in clear was close to 100% intelligible, and perception under noise was effortful but still largely intelligible (on average there was only a 9% reduction in accuracy as compared with outside the scanner). Furthermore, there was no difference in accuracy across masking conditions. However, we acknowledge that this does not rule out possible perceptual interactions between the continuous scanner noise and the different masking conditions (e.g., associated with modulation masking; Stone et al., 2011), and we note that this is a potentially more widespread problem for studies using continuous scanning with auditory stimuli (Peelle, 2014). Extensive further data collection would be required to definitively assess this.

Imaging Data

Masking and Intelligibility Networks

Activation in response to the clear (unmasked) target speaker, relative to the resting baseline [CL > Rest], was found within bilateral primary auditory cortex and extended to the anterior and posterior STG and middle

temporal gyrus (see Figure 2, white outline). By examining the response to the average of the masking conditions [(Sp + Rot + SMN)/3], we identified regions that responded more to masked than to clear speech (as each masking condition is implicitly the subtraction of the clear target speech from the masked conditions). Activation was found beyond the temporal lobe, in regions associated with cognitive control, in bilateral anterior cingulate, middle frontal gyri and insulae, as well as the left inferior and superior parietal lobule and superior orbital gyrus, and right inferior frontal gyrus (pars opercularis) and pallidum (Figure 2, blue; Table 1). The response plots at the peak voxels were similar for the three masking conditions suggesting that the masking conditions placed similar demands on this network (for simplicity, we plot the response of only two of the eight peak voxels, but the pattern of activity was similar across all peaks). At a reduced threshold (peak level $p < .001$ uncorrected, cluster level uncorrected), there was a small amount of activation observed in the left posterior STG (cluster level $p = .085$). The reverse contrast identified activation associated with the increased intelligibility of listening to the target speaker in clear as compared with during masking. Activation was found in regions associated with speech intelligibility: the bilateral STS extending from posterior to anterior in the left and from mid to anterior in the right hemisphere (see Figure 2, red; Table 1).

Individual Differences in the Comprehension of Masked Speech

A second-level covariate representing the accuracy of comprehension for each participant during the postscanning masking tasks (averaged across masking conditions) was regressed against neural activity associated with the response to the average of the masked conditions to identify regions in which activity was correlated with behavioral performance. At a whole-brain corrected level, a region of left mid-posterior STG exhibited a positive correlation with masking scores; that is, individuals who performed better on perception in noise tasks activated this region more (Figure 2, orange rendering and orange box; Table 1). To understand whether activity within this region predicted better accuracy for each individual masking condition, we correlated activity associated with each masking condition with behavioral performance associated with each masking condition within independent ROIs that were estimated using a whole-brain leave-one-subject-out correlation between

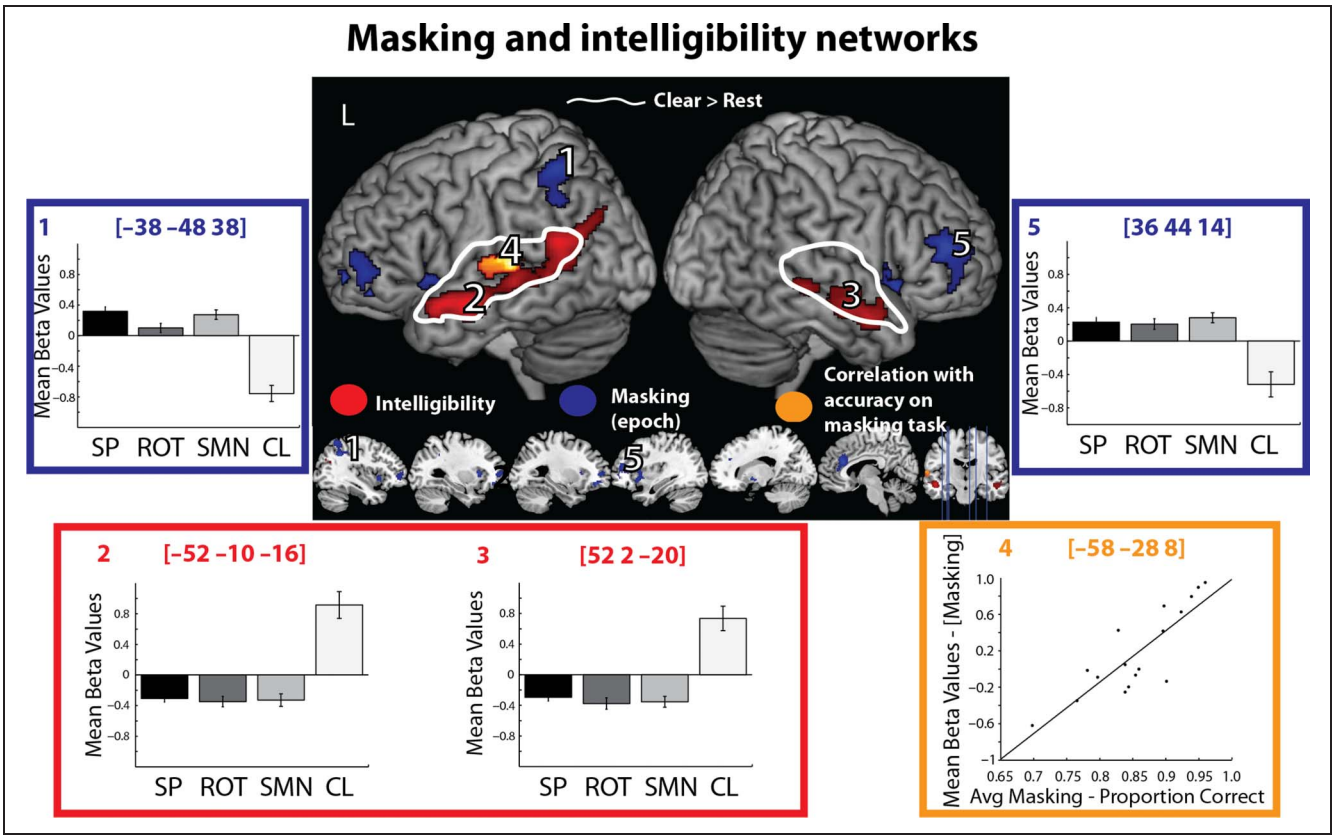

Figure 2. Masking and intelligibility networks. White: regions responding to clear speech as compared with the resting baseline. Red: regions responding more to clear than to masked speech. Blue: regions responding more to masked than to clear speech. Orange: regions in which activity correlated at a whole-brain level with accuracy in comprehension of speech in postscanning masking tasks (averaged across condition). Bar graphs of the mean beta value for each condition with within-subject error bars representing one standard error (Loftus & Masson, 1994). Scatter plot of the relationship between neural activity and comprehension of masked speech in postscanning masking tasks.

Table 1. Table of Activations

Location	MNI			Z-score	Number of Voxels
	x	*y*	*z*		
Masking (Epoch)					
Left inferior parietal lobule	−38	−48	38	4.58	417
Left middle orbital gyrus	−28	44	−10	4.43	1083
Right anterior cingulate	6	30	22	4.43	1013
Right pallidum	16	2	−4	4.38	230
Right middle frontal gyrus	36	44	14	4.27	563
Left insula	−30	16	10	4.15	271
Right insula	34	16	10	3.99	496
Left middle frontal gyrus	−28	48	6	3.86	331
Intelligibility					
Left anterior STS	−52	−10	−16	4.96	1754
Right anterior STS	52	2	−20	4.32	560
Covariate with Masking Performance					
Left mid-posterior STG	−58	−28	8	4.35	538
Modulation by Informational Content (Epoch)					
Right anterior STG	66	−12	−2	5.50	302
Left mid STG	−60	−16	2	5.02	524
Masking Onset (Scanner & Target Speech → Masking)					
Left posterior STG	−62	−30	14	6.39	5717
Right anterior STG	60	−4	−4	5.99	7588
Left middle cingulate	−4	−12	32	5.55	813
Right middle cingulate	6	26	36	4.84	674
Right SMA	2	12	56	4.76	195
Right precuneus	10	−66	34	4.27	331
Right middle frontal gyrus	46	44	2	4.12	445
Masking Onset (Scanner & Target Speech → Masking) Modulation by Informational Content					
Left superior parietal lobule	−6	−78	42	4.45	71
Clear Speech Onset (Scanner & Target Speech)					
Left middle STG	−50	−18	4	6.31	3998
Right posterior STG	50	−24	2	5.99	17401
Right middle orbital gyrus	10	42	−12	5.08	378
Left insula	−34	8	18	4.73	620

Table 1. (*continued*)

	MNI				
Location	*x*	*y*	*z*	*Z-score*	*Number of Voxels*
Right putamen	18	12	−4	4.41	335
Left insula	−32	24	6	4.38	563
Right thalamus	10	−8	12	4.36	197
Right thalamus	18	−26	−4	4.05	229

Coordinates reported for the peak maxima (at more than 8 mm apart) in MNI space.

the averaged masking response and averaged behavioral performance on the masking tasks (Esterman, Tamber-Rosenau, Chiu, & Yantis, 2010). That is, to identify an ROI for Participant 1, we reestimated the random effects whole-brain correlation between the averaged response to masking and the averaged behavioral performance of Participants 2–17. Within these ROIs, we found that neural activity in response to the most informational and energetic maskers was significantly correlated with behavioral performance on tasks involving those maskers (SP, p = .006; SMN, p = .026), but this was not the case for ROT (p = .132). It is unclear why activity within this region did not correlate with behavioral performance outside the scanner in the instance of rotated speech. This may reflect the fact that the behavioral measures were less reliable when considered individually than when averaged across condition. Alternatively, it may reflect the fact that perception of rotated speech selectively drives the right rather than the left STG (Evans et al., 2014; Scott, Rosen, et al., 2009). However, taken together, these results support the observation that the left STG supports perception in noise across multiple types of masking.

Effect of Varying Informational Content of the Masking Sounds

A one-way ANOVA investigating differences between the masking conditions identified clusters of activation in the bilateral mid to posterior STG, extending into the STS (see Figure 3 and Table 1). Plotting the response of these regions demonstrated that activation within these regions increased in response to the increasing informational content of the masking sounds (Figure 3, Plots 1 and 2). To test this observation, we conducted a follow-up contrast with the following contrast weights: [SP(discontinuous) × 0.5, SP(continuous) × 0.5, ROT × −0.25, SMN × −0.75], this confirmed that the response within these regions reflected a sensitivity to the informational content of masking sounds. The activation within the STG was situated within areas activated by a response to the target speaker in quiet relative to the "silent" baseline (Figure 3, white). It also significantly overlapped with the region in which activity positively correlated with behavioral masking scores (see Figure 4A, blue).

Response to the Intelligibility of the Masker

A conjunction null analysis (Nichols, Brett, Andersson, Wager, & Poline, 2005), showing regions commonly activated by [SP > ROT] and [SP > SMN], was conducted to more stringently identify regions modulated by the intelligibility of the masking stimulus. Note that, unlike the contrasts conducted above, this is a categorical contrast identifying regions in which the response to speech masking is significantly different to both the unintelligible

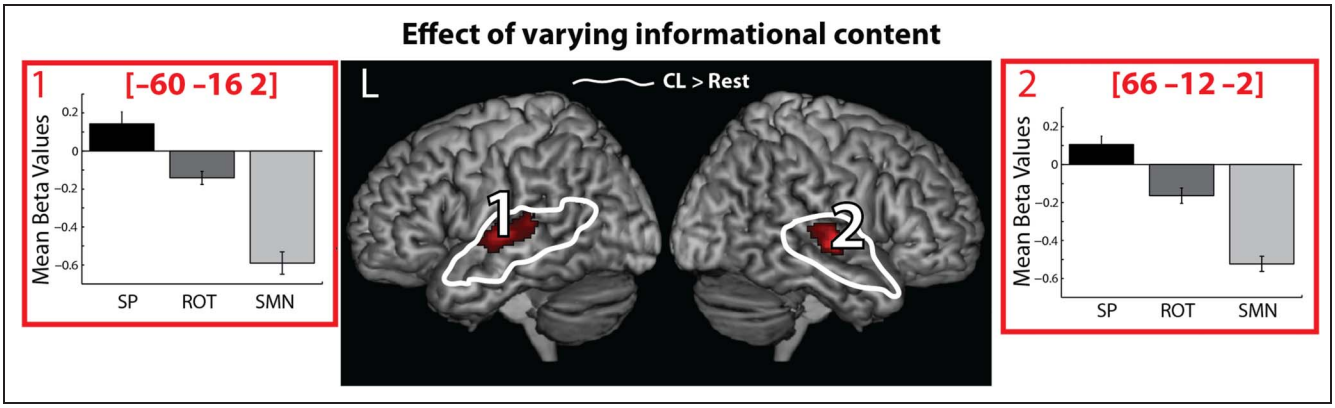

Figure 3. Regions showing increasing activation in response to masking sounds with increasing informational content. Bar graphs of the mean beta value for each condition with within-subject error bars representing one standard error (Loftus & Masson, 1994).

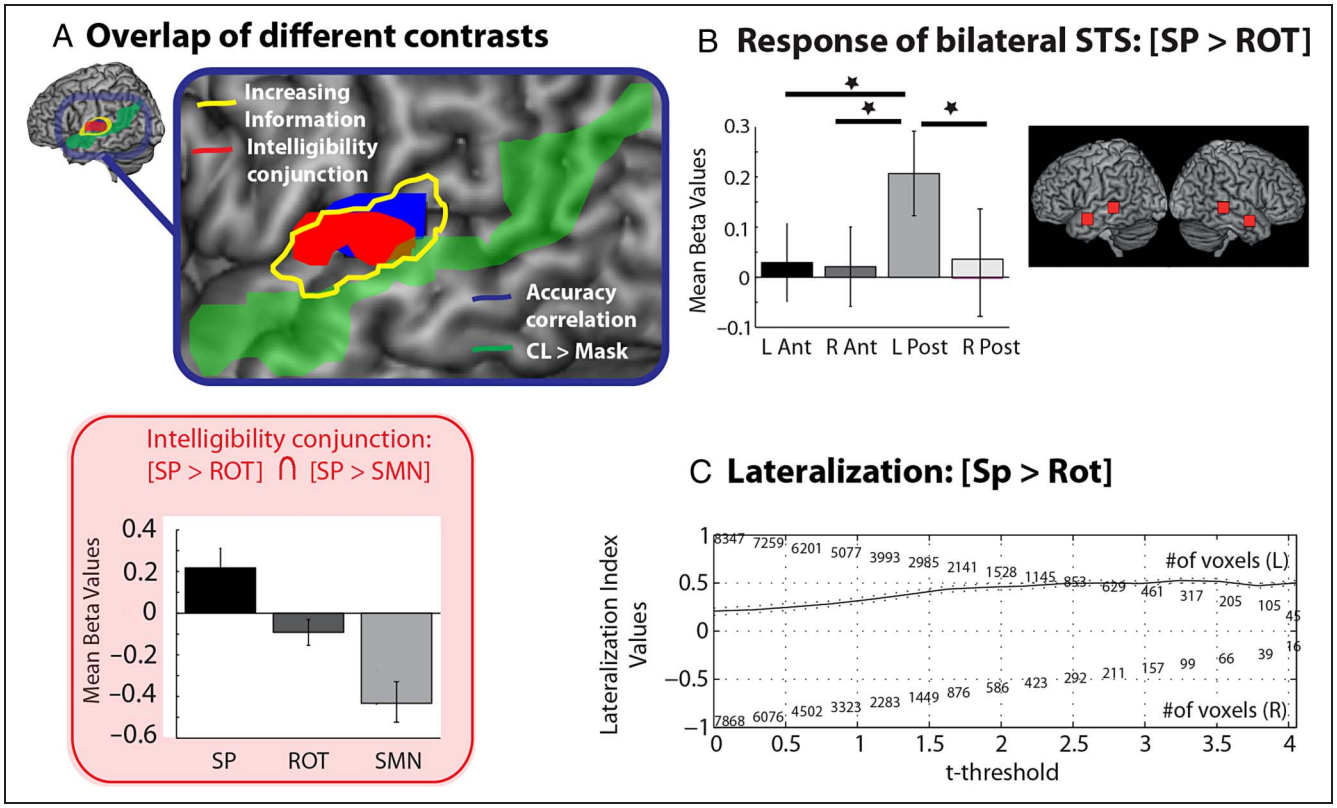

Figure 4. (A) Activation overlap map for the different contrasts, including the conjunction of [SP > ROT] ∩ [SP > SMN] in the red box with plot of the response. (B) ROI analyses comparing the neural response to the intelligibility of the masking stimulus [SP > ROT] for bilateral anterior and posterior STS. Plots show mean beta values for each condition with within-subject error bars representing one standard error (Loftus & Masson, 1994). (C) Lateralization curve for [SP > ROT] within temporal cortex.

masking conditions. This identified a cluster of activity in the left mid-posterior STG extending into the STS (see Figure 4A, red box). This region overlapped with cortical areas associated with better behavioral performance during masking (see Figure 4A, blue) and was found within the region responding to increasing informational content, as expected (Figure 4A, yellow). Note that this activity did not extend as far anterior or posterior within the STS as the response to CL (implicitly the subtraction of clear speech from masking; Figure 4A, green) or to [CL > Rest] (Figure 3, white), suggesting that intelligible masking speech was not processed equivalently to intelligible target speech. The response within the temporal cortex to the most closely controlled intelligibility contrast [SP > ROT] was submitted to laterality analyses. Reference to the weighted mean and laterality curve for the intelligibility of masked speech [SP > ROT] showed the response to be left lateralized in the temporal cortex (0.47; Figure 4C and Table 2).

We extracted the beta values in bilateral anterior and posterior STS for the contrast of [SP > ROT] to further address the extent to which intelligible competing sounds activated regions associated with the processing of intelligible target speech in quiet. The ROI locations were based on those used in Evans et al. (2014), albeit the anterior ROIs required a small change in location to account for the different field of view used in the two experiments. The ROIs were located in the left anterior ([−50 0 −16]), right anterior ([54 0 −18]), left posterior ([−62 −34 0]), and right posterior ([62 34 0]) STS. This analysis demonstrated that only the response in the left posterior STS was significantly modulated by the intelligibility of the masker, $t(16) = 2.386$, $p = .030$, and that the response in this region differed significantly to the left anterior, $t(16) = 2.419$, $p = .028$, right anterior, $t(16) = 2.542$, $p = .022$, and right posterior STS, $t(16) = 2.500$, $p = .024$ (Figure 4B).

Onset Responses

In an additional model (the Onset Model), we added events for the onset of clear speech and the onset of masking sounds which allowed us to identify activation associated with sound onsets. We began by examining the average effect of the onset of the different masking sounds. This analysis reflects neural activity associated with the onset of masking sounds in the context of an on-going background of scanner noise and target speech. Averaging over masking conditions, activation was found in the bilateral STG extending into the planum temporale and the inferior parietal lobule (Figure 5A, red rendering; Table 1). Activation was also found in the bilateral precuneus, the superior parietal lobule, SMA, middle and anterior cingulate, insula, the left inferior frontal gyrus

Table 2. Lateralization Index Values

Contrast and Region	*Trimmed Mean (±1 SD)*	*Weighted Mean*
SP > ROT		
Temporal	0.44 (0.06)	**0.47**
Masking (Epoch)		
Temporal	0.24 (0.17)	**0.44**
Frontal	0.03 (0.03)	0.06
Parietal	0.33 (0.23)	**0.60**
Masking Onset		
Temporal	0.12 (0.18)	**0.40**
Frontal	−0.33 (0.07)	**−0.41**
Parietal	−0.02 (0.07)	**0.20**
Clear Onset		
Temporal	−0.13 (0.03)	−0.10
Frontal	−0.32 (0.14)	**−0.45**
Parietal	−0.06 (0.04)	0.06

Mean values >0.2 or <−0.2 indicate a relative lateralization (in **bold**). Values expressed in radiological convention: Positive values represent a left lateralization, and negative values represent a right lateralization.

and pre- and postcentral gyrus, and right inferior and middle frontal gyrus, caudate nucleus and putamen. In an exploratory analysis, we shifted the onset response later in time by 1 sec to understand whether onset responses were altered. This did not change the results appreciably. To confirm these effects, we visualized the time course of masking responses, using a finite impulse response set (window length = 30 sec, order = 10) for peaks identified by the Epoch Model and those identified by the Onset Model. Plots from peaks associated with the onset of masking showed a phasic response, for example, a sharp increase at masking onset (which peaked ~6 sec) followed by a rapid decrease in activity as the epoch continued, by contrast plots from peaks identified by the epoch model evidenced a more sustained profile of activity (see onset peaks in Figure 5A, Plots 1 and 2, for comparison with the Epoch Plots 3 and 4). We then used the thresholded statistical map for the average effect of masking onset as a search volume to conduct an ANOVA differentiating activation between masking conditions. This revealed a cluster of activation in the superior parietal lobule ($p < .001$ voxelwise uncorrected, $q < 0.05$ FDR cluster-corrected). A plot from this region indicated increased activity to the onset of masking sounds with greater informational content (Figure 5B and Table 1). We tested this observation with a follow-up contrast [SP(discontinuous) × 0.5, SP(continuous) × 0.5, ROT × −0.25, SMN × −0.75]; this confirmed that the response within these regions reflected a response that was sensitive to increasing informational content.

We then assessed the effect of the onset of clear speech. Note that this analysis reflects neural activity associated with the onset of clear speech in the context of an on-going background of scanner noise. It therefore reflects, in a similar manner to masking onset, a response to the onset of an additional sound stream, however, rather than the onset of another sound in the context of scanner noise and target speech; it reflects the onset of an additional sound in the context of scanner noise alone. As expected, this gave rise to activation in similar regions to the onset of masking; clusters of activation were observed within bilateral STG, inferior frontal gyrus, SMA, inferior parietal lobule, anterior and middle cingulate, precuneus, insulae and right middle frontal gyrus, putamen, and thalamus. We further assessed the conjunction of masking onset and clear speech onset effects. This identified shared activity within bilateral superior parietal lobule, anterior and middle cingulate, STG, insulae, and right inferior frontal gyrus.

Finally, we examined the laterality of the epoch and onset responses. This showed that masking epoch responses were not lateralized in frontal regions (0.06) but were left lateralized in temporal (0.44) and parietal regions (0.60) (Table 2). By contrast, the masking onset responses were left lateralized in temporal (0.40) and parietal cortex (0.20) and right lateralized in the frontal cortex (−0.41) (Figure 4C and Table 2). The onset of clear speech was also right lateralized in the frontal cortex (−0.45) but showed no lateralization in temporal (−0.10) and parietal cortex (0.06).

DISCUSSION

In this study, we addressed how different kinds of masking sounds modulate neural responses. There were four main findings. First, we found that auditory attention and control networks were activated during attentive passive listening in noise. Second, competing speech was associated with left lateralized activity within the STG and was processed within the same processing pathway as speech in quiet but was not treated equivalently within that network. Third, increased activity in the left mid-posterior STG predicted performance on speech perception in noise tasks. Fourth, neural activity associated with the onset of sounds in the auditory environment engaged sensory and auditory attention and cognitive control networks—activation was right lateralized in frontal regions, and subregions within this wider network were modulated by the informational properties of these sounds. These findings, from young healthy adults, provide a basis for future work identifying how these systems

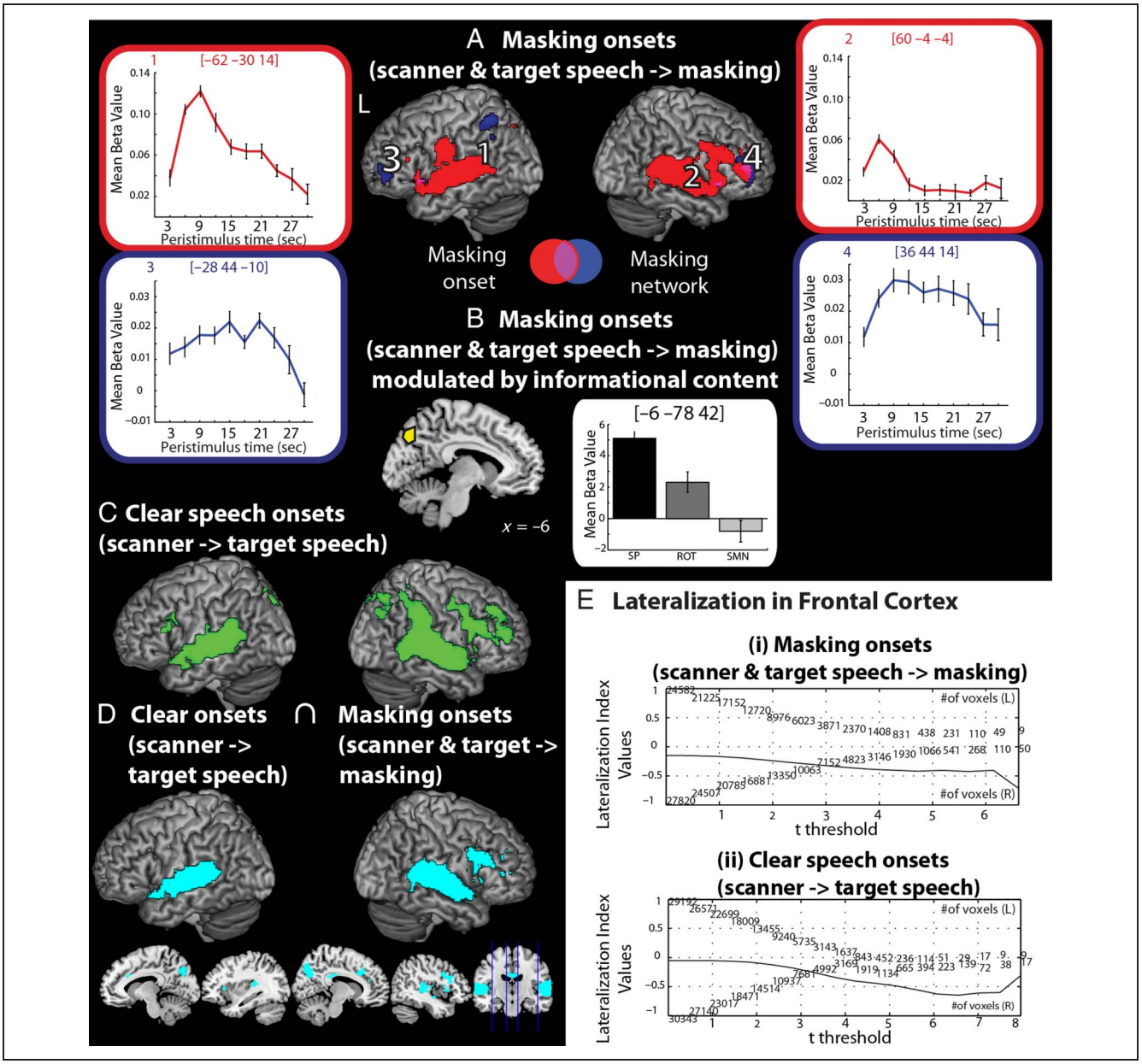

Figure 5. (A) Average effect of masking onsets—activation associated with the onset of masking sounds in the presence of on-going scanner noise and target speech. Red rendering shows the effect of masking onsets and the time course of the response in selected regions in Plots 1 and 2. The blue rendering shows regions responding more to masked epochs, as compared with clear speech, and the time course of the responses in selected regions in Plots 3 and 4. (B) Modulation of onset effects by informational content. (C) Clear speech onsets—activation associated with the onset of clear speech in the presence of on-going scanner. (D) Conjunction of clear speech and masking onsets. Bar graphs show the mean beta value for each condition with within-subject error bars representing one standard error (Loftus & Masson, 1994). (E) Lateralization curves for the frontal cortex for (i) masking onsets and (ii) clear speech onsets.

are impaired in individuals who find listening in noise difficult.

Sensory versus Executive Systems

Our results indicate a broad functional-anatomical delineation between sensory regions within bilateral STS that respond more to a single target speaker in the absence of additional masking as compared with the same speaker masked by other sounds, and regions beyond primary and secondary auditory cortex that show the opposite response. Greater activation within the STS to a single speaker is consistent with previous studies that have shown these regions to be sensitive to the intelligibility of heard speech (Evans et al., 2014; McGettigan et al., 2012; Okada et al., 2010; Awad et al., 2007; Scott, Rosen, Lang, & Wise, 2006; Spitsyna et al., 2006; Davis & Johnsrude, 2003). Elevated responses to masked speech in frontal, parietal, and cingulate cortex and the frontal operculum and insula are consistent with the association of these

regions with attentional and control processes (Vaden et al., 2013; Petersen & Posner, 2012; Wild et al., 2012; Dosenbach et al., 2008; Duncan & Owen, 2000) and auditory stream segregation (Teki et al., 2011; Cusack, 2005). Similar activation has been shown in previous studies of speech in noise and degraded speech perception more generally (Erb, Henry, Eisner, & Obleser, 2013; Adank, 2012; Hervais-Adelman, Carlyon, Johnsrude, & Davis, 2012; Osnes, Hugdahl, & Specht, 2011; Eisner, McGettigan, Faulkner, Rosen, & Scott, 2010; Wong et al., 2008, 2009; Zekveld et al., 2006; Scott et al., 2004). The recruitment of regions associated with cognitive control is striking in the absence of an overt active behavioral task. Recent work by Wild et al. (2012) suggests that frontal regions are engaged only when participants specifically attend to target speech in the presence of distractors (e.g., when they reflect on whether they understood a target sentence), rather than when they attend to the distractors instead of the target speech (e.g., when they monitor for a distractor stimulus). Our results replicate these findings by showing engagement of frontal regions when participants are asked to attend to target speech in the presence of distracting sounds. Furthermore, we extend them by showing that an active task, such as pressing a button to indicate the intelligibility of target speech on a trial-by-trial basis, is not essential in engaging effortful listening networks, provided that participants are asked to attend closely to target speech in the presence of distracting sounds. It seems unlikely given the Wild et al. result that we would have seen these frontal networks if we had not told participants that we would be asking them questions about what they had heard.

Modulation of Responses by Informational Content

Bilateral mid-posterior STG and the STS showed greater activity in response to the increasing informational content of masking sounds. Masking stimuli with increasing informational content have been argued to place greater emphasis on segregation and selection processes (Shinn-Cunningham, 2008), likely explaining this increased activity within regions associated with speech sound processing (Chang et al., 2010; Blumstein, Myers, & Rissman, 2005; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005). Extending our previous findings, an area within this wider region responded more to masking with speech than rotated speech, a non-speech baseline that is better matched to speech in acoustic complexity than modulated noise (Scott, Rosen, et al., 2009). This suggests that individuals are sensitive to the intelligibility of sounds that they are actively ignoring. The locus of response to intelligible speech measured in quiet extends along the length of the STS and is relatively left lateralized (Evans et al., 2014; McGettigan et al., 2012; Okada et al., 2010). Our results show that responses associated with the intelligibility of masked speech are also left lateralized, but found in posterior rather than anterior auditory fields. It is interesting to note that activation related to the intelligibility of the masker did not extend as far anterior or posterior as activation associated with the increased intelligibility of listening to a single speaker or to regions activated by clear speech as compared with "silence" (i.e., scanner noise). This suggests that, although additional speech streams are processed within broadly the same neural system as target speech in quiet, they are not processed equivalently, as the response does not enter the wider language processing system (Humphries, Love, Swinney, & Hickok, 2005). This may reflect the fact that the syntactic and other higher-order properties of masking speech are not actively processed. It is also consistent with the observation that "higher-order" regions of the auditory processing hierarchy track the target but not masking speakers (Golumbic et al., 2013), likely reflecting active suppression of unattended sounds in earlier regions of the processing hierarchy (Mesgarani & Chang, 2012).

Individuals who performed better at perceiving masked speech activated the left mid-posterior STG more. This region overlapped with areas modulated by increasing informational content in masked sounds and is in proximity to areas associated with speech sound representations. This may suggest that individuals who have better specified or more accessible representations of speech sounds are able to segregate target and competing sounds more effectively and may thus explain why individuals with language learning impairments perform poorly on masking tasks (Ziegler et al., 2005). Adjacent regions of STS responded more to the greater intelligibility of clear as compared with masked speech. This topographic organization of responses may reflect hierarchical processing within the ventral stream (Davis & Johnsrude, 2003), such that the STG plays an active role in separating out masked from target speech, with intelligible representations of target speech encoded in the adjacent sulcus. This may also explain why successful performance on masking tasks is paradoxically associated with stronger responses to masking sounds in the mid-posterior STG. Indeed, a number of individuals who performed poorly on masking tasks evidenced greater activity within this region in response to clear as compared with masked speech (i.e., beta values < 0 in response to Masking), suggesting that they may have found separating the masker from the target more difficult.

Masking Onset Responses

Regions of temporal, frontal, cingulate, and parietal cortex and the insula responded to the average effect of masking onset and the onset of clear speech. As scanner noise was always present, it is not possible to ascertain the extent to which these effects are specific to masked speech or reflect a more general sound onset response. Activation within auditory regions to sound onsets is consistent with previous studies that have used the vowel

continuity effect to identify sound onset effects (Heinrich, Carlyon, Davis, & Johnsrude, 2008, 2011). However, here we also observed additional activation in regions beyond the temporal lobes in areas associated with cognitive control and attention, consistent with likely modulation of attention. The transient right lateralized responses, which we observed in frontal regions, might be best described as a "phasic alerting" response—heightened arousal in readiness for subsequent stimulation. Phasic alerting is served by the neuromodulator norepinephrine and involves the locus coeruleus (the source of norepinephrine) and nodes in frontal and parietal areas (Petersen & Posner, 2012). Activity within this neural system is predominantly right lateralized with dorsolateral pFC, a key node in the phasic attention network (Périn, Godefroy, Fall, & de Marco, 2010; Sturm & Willmes, 2001). Indeed, right hemisphere regions play a crucial role in attention control. For example, hemispatial neglect, thought to result from damage to the intrinsic alerting system (Petersen & Posner, 2012), tends to be most severe and persistent following damage to the right hemisphere (Corbetta & Shulman, 2011) with the induction of phasic alertness shown to transiently improve neglect during visual tasks (Robertson, Mattingley, Rorden, & Driver, 1998) and to improve recognition accuracy in healthy participants during speech in noise tasks (Best, Ozmeral, & Shinn-Cunningham, 2007).

Onset effects in the superior parietal lobule were modulated by the informational properties of the masking sounds. Increased neural responses to the onset of masking speech (which is more highly confusable with target speech) as compared with noise masking may be suggestive that this region is functionally involved in stream segregation in the context of masked speech. Indeed, this would be consistent with previous studies implicating parietal regions (albeit more lateral and inferior than those described here) in the processing of multiple auditory streams (Teki et al., 2011; Cusack, 2005). However, further work in which neural responses to speech presented in absolute quiet are recorded is necessary to delimit the extent to which these processes are specific to the onset of speech masking rather than speech more generally. It may be the case that other imaging modalities are better suited to this endeavor given the ubiquitous presence of noise in fMRI scanning (Peelle, 2014).

Conclusions

To conclude, we have shown that auditory attention and control networks are activated during attentive listening in the absence of an overt behavioral task. We have provided evidence that the informational content of masking sounds modulates activity in the superior temporal cortex and have shown that, although competing speech is processed within the same pathway as speech in quiet, it is not treated equivalently within that system. We have also shown evidence for neural responses associated with sound onsets that are consistent with a phasic alerting response. These results provide a basis for describing the neural contribution of sensory and control processes in perception in noise in healthy adults, which may in turn inform our understanding of how these same processes are impaired in special populations.

Acknowledgments

We would like to thank Jonathan Peelle for advice on data analysis and Abigail Evans and Annabel Port for lending their voices to the audio recordings. In addition, we would like to thank the staff at the Birkbeck-UCL Centre for Neuroimaging for technical advice and assistance. This work was supported by the Wellcome Trust (WT074414MA to S. K. S.) and the Economic and Social Research Council (studentship to S. E.). We would also like to thank our anonymous reviewers for greatly improving the manuscript.

Reprint requests should be sent to Samuel Evans, Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, United Kingdom, WC1N 3AR, or via e-mail: samuel.evans@ucl.ac.uk.

REFERENCES

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: Two activation likelihood estimation (ALE) meta-analyses. *Brain and Language, 122,* 42–54.

Awad, M., Warren, J. E., Scott, S. K., Turkheimer, F. E., & Wise, R. J. S. (2007). A common system for the comprehension and production of narrative speech. *Journal of Neuroscience, 27,* 11455–11464.

Azadpour, M., & Balaban, E. (2008). Phonological representations are unconsciously used when processing complex, non-speech signals. *PloS One, 3,* e1966.

Badcock, N. A., Bishop, D. V. M., Hardiman, M. J., Barry, J. G., & Watkins, K. E. (2012). Co-localisation of abnormal brain structure and function in specific language impairment. *Brain and Language, 120,* 310–320.

Ben-David, B. M., Tse, V. Y. Y., & Schneider, B. A. (2012). Does it take older adults longer than younger adults to perceptually segregate a speech target from a background masker? *Hearing Research, 290,* 55–63.

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology, 8,* 294–304.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (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,* 5–41.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience, 17,* 1353–1366.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *Journal of the Acoustical Society of America, 137,* 378–387.

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, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain, Japan.
Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America, 109,* 1101–1109.
Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13,* 1428–1432.
Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119,* 1562–1573.
Corbetta, M., & Shulman, G. L. (2011). Spatial neglect and attention networks. *Annual Review of Neuroscience, 34,* 569–599.
Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience, 17,* 641–651.
Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23,* 3423–3431.
Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top–down influences on the interface between audition and speech perception. *Hearing Research, 229,* 132–147.
Deike, S., Gaschler-Markefski, B., Brechmann, A., & Scheich, H. (2004). Auditory stream segregation relying on timbre involves left auditory cortex. *NeuroReport, 15,* 1511–1514.
Dole, M., Hoen, M., & Meunier, F. (2012). Speech-in-noise perception deficit in adults with dyslexia: Effects of background type and listening configuration. *Neuropsychologia, 50,* 1543–1552.
Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top–down control. *Trends in Cognitive Sciences, 12,* 99–105.
Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23,* 475–483.
Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage, 25,* 1325–1335.
Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., & Scott, S. K. (2010). Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. *Journal of Neuroscience, 30,* 7179–7186.
Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *Journal of Neuroscience, 33,* 10688–10697.
Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *Neuroimage, 50,* 572–576.
Evans, S., Kyong, J. S., Rosen, S., Golestani, N., Warren, J. E., McGettigan, C., et al. (2014). The pathways for intelligible speech: Multivariate and univariate perspectives. *Cerebral Cortex (New York, N.Y.: 1991), 24,* 2350–2361.
Ezzatian, P., Li, L., Pichora-Fuller, M. K., & Schneider, B. A. (2012). The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues. *Language and Cognitive Processes, 27,* 1056–1088.
Freyman, R. L., Helfer, K. S., Mccall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *Journal of the Acoustical Society of America, 106,* 3578–3588.
Friederici, A. D., Kotz, S. A., Scott, S. K., & Obleser, J. (2010). Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping, 31,* 448–457.
Gelinas, J. N., Fitzpatrick, K. P. V., Kim, H. C., & Bjornson, B. H. (2014). Cerebellar language mapping and cerebral language dominance in pediatric epilepsy surgery patients. *Neuroimage. Clinical, 6,* 296–306.
Golestani, N., Hervais-Adelman, A., Obleser, J., & Scott, S. K. (2013). Semantic versus perceptual interactions in neural processing of speech-in-noise. *Neuroimage, 79,* 52–61.
Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77,* 980–991.
Green, T., Rosen, S., Faulkner, A., & Paterson, R. (2013). Adaptation to spectrally-rotated speech. *Journal of the Acoustical Society of America, 134,* 1369–1377.
Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., & Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. *Journal of Neuroscience, 27,* 13074–13081.
Hayasaka, S., Phan, K. L., Liberzon, I., Worsley, K. J., & Nichols, T. E. (2004). Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage, 22,* 676–687.
Heinrich, A., Carlyon, R. P., Davis, M. H., & Johnsrude, I. S. (2008). Illusory vowels resulting from perceptual continuity: A functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience, 20,* 1737–1752.
Heinrich, A., Carlyon, R. P., Davis, M. H., & Johnsrude, I. S. (2011). The continuity illusion does not depend on attentional state: fMRI evidence from illusory vowels. *Journal of Cognitive Neuroscience, 23,* 2675–2689.
Hervais-Adelman, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes, 27,* 1145–1166.
Hickok, G., & Poeppel, D. (2007). Opinion—The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393–402.
Huang, Y., Xu, L., Wu, X., & Li, L. (2010). The effect of voice cuing on releasing speech from informational masking disappears in older adults. *Ear and Hearing, 31,* 579–583.
Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human Brain Mapping, 26,* 128–138.
Lebel, C., & Beaulieu, C. (2009). Lateralization of the arcuate fasciculus from childhood to adulthood and its relation to cognitive abilities in children. *Human Brain Mapping, 30,* 3563–3573.
Lidzba, K., Schwilling, E., Grodd, W., Krägeloh-Mann, I., & Wilke, M. (2011). Language comprehension vs. language production: Age effects on fMRI activation. *Brain and Language, 119,* 6–15.
Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15,* 1621–1631.
Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. (2006). Speech perception problems of the hearing impaired

reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences, U.S.A., 103,* 5–8.
McGettigan, C., Agnew, Z. K., & Scott, S. K. (2010). Are articulatory commands automatically and involuntarily activated during speech perception? *Proceedings of the National Academy of Sciences, U.S.A., 107,* E42; author reply E43.
McGettigan, C., Evans, S., Rosen, S., Agnew, Z. A., Shah, P., & Scott, S. K. (2012). An application of univariate and multivariate approaches in fMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. *Journal of Cognitive Neuroscience, 24,* 636–652.
Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485,* 233–236.
Nagel, B. J., Herting, M. M., Maxwell, E. C., Bruno, R., & Fair, D. (2013). Hemispheric lateralization of verbal and spatial working memory during adolescence. *Brain and Cognition, 82,* 58–68.
Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., et al. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex, 13,* 1362–1368.
Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage, 25,* 653–660.
Norrelgen, F., Lilja, A., Ingvar, M., Åmark, P., & Fransson, P. (2015). Presurgical language lateralization assessment by fMRI and dichotic listening of pediatric patients with intractable epilepsy. *Neuroimage. Clinical, 7,* 230–239.
Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., et al. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex, 20,* 2486–2495.
Osnes, B., Hugdahl, K., & Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage, 54,* 2437–2445.
Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. (2003). Informational masking and musical training. *Journal of the Acoustical Society of America, 114,* 1543.
Pahs, G., Rankin, P., Helen Cross, J., Croft, L., Northam, G. B., Liegeois, F., et al. (2013). Asymmetry of planum temporale constrains interhemispheric language plasticity in children with focal epilepsy. *Brain: A Journal of Neurology, 136,* 3163–3175.
Peelle, J. E. (2014). Methodological challenges and solutions in auditory functional magnetic resonance imaging. *Frontiers in Neuroscience, 8,* 1–13.
Peelle, J. E., Johnsrude, I., & Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience, 4,* 1–3.
Périn, B., Godefroy, O., Fall, S., & de Marco, G. (2010). Alertness in young healthy subjects: An fMRI study of brain region interactivity enhanced by a warning signal. *Brain and Cognition, 72,* 271–281.
Peters, R. W., Hill, C., Carolina, N., Moore, B. C. J., & Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *Journal of the Acoustical Society of America, 103,* 577–587.
Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience, 35,* 73–89.
Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience, 12,* 718–724.
Robertson, I. H., Mattingley, J. B., Rorden, C., & Driver, J. (1998). Phasic alerting of neglect patients overcomes their spatial deficit in visual awareness. *Nature, 395,* 169–172.
Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.
Scott, S. K., & McGettigan, C. (2013). The neural processing of masked speech. *Hearing Research, 303,* 58–66.
Scott, S. K., McGettigan, C., & Eisner, F. (2009). OPINION A little more conversation, a little less action—Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience, 10,* 295–302.
Scott, S. K., Rosen, S., Beaman, P., Davis, J., & Wise, R. J. S. (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, S. K., Rosen, S., Lang, H., & Wise, R. J. S. (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, S. K., Rosen, S., Wickham, L., & Wise, R. J. S. (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.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12,* 182–186.
Spitsyna, G., Warren, J. E., Scott, S. K., Turkheimer, F. E., & Wise, R. J. S. (2006). Converging language streams in the human temporal lobe. *Journal of Neuroscience, 26,* 7328–7336.
Stone, M. A., Füllgrabe, C., Mackinnon, R. C., & Moore, B. C. J. (2011). The importance for speech intelligibility of random fluctuations in “steady” background noise. *Journal of the Acoustical Society of America, 130,* 2874–2881.
Studebaker, G. A. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research, 28,* 455–462.
Sturm, W., & Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *Neuroimage, 14,* S76–S84.
Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. *Journal of Neuroscience, 31,* 164–171.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.
Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *Journal of Neuroscience, 33,* 18979–18986.
Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience, 13,* 1–7.
Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *Journal of Neuroscience, 32,* 14010–14021.
Wilke, M., & Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. *Journal of Neuroscience Methods, 163,* 128–136.

Wilke, M., Lidzba, K., Staudt, M., Buchenau, K., Grodd, W., & Krägeloh-Mann, I. (2006). An fMRI task battery for assessing hemispheric language dominance in children. *Neuroimage, 32,* 400–410.
Wilke, M., & Schmithorst, V. J. (2006). A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data. *Neuroimage, 33,* 522–530.
Wilson, E. C., Melcher, J. R., Micheyl, C., Gutschalk, A., & Oxenham, A. J. (2007). Cortical fMRI activation to sequences of tones alternating in frequency: Relationship to perceived rate and streaming. *Journal of Neurophysiology, 97,* 2230–2238.
Wong, P. C. M., Jin, J. X. M., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia, 47,* 693–703.
Wong, P. C. M., Uppunda, A. K., Parrish, T. B., & Dhar, S. (2008). Cortical mechanisms of speech perception in noise. *Journal of Speech Language and Hearing Research, 51,* 1026–1041.
Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage, 32,* 1826–1836.
Ziegler, J. C., Pech-Georgel, C., George, F., Alario, F. X., & Lorenzi, C. (2005). Deficits in speech perception predict language learning impairment. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 14110–14115.