Distinct neural systems recruited when speech production is modulated by different masking sounds

Meekings, S., Evans, S., Lavan, N., Boebinger, D., Krieger-Redwood, K., Cooke, M. and Scott, S.K.

This is a copy of the final version of an article published in The Journal of the Acoustical Society of America 140, 8 (2016); doi: 10.1121/1.4948587. It is available from the publisher at:

http://dx.doi.org/10.1121/1.4948587

Distinct neural systems recruited when speech production is modulated by different masking sounds

Sophie Meekings, Samuel Evans, Nadine Lavan, Dana Boebinger, and Katya Krieger-Redwood
Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, United Kingdom

Martin Cooke
University of the Basque Country, Facultad de Letras, Universidad del País Vasco/EHU, Paseo de la Universidad 5, Vitoria, Alava 01006, Spain

Sophie K. Scott[a)]
Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT, United Kingdom

(Received 2 October 2015; revised 14 April 2016; accepted 21 April 2016; published online 5 July 2016)

When talkers speak in masking sounds, their speech undergoes a variety of acoustic and phonetic changes. These changes are known collectively as the Lombard effect. Most behavioural research and neuroimaging research in this area has concentrated on the effect of energetic maskers such as white noise on Lombard speech. Previous fMRI studies have argued that neural responses to speaking in noise are driven by the quality of auditory feedback—that is, the audibility of the speaker's voice over the masker. However, we also frequently produce speech in the presence of informational maskers such as another talker. Here, speakers read sentences over a range of maskers varying in their informational and energetic content: speech, rotated speech, speech modulated noise, and white noise. Subjects also spoke in quiet and listened to the maskers without speaking. When subjects spoke in masking sounds, their vocal intensity increased in line with the energetic content of the masker. However, the opposite pattern was found neurally. In the superior temporal gyrus, activation was most strongly associated with increases in informational, rather than energetic, masking. This suggests that the neural activations associated with speaking in noise are more complex than a simple feedback response. [http://dx.doi.org/10.1121/1.4948587]

[JFL] Pages: 8–19

I. INTRODUCTION

When two people try to strike up a conversation at a loud party, the background noise "masks" the sound of the talker's own voice, either by physically occluding the signal or by acting as a distractor, and leading to central competition for resources. In such a situation, the talker usually responds by changing the intensity, pitch, and spectral properties of her voice to make it more intelligible—a partly automatic response known as the Lombard effect (Lombard, 1911). Most neural research so far has assumed that the brain response to speaking in noise is driven by the energetic masking potential of the noise. However, there is behavioural evidence that suggests talkers are influenced differently by sounds with informational masking potential (Cooke and Lu, 2010). Here, we aimed to investigate if and how the presence of informational masking changes the way the brain responds to speaking in masking sound.

There are at least two properties of masking sound that influence the way that we speak over it. The first is its energetic potential. This describes how effectively the masker's acoustic properties interact with those of the signal, resulting in overlapping patterns of excitation at the periphery of the auditory system over time (Festen and Plomp, 1990; Stone *et al.*, 2012). Thus, the energetic masking potential of a noise is determined by acoustic properties such as its frequency spectrum and intensity relative to the signal (Brungart, 2001) and properties of random amplitude fluctuations (Stone *et al.*, 2011). Meanwhile, masking properties that cannot be explained by the energetic properties of the masking noise are described as its informational masking potential. An informational masker creates competition for more central cognitive, rather than peripheral resources, often because the sound contains some kind of salient or meaningful content that could distract the listener (Carhart *et al.*, 1969). Functional imaging studies of speech perception have established that informational and energetic maskers activate different neural systems. Consistent with the notion that informational masking is associated with greater competition for central resources, trying to understand speech masked by another talker results in bilateral activation of the superior temporal gyrus (STG) (Scott *et al.*, 2009). In contrast, listening to speech against an energetic masker is associated with

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

activations in prefrontal and posterior parietal cortex, which implies an increase in attentional rather than linguistic resources (Scott *et al.*, 2004).

The distinction between energetic and informational masking has not been widely studied in speech production, where research has tended to focus on the effects of speaking over energetic sounds such as white noise. This research has established that talkers respond to energetic masking by increasing their vocal intensity (Lombard, 1911), raising the pitch of their voice (Lu and Cooke, 2008; Schell, 2008), increasing word or vowel duration (Junqua, 1993; Summers *et al.*, 1988), and shifting energy to higher frequencies (Lu and Cooke, 2008; Varadarajan and Hansen, 2006). These changes effectively reduce the acoustic overlap between produced speech and the masking noise, and improve its intelligibility to others (Summers *et al.*, 1988). However, more recently, Cooke and Lu (2010) demonstrated that the Lombard effect is also influenced by the informational properties of the masker. Talkers are better at retiming their voices to accommodate spectral and amplitude dips in a speech masker, as compared to speech modulated noise (which has the same kind of amplitude dips, but no intelligible content). Although speaking in noise reliably causes vocal adaptation, the degree to which talkers change their voice is highly situation-dependent, with the greatest response always evoked by communicative contexts (Cooke and Lu, 2010; Garnier *et al.*, 2010).

Neuroscientific studies of the effect of talking over noise have typically characterised the effect of masking noise as reducing auditory feedback rather than as causing a communication problem. This approach equates speaking in noise with other altered-feedback approaches, such as delayed auditory feedback and pitch-shifted feedback. Functional neuroimaging research has found that when talkers hear their voice changed in these ways, they show increased activation in superior temporal cortex (Hashimoto and Sakai, 2003; Tourville *et al.*, 2008; Toyomura *et al.*, 2007). Two prominent models of speech production, DIVA and the Hierarchical State Feedback Model (Tourville *et al.*, 2008; Hickok, 2012), have interpreted such activation as an indication that this area, specifically posterior STG and planum temporale, is a critical site for sensorimotor integration and error detection. Although conceptually quite different (for example, in Hickok's model feedback is compared not to auditory goals directly, but to an internal model of the predicted consequences of motor commands), both models predict the same end result in terms of brain activation: that neurons in superior temporal cortex are less active when an auditory target is met, and excited when it is not; the greater the mismatch between target and feedback, the greater the activation. Specifically, both models incorporate a feedback loop where the talker's auditory feedback is compared to a target. When motor plans are sent to the articulators, a forward prediction (Hickok, 2012) or efference copy (Tourville *et al.*, 2008) is projected as an inhibitory signal to the sensory regions. This region then also receives excitatory input from sensory "state maps" (Tourville and Guenther, 2011), or from the activated auditory target (Hickok, 2012). Any mismatch between the signals is seen as excitation in sensory regions. If, on the other hand, the expected signal matches the actual sensory state, the two projections effectively cancel each other out.

This is thought to explain the "speech-induced suppression" response, in which temporal cortex responds less to speaking aloud than to hearing a recording of an equivalent vocalization. This response has been found in several studies of voice production in humans and non-human primates (Eliades and Wang, 2003; Flinker *et al.*, 2010; Houde *et al.*, 2002; Wise *et al.*, 1999), although a recent study (Agnew *et al.*, 2013) clarified that suppression was only clarified in anterior temporal regions, rather than the posterior temporal fields identified by speech models as the critical site for processing feedback.

One way of interpreting the Lombard response according to these models is as a response to reduced auditory feedback. The less well you can hear your own voice over the noise, the greater the difference between the auditory feedback you receive and your "auditory target," and so the more you change your voice (Christoffels *et al.*, 2007; Christoffels *et al.*, 2011). The map of auditory targets is suggested to lie in the posterior superior temporal cortex; therefore, the more effective the masker is at preventing you from hearing yourself (i.e., the greater its energetic masking potential), the greater the error signal and therefore activation within this region. Christoffels *et al.* (2011) tested this by asking participants to speak in successively louder levels of pink noise, and found that speaking over but not listening to higher levels of noise correlated with higher activity in the STG. However, these findings are potentially complicated by the nature of the task. Speaking in noise naturally prompts the Lombard response, which as we have previously noted improves intelligibility, presumably therefore reducing feedback mismatch. Christoffels *et al.* (2011, 2007) addressed this by asking participants not to raise their voices; another study by Zheng and colleagues (Zheng *et al.*, 2010) asked subjects to whisper. But the Lombard response is difficult to suppress (Pick *et al.*, 1989) and neither study considers or accounts for the costs involved in following these task instructions. Consequently, any activity seen may result from the cognitive effort associated with suppressing the participants' natural vocal response rather than from their response to feedback.

At present, therefore, our understanding of the neural underpinnings of typical human speech behaviour in noise rests on studies that asked their subjects to suppress that same speech behaviour. These studies are further limited by the fact that they looked only at single-syllable utterances made in steady-state noise. Since we rarely have to utter words in isolation, a study that strives for ecological validity should use connected speech, especially as this may be processed differently to single words. In speech perception, unconnected speech largely activates a less widely left-lateralized fronto-temporal network than connected speech (Peelle, 2012); it is possible that there is a similar distinction in speech production.

We therefore aimed to build on these speech production studies (Christoffels *et al.*, 2011; Zheng *et al.*, 2010), as well as our work on the perception of speech in masking sounds

(Scott *et al.*, 2006, 2009), by asking participants to read sentences aloud in different acoustic environments. We chose maskers that varied in their energetic and informational content to differentiate the neural effects of speaking over these different types of sounds. We recorded participants' voices without instructing them to change or suppress their responses to masking sound, and used this data to supplement our interpretation of the neural activation, thus enhancing the ecological validity of our speech production task. In addition, this experiment may help us better understand how speaking in noise relates to forward models of speech. If the brain is constantly evaluating match and mismatch based on the audibility of our voices as we speak in noise, we would expect activity in superior temporal cortex to be modulated primarily by the energetic content of the masker. If, by contrast, activity is affected by the informational content of the masker, this might indicate that linguistic content of competing sounds, rather than their audibility, is important in predicting neural responses.

II. METHODS: STIMULUS PREPARATION

Four different maskers were constructed: continuous white noise (WH), speech modulated (SM) noise, rotated speech (RO), and intelligible speech (SP) (see Fig. 1). These were intended to represent points on a continuum from strongly energetic, weakly informational masking to strongly informational, weakly energetic masking. White noise, which has equal energy across the range of frequencies, is an extremely effective energetic masker, but shares neither the spectral nor the amplitude profile of speech. SM stimuli were derived 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 SM was given the same long term average spectrum (LTAS) as the original speech. This was achieved by subjecting the speech signal to a spectral analysis using a fast Fourier transform (FFT) of length 512 sample points (23.22 ms) with windows overlapping by 256 points, giving a value for the LTASs at multiples of 43.1 Hz. This spectrum was then smoothed in the frequency domain with a 27-point Hamming window that was 2 octaves wide, over the frequency range 50–7000 Hz. The smoothed spectrum was then used to construct an amplitude spectrum for an inverse FFT with component phases randomized with a uniform distribution over the range 0–2π. The resulting signal, which sounds like a rhythmic rustling noise, has similar amplitude modulations as the speech signal used to derive it. Low amplitude sections ensure that SM is a less effective energetic masker than white noise (Cooke, 2006); however, it does not contain any phonetic information and is completely unintelligible; whilst it provides participants with some informational content (Bashford *et al.*, 1996) subjects did not identify this during the experiment. Next, RO was created by inverting the frequency spectrum around a centre frequency of 2 kHz (Blesser, 1972). As natural and spectrally inverted signals have different long-term spectra, the signal was equalized with a filter giving the RO approximately the same long-term spectrum as the original speech. Since RO can only contain energy up to twice the rotation frequency, all stimuli were low-pass filtered at 3.8 kHz, including the speech, to ensure a similar distribution of spectral energy across all the conditions. Rotated speech retains the spectral and amplitude modulations of the original speech signal but is unintelligible without extensive training (Blesser, 1972). It sounds like an "alien language" and has some phonetic features, a quasi-harmonic structure, and generates a sense of pitch. Rotated speech is a poorer

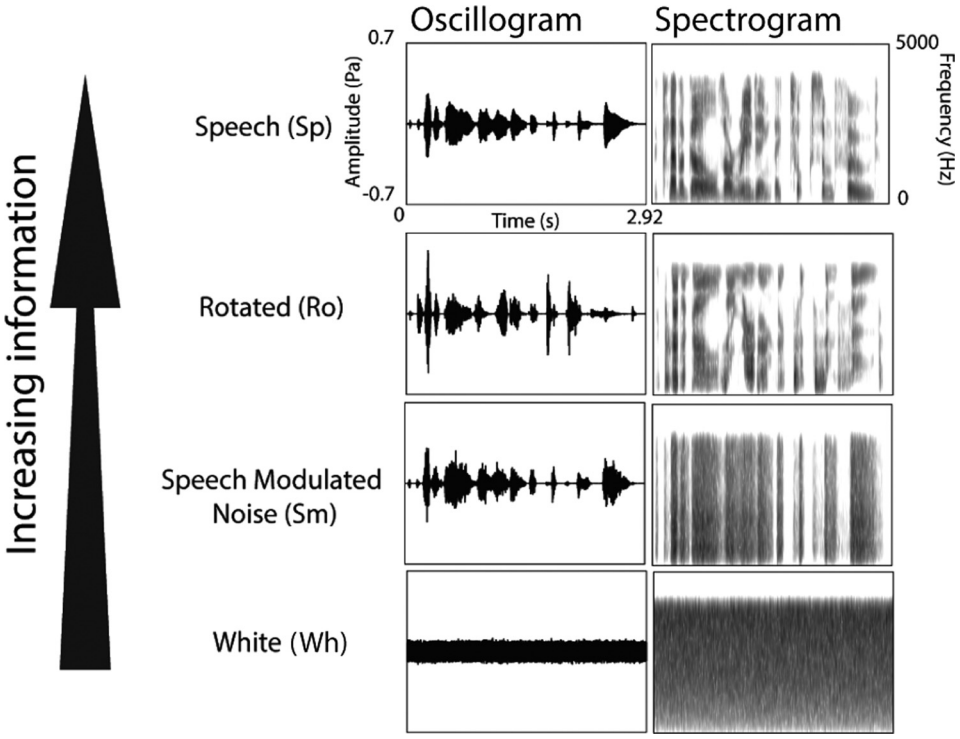

FIG. 1. Spectrograms and oscillograms of auditory stimuli.

energetic masker than SM as it contains spectral and amplitude modulations. SM by comparison has a relatively constant spectrum equal to the average long term spectrum of the speech stimuli. Finally, SP has high informational masking potential (including semantic and syntactic information) but contains spectral and amplitude modulations that render it a poor energetic masker. The resulting maskers are not intended to represent equal steps along the scale from high to low energy/information (for example, the difference in energetic masking potential between white noise and SM is likely to be much greater than that between SM and RO). Rather, the intention was to co-vary the energetic and informational properties of the four sounds, such that generally, the greater the sound's energetic masking potential, the lower its informational masking potential, and vice versa. We note, however, that theoretically RO has the same energetic properties as speech.

The SP maskers were 20 digital recordings (sampled originally at 22.05 kHz with 16-bit quantization) of a male and female talker reading from the BKB sentence lists (Bench *et al.*, 1979). These sentences were chosen as they contained simple vocabulary and syntax making it easier for talkers to comprehend and produce these sentences within the scanner in the interval between brain acquisitions. The BKB sentence lists consist of short sentences (maximum seven syllables) based on utterances from a language sample produced by young hearing-impaired children. The sentences are designed to be reasonably consistent in structure and complexity, with phrase structure constrained to the ten most commonly used structures in the language sample, and similar restrictions for morphology and vocabulary (Bench *et al.*, 1979). We included both male and female speakers to control for a possible gender effect, since in speech perception, same-gender maskers are more effective than opposite-gender maskers (Festen and Plomp, 1990). All the other maskers, with the exception of white noise, were derived from the SP stimuli, ensuring that all conditions were matched as closely as possible. All the stimuli were also root-mean-square (RMS) equalized.

Each experimental trial consisted of 2 consecutive BKB sentences (or manipulations thereof) with a silent interval of less than 30 ms between sentences. The duration of the white noise and silent trials was fixed to the mean duration of the other maskers (3.2 s). Behavioural piloting confirmed that 3.2 s was enough time for participants to respond and did not result in long silent periods. Concurrently with the auditory stimuli, subjects were visually presented with a sentence from the Institute of Hearing Research (IHR) lists (MacLeod and Summerfield, 1990) (examples of the IHR sentences are in the Appendix). The IHR sentences are based on the BKB sentence lists with similar syntax, vocabulary, and ratio of key words to function words. The words were presented in the middle of the screen in a large and clearly readable font. Participants always saw sentences regardless of whether they were being presented with a masker or not. The baseline condition was therefore reading silently in quiet. This was intended to control for higher order processes such as semantic processing involved in reading.

III. fMRI SCANNING—BEHAVIOURAL TASKS IN THE SCANNER

In the scanner, visual and auditory stimuli were displayed using MATLAB R2010b (Mathworks, Natick, MA) with the Psychophysics Toolbox extension (Brainard, 1997). Subjects listened to sounds presented through Sensimetrics S14 fMRI-compatible insert earphones (Sensimetrics Corp., Malden, MA), and spoke into an OptoAcoustics FOMRI-III noise-canceling optical microphone (OptoAcoustics Ltd., Israel), while viewing sentences projected onto an in-bore screen, using a specially-configured video projector (Eiki International, Japan). All the sounds were played at 84 dB sound pressure level as measured by a Brüel & Kjær 4153 artificial ear (Brüel & Kjær Sound & Vibration, Nærum, Denmark). Subjects practised the experiment outside the scanner on a laptop until they were comfortable with the task and were able to respond accurately and quickly.

Participants were trained to read aloud or silently, depending on the colour of the text presented on-screen. If the text was red, they spoke the sentence aloud. If it was black, they read it silently to themselves. At the same time, they heard one of the masking sounds, or silence. This gives us four main experimental tasks: reading silently, hearing nothing (Rest); reading silently, hearing sounds (Listen); reading aloud while hearing nothing (SpeakQuiet); and reading aloud while hearing sounds (SpeakNoise) (see Fig. 2). The SpeakNoise condition consisted of four separate conditions, one for each of the masking noises: SP, RO, SM, and WH. The Listen task was one condition composed of a combination of sounds from the four masking conditions. Because of constraints on experiment duration and participants' attention, we made the choice to include one listening condition containing all of the maskers, rather than four separate listening conditions, one for each of the maskers. This was intended as an approximate control for activation caused by auditory processing in the SpeakNoise condition (caused by hearing the different masking sounds).

In SpeakNoise trials, participants spoke for the duration of the masking sound; if they spoke after the noise had finished these trials were excluded from acoustic analysis and were recoded in the design matrix (see Fig. 2). SpeakQuiet trials were excluded if participants continued to speak for longer than 3.2 s (the average trial length for the noise), or if they failed to obey the task instructions (speaking when they were meant to remain silent or being silent when they were meant to speak). These errors occurred very infrequently (mean number errors per participant = 3 of 270 trials, min = 0/270, max = 10/270) except in the case of two excluded participants.

Participants were told to speak as clearly as possible when reading aloud as someone within the console room would be scoring their speech intelligibility, as heard over the intercom. They were not specifically prompted to speak loudly.

IV. fMRI SCANNING

A. Participants

Ethical approval was granted by the UCL Psychology Research Ethics Committee. Written consent was obtained

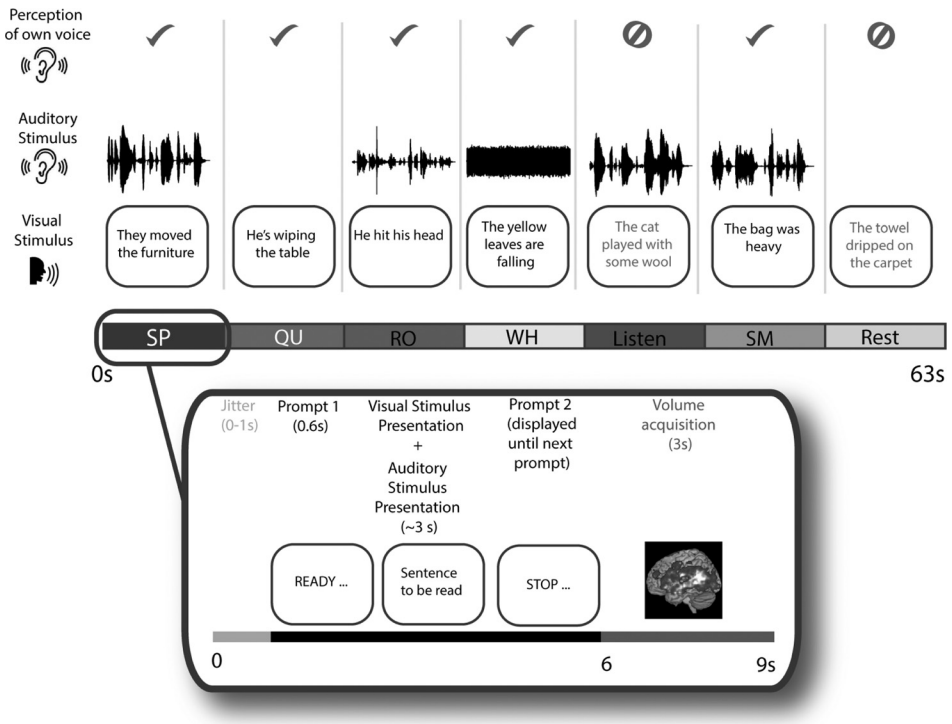

FIG. 2. Experimental procedure and fMRI time sequence.

from 16 right-handed native British English talkers (7 females, 9 males; aged 21–38; mean age 29). All participants spoke with a Southern British English accent and reported no history of hearing or language impairment. Two participants (one male and one female) did not consistently follow the task instructions (i.e., remained silent when they were meant to speak or spoke when they were meant to listen) and were excluded. The analysis was conducted on the remaining 14 subjects (6 females, 8 males).

B. Image acquisition

Participants took part in 2 functional runs, each consisting of 20 trials per condition (SP, RO, SM, WH, SpeakQuiet, Listen) and 15 ReadSilently baseline trials, making a total of 135 trials per subject. Every trial consisted of 2 sounds (or a silent period) lasting about 3.2 s on average with 1 sentence presented on the screen for the subject to read. Masking stimuli were repeated across runs, but the visually presented sentences were all unique. The conditions were randomly permuted in sets of six such that each condition was represented once every six trials. This ensured that at most there could be a single consecutive repetition of a particular condition type. The 15 silent trials, which constituted an implicit resting baseline, were distributed at regular but unpredictable intervals throughout each run.

To ensure that the stimuli were presented in silence and to minimize movement and susceptibility artefacts caused by the subjects speaking, slow sparse acquisition was used. Each trial was randomly jittered by 0, 0.5, or 1 s. Participants then saw a visual prompt "READY…" which lasted 0.6 s, followed by the presentation of a sentence displayed on screen for the participant to read for the duration of the masking sound (or 3.2 s in the case of the quiet and listen conditions). A "STOP" prompt was displayed following the offset of the sentence and was displayed during the volume acquisition until the subsequent "READY…" prompt.

Subjects were scanned on a 1.5T MRI scanner (Siemens Avanto, Siemens Medical Systems, Erlangen, Germany) with a 32- channel head coil. Functional MRI images were acquired using a T2-weighted gradient-echo planar imaging sequence, which covered the whole brain ($TR = 10$ s, $TA = 3$ s, $TE = 50$ ms, flip angle 90°, 35 axial slices, matrix size $= 64 \times 64 \times 35$, $3 \times 3 \times 3$ mm in-plane resolution). High-resolution anatomical volume images (HIRes MP-RAGE, 160 sagittal slices, matrix size: $224 \times 256 \times 160$, voxel size $= 1\,\text{mm}^3$) were also acquired for each subject. The field of view was oblique angled away from the eyes (to avoid ghosting artefacts from eye movements) and included the frontal and parietal cortex at the expense of the inferior temporal cortex and inferior cerebellum.

C. fMRI preprocessing and whole-brain analysis

Functional and structural images were analysed using Statistical Parametric Mapping (SPM 8).

The first three functional volumes of each run were discarded to allow for T_1 saturation effects. Scans were realigned to the first volume by six-parameter rigid-body spatial transformation. The mean functional image was written out and co-registered with the $T1$ structural image. The estimated translation (x, y, z) and rotation (roll, pitch, yaw) parameters that resulted from motion correction were inspected. These did not exceed 3 mm or 3° in any direction.

Scans were spatially normalized into MNI space at $2\,\text{mm}^3$ isotropic voxels using the parameters derived from the segmentation of each participant's $T1$-weighted scan, and smoothed using a Gaussian kernel of $8\,\text{mm}^3$ at full-width-

half-maximum to ameliorate differences in intersubject localization.

First-level analysis was carried out modeling the conditions of interest: Speech in noise: (1) SP, (2) RO, (3) SM, (4) WH, (5) SpeakQuiet (QU), and (6) Listen (LI), all with silent trials as an implicit baseline. In addition, first-level contrasts were generated for each of the speech production conditions (SP, RO, SM, WH, QU) with Listen as the baseline. The model also included 11 motion parameters of no interest and a Volterra expansion of those parameters, shown previously to reduce movement related artefact (Lund *et al.*, 2005). Events were modelled from the coincident presentation of the written text with sound using a canonical hemodynamic response function. For each condition in which spoken output was required, a parametric regressor modelled variation in RMS amplitude of the speech produced on each trial, measured *post hoc* using the within scanner recordings. As a proxy for vocal change induced by speaking in noise, this removed neural activity associated with within condition variance in vocal loudness (Wood *et al.*, 2008). This was likely to be greater in the speaking in quiet condition (in which participants could vary their voice unsystematically) than the speaking in noise condition (in which participants altered their voice specifically in response to masking sounds). Hence, by modeling out *within* condition variance in neural responses using parametric regressors we hoped to more sensitively identify differences in mean activity *between* conditions. Errors were coded as an additional regressor and the event was removed from the appropriate condition regressor.

These contrasts were taken up to a second level random effects model to create two repeated measures analysis of variances (ANOVAs): one looking at the difference between BOLD responses during the three different tasks (SpeakNoise, SpeakQuiet, and Listen) with Rest as the baseline, and another looking at differences between responses to speaking in the different masking conditions (SP, RO, SM, and WH) relative to Listen (as an attempt to control for the fact that when participants spoke in masking sound, they were hearing more than just their own voice). At the group level, contrasts were thresholded using a voxel wise familywise error (FWE) rate correction for multiple comparisons at $p < 0.05$. Statistical images were rendered on the normalized mean functional image for the group of participants.

V. RESULTS

A. Behavioural results

Audio recordings from the scanner were edited to remove silent periods at the start and end of each trial, and analysed using Praat (Boersma and Weenink, 2008). There was a very quiet repetitive noise in the background from the scanner helium pump, which was filtered out using the method described by Rafii and Pardo (2008). Any residual noise that survived the filter was distributed equally across conditions so should not affect interpretation of the data. The data extracted were evaluated using IBM SPSS Statistics (version 20).

The following acoustic parameters were extracted: mean intensity (measured in dB relative to the auditory threshold), median *F*0, spectral centre of gravity (CoG), mean harmonic-to-noise ratio (HNR), mean duration, and spectral standard deviation.

*F*0 was computed using the auto-correlation method, with pitch floor set at 75 Hz and pitch ceiling at 1000 Hz. Changes in pitch were assessed using the median, as the pitch estimation was less affected by outliers caused by occasional failure to accurately track the pitch of the utterances using the automated pitch tracking algorithm within Praat. Spectral CoG and standard deviation (calculated using the power spectrum) were used to track changes in the distribution of energy across the spectrum. Spectral dispersion, or standard deviation, measures whether the energy is concentrated mainly around the CoG, or spread out over a range of frequencies. The spectral CoG is the frequency which divides the spectrum into two, such that the amount of energy in both parts is equal. Previous studies (Lu and Cooke, 2008; Varadarajan and Hansen, 2006) have found that Lombard speech is characterized by an energy shift to higher frequencies, meaning that in this study we would expect to see a higher CoG in speech produced in masking noise compared to speech in quiet. Mean HNR was the mean ratio of quasi-periodic to non-periodic signal across time segments. Increases in HNR are associated with a perceptually "clear" voice (Warhurst *et al.*, 2012). Mean duration was evaluated after the sentences had been manually trimmed for silence at the beginning and end of an utterance. Talkers sometimes exhibit a slower duration or speech rate in Lombard speech (Pittman and Wiley, 2001; but cf. Varadarajan and Hansen, 2006), and have likewise been found to slow their speech rate in studies of clear speech produced to counter adverse listening conditions (Picheny *et al.*, 1986).

We used a linear mixed model to investigate the relationship between masking condition and acoustic properties of speech, with condition as a fixed effect, crossed random effects for subjects and sentences read, and a by-subjects random slope for the effects of condition. This was intended to handle the correlated subject data and address the fact that both subjects and sentences are sampled from a larger population (Barr *et al.*, 2013; Clark, 1973).

This model showed no effect of masking condition on spectral CoG ($F(4,61) = 1.51$, $p = 0.209$), mean HNR ($F(4,53.8) = 1.85$, $p = 0.132$) or median pitch ($F(4, 2454) = 0.476$, $p = 0.754$). However, intensity was significantly affected by masking condition ($F(4, 54) = 24.15$, $p < 0.001$). Sidak-corrected *post hoc* comparisons revealed that intensity was significantly greater in ROT, SM, and WH than SP or QU ($p < 0.001$). There were no significant differences between SP and QU ($p = 0.989$). There was a statistically significant linear trend ($F(1,13) = 7.85$, $p = 0.015$, $\eta_p^2 = 0.377$) in which intensity increased as the energetic content of the masker increased (see Fig. 3). There was also a significant effect of masking condition on spectral standard deviation ($F(4,60.17) = 3.50$, $p = 0.012$), caused by a significant decrease in spectral standard deviation in the SM condition compared to SP (see Fig. 3). There were no other significant differences between conditions. A significant effect of masker on mean duration ($F(4,58.4) = 2.208$,

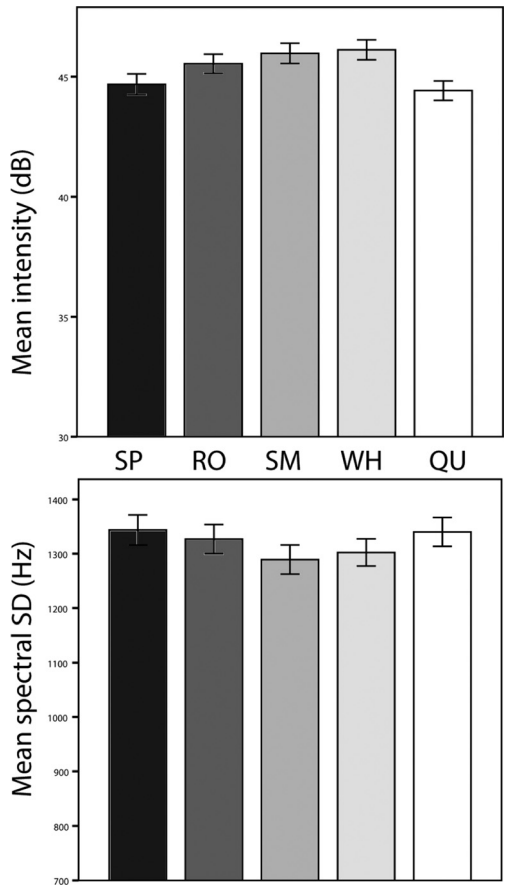

FIG. 3. Intensity and spectral CoG in each of the four masking conditions and quiet. Error bars indicate 95% confidence intervals.

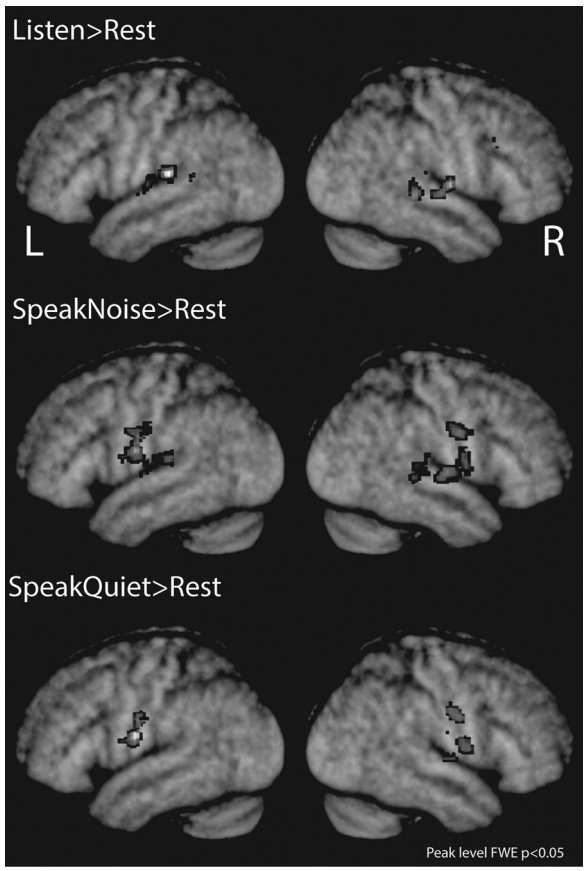

FIG. 4. Each of the three task conditions (Listen, SpeakQuiet, SpeakNoise) contrasted with silent reading. Contrasts shown on the mean normalised brain image of all participants at FWE $p < 0.05$.

$p = 0.016$) was driven by a trend toward increased duration in the masking conditions compared to quiet, but these differences did not survive Sidak correction for multiple comparisons.

B. fMRI results

The perception of sounds (speech, rotated speech, speech modulated noise, and white noise) was associated with activation of the dorsolateral temporal lobes (including superior temporal gyri) (see Fig. 4). In contrast speech production (in silence and in masking sound) was associated with activation in auditory and sensorimotor cortical fields (see Fig. 4). To look more specifically at the differences between tasks, we conducted an F-test, FWE-corrected at $p < 0.05$ (whole brain level) (see Fig. 5). This confirmed that activation in the bilateral postcentral gyri was significantly greater in the two speaking conditions than in the Listen condition, with no significant differences between SpeakQuiet and SpeakNoise. In temporal cortex, activation was seen bilaterally in regions covering most of the STG with peaks at [−52 −28 10] and [−60 −30 18] in the left, and [50 −28 12] and [54 −18 8] in the right (see Table I). Across these regions, the response to the SpeakNoise condition was significantly greater than to SpeakQuiet or Listen.

We saw a response that could be characterised as speaking-induced suppression in bilateral STG, where speaking in quiet resulted in a reduction of activity relative to passive listening. Although this difference was only statistically significant in the left hemisphere a comparison of the activation at peak voxels in STG identified by the whole brain analysis using a two-way repeated measures ANOVA revealed no significant effect of hemisphere ($F(1,13) = 0.188$, $p = 0.67$, $\eta_p^2 = 0.014$), or any significant task*hemisphere interaction ($F(2,26) = 2.45$, $p = 0.106$, $\eta_p^2 = 0.159$), indicating that there was no significant lateralization of brain response to speech in quiet vs listening at these locations in the STG.

Activation was also seen bilaterally in postcentral gyri and in cerebellar lobule VI (see Fig. 5). In these regions, responses were significantly greater in the two speaking conditions than in the listening condition; there were no significant differences between the two speaking conditions.

Next, to establish modulation of brain activity associated with speaking in the different maskers, we conducted an F-test at the whole brain level (FWE corrected at $p < 0.05$) looking at the differences between each of the speech production conditions (SP, RO, SM, WH, and QU), contrasted with listening as a baseline (Fig. 6). This was intended to factor out activation in auditory areas caused by just hearing the masking noise, and reveal only areas that were associated with the act of speaking in noise.

The analysis revealed activation in the bilateral superior temporal cortices and left middle temporal gyrus (see Table II). In both left and right temporal cortices the response was greatest for talking over speech, with activation decreasing in line with the amount of informational content in the masker.

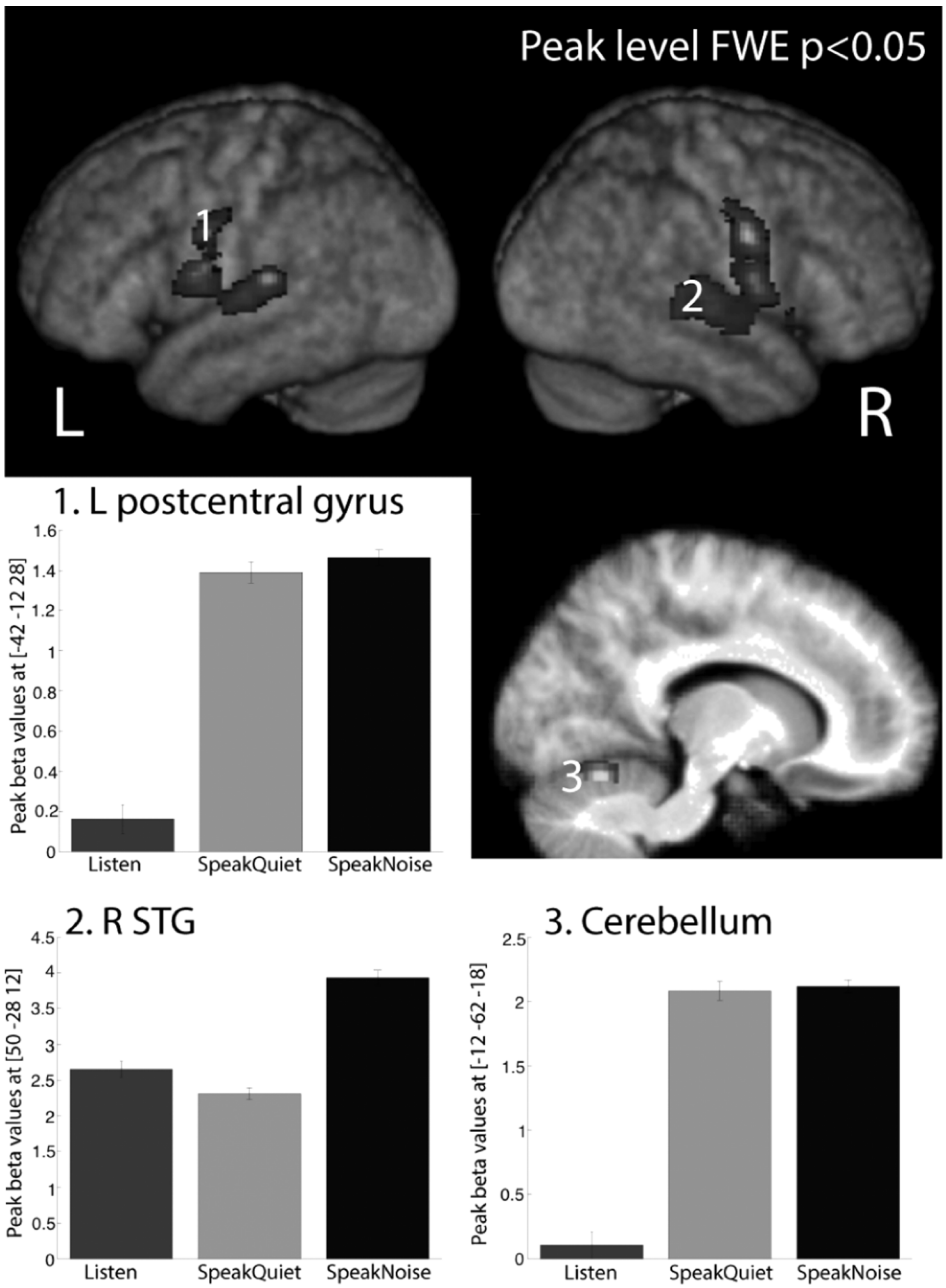

FIG. 5. Differences between the three task conditions (Listen, SpeakQuiet, SpeakNoise), shown on the mean normalised brain image of all participants at FWE $p < 0.05$. Bar graphs show beta values at peak co-ordinates. Error bars represent 95% confidence intervals.

At peak [−58 −12 2] in the left STG, a one-way repeated measures ANOVA revealed a significant effect of masking condition ($F(1.5, 19.6) = 61.8$, $p < 0.001$, $\eta_p^2 = 0.826$). Sidak-corrected *post hoc* tests showed that responses in the QU and WH conditions were not significantly different from each other ($p = 1.0$), and there was also no significant difference between responses in the QU and SM conditions, though this was marginal ($p = 0.053$). One-sample *t*-tests with a test value of 0 (representing the listening baseline) showed that activity in the QU and WH conditions were not significantly different from baseline; all other conditions were significantly different from the baseline and from each other ($p < 0.05$). In the right hemisphere, at peak [62 −16 6] in the STG, a similar pattern of activation was seen. Neither WH nor QU were significantly different from baseline. However, there was a significant effect of masking ($F(1.6, 20.8) = 63.7$, $p < 0.001$, $\eta_p^2 = 0.831$), and Sidak-corrected *post hoc* tests confirmed that all conditions were significantly different to each other ($p < 0.05$).

At the whole brain level we did not see any regions that responded most to energetic masking. To more sensitively address the response at locations in which speech induced suppression was identified, we conducted a region of interest (ROI) analysis at peaks in which less activation was seen in the SpeakQuiet condition relative to Listen and SpeakNoise. From the task ANOVA two peaks were identified as fitting this profile, one in the left STG at [−52 −28 10] and one in the right STG at [52 −28 10]. A spherical ROI of radius 8 mm (the size of the smoothing kernel) was built around each of these points using the MarsBaR toolbox for SPM (Brett *et al.*, 2002). Within each of the two ROIs an ANOVA was carried out to evaluate differences between the SpeakNoise conditions (SP, ROT, SMN, WH) relative to the baseline of silent reading.

TABLE I. Peak voxel co-ordinates revealed by an ANOVA comparing the three task conditions (SpeakNoise, SpeakQuiet, and Listen), with the Rest condition as a baseline. Corrected for multiple comparisons at FWE $p < 0.05$.

Anatomy	Voxels (k)	Z-score	X	y	z
Cerebellum Lobule VI	726	7.36	−12	−62	−18
Cerebellum Lobule VI		7.11	12	−64	−16
Left postcentral gyrus	2747	6.85	−42	−12	28
Left STG		6.65	−52	−28	10
Left STG		6.53	−60	−30	18
Right STG	2751	6.74	50	−28	12
Right postcentral gyrus		6.64	58	−4	36
Right STG		6.23	54	−18	8
	13	5.42	10	−28	−6
Left Insula	27	5.37	−34	8	4
Right Pallidum	57	5.34	28	−4	−6
Right Pallidum		5.18	28	−12	−2
Right Insula	32	5.29	40	12	6
Thalamus- parietal	3	4.96	−12	−26	−4
Right inferior frontal gyrus	8	4.95	54	14	0

In the left STG ROI, one-way repeated measures ANOVAs revealed a significant effect of masking condition ($F(3,39) = 35.424$, $p < 0.001$, $\eta_p^2 = 0.732$); Sidak-corrected *post hoc* tests showed significant differences between all conditions except for SM and WH. There was a statistically significant linear trend in which greater BOLD responses were seen for maskers with more informational content ($F(1,13) = 54.65$, $p < 0.001$, $\eta_p^2 = 0.808$). There was also a significant effect of masking condition in the right STG ROI ($F(3,39) = 17.428$, $p < 0.001$, $\eta_p^2 = 0.573$). *Post hoc* Sidak-corrected t-tests showed that while there were no significant differences between responses to SP and ROT, or between SM and WH, all other conditions were significantly different from each other ($p < 0.05$). There was also a statistically significant linear trend in the data ($F(1,13) = 31.194$, $p < 0.001$, $\eta_p^2 = 0.706$), with BOLD responses increasing in line with the informational content of the masker.

In this analysis, we found no neural profiles that correlated with the direction of behavioural vocal modification, i.e., where the greatest response was to talking in continuous noise, and the weakest response was to speaking against another talker. The contrast WH > SP, designed to test for regions that responded more to speaking in energetic than informational masking, also revealed no activation even at a weak threshold of uncorrected $p < 0.0005$.

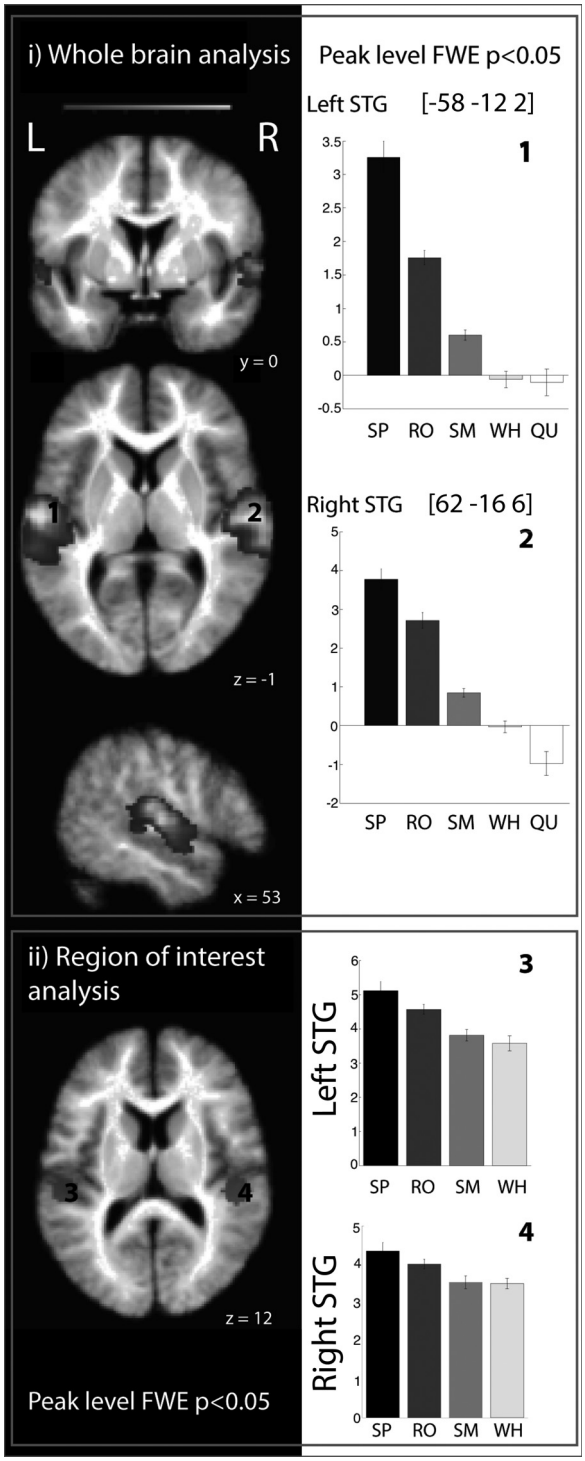

FIG. 6. Neural difference between the four masking conditions compared to Listen as a baseline, projected on group mean brain image. Bar charts show beta values at peak co-ordinates; error bars represent 95% confidence intervals.

VI. DISCUSSION

Contemporary neural accounts of speech production propose that superior temporal cortex acts as an auditory error monitor during talking. When what we hear does not match up with what we intended to say, the error monitor registers this and sends a corrective signal; conversely, if there is no mismatch, this activation is suppressed. Previous studies have found increased activation in superior temporal cortex when subjects speak in continuous noise compared to speaking in quiet, which has been interpreted as supporting this theory. In this study, we aimed to interrogate this response further. Specifically, we were interested in whether the type of background noise would have an effect on neural responses—and in which direction. If the brain cares more about the audibility of auditory feedback, we would expect to see the greatest response to sounds with high energetic masking potential, as these are the most effective at

TABLE II. Peak voxel co-ordinates revealed by an ANOVA comparing the five speech conditions (QU, SP, RO, SM, WH) with the Listen condition as a baseline. Corrected for multiple comparisons at FWE $p < 0.05$.

Anatomy	Voxels (k)	Z-score	x	Y	z
Left STG	2302	Inf	−58	−12	2
Left STG		6.56	−44	−30	12
Middle temporal gyrus		6.52	−60	−32	8
Right STG	2289	Inf	62	−16	6
Right STG		7.77	64	−6	0
Right STG		7.37	52	−24	14
Right STG	7	5.07	50	−46	16

occluding your voice. If, however, the greatest response were to sounds with informational masking potential, this might reflect mechanisms for monitoring and using linguistic information implied by behavioural studies showing that we adopt different strategies when talking over intelligible background noise. Consistent with other studies, we found that overall, responses in bilateral STG were greatest for speaking in masking sounds compared with listening and speaking in quiet, with a suppression response for speaking in quiet relative to listening. However, when the differences between masking conditions were examined, it became apparent that the speech-in-noise response was driven by the informational rather than the energetic masking potential of the background noise. Responses to white noise were not significantly greater than listening, and there was a linear relationship between the degree of activation and the informational content of the masker.

The STG is a functionally heterogeneous region so it is possible that the peaks in the condition ANOVA do not represent areas involved with feedback processing. To investigate this we constructed ROIs in left and right temporal lobes centred on areas that showed the feedback response profile of suppression when speaking in quiet compared to speaking in noise and to listening. These regions also demonstrated an enhanced response to informational content, with the speaking in white noise condition not significantly different to the Listen condition, and increasingly greater activation seen for maskers with more informational content. This makes the simple interpretation of a suppression effect as a feedback response hard to sustain. The relative deactivation in white noise compared to other maskers might be explained by the behavioural data—on average, talkers increased their vocal level most in white noise. This increased amplitude will have improved the signal-to-noise ratio, potentially causing a move back toward the activation patterns seen in quiet, as has been observed in macaques (Eliades and Wang, 2012). Although talkers also change their voices in the other masking conditions, they do so less than they do in the white noise condition, but show more neural activation, in a manner linked to the informational content of the masker. This pattern is similar to that found in studies of speech perception during informational and energetic masking (Scott *et al.*, 2012; Scott *et al.*, 2004), so this may indicate a similar route for central auditory processing of informational maskers in production and perception. Unattended words can prime a semantically related attended target (Aydelott *et al.*, 2015; Rivenez *et al.*, 2006), and can intrude into speech production (Saito and Baddeley, 2004). This suggests both that there is considerable central processing of "unattended" information (consistent with information masking accounts) and also that there is considerable competition between activated lexical items when a talker is speaking: both of these factors likely contribute to this enhanced STG activation when a talker speaks against the sound of another's speech.

Behaviourally, we found that talkers increased the RMS amplitude of their voice in masking sounds compared to quiet, and there were also differences between adaptations to different conditions. Notably, several acoustic responses to speaking in noise relative to quiet that have been observed in other studies (Cooke and Lu, 2010; Lu and Cooke, 2008) such as increased spectral CoG and increased pitch, were not seen here. This may be because of physiological considerations—the subjects were lying supine in the scanner, which affects vocal tract shape and articulator positions (Kitamura *et al.*, 2005). Alternatively, participants may not have been motivated to maximize their communicative efforts (despite being told they were being scored for intelligibility) because they were vocalizing on their own in a darkened room. Although Lombard speech occurs in the absence of a conversational partner, it is significantly modulated by communicative intent (Garnier *et al.*, 2010). Since exploring communicative adaptations is of critical interest here, it is important to develop more interactive experimental paradigms—perhaps allowing the participant to directly speak to a partner in the control room via audio or video link-up.

These findings demonstrate that masking sounds do not solely affect speech production mechanisms by reducing the talker's ability to self-monitor. Instead, these data suggest a dominant cortical effect of informational masking during speech production: talkers process unattended speech to a high cortical level. This is highly congruent with the pattern seen during speech perception, where masking speech leads to extensive activation in bilateral superior temporal lobes, in addition to the activation seen to attended speech. This strong cortical effect of informational masking may underlie the kind of intrusions from the unattended masking speech that is seen in both speech perception (Brungart and Simpson, 2001) and speech production (Cherry, 1953) paradigms, as well as the more specific ways that speech production can be affected by concurrent masking sounds (Cooke and Lu, 2010). Instead of the emphasis on self-monitoring seen in many studies of speech production (Christoffels *et al.*, 2007; Lind *et al.*, 2014), perceptual systems are also processing information in our acoustic surroundings, such that there is a route for meaningful elements in unattended auditory streams to be processed centrally. Indeed, auditory streams that are high in informational content (or semantic content) are processed centrally even when the task at hand requires that we actively disregard it. Further studies with more sensitive analysis techniques may be able to establish whether we are seeing a role for multiple auditory streams of information in STG associated with both production and

perception mechanisms, as has been previously suggested for perception (Rauschecker and Scott, 2009; Zatorre *et al.*, 2002). It would also be important to investigate the precise nature of the kinds of relevant informational content—both phonetic and semantic—and the ways that this can affect the cortical responses. Meanwhile, this study emphasises the importance of not assuming that the STG is solely focused on error detection and audibility during speech production—and not underestimating the effect that informational content has on us when we attempt to speak in background noise.

ACKNOWLEDGMENTS

S.M. and S.E. contributed equally to this work. This research was funded by Wellcome Trust Grant No. WT090961MA to S.K.S. and by an ESRC studentship awarded to S.M.

APPENDIX: EXAMPLE LISTS OF STIMULI SENTENCES READ BY PARTICIPANTS

They moved the furniture.
He's wiping the table.
He hit his head.
The yellow leaves are falling.
The cat played with some wool.
The bag was very heavy.
The towel dripped on the carpet.
The bull chased the lady.
The man dug his garden.
The room has a lovely view.
The girl helped in the kitchen.
The old shoes were muddy.
Father's hiding the presents.
The milk boiled over.
The neighbour knocked at the door.
He tore his shirt.
They finished the jigsaw.
She brought her camera.
The lady watered her plants.
The salt cellars full.

Agnew, Z. K., McGettigan, C., Banks, B., and Scott, S. K. (**2013**). "Articulatory movements modulate auditory responses to speech," NeuroImage **73**, 191–199.
Aydelott, J., Jamaluddin, Z., and Nixon Pearce, S. (**2015**). "Semantic processing of unattended speech in dichotic listening," J. Acoust. Soc. Am. **138**(2), 964–975.
Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Memory Lang. **68**(3), 255–278.
Bashford, J. A., Warren, R. M., and Brown, C. A. (**1996**). "Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration," Percept. Psychophys. **58**(3), 342–350.
Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**(3), 108–112.
Blesser, B. (**1972**). "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," J. Speech Lang. Hear. Res. **15**(1), 5–41.
Boersma, P., and Weenink, D. (**2008**). Praat: doing phonetics by computer [Computer program], Version 6.0.17, retrieved 21 April 2016 from http://www.praat.org/.
Brainard, D. H. (**1997**). The Psychophysics Toolbox. Spatial Vision **10**(4), 433–436.
Brett, M., Anton, J-L., Valabregue, R., and Poline, J-B. (**2002**). "Region of interest analysis using an SPM toolbox" [abstract], in *8th International Conference on Functional Mapping of the Human Brain*, June 2–6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage, Vol 16, No 2.
Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.
Brungart, D. S., and Simpson, B. D. (**2001**). "Contralateral masking effects in dichotic listening with two competing talkers in the target ear," J. Acoust. Soc. Am. **109**(5), 2486–2486.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**(3), 694–703.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**(5), 975–979.
Christoffels, I. K., Formisano, E., and Schiller, N. O. (**2007**). "Neural correlates of verbal feedback processing: An fMRI study employing overt speech," Human Brain Map. **28**(9), 868–879.
Christoffels, I. K., van de Ven, V., Waldorp, L. J., Formisano, E., and Schiller, N. O. (**2011**). "The sensory consequences of speaking: Parametric neural cancellation during speech in auditory cortex," PloS One **6**(5), e18307.
Clark, H. H. (**1973**). "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research," J. Verbal Learn. Verbal Behav. **12**(4), 335–359.
Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.
Cooke, M., and Lu, Y. (**2010**). "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," J. Acoust. Soc. Am. **128**(4), 2059–2069.
Eliades, S. J., and Wang, X. (**2003**). "Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations," J. Neurophysiol. **89**(4), 2194–2207.
Eliades, S. J., and Wang, X. (**2012**). "Neural correlates of the Lombard effect in primate auditory cortex," J. Neurosci. **32**(31), 10737–10748.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E., and Knight, R. T. (**2010**). "Single-trial speech suppression of auditory cortex activity in humans," J. Neurosci. **30**(49), 16643–16650.
Garnier, M., Henrich, N., and Dubois, D. (**2010**). "Influence of sound immersion and communicative interaction on the Lombard effect," J. Speech, Lang., Hear. Res. **53**, 588–608.
Hashimoto, Y., and Sakai, K. L. (**2003**). "Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: An fMRI study," Human Brain Mapping **20**(1), 22–28.
Hickok, G. (**2012**). "Computational neuroanatomy of speech production," Nature Rev. Neurosci. **13**(2), 135–145.
Houde, J. F., Nagarajan, S. S., Sekihara, K., and Merzenich, M. M. (**2002**). "Modulation of the auditory cortex during speech: An MEG study," J. Cognit. Neurosci. **14**(8), 1125–1138.
Junqua, J. (**1993**). "The Lombard reflex and its role on human and automatic speech recognizers," J. Acoust. Soc. Am. **93**(1), 510–524.
Kitamura, T., Takemoto, H., Honda, K., Shimada, Y., Fujimoto, I., Syakudo, Y., Masaki, S., Kuroda, K., Oku-uchi, N., and Senda, M. (**2005**). "Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner," Acoust. Sci. Technol. **26**(5), 465–468.
Lind, A., Hall, L., Breidegard, B., Balkenius, C., and Johansson, P. (**2014**). "Auditory feedback of one's own voice is used for high-level semantic monitoring: The 'self-comprehension' hypothesis," Front. Human Neurosci. **8**, 166.
Lombard, E. (**1911**). "Le signe de l'elevation de la voix" ("The sign of the elevation of the voice"), Annales Des Maladies de L'Oreille et Du Larynx **37**, 101–119.
Lu, Y., and Cooke, M. (**2008**). "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am. **124**(5), 3261–3275.
Lund, T. E., Nørgaard, M. D., Rostrup, E., Rowe, J. B., and Paulson, O. B. (**2005**). "Motion or activity: Their role in intra- and inter-subject variation in fMRI," NeuroImage **26**(3), 960–964.

MacLeod, A., and Summerfield, Q. (**1990**). "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use," Br. J. Audiol. **24**(1), 29–43.
Peelle, J. E. (**2012**). "The hemispheric lateralization of speech processing depends on what 'speech' is: A hierarchical perspective," Front. Human Neurosci. **6**, 3091–3094.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**(4), 434–446.
Pick, H. L., Siegel, G. M., Fox, P. W., and Kearney, J. K. (**1989**). "Inhibiting the Lombard effect," J. Acoust. Soc. Am. **5**(2), 894–900.
Pittman, A. L., and Wiley, T. L. (**2001**). "Recognition of speech produced in noise," J. Speech, Lang., Hear. Res. **44**(3), 487–496.
Rafii, Z., and Pardo, B. (**2011**). "A simple music/voice separation method based on the extraction of the repeating musical structure," in *36th International Conference on Acoustics, Speech and Signal Processing*, pp. 1–4.
Rauschecker, J. P., and Scott, S. K. (**2009**). "Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing," Nature Neurosci. **12**(6), 718–724.
Rivenez, M., Darwin, C. J., and Guillaume, A. (**2006**). "Processing unattended speech," J. Acoust. Soc. Am. **119**(6), 4027–4040.
Saito, S., and Baddeley, A. (**2004**). "Irrelevant sound disrupts speech production: Exploring the relationship between short-term memory and experimentally induced slips of the tongue," Q. J. Exp. Psychol. A **57**(7), 1309–1340.
Schell, K. W. (**2008**). "The influence of linguistic content on the Lombard effect," J. Speech, Lang., Hear. Res. **51**, 209–220.
Scott, S., Evans, S., McGettigan, C., and Rosen, S. (**2012**). "The neural basis for energetic and informational masking effects in speech perception," J. Acoust. Soc. Am. **131**(4), 3341–3341.
Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., and Wise, R. J. S. (**2009**). "The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes," J. Acoust. Soc. Am. **125**(3), 1737–1743.
Scott, S. K., Rosen, S., Lang, H., and Wise, R. J. S. (**2006**). "Neural correlates of intelligibility in speech investigated with noise vocoded speech—A positron emission tomography study," J. Acoust. Soc. Am. **120**(2), 1075–1083.
Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. S. (**2004**). "A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception," J. Acoust. Soc. Am. **115**(2), 813–821.
Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (**2011**). "The importance for speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am. **130**(5), 2874–2881.
Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**(1), 317–326.
Summers, W. Van, Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Michael, A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses after date," NIH Public Access **84**(3), 917–928.
Tourville, J. A., and Guenther, F. H. (**2011**). "The DIVA model: A neural theory of speech acquisition and production," Lang. Cognit. Process. **26**(7), 952–981.
Tourville, J. A., Reilly, K. J., and Guenther, F. H. (**2008**). "Neural mechanisms underlying auditory feedback control of speech," NeuroImage **39**(3), 1429–1443.
Toyomura, A., Koyama, S., Miyamaoto, T., Terao, A., Omori, T., Murohashi, H., and Kuriki, S. (**2007**). "Neural correlates of auditory feedback control in human," Neuroscience **146**(2), 499–503.
Varadarajan, V. S., and Hansen, J. H. L. (**2006**). "Analysis of Lombard effect under different types and levels of noise with application to In-set Speaker ID systems 2. The UT-SCOPE database 3," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2006), Vol. 2, pp. 937–940.
Warhurst, S., Madill, C., McCabe, P., Heard, R., and Yiu, E. (**2012**). "The vocal clarity of female speech-language pathology students: An exploratory study," J. Voice **26**(1), 63–68.
Wise, R., Greene, J., Büchel, C., and Scott, S. (**1999**). "Brain regions involved in articulation," The Lancet **353**(9158), 1057–1061.
Wood, G., Nuerk, H.-C., Sturm, D., and Willmes, K. (**2008**). "Using parametric regressors to disentangle properties of multi-feature processes," Behavior. Brain Funct. **4**(1), 38.
Zatorre, R. J., Bouffard, M., Ahad, P., and Belin, P. (**2002**). "Where is 'where' in the human auditory cortex?," Nat. Neurosci. **5**(9), 905–909.
Zheng, Z., Munhall, K., and Johnsrude, I. (**2010**). "Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production," J. Cognit. Neurosci. **22**(8), 1770–1781.