Neuropsychologia 50 (2012) 762–776

Contents lists available at SciVerse ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Speech comprehension aided by multiple modalities: Behavioural and neural interactions

Carolyn McGettigan[a,*], Andrew Faulkner[b], Irene Altarelli[b,c], Jonas Obleser[d], Harriet Baverstock[e], Sophie K. Scott[a]

[a] Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK
[b] Department of Speech, Hearing & Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK
[c] Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France
[d] Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, 04103 Leipzig, Germany
[e] School of Psychological Sciences, The University of Manchester, Coupland 1 Building, Coupland Street, Oxford Road, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 24 August 2011
Received in revised form
30 December 2011
Accepted 8 January 2012
Available online 17 January 2012

Keywords:
Speech
fMRI
Auditory cortex
Individual differences

ABSTRACT

Speech comprehension is a complex human skill, the performance of which requires the perceiver to combine information from several sources – e.g. voice, face, gesture, linguistic context – to achieve an intelligible and interpretable percept. We describe a functional imaging investigation of how auditory, visual and linguistic information interact to facilitate comprehension. Our specific aims were to investigate the neural responses to these different information sources, alone and in interaction, and further to use behavioural speech comprehension scores to address sites of intelligibility-related activation in multifactorial speech comprehension. In fMRI, participants passively watched videos of spoken sentences, in which we varied Auditory Clarity (with noise-vocoding), Visual Clarity (with Gaussian blurring) and Linguistic Predictability. Main effects of enhanced signal with increased auditory and visual clarity were observed in overlapping regions of posterior STS. Two-way interactions of the factors (auditory × visual, auditory × predictability) in the neural data were observed outside temporal cortex, where positive signal change in response to clearer facial information and greater semantic predictability was greatest at intermediate levels of auditory clarity. Overall changes in stimulus intelligibility by condition (as determined using an independent behavioural experiment) were reflected in the neural data by increased activation predominantly in bilateral dorsolateral temporal cortex, as well as inferior frontal cortex and left fusiform gyrus. Specific investigation of intelligibility changes at intermediate auditory clarity revealed a set of regions, including posterior STS and fusiform gyrus, showing enhanced responses to both visual and linguistic information. Finally, an individual differences analysis showed that greater comprehension performance in the scanning participants (measured in a post-scan behavioural test) were associated with increased activation in left inferior frontal gyrus and left posterior STS. The current multimodal speech comprehension paradigm demonstrates recruitment of a wide comprehension network in the brain, in which posterior STS and fusiform gyrus form sites for convergence of auditory, visual and linguistic information, while left-dominant sites in temporal and frontal cortex support successful comprehension.

1. Introduction

1.1. Background

Everyday speech comprehension is multi-faceted: in face-to-face conversation, the listener receives information from the voice and face of a talker, the accompanying gestures of the hands and body and the overall semantic context of the discussion, which can all be used to aid comprehension of the spoken message. Behaviourally, auditory speech comprehension is enhanced by simultaneous presentation of a face or face-like visual cues (Bernstein, Auer, & Takayanagi, 2004; Bishop & Miller, 2009; Girin, Schwartz, & Feng, 2001; Grant & Seitz, 2000b; Hazan, Kim, & Chen, 2010; Helfer & Freyman, 2005; Kim & Davis, 2004; Kim, Davis, & Groot, 2009; Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Schwartz, Berthommier, & Savariaux, 2004; Sumby & Pollack, 1954; Thomas & Pilling, 2007). Higher-order linguistic information can also benefit intelligibility: words presented in a sentence providing a rich semantic context are more intelligible than words in isolation or in an abstract sentence, particularly when auditory clarity is compromised (Dubno, Ahlstrom, & Horwitz, 2000; Grant & Seitz, 2000a; Kalikow,

* Corresponding author. Tel.: +44 020 7679 7529; fax: +44 0020 7813 2835.
E-mail address: c.mcgettigan@ucl.ac.uk (C. McGettigan).

0028-3932/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2012.01.010

Stevens, & Elliott, 1977; Miller & Isard, 1963; Obleser, Wise, Dresner, & Scott, 2007; Pichora-Fuller, Schneider, & Daneman, 1995; Stickney & Assmann, 2001).

1.2. Factors affecting speech intelligibility: degraded speech in the brain

The use of signal degradation (e.g. noise-vocoding; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) in neuroimaging research has contributed considerably to our knowledge of the neural underpinnings of auditory speech comprehension (Davis & Johnsrude, 2003; Eisner, McGettigan, Faulkner, Rosen, & Scott, 2010; Narain et al., 2003; Obleser & Kotz, 2010; Obleser et al., 2007; Scott, Blank, Rosen, & Wise, 2000; Scott, Rosen, Lang, & Wise, 2006). By titrating the auditory clarity of noise-vocoded speech against comprehension, several studies have identified intelligibility-specific responses to speech in anterior and posterior sites in the superior temporal sulcus (STS), often lateralized to the left hemisphere (Eisner et al., 2010; Narain et al., 2003; Scott et al., 2000).

Several studies have also implicated frontal, particularly motor and premotor, sites in the comprehension of degraded speech (Adank & Devlin, 2010; Davis & Johnsrude, 2003; Eisner et al., 2010; Obleser et al., 2007; Osnes, Hugdahl, & Specht, 2011; Scott, Rosen, Wickham, & Wise, 2004). Davis and Johnsrude (2003) found elevated responses to degraded speech in the frontal operculum that were modulated by the intelligibility of degraded speech, but insensitive to its acoustic form. They implicated these frontal activations in higher-order syntactic and semantic aspects of linguistic comprehension in the presence of incomplete acoustic information. Adank and Devlin (2010) related premotor activation in the inferior frontal gyrus to perceptual adaptation to time-compressed speech, while Eisner et al. (2010) demonstrated that activating in overlapping regions of left posterior inferior frontal gyrus correlated with individual differences in perceptual learning of degraded speech and working memory capacity. Scott et al. (2004) observed parametrically increasing activation in dorsomedial premotor (in the region of the supplementary motor area; SMA) cortex as speech became more difficult to understand (through the addition of noise), which they interpreted as the recruitment of an articulatory strategy to support the performance of a difficult speech comprehension task. Osnes et al. (2011) elaborated on this view by showing that premotor involvement in speech perception is most pronounced when speech is degraded, but still intelligible.

1.2.1. Visual cues and cross-modal integration

Some early neuroimaging studies of audiovisual perception identified superadditive responses to congruent audiovisual speech (where responses to audiovisual speech are greater than the summed responses to each modality alone; AV > A + V) and subadditive response to incongruent (AV < A + V) as reflective of multimodal integration, and linked such responses to sites on the superior temporal sulcus and superior temporal gyrus (STS and STG, respectively: Calvert, Campbell, & Brammer, 2000; Calvert, 2001; though note that this approach is not without controversy in fMRI: see Beauchamp, 2005; James & Stevenson, 2011; Laurienti, Perrault, Stanford, Wallace, & Stein, 2005; Love, Pollick, & Latinus, 2011; Stevenson, Kim, & James, 2009). While these, and other, studies have explored audiovisual integration via manipulation of temporal (Miller & D'Esposito, 2005; Stevenson, Altieri, Kim, Pisoni, & James, 2010) or content congruency (Bernstein, Auer, Wagner, & Ponton, 2008a; Bernstein, Lub, & Jianga, 2008b; Calvert et al., 2000; Nath and Beauchamp, 2012) across the auditory and visual streams, a smaller number of studies have focused on the effects of signal degradation in one or other of the input modalities on neural activation during speech comprehension (Bishop & Miller, 2009; Callan et al., 2003; Nath & Beauchamp, 2011; Scott et al., 2002; Sekiyama, Kanno, Miura, & Sugita, 2003; Stevenson et al., 2009). These studies of speech intelligibility using audiovisual stimuli have elaborated on the studies of integration and shown that visual information can enhance intelligibility-related responses to auditory speech in similar areas of temporal cortex (Callan et al., 2003; Scott et al., 2002; Sekiyama et al., 2003).

Sekiyama and colleagues (2003) used a McGurk-type syllable identification paradigm in both PET and fMRI, in which they presented participants with videos of spoken syllables (/ba/,/da/,/ga/) where the concurrent audio stream contained a mismatching syllable. Using a further manipulation of the amplitude of the auditory stimulation (by varying signal-to-noise ratio against background scanner noise (fMRI experiment) or an applied white noise (PET experiment)), the authors were able to bias the participants' perception toward the visual content when the auditory intelligibility was low. During these conditions, the authors found greater activation in left posterior STS (fMRI and PET) and right temporal pole (PET) compared with responses when the auditory intelligibility was high. In the PET study, they identified further areas outside temporal cortex for the low > high intelligibility comparison, in right thalamus and cerebellum.

Kawase et al. (2005) factorially applied stimulus degradation to audio (low-pass filtering at 500 Hz) and visual (application of Gaussian blurring) channels in a 2 × 2 design during the presentation of the spoken syllable 'Be' in fMRI. However, the nature of the manipulations was such that the blurring rendered visual stream totally unintelligible while the auditory filtering only effected a partial disruption to auditory intelligibility. The authors report significant increases in signal for contrasts of high > low visual intelligibility where the auditory clarity was held constant. They observed increased activation in bilateral visual cortex for both clear and filtered speech, but with additional activation in right fusiform gyrus when the auditory speech was filtered. The authors interpret this as evidence of additional face processing to support speech perception in the presence of an unreliable auditory signal. More recently, Nath and Beauchamp (2011) adopted a similar design in fMRI, using noise-vocoding to degrade the auditory speech (by reducing the spectral detail; Shannon et al., 1995) and a combination of reduced contrast and Gaussian blurring to reduce the clarity of the visual stream. They played participants congruent audiovisual syllable and word tokens in fMRI and used functional connectivity analyses to demonstrate increased connection strength from fusiform gyrus to STS when the visual information was more 'reliable' than the auditory, and increased connectivity from Heschl's gyrus to STS when the auditory information was more reliable. Importantly, they were able to show these effects were present regardless of whether the participant was instructed to attend to the visual or auditory signal.

Other studies of audiovisual perception have explored the use of parametric designs to assess the interaction of the two modalities. Stevenson et al. (2009) argued that the choice of baseline condition would have strongly influenced the results of previous studies that identified sites of cross-modal integration by measuring super- and sub-additivity in the BOLD response. By modulating the intelligibility of auditory and visual representations of tool use across several levels in an additive factors approach, Stevenson et al. (2009) identified regions showing evidence for neuronal convergence of auditory and visual information as those exhibiting *inverse effectiveness* – a progressively greater gain for audiovisual stimuli as the quality of the individual streams is reduced. These regions covered a wider network than identified in previous studies, and included bilateral medial frontal gyrus, anterior and posterior cingulate cortex, parahippocampal gyrus, insula, caudate nucleus, left inferior temporal gyrus and left inferior parietal cortex.

Scott et al. (2002) also employed a parametric design, in order to explore neural responses to the intelligibility of audio-visual

sentences. Auditory stimuli were noise-vocoded to four different levels of intelligibility, while the videos were manipulated using three levels of Gaussian blurring. Behavioural sentence report scores showed that facial information was most effective in enhancing speech intelligibility at intermediate levels of auditory clarity. The authors used PET to probe the neural responses to the stimuli and found that, while extensive portions of the dorsolateral temporal lobes in both hemispheres showed enhanced signal to the speech when the clarity of the face improved, the sites of greatest visual enhancement were located in bilateral anterior STS.

While the focus of the current paper is on speech comprehension, it is important to point out that audiovisual integration is not a speech-specific process. Many other studies have explored mechanisms of multisensory integration for non-speech stimuli, also identifying STS as a key site in this (Beauchamp, Lee, Argall, & Martin, 2004a; Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004b; Stevenson & James, 2009; Stevenson et al., 2010; Werner & Noppeney, 2009). One study identified different spatial locations for sites showing inverse effectiveness in the perception of audiovisual speech and tool use stimuli, but an identical mechanism within each site indicated a lack of speech-specificity in the process of cross-modal integration (Stevenson & James, 2009).

1.2.2. Linguistic factors

A number of recent studies have explored the neural correlates of semantic predictability/expectancy in the context of the comprehension of degraded speech (Obleser & Kotz, 2010, 2011). Obleser et al. (2007) presented participants with auditory sentences at three levels of auditory clarity (noise-vocoded with 2, 8 and 32 channels), and additionally varied the semantic predictability of the items. In a behavioural experiment, the authors showed that sentences of greater semantic predictability were significantly more intelligible than low predictability items, and that this linguistic enhancement was most marked at intermediate auditory clarity (8 channels). In fMRI, the response in bilateral superior temporal cortex and inferior frontal gyri increased with improved auditory clarity. However, a direct comparison of the response to high and low predictability items (at 8 channels) showed activations that were confined to regions outside temporal cortex, including left angular gyrus, left inferior frontal gyrus (pars orbitalis), superior frontal gyrus, and posterior cingulate cortex. Obleser and Kotz (2010) carried out a similar study using noise-vocoded sentences in fMRI, where the linguistic manipulation was one of cloze probability – that is, they compared responses to highly-controlled sentences in which the final word (e.g. 'beer') was highly probable given the sentential context (e.g. 'He *drinks* the beer'), with those in which the key word was less obvious (e.g. 'He *sees* the beer'). As in the previous study, the greatest behavioural effect of cloze probability (i.e. enhanced intelligibility when the expectancy constraints were larger) was seen at an intermediate level of acoustic clarity. The authors found some evidence for effects of cloze probability in superior temporal cortex, where the spatial extent of activation showing a positive relationship with auditory clarity was greater for low than high cloze items. An overall effect of expectancy was observed in left inferior frontal gyrus (BA 44), where a low > high cloze effect became more pronounced as the auditory clarity of the signal improved. As in Obleser et al. (2007), the left angular gyrus was implicated in an expectancy enhancement (i.e. high cloze > low cloze) that became more marked at the intermediate levels of auditory clarity where this effect had been seen behaviourally.

1.3. Current study

The existing data from studies of facial and linguistic effects on auditory speech intelligibility have implicated superior temporal cortex, in particular the posterior STS extending to inferior parietal cortex, as a site of integration of multiple sources of information in speech. Beyond this region, the previous work has identified a wide range of regions in the combination of multiple inputs, but some recurrent findings include the involvement of the fusiform gyrus for facial inputs, and the angular gyrus and inferior frontal gyrus in the processing of linguistic influences. It remains to be empirically demonstrated how facial and linguistic information would interact neutrally when both factors are varied, as is the case in face-to-face communication. Speech-reading is possible from visual cues alone, while in contrast linguistic manipulations in speech require a carrier (i.e. the visual or auditory signal) in order to be detectable. Therefore, we might expect differences in the expression and interaction of these types of cue with auditory speech in the brain – only by studying these factors in combination can we identify any neural sites that might be responsive to both.

The aim of the current study was to investigate the interactions of low-level visual and higher-order linguistic manipulations with the neural responses to auditory speech. Using an audiovisual sentence comprehension paradigm, we manipulated the information in auditory (noise-vocoding with 2, 4 and 6 channels), visual (Gaussian blurring at two levels of clarity) and linguistic streams (high and low linguistic predictability) to identify the behavioural effects of these factors on intelligibility, and then used these results to probe neural responses in a passive comprehension task in fMRI. Specifically, we aimed to harness behavioural data to investigate the neural responses related to speech intelligibility, and to further identify the correlates of individual differences in speech comprehension performance in the brain.

Although no previous study of speech intelligibility has combined these three factors, research on the neural processing of co-speech gesture offers some basis for the formation of predictions. Much of this work has investigated the effect of iconic gestures (which illustrate physical properties of the world) and metaphoric gestures (which illustrate abstract information) on the processing of speech (Straube, Green, Weis, Chatterjee, & Kircher, 2009; Straube, Green, Bromberger, & Kircher, 2011; Green et al., 2009; Holle, Obleser, Rueschmeyer, & Gunter, 2010). A consistent finding across many of these studies is that responses in bilateral posterior superior temporal cortex (in particular, the STS) are enhanced by the addition of gestures during speech. However, higher-order aspects of speech-gesture integration, such as those required for metaphorical gesture, tend to recruit structures beyond sensory cortex, such as inferior frontal gyrus and premotor cortex (Kircher et al., 2009; Straube et al., 2009, 2011), with Straube and colleagues (2011) concluding that the IFG supports higher-order relational processes in speech-gesture integration, while the posterior temporal cortex performs perceptual matching of auditory and visual information. Holle and collegues (2010) found inverse effectiveness only in left posterior STS for the integration of iconic gesture with speech in noise, which they interpret as evidence for stronger semantic integration in the left hemisphere. Similarly to Holle et al.'s study, we manipulate auditory, visual and semantic/linguistic cues in the context of a speech perception experiment, but employing higher-order manipulations are intrinsic to the audiovisual speech stimulus.

In the current experiment, we expected that improved facial clarity should lead to enhanced neural responses to auditory speech in superior temporal cortex, and in particular, the posterior STS. As others have argued that multimodal responses in STS are not speech-specific, we predict that visual enhancements will not necessarily show a strong left-lateralization in the current experiment. In contrast, we do expect overt linguistic manipulations of the speech stimuli to generate left-dominant responses. Further, we expect to find left-dominant structures to be most sensitive to item intelligibility (Eisner et al., 2010; McGettigan et al., 2011; Narain

et al., 2003; Rosen, Wise, Chadha, Conway, & Scott, 2011; Scott et al., 2000), and to individual differences in task performance (Eisner et al., 2010; Nath, Fava, & Beauchamp, 2011). Based on the findings of this study's direct predecessor (Obleser et al., 2007), we expect that manipulations of predictability should engage higher-order language regions in fronto-parietal cortex, and that we should also find greatest evidence for neural interaction of the three factors at intermediate levels of stimulus intelligibility.

2. Methods

2.1. Materials

The stimulus material used comprised 400 sentences from the "speech intelligibility in noise" or SPIN test (Kalikow et al., 1977), half of which were characterized by high predictability, the other half being of lower predictability (depending on the strength of the association between their key words, e.g. High: The boat sailed across the bay vs. Low: The old man discussed the dive). The SPIN sentence lists are matched for phonetic and linguistic variables such as phonemic features, number of syllables and content words. The sentences were read by a female speaker of British English in a quiet room. The speaker was forward-facing and positioned against a mid-blue background. Video and audio recordings were made using a digital video camera (Canon XL-1 mini-DV camcorder; Canon (UK) Ltd., Reigate, UK), with the speaker's face positioned centrally and the field of view incorporating the whole head, neck and shoulders. Audio was recorded from a Brüel & Kjær type 4165 microphone (Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark). Video data were recorded at 720 × 576 pixels and 25 frames per second, and these parameters were maintained in subsequent image processing.

The video data were transferred to a PC, edited into separate files for each sentence and then split into separate audio and video channels for further processing, using Adobe Premiere version 6.5 (Adobe Systems Europe Ltd., Uxbridge, UK). The audio files were processed using a noise-vocoding technique (after Shannon et al., 1995) to 2, 4 and 6 channels, in MATLAB Version 7.0 (The MathWorks, Inc., Natick, MA). Behavioural pre-testing was used to assess the intelligibility of items (via typed report, and using the visual manipulations described below) for two ranges of auditory degradation – 2, 3 and 4 channels (16 participants), and 2, 4 and 6 channels (6 participants). There was little difference in the mean intelligibility at 3 and 4 channels (with mean report scores of 43% and 48%, respectively, across 16 participants), so it was decided to use 2, 4 and 6 in the fMRI experiment (as these levels covered a fuller range of intelligibility performance). Briefly, the vocoding technique involved the following steps. For each sentence, the speech waveform was passed through a bank of analysis filters (2, 4 or 6) spanning the range 70–9000 Hz. The filter bandwidths were set to represent equal distances along the basilar membrane (according to the Greenwood (1990) equation relating filter position to best frequency). Amplitude envelope extraction (via half-wave rectification and low-pass filtering at 1000 Hz) occurred at the output of each analysis filter. The envelopes were each then multiplied with a band-limited white noise carrier, filtered and summed together. Finally, the re-summed stimulus was low-pass filtered at 9000 Hz. The sentences were normalized for root-mean-squared amplitude in PRAAT (Boersma & Weenink, 2008).

For the fMRI experiment, an unintelligible auditory baseline was created by multiplying a wideband noise (with flat amplitude; band-limited to 70 and 9000 Hz) with the long-term average spectrum from all the vocoded items (Scott et al., 2004).

Videos were blurred using an image-processing filter in Matlab with the image processing toolbox extension (Matlab Version 6.5). The effect of the blur was to redistribute the luminance of a single pixel over space according to a 2-dimensional symmetric Gaussian distribution. Here the degree of blur is expressed as the diameter in pixels that corresponded to two standard deviations of the Gaussian distribution. Informed by pilot data, two different levels of blurring were chosen: parameter 15 (standard deviation of 7.5 pixels) and parameter 45 (standard deviation of 22.5 pixels). Static examples of these levels can be seen in Fig. 1a. The more degraded stimuli (parameter 45) reduced the visual resolution such that only mouth aperture and head movements could be distinguished. In the clearer stimuli, more complex articulatory movements (e.g. lip shaping, some tongue movements) were visible.

Video and audio streams were put back together in Virtualdub (http://www.virtualdub.org), such that they were temporally congruent. In the final stimulus set, each of the 400 SPIN items was available in 12 conditions: 3 levels of Auditory Clarity (2, 4 and 6 channels) × 2 levels of Linguistic Predictability (low, high) × 2 levels of Visual Clarity (45, 15). The baseline condition comprised videos with blur parameter 45, accompanied by the wideband noise auditory baseline. The mean duration of the video files was 2.6 s (incorporating a few hundred milliseconds before and after the sentence to allow the video to begin and end with a neutral, closed-mouth expression). In the fMRI study, different sub-sets from this pool were used in the fMRI experiment and the behavioural post-test.

Example stimuli from each of the 13 conditions can be found in the Supplementary Material.

2.2. Behavioural experiment

2.2.1. Participants

Twelve normal-hearing, adult speakers of English (7 female, 5 male; aged 18–40) participated in a behavioural experiment. All the participants had normal hearing and reported no neurological history, nor any problems with speech or language. The study was approved by the UCL Department of Psychology Ethics Committee.

2.2.2. Design and procedure

Twenty stimuli from each of the 12 speech conditions were played only once, via a PC laptop (12.1″ monitor) and headphones (Sennheiser HD201; Sennheiser U.K., High Wycombe, Buckinghamshire, UK) after which the participant attempted spoken report of the sentence content. Order of presentation of the stimuli was pseudorandomized for each participant, such that each 'mini-block' of 12 stimuli contained one example from each condition. Assignment of SPIN sentences to each condition was fully randomized for each participant.

2.3. Functional imaging (fMRI) experiment

2.3.1. Participants

26 adult speakers of English (13 female, 13 male; mean age 24 years 7 months, range 19–35 yrs) participated in the study. All were selected and recruited as described for the behavioural study. The study was approved by the UCL Department of Psychology Ethics Committee.

2.3.2. Design and procedure

Functional imaging data were acquired on a Siemens Avanto 1.5 Tesla scanner (Siemens AG, Erlangen, Germany) with a 12-channel birdcage headcoil, in two runs of 133 echo-planar whole-brain volumes (TR = 9 s, TA = 3 s, TE = 50 ms, flip angle 90°, 35 axial slices, 3 mm × 3 mm × 3 mm in-plane resolution). A sparse-sampling routine (Edmister, Talavage, Ledden, & Weisskoff, 1999; Hall et al., 1999) was employed, in which each stimulus was presented 4.5 s (with jittering of ±500 ms) before acquisition of the next scan commenced.

In the scanner, trigger pulses from the first three volumes were used to engage and prepare stimulus randomization routines in a presentation script, run in MATLAB with the Psychophysics Toolbox extension (Brainard, 1997), via a Denon amplifier (Denon UK, Belfast, UK) and electrodynamic headphones worn by the participant (MR Confon GmbH, Magdeburg, Germany). Videos were projected from a specially configured video projector (Eiki International, Inc., Rancho Santa Margarita, CA) onto a custom-built front screen, which the participant viewed via a mirror placed on the head coil. A set of 260 different items (130 high-predictability, 130 low-predictability) from the SPIN corpus was used in the experiment. In each functional run, the participant viewed a total of 130 videos, comprising 10 from each of the 12 experimental conditions and a further 10 from the baseline condition. Participants were requested to look at the videos and try to understand what was being said, without speaking or moving their mouths. The order of presentation of stimuli from the different conditions was pseudorandomized to allow a relatively even distribution of the conditions across the run without any predictable ordering effects. Within the high- and low-predictability conditions, assignment of SPIN items to conditions was fully randomized. After the functional run was complete, a high-resolution T1-weighted anatomical image was acquired (HIRes MP-RAGE, 160 sagittal slices, voxel size = 1 mm^3). The total time in the scanner was around 55 min.

2.3.3. Analysis of fMRI data

Data were preprocessed and analysed in SPM5 (Wellcome Trust Centre for Neuroimaging, London, UK). Functional images were realigned and unwarped, coregistered with the anatomical image, normalised to Montreal Neurological Institute stereotactic space using parameters obtained from segmentation of the anatomical image, and smoothed using a Gaussian kernel of 8 mm FWHM. Event-related responses for each event type were modelled using the canonical haemodynamic response function in SPM5. For one participant, the second functional run had to be discarded due to excessive head movement in the scanner (i.e. greater than 3 mm translation or 3° rotation in any direction). Each condition was modelled as a separate regressor in a generalised linear model (GLM), with event onsets modelled 0.5 s after the stimulus onset. Six movement parameters (3 translations, 3 rotations) were included as regressors of no interest. Twelve contrast images were created for the comparison of each individual experimental condition with the unintelligible baseline. These 12 images from each participant were entered into a second-level 3 × 2 × 2 full factorial ANOVA model in SPM5 (with a non-sphericity correction for non-independence across factor levels) with factors Auditory Clarity (2, 4, 6 channels of noise-vocoding), Visual Clarity (parameter 45, parameter 15) and Linguistic Predictability (low, high). Within this second-level model, F contrasts were constructed to explore the main effects and interactions of the three factors. A weighted T-contrast for Intelligibility (all conditions) was also constructed using mean scores from each condition in the behavioural experiment (see Fig. 1b). A second full factorial ANOVA (2 × 2) was run on conditions with 4 channels of auditory clarity, employing factors Visual Clarity (parameter 45, parameter 15) and Linguistic Predictability (low, high). A T-contrast in this model was set up to describe the profile of behavioural intelligibility scores across the four conditions.

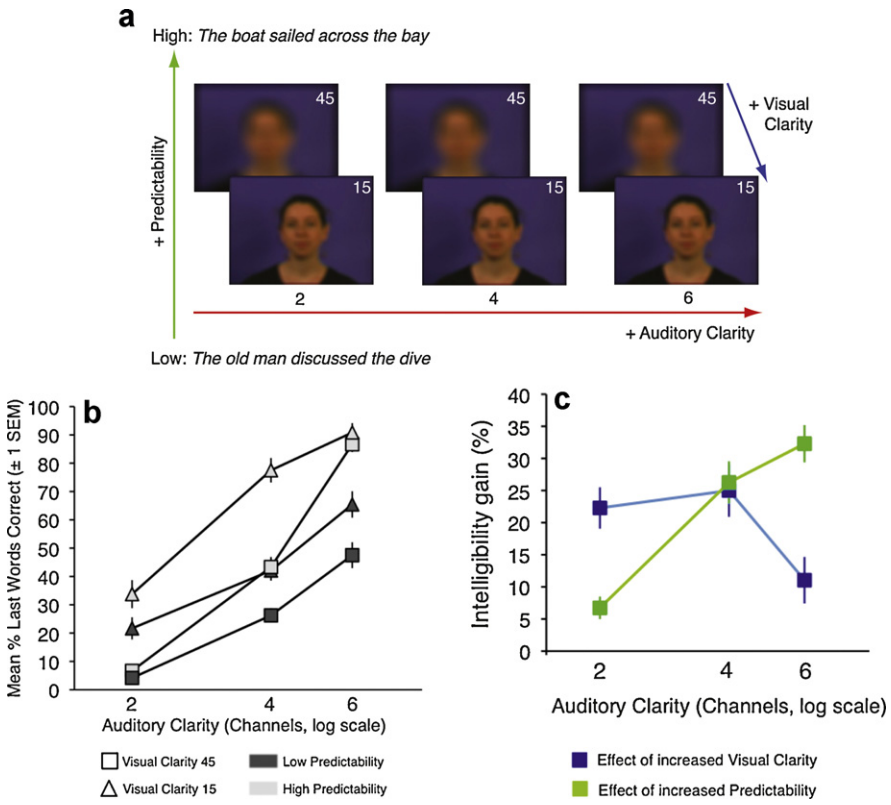

Fig. 1. (a) Conditions used in the current study. (b) Plot showing sentence report accuracy by condition in the behavioural experiment. (c) Plot showing intelligibility profiles describing the significant 2-way interactions observed in the behavioural experiment.

Group intelligibility scores from the 12 experimental conditions in the behavioural experiment were used to generate an additional T-contrast image at the single-subject level – these contrasts were used in a second-level investigation of individual differences.

Coordinates of peak activations were labelled using the SPM5 anatomy toolbox (Eickhoff et al., 2005).

All reported contrasts were thresholded at $p < .005$ (uncorrected), with a cluster extent threshold of 20 voxels (corrected to a whole-brain alpha of $p < .000$ (calculated to 3 decimal places) using a Monte Carlo simulation in Matlab with 10,000 iterations; Slotnick, Moo, Segal, & Hart, 2003) – statistical and significance values for activations in individual peak and sub-peak voxels are listed in the results tables.

2.3.4. Behavioural post-test

At the end of the scanning session, each of the participants took part in a short behavioural post-test. The test was designed to obtain a single, independent behavioural measure of speech comprehension performance from each participant for use in an individual differences regression analysis on the fMRI data. For this reason, only a small number of trials were employed. Thirty-nine stimuli (3 new examples from each of the 12 speech conditions, plus 3 baseline trials) were shown via a PC laptop (12.1″ monitor) and headphones (Sennheiser HD201). Participants were asked to repeat what they could understand from the sentences, with each item being scored according to whether the participant accurately identified the last word in the sentence. Each participant was then assigned an overall score, reflecting the total number of words they repeated correctly.

3. Results

3.1. Behavioural experiment

Performance accuracy was scored in terms of whether the participant correctly reported the last word in the sentence – Fig. 1b shows the group results. Proportion scores on each condition were entered into the repeated-measures ANOVA in SPSS (v.16.0; SPSS Inc., Chicago, IL). There were significant main effects of Auditory Clarity ($F(2, 22) = 258.76$, $p < 0.0001$), Visual Clarity ($F(1, 11) = 34.31$, $p < 0.0001$) and Predictability ($F(1, 11) = 283.68$, $p < 0.0001$), reflecting increases in intelligibility with increases in information from the three factors. There were significant 2-way interactions for Auditory Clarity × Visual Clarity ($F(2, 22) = 10.26$, $p < 0.005$) and Auditory Clarity × Predictability ($F(2, 22) = 18.09$, $p < 0.0001$. Fig. 1c illustrates how the behavioural enhancement from the facial and linguistic factors was modulated by the quality of the auditory signal. A significant three-way interaction was identified (Auditory Clarity × Visual Clarity × Predictability: $F(2, 22) = 17.30$, $p < 0.0001$), reflecting a relatively greater visual enhancement at lower levels of Auditory Clarity and a greater effect of Predictability at higher levels.

3.2. fMRI experiment

3.2.1. Second-level factorial ANOVA: main effects

Fig. 2 shows the main effects of the factors Auditory Clarity (Fig. 2a), Visual Clarity (Fig. 2b), and Linguistic Predictability (Fig. 2c), with plots of mean parameter estimates taken from spherical regions of interest (ROIs; 4 mm in diameter, to align with the 8 mm smoothing FWHM used in pre-processing) built around the peak activations using the MarsBaR toolbox (version 0.42; Brett, Anton, Valabregue, & Poline, 2002) in SPM. Increasing Auditory Clarity resulted in increased activation along the length of bilateral superior temporal sulcus/gyrus (STS/STG), and in left inferior frontal gyrus (IFG), but a decrease in cingulate cortex, middle frontal gyrus, inferior occipital cortex and parietal cortex bilaterally. Increasing Visual Clarity gave increased signal in posterior regions of STS/STG bilaterally, primary visual cortex, bilateral fusiform gyrus and right amygdala, and a decrease in bilateral middle occipital gyrus, left anterior cingulate, right inferior occipital gyrus, bilateral middle frontal gyrus, left rolandic operculum, right inferior parietal cortex, left lingual gyrus and left precuneus.

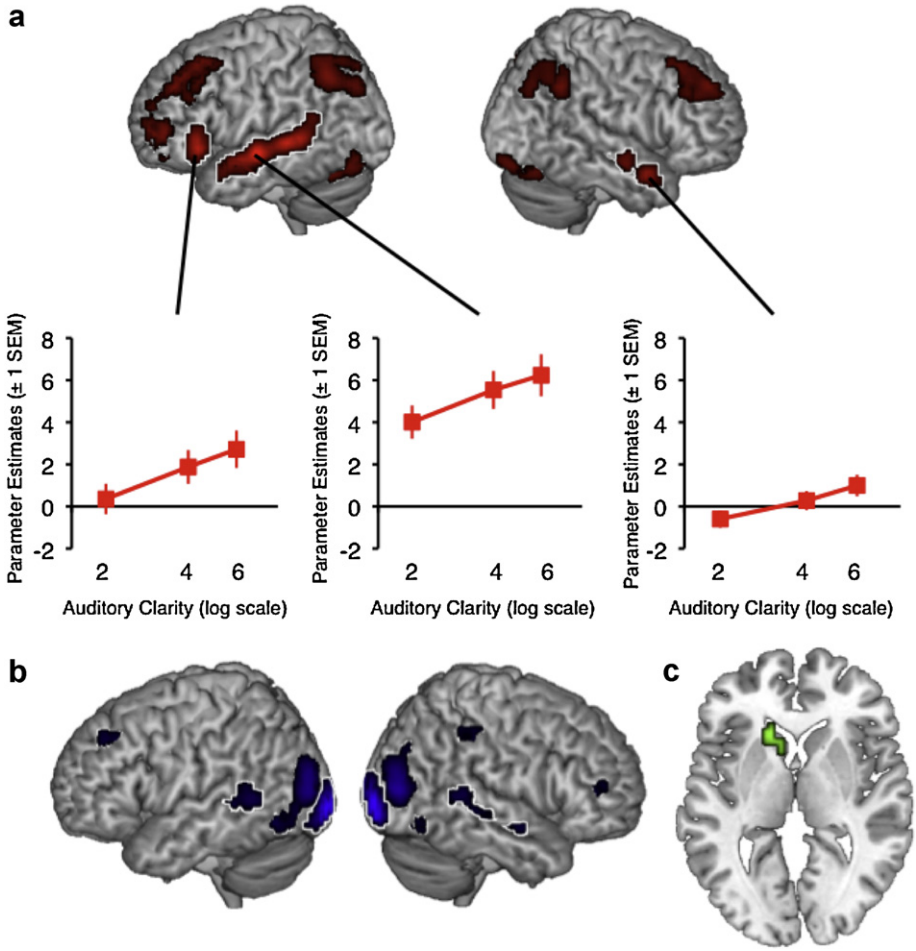

Fig. 2. Surface-rendered images of the Main Effects of Auditory Clarity (a), Visual Clarity (b) and Linguistic Predictability (c). Areas of activation outlined in white indicate regions where fMRI signal increased with increasing visual, auditory or linguistic information. All other activations revealed a negative correlation. Plots of mean parameter estimates extracted from ROIs around the three peak activations illustrate the main effect of Auditory Clarity in those regions. Images are shown at a height threshold $p < .005$ (uncorrected) and a cluster extent threshold of 20 voxels.

The main effect of Linguistic Predictability revealed a single cluster of activation in left caudate nucleus, which showed an overall decrease in activation for sentences of higher predictability.

Significant peak voxels (as well as sub-peak voxels if more than 8 mm apart) for the Main Effects are listed in Table 1.

3.2.2. *Second-level factorial ANOVA: interactions*

The factorial design allowed us to explore interactions between the three sources of information in the stimuli. Only those interactions that emerged as significant in the behavioural experiment were explored in the functional analysis – that is, the two-way interactions of Auditory Clarity × Visual Clarity and Auditory Clarity × Linguistic Predictability, and the three-way interaction of Auditory Clarity × Visual Clarity × Linguistic Predictability.

The three-way interaction did not yield any significant activation at the chosen threshold. However, the two-way interaction of Auditory Clarity × Visual Clarity gave a single cluster in left supramarginal gyrus, and the interaction of Auditory Clarity × Linguistic Predictability gave clusters in left supplementary motor area (SMA) and a region encroaching on left cuneus (Fig. 3). Plots of mean parameter estimates taken from 4-mm ROIs around the peak voxels in the interactions (Fig. 3) showed a consistent enhancement of signal by increased visual/linguistic information at intermediate auditory clarity (i.e. 4 channels of noise vocoding). This suggested, as observed by Obleser et al. (2007) and Ross et al. (2007), that additional sources of information supporting speech comprehension are most effective at intermediate levels of intelligibility (though we note that this stands at odds with other studies demonstrating the property of inverse effectiveness; Senkowski, Saint-Amour, Hoefle, & Foxe, 2011; Stevenson et al., 2010; Werner & Noppeney, 2009).

Significant peak voxels (as well as sub-peak voxels if more than 8 mm apart) for the two-way interactions are listed in Table 1.

3.2.3. *Effects of intelligibility*

In the second-level ANOVA model, condition-by-condition group mean scores from the behavioural pre-test were used to generate a group T-contrast measuring correlates of increasing stimulus intelligibility in the experiment. This gave activity in bilateral STS/STG, bilateral IFG and a cluster on left fusiform gyrus (Table 2 and Fig. 4a). Plots of mean contrast estimates across the conditions (Fig. 4a) showed that several of the peak voxels exhibited an activation profile strongly representative of the behavioural effects of all three factors on stimulus intelligibility (Fig. 1b), more clearly so for the conditions at intermediate auditory clarity (4 channels of noise-vocoding). Thus, although these activations do not express the full statistical interaction of the three factors in the neural data, there are clear indications that the method is sufficiently sensitive to capture responses tracking intelligibility changes across the conditions.

Informed by the findings of the ANOVA two-way interaction contrasts, and based on the findings of Obleser et al. (2007) in

Table 1
Results of the main effects contrasts for Auditory Clarity, Visual Clarity and Predictability (all conditions). Coordinates and statistics are given for peak voxels and local maxima more than 8 mm apart. Asterisks indicate activation sites where the signal showed a negative relationship with increased information from the input factor. The last column shows the voxelwise significance level, to 3 decimal places. Coordinates are given in Montreal Neurological Institute (MNI) stereotactic space. STS = superior temporal sulcus, STG = superior temporal gyrus, PT = planum temporale, IFG = inferior frontal gyrus, MOG = middle occipital gyrus, SPL = superior parietal lobule, SMA = supplementary motor area, MFG = middle frontal gyrus, SFG = superior frontal gyrus, IPL = inferior parietal lobe, IOG = inferior occipital gyrus, ITG = inferior temporal gyrus.

Contrast	No of voxels	Region	Coordinates			*z*	Sig (voxel)
			x	*y*	*z*		
Main Effect of Auditory Clarity	446	Left STS	−57	−12	−6	27.27	.000
		Left STS	−51	−42	3	22.76	.000
	125	Left STS	−60	−33	3	22.62	.000
	123	Left STG/STS	−57	−3	−9	16.17	.000
		Left STG (Temporal Pole)	−51	9	−18	9.34	.000
	58	Posterior STS	−60	−51	18	8.14	.000
		Posterior medial PT	−42	−48	21	7.40	.001
	352	Left IFG (pars orbitalis)	−48	27	−3	19.63	.000
		Right STS (Temporal Pole)	51	6	−21	17.20	.000
	109	Right STS	51	−18	−9	11.24	.000
		Right STS	54	−9	−12	10.39	.000
	175	Right Heschl's gyrus	36	−33	18	14.06	.000
		Right Heschl's gyrus	42	−27	9	6.93	.005
	143	Left precuneus*	−6	−63	51	13.72	.000
		Left angular gyrus*	−45	−60	45	12.13	.000
	41	Left MOG*	−36	−78	39	9.06	.000
	228	Left SPL*	−24	−66	54	7.68	.001
		Left SPL*	−12	−81	48	6.97	.005
	27	Left mid cingulate cortex*	−3	−30	36	12.27	.000
	96	Right mid cingulate cortex*	6	−27	39	10.12	.000
		Right SMA*	9	−27	51	8.45	.000
	26	Left MFG*	−27	24	51	12.23	.000
	95	Left MFG*	−30	42	36	11.83	.000
		Left MFG*	−42	27	42	8.37	.000
	46	Left MFG*	−39	33	33	6.88	.005
		Left MFG*	−39	12	54	6.73	.005
	25	Left MFG *	−36	51	3	11.78	.000
	26	Left MFG*	−27	57	6	10.37	.000
		Left middle orbital gyrus*	−33	51	−9	8.14	.000
	69	Left SFG*	−24	57	24	7.05	.005
		Right IPL*	57	−51	42	11.42	.000
		Right SFG*	27	48	39	11.10	.000
		Right SFG*	24	27	48	10.49	.000
		Right MFG*	33	36	42	8.42	.000
		Right MFG*	42	33	36	7.52	.001
		Right MFG*	27	18	39	6.19	.005
		Left calcarine gyrus*	−3	−99	0	9.66	.000
		Right IOG*	33	−90	−15	9.39	.000
		Right cerebellum (lobule VI)*	33	−69	−21	8.22	.000
		Left IFG (pars opercularis)	−42	15	21	9.31	.000
		Left fusiform gyrus*	−36	−75	−18	8.96	.000
		Left cerebellum (lobule VI)*	−36	−66	−21	8.71	.000
		Left IOG*	−45	−78	−12	8.60	.000
		Left Rolandic operculum*	−33	−30	18	8.87	.000
		Left Rolandic operculum*	−36	−39	18	8.72	.000
		Right STG*	60	−30	15	8.79	.000
		Right IFG (pars orbitalis)*	54	30	−3	7.73	.001
		Right IFG (pars triangularis)*	57	24	3	7.48	.001
		Right IFG (pars orbitalis)*	45	30	−6	7.19	.001
		Right angular gyrus*	42	−72	42	7.45	.001
		Right angular gyrus*	42	−60	45	6.82	.005
		Right SPL*	39	−60	54	6.57	.005
Main Effect of Visual Clarity	149	Right calcarine gyrus	27	−96	0	64.82	.000
	117	Left MOG	−21	−99	−3	63.32	.000
	243	Right MOG*	39	−81	12	37.15	.000
	159	Right STS	48	−39	9	30.25	.000
		Right STG	57	−3	−9	12.70	.000
	276	Right STS	54	−15	−9	9.12	.005
		Left MOG *	−36	−90	12	29.33	.000
	58	Left MOG*	−30	−87	18	28.01	.000
	29	Left IOG*	−39	−66	−9	20.73	.000
		Left MOG*	−48	−81	0	12.47	.000
	101	Left IOG*	−51	−72	−9	10.27	.005
		Left anterior cingulate*	−9	51	0	26.29	.000
	46	Right fusiform gyrus	45	−48	−18	18.22	.000
		Right fusiform gyrus	39	−39	−21	11.78	.001
	57	Left STS	−51	−51	6	17.45	.000
	36	Left STS	−54	−42	6	12.81	.000
		Left STG	−60	−48	15	11.74	.001
	22	Right IOG*	36	−69	−9	17.21	.000

Table 1 (*Continued*)

Contrast	No of voxels	Region	Coordinates			*z*	Sig (voxel)
			x	*y*	*z*		
	23	Right fusiform gyrus	27	−66	−12	12.88	.000
	26	Right ITG	48	−69	−9	9.19	.005
	27	Left MFG*	−27	33	39	13.97	.000
	40	Left Rolandic operculum*	−45	−6	15	13.88	.000
		Left insula	−36	−9	9	12.90	.000
	21	Right MFG*	36	45	12	13.60	.000
		Left fusiform gyrus	−39	−48	−18	13.41	.000
		Right IPL*	42	−42	48	13.28	.000
		Left lingual gyrus*	−21	−69	−6	12.70	.000
		Left precuneus*	−6	−72	39	12.22	.001
		Left precuneus	−15	−66	36	8.71	.005
		Right hippocampus (amygdala)	18	−6	−12	11.83	.001
Main Effect of Linguistic Predictability	31	Left caudate nucleus*	−15	21	3	13.08	.000
		Left caudate*	−6	12	0	9.48	.005
Interaction: Auditory Clarity × Visual Clarity	28	Left supramarginal gyrus	−57	−33	36	11.02	.000
Interaction: Auditory Clarity × Linguistic Predictability	29	Left SMA	−3	−3	51	7.76	.001
			−12	−24	48	7.12	.001
		Left calcarine gyrus	−6	−15	54	6.83	.005
			−3	−66	21	7.31	.001
			−18	−60	18	6.67	.005

which maximal effects of predictability were observed at intermediate auditory clarity, a second full factorial (2 × 2) ANOVA model was constructed at the second level to explore effects of visual and predictability information at 4 channels only. A 2 × 2 repeated-measures ANOVA of the behavioural data for conditions with 4 channels of auditory clarity gave significant main effects of Visual Clarity ($F(1, 11) = 34.38$, $p < 0.001$) and Linguistic Predictability ($F(1, 11) = 58.45$, $p < 0.0001$), and a significant interaction of both factors ($F(1, 11) = 21.47$, $p < 0.001$) reflecting a more pronounced intelligibility enhancement for increased visual clarity with high linguistic predictability. In the neural data, a T-contrast describing this intelligibility profile revealed activations in bilateral posterior STS, bilateral fusiform gyrus, bilateral calcarine gyrus, left inferior frontal gyrus (Brodmann areas 45 and 44), left SMA and right anterior STG/STS (Table 3 and Fig. 4b). Several of these sites of activation – bilateral posterior STS, bilateral calcarine gyrus and right fusiform gyrus – overlapped with activations showing a main effect of Visual Clarity (see Table 3). A significant main effect of Linguistic Predictability was found in a site encroaching on left angular gyrus, while there were no sites of significant interaction of the two factors at the chosen threshold. Nonetheless, the activation profiles in the peak activations for the intelligibility contrast indicated increased responses to greater predictability in all sites, with several regions (posterior STS, fusiform gyrus, IFG, SMA) exhibiting a

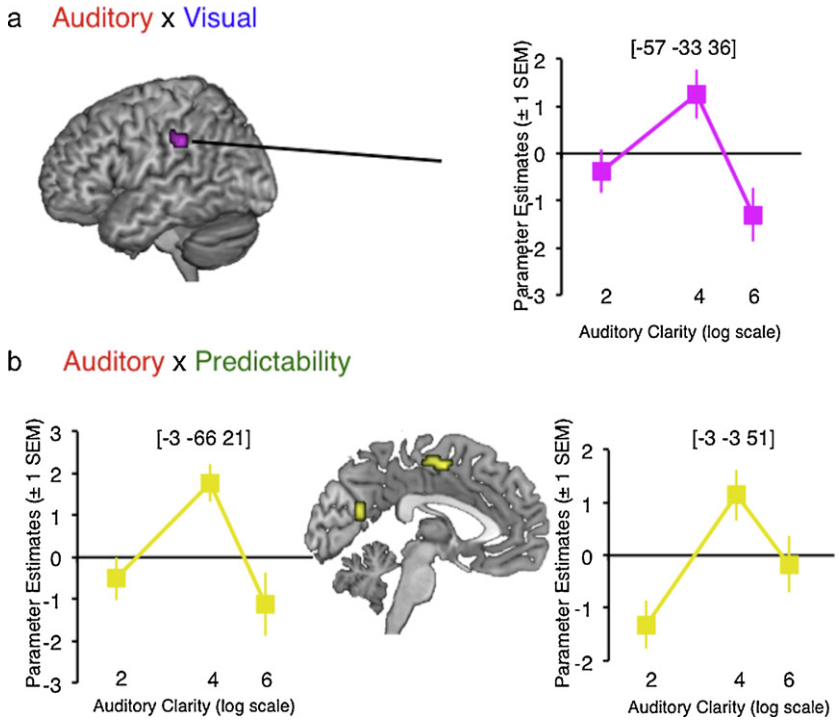

Fig. 3. Surface-rendered images of the interaction of Auditory Clarity × Visual Clarity (a) and Auditory Clarity × Linguistic Predictability (b). Plots of mean parameter estimates illustrate the activation profiles in the interactions (i.e. the change in mean parameter estimate associated with increased Visual Clarity/Linguistic Predictability at each level of Auditory Clarity). Images are shown at a height threshold $p < .005$ (uncorrected) and a cluster extent threshold of 20 voxels.

Table 2
Results of contrasts based on Intelligibility. Coordinates and statistics are given for peak voxels and local maxima more than 8 mm apart. The last column shows the voxelwise significance level, to 3 decimal places. Coordinates are given in Montreal Neurological Institute (MNI) stereotactic space. STS = superior temporal sulcus, STG = superior temporal gyrus, IFG = inferior frontal gyrus.

Contrast	No of voxels	Region	Coordinates			*z*	Sig (voxel)
			x	*y*	*z*		
Increasing intelligibility (all conditions)	558	Left STS	−57	−36	3	6.36	.000
		Left STS	−57	−12	−6	5.96	.000
	195	Posterior STS/STG	−54	−51	18	4.80	.000
		Left STG	−54	0	−9	4.32	.000
	236	Left STG (Temporal Pole)	−51	9	−18	3.65	.000
		Left STS (medial Temporal Pole)	−45	15	−27	3.31	.001
	35	Left STS	−48	−18	−12	3.21	.001
		Left IFG (pars triangularis)	−51	30	3	6.29	.001
	50	Left IFG (pars opercularis)	−42	15	21	3.94	.000
		Left IFG (pars opercularis)	−51	15	18	3.90	.000
		Right STS (Temporal Pole)	48	6	−21	5.01	.000
		Right STS	48	−36	3	4.55	.000
		Right STG	57	−6	−9	4.39	.000
		Right STS	51	−18	−9	3.83	.000
		Right STS	60	−27	3	3.01	.005
		Left fusiform gyrus	−42	−39	−18	3.96	.000
		Left fusiform gyrus	−39	−30	−18	2.99	.005
		Right IFG (pars orbitalis)	45	30	−6	3.83	.000

marked enhancement in signal for the items with both higher visual clarity and predictability, reflective of the interaction observed in the behavioural data.

3.2.4. Individual differences

The behavioural post-test showed a wide range of variability in performance of the post-test (mean: 12.8/36 (36%) correct, SD: 3.4, range: 7–19 items correct). As expected, stimuli from the baseline condition were completely unintelligible. The overall mean score for the scanning participants was slightly lower than the grand mean obtained in the full behavioural experiment (45%). This difference likely reflects item effects: the post-test was designed to be a brief post-scan assessment of performance, and so used a short, fixed sentence list to allow direct comparison across individuals. In

Table 3
Results of a 2 × 2 ANOVA exploring conditions of intermediate Auditory Clarity. Coordinates and statistics are given for peak voxels and local maxima more than 8 mm apart. Asterisks indicate activation sites where the signal showed a negative relationship with increased information from the input factor. The last column shows the voxelwise significance level, to 3 decimal places. Coordinates are given in Montreal Neurological Institute (MNI) stereotactic space. STS = superior temporal sulcus, STG = superior temporal gyrus, IFG = inferior frontal gyrus, MOG = middle occipital gyrus, SMA = supplementary motor area, IOG = inferior occipital gyrus.

Contrast	No of voxels	Region	Coordinates			*z*	Sig (voxel)
			x	*y*	*z*		
Main Effect of Visual Clarity	91	Right calcarine gyrus	27	−96	0	5.47	.000
		Right inferior occipital gyrus	15	−99	9	3.59	.000
	88	Left calcarine gyrus	−12	−102	−6	5.25	.000
		Left IOG	−24	−96	−9	4.88	.000
	67	Left MOG*	−24	−81	18	4.04	.000
		Left MOG*	−30	−90	21	3.77	.000
	64	Left STS	−54	−51	6	4.00	.000
	67	Right MOG*	36	−81	12	3.99	.000
	28	Right fusiform gyrus	39	−39	−21	3.70	.000
	73	Right STS	51	−45	12	3.64	.000
		Right STS	51	−36	0	3.33	.000
	23	Right STS	54	−3	−12	3.29	.001
Main Effect of Linguistic Predictability	23	Left middle occipital gyrus	−45	−78	27	3.49	.000
		Left angular gyrus	−42	−69	33	2.67	.005
Increasing Intelligibility	74	Right calcarine gyrus	27	−96	0	4.81	.000
		Right cuneus	12	−99	9	2.94	.005
	71	Left MOG	−21	−99	−3	4.69	.000
		Left calcarine gyrus	−12	−93	−3	3.30	.000
	93	Right STS	51	−45	9	4.35	.000
		Right STS	60	−33	0	2.83	.005
	26	Right STG	63	−36	12	2.81	.005
		Left IFG (pars triangularis)	−51	30	3	4.03	.000
	61	Left IFG (pars triangularis)	−54	27	15	2.82	.005
		Left fusiform gyrus	−42	−54	−18	3.83	.000
	42	Left fusiform gyrus	−39	−45	−24	3.38	.000
		Left cerebellum	−33	−63	−24	2.98	.005
	32	Left SMA	−6	−18	57	3.58	.000
		Left SMA	−3	−3	54	2.74	.005
	31	Right fusiform gyrus	42	−51	−18	3.43	.000
		Right fusiform gyrus	33	−45	−18	2.85	.005
	20	Right STS	57	−3	−9	3.39	.000
		Right STS	57	−9	−15	3.14	.000
		Left STS	−51	−51	15	3.06	.005

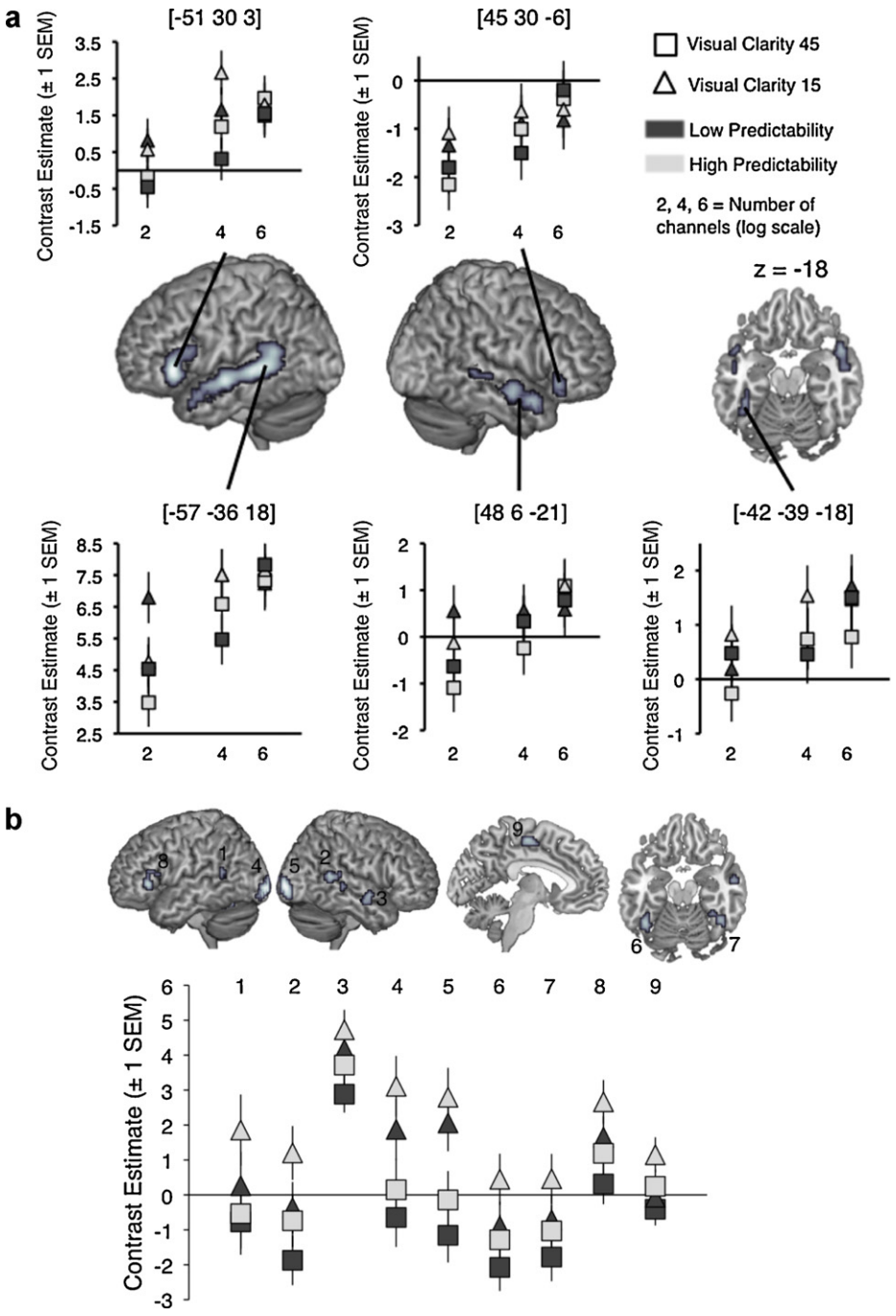

Fig. 4. (a) Results of a contrast describing the intelligibility scores across all 12 conditions, with plots of contrast estimates for each condition (compared with baseline) taken from local peak activations in bilateral STS, IFG and left fusiform gyrus. (b) Results of a contrast describing the intelligibility profile across conditions at 4 channels (Auditory Clarity) only. Images are shown at a height threshold $p < .005$ (uncorrected) and a cluster extent threshold of 20 voxels.

Table 4
Results of an individual differences analysis. Coordinates and statistics are given for peak voxels and local maxima more than 8 mm apart. The last column shows the voxelwise significance level, to 3 decimal places. Coordinates are given in Montreal Neurological Institute (MNI) stereotactic space. IFG = inferior frontal gyrus, STS = superior temporal sulcus.

Contrast	No of voxels	Region	Coordinates			*z*	Sig (voxel)
			x	*y*	*z*		
Positive correlation with post-test performance	39	Left IFG (pars triangularis)	−51	27	6	5.62	.000
		Left IFG (pars triangularis)	−51	18	18	3.22	.005
	23	Left STS	−54	−54	18	3.74	.001

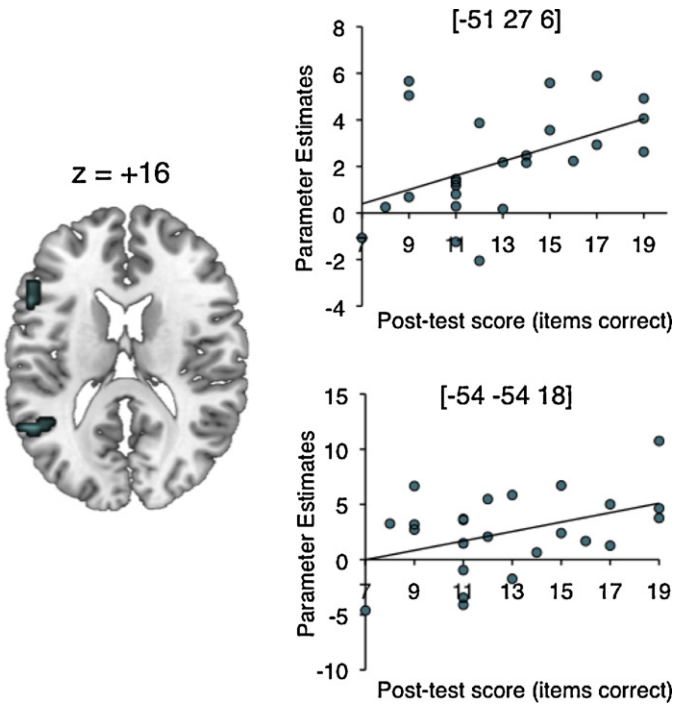

Fig. 5. Results of the individual differences analysis showing brain areas where activity across all speech conditions correlated positively with performance on the sentence comprehension post-test. Scatter-plots show the correlations between signal and behaviour in the peak voxels, with linear trend lines of best fit. The brain image is shown at a height threshold of $p < .005$ (uncorrected) and a cluster extent threshold of 20 voxels.

a second-level *t*-test model, each participant's contrast image for Intelligibility was masked inclusively ($p < .05$ uncorrected) with a covariate containing individual scores from the behavioural post-test. The resulting contrast image was thresholded at $p < .005$ with a cluster extent threshold of 20 voxels. This revealed clusters on the left inferior frontal gyrus ([−51 27 6]) and left posterior STS ([−51 −54 18]) – see Table 4 and Fig. 5.

4. Discussion

The results indicate interactivity of sensory and linguistic information, both behaviourally and neurally, in the perception of multimodal speech. Enhanced responses to improved Auditory Clarity were seen along the length of left dorsolateral temporal cortex, and in the right anterior dorsolateral temporal lobe, as well as bilateral IFG. These results are consistent with previous studies using parametric manipulations of stimulus intelligibility (Bishop & Miller, 2009, Stevenson & James, 2009; Davis & Johnsrude, 2003; Obleser et al., 2007; Scott et al., 2006, 2002). There were also numerous sites of reduced signal for increasing Auditory Clarity, in medial prefrontal cortex, cingulate cortex and parietal sites including bilateral angular gyrus. Several of these regions have been consistently implicated in the default-state network and the observed modulations may indicate a relative disengagement of attention from the task as the speech becomes more degraded (Buckner, Andrews-Hanna, & Schacter, 2008). However, while the activations in some regions, (e.g. cingulate cortex), showed overall 'de-activation' relative to the baseline condition, others (e.g. left planum temporale/Rolandic operculum) showed relative decreases (with increasing auditory clarity) in the context of greater activity for the speech trials compared with baseline. This is indicative of several mechanisms at play within this network during speech comprehension (Leech, Kamourieh, Beckmann, & Sharp, 2011; Lin, Hasson, Jovicich, & Robinson, 2011).

Increased signal in response to greater Visual Clarity was observed in visual areas associated with object recognition (the visual 'what' stream: Goodale & Milner, 1992), including primary and secondary visual cortex bilaterally, and bilateral fusiform gyrus. There was also activation in right amygdala, which has previously been associated with face processing (Haxby, Hoffman, & Gobbini, 2000). Activations in bilateral posterior STS, in regions overlapping with those responding to increases in Auditory Clarity, support the implication of this region in multisensory integration (Beauchamp et al., 2004a, 2004b; Beauchamp, Nath, & Pasalar, 2010; Calvert et al., 2000; Calvert, 2001; Reale et al., 2007; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003) and speech reading (Bernstein, Jiang, Pantazis, Lu, & Joshi, 2011; Calvert & Campbell, 2003; Hall, Fussell, & Summerfield, 2005; MacSweeney et al., 2002). The extent of superior temporal activation was greater on the right, and extended anterior to primary auditory cortex. Brefczynski-Lewis, Lowitszch, Parsons, Lemieux, & Puce, 2009 found a right-hemisphere dominance in fMRI and EEG data for the visual and audiovisual processing of non-verbal stimuli such as coughs and sneezes – they related this to a tendency for right-lateralized responses to social stimuli. There may be an alternative explanation, however – Stevenson et al. (2010) found that peak AV integration sites in left pSTS were showed greater spatial variability in the left hemiphere than in the right, leading to a robust right-lateralized effect emerging in the group analysis. Overall, including facial information may lead to more bilateral distribution of temporal lobe responses for audiovisual speech (Scott et al., 2002), as opposed to the left-dominant intelligibility effects seen for auditory-only speech (Eisner et al., 2010; Narain et al., 2003; Scott et al., 2000, 2006). The current results clearly indicate bilateral streams of processing in superior temporal sulcus (STS) for auditory and visual information in speech, with the implication of shared substrates for the effects of the two modalities. This is consistent with evidence from non-human primates suggesting that interactions between early auditory areas and the STS facilitate the integration of face and voice information (Ghazanfar, Maier, Hoffman, & Logothetis, 2005).

Previous studies have found enhanced responses associated with greater linguistic predictability in sites outside temporal cortex (Obleser & Kotz, 2010; Obleser et al., 2007). In the current experiment, a contrast exploring the main effect of Linguistic Predictability gave only one cluster in left caudate nucleus, which exhibited greater signal for *lower* predictability sentences. The caudate has been previously implicated in language tasks, including a study of language-switching in bilingual participants (Crinion et al., 2006). Friederici (2006) interprets caudate activation in language studies as evidence for the recruitment of control processes in language comprehension. Consistent with this, Stevenson et al. (2009) found that the caudate showed greater activation in response to lower-quality stimuli in their study of multisensory integration. The current data indicate that perceiving less predictable items required recruitment of the caudate to support task performance.

A targeted analysis of the main effect of Linguistic Predictability at intermediate Auditory Clarity (4 channels) in the current experiment revealed an activation in left inferior parietal cortex that encroached upon angular gyrus, with a local peak in a very similar location to sites observed by both Obleser et al. (2007) and Obleser and Kotz (2010) as exhibiting an enhanced response to greater predictability in sentences at intermediate auditory intelligibility. Obleser and Kotz (2010) identify the inferior parietal cortex as "a postsensory interface structure that taps long-term semantic knowledge/memory" (p. 638). Our results suggest an important replication of these previous findings, despite the limited power in analyzing only a subset of conditions in the current study.

The employment of a fully factorial design in the current experiment allowed us to explore statistical interactions of the

factors in the neural signal. Previous studies exploring the effects of facial information (Scott et al., 2002) and linguistic manipulations (Obleser & Kotz, 2010) on auditory speech intelligibility have reported modulation of responses in superior temporal cortex that reflect the varying influences of these additional information sources at different levels of acoustic clarity. Obleser et al. (2007) took advantage of a behavioural interaction of auditory clarity and linguistic predictability to explore neural responses to predictability at the level of auditory degradation where these effects were most pronounced. However, none of these studies reported full statistical interactions in the neural data. In the current experiment, we report statistically significant interactions between Auditory Clarity and Visual Clarity (in left supramarginal gyrus), and between Auditory Clarity and Linguistic Predictability (in left SMA and cuneus). In both cases, the regions showing the interaction exhibited a profile of activation across conditions that showed an increase of the BOLD response for greater visual clarity/predictability at 4 channels of noise vocoding, but decreases at 2 and 6 channels. This pattern of results departs from the behavioural data, where increased facial clarity and predictability enhanced the intelligibility of sentences at all levels of auditory clarity, and where predictability enhancements were slightly greater at 6 channels than at 4 channels – thus, in the SMG, SMA and cuneus, there is not a monotonic relationship between intelligibility and the neural signal. The profile of activation in these regions might reflect the engagement of specific integration processes to support speech understanding at 4 channels, when both the auditory and visual streams contain sufficient detail for partial comprehension, and no source of information is redundant. Obleser and colleagues (2007) observed that additional brain regions outside Broca's and Wernicke's areas were engaged by greater semantic predictability at an intermediate level of auditory clarity – we take this finding a step further by describing the contribution of visual and linguistic information in the context of statistical interactions observed across all tested levels of auditory intelligibility.

The SMG and inferior parietal cortex have been repeatedly implicated in audiovisual integration (Bernstein et al., 2008a, 2008b; Calvert & Campbell, 2003; Calvert et al., 2000; Dick, Solodkin, & Small, 2010; Jones & Callan, 2003; Miller & D'Esposito, 2005; Pekkola et al., 2006; Skipper, Nusbaum, & Small, 2005). The specific role for this general region has been rather dependent on the task design in each study. Several authors have reported enhancement of signal in SMG for visual sequences of stills suggesting speech movements, compared with images of a motionless face (Calvert et al., 2000; Ojanen et al., 2005), while another study reported increased signal in this region for audiovisual speech compared with visual or auditory signals alone (Skipper et al., 2005). Some studies manipulating the auditory and visual content of AV stimuli have reported increased signal in right supramarginal gyrus and inferior parietal lobe when the auditory and visual signals are mismatched (in timing or content; Jones & Callan, 2003; Miller & D'Esposito, 2005; Pekkola et al., 2006) compared with matched, while other authors have reported these effects in left SMG when the participants' susceptibility to the 'fused' illusion percept is taken into account in analyses (Bernstein et al., 2008a, 2008b; Hasson, Skipper, Nusbaum, & Small, 2007). Bernstein et al. (2008a) presented EEG data showing a prominent role for left SMG and angular gyrus (AG) in audiovisual processing, in which early activations in these areas persisted to a late phase response and showed differential sensitivity to congruent and incongruent stimuli – Bernstein et al. (2008b) argue that this "sensitivity [in SMG] to multisensory incongruity could be implemented as a comparison of bottom-up auditory and visual perceptual representations with stored knowledge of the normal relationships between auditory and visual patterns" (p. 179).

The supplementary motor area is strongly implicated in motor control aspects of speech production (Alario, Chainay, Lehericy, & Cohen, 2006; Blank, Scott, Murphy, Warburton, & Wise, 2002). In the context of speech perception, establishing the precise role for motor cortex is controversial topic (Lotto, Hickok, & Holt, 2009; Scott, McGettigan, & Eisner, 2009), however some authors have put forward the hypothesis that motor processes or representations may be engaged to support auditory speech comprehension processes in the temporal lobe, when the speech perception task is difficult (e.g. if the speech is degraded) or loads specifically on motoric processes such as phonemic segmentation (Sato, Tremblay, & Gracco, 2009; Osnes et al., 2011). Osnes et al. (2011) presented listeners with a morphed continuum of sounds between a noise and a spoken syllable. They found that premotor cortex was activated at an intermediate point on the continuum, where the speech was noisy but could still be recognized. Evidence for a motoric involvement in speech perception could reflect the comparison of ambiguous/degraded incoming auditory signals with articulatory representations to facilitate comprehension (Davis & Johnsrude, 2007). In the current study, however, SMA was implicated in an interaction with linguistic predictability. As pointed out by Obleser and Kotz (2010), manipulations of linguistic expectancy in sentences likely bring associated expectancies at the acoustic–phonetic level (i.e. expectancies related to the phonetic/phonemic content of upcoming words). If so, the current observed interaction in SMA may reflect an articulatory strategy supporting speech comprehension at 4 channels, where the signal contains sufficient detail to build expectancies at the segmental level that are more likely to be correct for high predictability items. Given our instructions to the participants that they should not overtly articulate during the experiment, we assume this process to have occurred covertly, but not necessarily consciously. This strategy at 4 channels may be distinct from the process supporting the large behavioural predictability effects at 6 channels, which could reflect a higher-order lexical/syntactic/semantic listening strategy when the auditory signal is more intelligible. Thus, with our interaction analyses, we uncover a possible fractionation of a complex speech comprehension system in which multiple overlapping processes contribute to the overall outcomes measured by the behavioural task.

The results of the behavioural experiment clearly show that the benefits of increased Visual Clarity and Linguistic Predictability for speech intelligibility varied with Auditory Clarity. Visual manipulations were more helpful than enhanced linguistic information when Auditory Clarity was poor, and increased Linguistic Predictability was more effective than visual information when Auditory Clarity was greatest. When the behavioural intelligibility profile for all 12 conditions was modelled as a contrast in the fMRI second-level ANOVA, activation in bilateral STS/STG and IFG and the left fusiform gyrus was revealed to be positively correlated with increases in intelligibility Activity in STS and IFG has been previously identified in parametric responses to speech intelligibility (Davis and Johnsrude, 2003; Obleser et al., 2007; Scott et al., 2006), while the fusiform has been implicated in audiovisual speech perception (Kawase et al., 2005; Nath & Beauchamp, 2011). Motivated by the observed interactions in the behavioural test and the functional interactions described above, which indicate that the contribution of visual and linguistic cues is maximal at intermediate auditory clarity, a further analysis explored activations in response to conditions at 4 channels of auditory clarity only. A contrast describing the behavioural intelligibility profile at this level of auditory clarity revealed activation in bilateral calcarine gyrus, posterior STS and fusiform gyrus, left IFG (pars triangularis), left SMA and right anterior STS. The basal language area (Luders, Lesser, & Hahn, 1986) has been implicated in responses to intelligible speech in a number of previous studies (Awad,

Warren, Scott, Turkheimer, & Wise, 2007; Spitsyna, Warren, Scott, Turkheimer, & Wise, 2006). Ours is a novel demonstration of sensitivity in this region to both visual and linguistic intelligibility modulations in speech, although parts of ventral temporal cortex have long been linked to aspects of semantic (Balsamo, Xu, & Gaillard, 2006; Binder et al., 1997; Chee, O'Craven, Bergida, Rosen, & Savoy, 1999; Kuperberg et al., 2000; Spitsyna et al., 2006) and facial (reviewed in Haxby et al., 2000 and Kanwisher, 2010) processing. In a sentence-processing experiment with EEG, Dien, Frishkoff, Cerbone, & Tucker (2003) localized a semantic expectancy ERP (a negativity around 200 ms post-stimulus) to left fusiform gyrus. In fMRI, Kuperberg et al. (2000) found common left fusiform/inferior temporal gyrus activation in response to violations of pragmatic, syntactic and semantic content of auditory sentences. The basal language region has been infrequently implicated in neuroimaging studies of auditory speech intelligibility (cf in aphasia: Sharp, Scott, & Wise, 2004), although authors who have shown selective enhancement of responses in the basal language area during attention to the linguistic content of speech have interpreted these findings in terms of a role for visual representations in the processing of auditory speech (von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Yoncheva, Zevin, Maurer, & McCandliss, 2010). The fusiform gyrus has previously been described as a multimodal 'convergence zone' (Büchel, Price, Frackowiak, & Friston, 1998). Büchel et al. ascribed a rather more low-level combinatorial role to their site. However, Kassuba and colleagues have identified the left fusiform gyrus as a critical structure in multimodal object recognition (Kassuba et al., 2011), and other studies have associated the fusiform gyrus with aspects of audiovisual speech perception (Nath & Beauchamp, 2011; Stevenson et al., 2010; Wyk et al., 2010). Our data are novel in demonstrating evidence for audiovisual speech processing in fusiform gyrus that is sensitive to the overall intelligibility of speech, including higher-order linguistic manipulations at sentence level.

In sum, the data indicate strong interactivity of the experimental factors, where responses to the visual and linguistic manipulations were most pronounced for 4-channel noise-vocoded speech. This is in line with previous studies exploring the effects of visual (Scott et al., 2002) and linguistic (Obleser et al., 2007) cues on auditory sentence comprehension. More recently, an MEG study of audiovisual speech revealed prestimulus engagement of a fronto-parietal network that was dependent on the extent to which participants experienced the McGurk illusion for mismatched audiovisual syllables (Keil, Muller, Ihssen, & Weisz, 2012). Keil et al. highlight the importance of brain state in the perceptual 'fusion' of auditory and visual signals. It may be that a process of integration of the multiple information sources is used to generate an illusory percept of sorts, where participants 'hear' a sentence despite that fact that degradation has removed many of the normal cues to speech identification. However, this integrative process only occurs when it is maximally useful, i.e. when there is sufficient information in the signal to extract some, but not all of the meaning from one source alone. Further work concentrating on conditions supporting maximal integration of auditory, visual and linguistic factors will allow for more detailed description of this wider integration or 'repair' network.

There has been extensive debate about whether the processing of intelligibility in auditory speech is left-dominant or bilateral (Hickok & Poeppel, 2007; Obleser, Eisner & Kotz, 2008; Okada et al., 2010; Scott & Wise, 2004; Scott et al., 2000). Although the observed main effects of sensory information showed strongly bilateral responses in temporal cortex, an individual differences analysis found the strongest correlations with performance in the behavioural post-test in the left hemisphere, in inferior frontal gyrus and posterior STS. This lateralized response remained at reduced thresholds. This is strong indication that the processes most critical to supporting successful comprehension of degraded speech are performed in the left hemisphere. The peak activation in this analysis was in left IFG, in a similar site to that observed by Eisner et al. (2010) as a correlate of individual differences in learning to understand a cochlear implant simulation (spectrally shifted, noise-vocoded speech). Eisner et al. linked activation in posterior IFG to individual variability in working memory processes. The IFG activations observed in the current study implicated both Brodmann areas 44 and 45, with the more anterior involvement perhaps reflecting the involvement of higher-order linguistic manipulations (see Obleser et al., 2007). We also found performance-related variability in posterior STS in the left hemisphere. Both sites showed evidence of intelligibility-related enhancements in response to auditory, visual and linguistic manipulations, therefore we think it is unlikely that variability in these sites is reflective of differences in basic combinatorial sensory processing but rather some higher-order, post-perceptual response. This interpretation is supported by recent findings from Nath and colleagues, who found individual variability in left posterior STS underlying susceptibility to the McGurk illusion in adults and children (Nath and Beauchamp, 2012; Nath et al., 2011). In the context of growing evidence for posterior STS as a key site in the processing of socially relevant cues such as eye gaze direction (Haxby et al., 2000), our finding presents an important role for this site in naturalistic, multimodal spoken communication. Further work using the current stimuli will focus on the timecourse of processing, with particular emphasis on the relationship between left inferior frontal gyrus, STS and fusiform gyrus.

In conclusion, the manipulation of three different factors in an fMRI study of speech comprehension – one auditory, one visual, one linguistic – engaged a wide network of cortical and subcortical activations. Interaction analyses in the behavioural and functional data indicated a strong modulation of responses as the quality of the auditory signal was altered, with neural enhancement in inferior parietal and premotor cortex for visual and linguistic cues that was most pronounced at the intermediate level of auditory intelligibility. Crucially, despite strong bilateral responses to the main factors, there was clear evidence from an individual differences analysis that it is the left hemisphere that is most crucial in supporting speech comprehension performance. Our study is a novel demonstration of the power of multifactorial designs to investigate the speech comprehension network at a system level, within a richer, more naturalistic behavioural context.

Acknowledgements

C.M. and S.K.S. are funded by Wellcome Trust Grant WT074414MA awarded to S.K.S. J.O. is funded by the Max Planck Society. The authors would like to thank Kate Wakeling for assistance in stimulus preparation and the staff at the Birkbeck-UCL Centre for Neuroimaging for technical advice and support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuropsychologia.2012.01.010.

References

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*, 1124–1132.

Alario, F. X., Chainay, H., Lehericy, S., & Cohen, L. (2006). The role of the supplementary motor area (SMA) in word production. *Brain Research*, *1076*, 129–143.

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.

Balsamo, L. M., Xu, B., & Gaillard, W. D. (2006). Language lateralization and the role of the fusiform gyrus in semantic processing in young children. *NeuroImage*, *31*, 1306–1314.
Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics*, *3*, 93–113.
Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823.
Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unravelling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*, 1190–1192.
Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience*, *30*, 2414–2417.
Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*, 5–18.
Bernstein, L. E., Auer, E. T., Jr., Wagner, M., & Ponton, C. W. (2008). Spatio-temporal dynamics of audiovisual speech processing. *NeuroImage*, *39*, 423–435.
Bernstein, L. E., Lub, Z., & Jianga, J. (2008). Quantified acoustic optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Research*, *1242*, 172–184.
Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z. L., & Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Human Brain Mapping*, *32*, 1660–1676.
Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*, 353–362.
Bishop, C. W., & Miller, L. M. (2009). A multisensory cortical network for understanding speech in noise. *Journal of Cognitive Neuroscience*, *21*, 1790–1804.
Blank, S. C., Scott, S. K., Murphy, K., Warburton, E., & Wise, R. J. S. (2002). Speech production: Wernicke, Broca and beyond. *Brain*, *125*, 1829–1838.
Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer. Downloaded from: http://www.fon.hum.uva.nl/praat.
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.
Brefczynski-Lewis, J., Lowitszch, S., Parsons, M., Lemieux, S., & Puce, A. (2009). Audiovisual non-verbal dynamic faces elicit converging fMRI and ERP responses. *Brain Topography*, *21*, 193–206.
Brett, M., Anton, J., Valabregue, R., & Poline, J. (2002). Region of interest analysis using an SPM toolbox. In *Presented at the 8th international conference on functional mapping of the human brain* June 2–6, Sendai, Japan,
Büchel, C., Price, C., Frackowiak, R. S. J., & Friston, K. (1998). Different activation patterns in the visual cortex of late and congenitally blind subjects. *Brain*, *121*, 409–419.
Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38.
Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport*, *14*, 2213–2218.
Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*, 1110–1123.
Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57–70.
Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.
Chee, M. W. L., O'Craven, K. M., Bergida, R., Rosen, B. R., & Savoy, R. L. (1999). Auditory and visual word processing studied with fMRI. *Human Brain Mapping*, *7*, 15–28.
Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J. T., et al. (2006). Language control in the bilingual brain. *Science*, *312*, 1537–1540.
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.
Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language*, *114*, 101–114.
Dien, J., Frishkoff, G. A., Cerbone, A., & Tucker, D. M. (2003). Parametric analysis of event-related potentials in semantic comprehension: Evidence for parallel brain mechanisms. *Cognitive Brain Research*, *15*, 137–153.
Dubno, J. R., Ahlstrom, J. B., & Horwitz, A. R. (2000). Use of context by young and aged adults with normal hearing. *Journal of the Acoustical Society of America*, *107*, 538–546.
Edmister, W. B., Talavage, T. M., Ledden, P. J., & Weisskoff, R. M. (1999). Improved auditory cortex imaging using clustered volume acquisitions. *Human Brain Mapping*, *7*, 89–97.
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.
Friederici, A. D. (2006). What's in control of language? *Nature Neuroscience*, *9*, 991–992.
Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience*, *25*, 5004–5012.
Girin, L., Schwartz, J. L., & Feng, G. (2001). Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America*, *109*, 3007–3020.
Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*, 20–25.
Grant, K. W., & Seitz, P. F. (2000a). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *Journal of the Acoustical Society of America*, *107*, 1000–1011.
Grant, K. W., & Seitz, P. F. (2000b). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*, 1197–1208.
Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K., et al. (2009). Neural integration of iconic and unrelated coverbal gestures: A functional MRI study. *Human Brain Mapping*, *30*, 3309–3324.
Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *Journal of the Acoustical Society of America*, *87*, 2592–2605.
Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). Sparse temporal sampling in auditory fMRI. *Human Brain Mapping*, *3*, 213–223.
Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience*, *17*, 939–953.
Hasson, U., Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2007). Abstract coding of audiovisual speech: Beyond sensory representation. *Neuron*, *56*, 1116–1126.
Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–233.
Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*, 996–1009.
Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *Journal of the Acoustical Society of America*, *117*, 842–849.
Hickok, G., & Poeppel, D. (2007). Opinion – The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402.
Holle, H., Obleser, J., Rueschmeyer, S. A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*, 875–884.
James, T. W., & Stevenson, R. A. (2011). The use of fMRI to assess multisensory integration. In M. H. Wallace, & M. M. Murray (Eds.), *Frontiers in the neural basis of multisensory processes*. London: Taylor & Francis.
Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *Neuroreport*, *14*, 1129–1133.
Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech-intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, *61*, 1337–1351.
Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences United States of America*, *107*, 11163–11170.
Kassuba, T., Klinge, C., Holig, C., Menz, M. M., Ptito, M., Roder, B., et al. (2011). The left fusiform gyrus hosts trisensory representations of manipulable objects. *NeuroImage*, *56*, 1566–1577.
Kawase, T., Yamaguchi, K., Ogawa, T., Suzuki, K., Suzuki, M., Itoh, M., et al. (2005). Recruitment of fusiform face area associated with listening to degraded speech sounds in auditory–visual speech perception: A PET study. *Neuroscience Letters*, *382*, 254–258.
Keil, J., Muller, N., Ihssen, N., & Weisz, N. (2012). On the variability of the McGurk effect: Audiovisual integration depends on prestimulus brain states. *Cerebral Cortex*, *22*, 221–231.
Kim, J., & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication*, *44*, 19–30.
Kim, J., Davis, C., & Groot, C. (2009). Speech identification in noise: Contribution of temporal, spectral, and visual speech cues. *Journal of the Acoustical Society of America*, *126*, 3246–3257.
Kircher, T., Straube, B., Leube, D., Weis, S., Sachs, O., Wilmes, K., et al. (2009). Neural interaction of speech and gesture: Differential activations of metaphoric coverbal gestures. *Neuropsychologia*, *47*, 169–179.
Kuperberg, G. R., McGuire, P. K., Bullmore, E. T., Brammer, M. J., Rabe-Hesketh, S., Wright, I. C., et al. (2000). Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: An fMRI Study. *Journal of Cognitive Neuroscience*, *12*, 321–341.
Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, *166*, 289–297.
Leech, R., Kamourieh, S., Beckmann, C. F., & Sharp, D. J. (2011). Fractionating the default mode network: Distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *Journal of Neuroscience*, *31*, 3217–3224.
Lin, P., Hasson, U., Jovicich, J., & Robinson, S. (2011). A neuronal basis for task-negative responses in the human brain. *Cerebral Cortex*, *21*, 821–830.
Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*, 110–114.
Love, S. A., Pollick, F. E., & Latinus, M. (2011). Cerebral correlates and statistical criteria of cross-modal face and voice integration. *Seeing and Perceiving*, *24*, 351–367.
Luders, H., Lesser, R. P., Hahn, J., et al. (1986). Basal temporal language area demonstrated by electrical stimulation. *Neurology*, *36*, 505–510.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS One*, *4*. doi:10.1371/journal.pone.0004638

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C. R., et al. (2002). Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain*, *125*, 1583–1593.

McGettigan, C., Evans, S., Rosen, S., Agnew, Z. K., Shah, P., & Scott, S. K. (2011). An application of univariate and multivariate approaches in fMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. *Journal of Cognitive Neuroscience* [Epub ahead of print].

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience*, *25*, 5884–5893.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behaviour*, *2*, 217–228.

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.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, *31*, 1704–1714.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, *59*, 781–787.

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience*, *31*, 1963–1971.

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, *20*, 633–640.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*, 713–723.

Obleser, J., Wise, R. J. S., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*, 2283–2289.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *Journal of Neuroscience*, *28*, 8116–8123.

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage*, *25*, 333–338.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I., 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.

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, I. P., Kujala, T., et al. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: An fMRI study at 3 T. *NeuroImage*, *29*, 797–807.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, *97*, 593–608.

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., et al. (2007). Auditory–visual processing represented in the human superior temporal gyrus. *Neuroscience*, *145*, 162–184.

Rosen, S., Wise, R. J. S., Chadha, S., Conway, E. J., & Scott, S. K. (2011). Hemispheric asmmetries in speech perception: Sense, nonsense and modulations. *PLoS One*, *6*, e24672 [Epub].

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. *Cerebral Cortex*, *17*, 1147–1153.

Sato, M., Tremblay, P., & Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language*, *111*, 1–7.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*, B69–B78.

Scott, S. K., & Wise, R. J. S. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*, 13–45.

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., Rosen, S., Wickham, L., & Wise, R. J. S. (2004). A positron emission tomography study of the neural basis of informational and energetic masking efforts in speech perception. *Journal of the Acoustical Society of America*, *115*, 813–821.

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., McGettigan, C., & Eisner, F. (2009). 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., Spitsyna, G., Faulkner, A., Neville, L., Wise, R.J.S. (2002). The neural basis of cross modal enhancement in speech perception: A pet study. Society for Neuroscience Abstract Viewer and Itinerary Planner, Abstract No. 354.17.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory–visual speech perception examined by fMRI and PET. *Neuroscience Research*, *47*, 277–287.

Senkowski, D., Saint-Amour, D., Hoefle, M., & Foxe, J. J. (2011). Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness. *NeuroImage*, *56*, 2200–2208.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Sharp, D. J., Scott, S. K., & Wise, R. J. S. (2004). Retrieving meaning after temporal lobe infarction: The role of the basal language area. *Annals of Neurology*, *56*, 836–846.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage*, *25*, 76–89.

Slotnick, S. D., Moo, L. R., Segal, J. B., & Hart, J., Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Cognitive Brain Research*, *17*, 75–82.

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.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage*, *44*, 1210–1223.

Stevenson, R. A., Kim, S., & James, T. W. (2009). An additive-factors design to disambiguate neuronal and areal convergence: Measuring multisensory interactions between audio, visual, and haptic sensory streams using fMRI. *Experimental Brain Research*, *198*, 183–194.

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., & James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *NeuroImage*, *49*, 3308–3318.

Stickney, G. S., & Assmann, P. F. (2001). Acoustic and linguistic factors in the perception of bandpass-filtered speech. *Journal of the Acoustical Society of America*, *109*, 1157–1165.

Straube, B., Green, A., Weis, S., Chatterjee, A., & Kircher, T. (2009). Memory effects of speech and gesture binding: Cortical and hippocampal activation in relation to subsequent memory performance. *Journal of Cognitive Neuroscience*, *21*, 821–836.

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping*, *32*, 520–533.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215.

Thomas, S. M., & Pilling, M. (2007). Benefits of facial and textual information in understanding of vocoded speech. In *Auditory–visual speech processing 2007 (AVSP2007)* Hilvarenbeek, The Netherlands, August 31–September 3,. Accessed from: http://www.isca-speech.org/archive

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

Werner, S., & Noppeney, U. (2009). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*, *20*, 1829–1842.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*, 1034–1043.

Wyk, B. C. V., Ramsay, G. J., Hudac, C. M., Jones, W., Lin, D., Klin, A., et al. (2010). Cortical integration of audio-visual speech and non-speech stimuli. *Brain and Cognition*, *74*, 97–106.

Yoncheva, Y. N., Zevin, J. D., Maurer, U., & McCandliss, B. D. (2010). Auditory selective attention to speech modulates activity in the visual word form area. *Cerebral Cortex*, *20*, 622–632.