NeuroImage 245 (2021) 118764

Contents lists available at ScienceDirect

NeuroImage

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

Dissociating the functions of three left posterior superior temporal regions that contribute to speech perception and production

Justyna O. Ekert[a,*], Andrea Gajardo-Vidal[a,b], Diego L. Lorca-Puls[a], Thomas M.H. Hope[a], Fred Dick[d,e], Jennifer T. Crinion[c], David W. Green[d], Cathy J. Price[a]

[a] Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, 12 Queen Square, London WC1N 3AR, United Kingdom
[b] Faculty of Health Sciences, Universidad del Desarrollo, Concepcion, Chile
[c] Institute of Cognitive Neuroscience, University College London, London, United Kingdom
[d] Department of Experimental Psychology, University College London, London, United Kingdom
[e] Department of Psychological Sciences, Birkbeck University of London, London, United Kingdom

ARTICLE INFO

Keywords:
fMRI
Left posterior superior temporal sulcus
Reading
Naming
Repetition

ABSTRACT

Prior studies have shown that the left posterior superior temporal sulcus (pSTS) and left temporo-parietal junction (TPJ) both contribute to phonological short-term memory, speech perception and speech production. Here, by conducting a within-subjects multi-factorial fMRI study, we dissociate the response profiles of these regions and a third region – the anterior ascending terminal branch of the left superior temporal sulcus (atSTS), which lies dorsal to pSTS and ventral to TPJ. First, we show that each region was more activated by (i) 1-back matching on visually presented verbal stimuli (words or pseudowords) compared to 1-back matching on visually presented non-verbal stimuli (pictures of objects or non-objects), and (ii) overt speech production than 1-back matching, across 8 types of stimuli (visually presented words, pseudowords, objects and non-objects and aurally presented words, pseudowords, object sounds and meaningless hums). The response properties of the three regions dissociated within the auditory modality. In left TPJ, activation was higher for auditory stimuli that were non-verbal (sounds of objects or meaningless hums) compared to verbal (words and pseudowords), irrespective of task (speech production or 1-back matching). In left pSTS, activation was higher for non-semantic stimuli (pseudowords and hums) than semantic stimuli (words and object sounds) on the dorsal pSTS surface (dpSTS), irrespective of task. In left atSTS, activation was not sensitive to either semantic or verbal content. The contrasting response properties of left TPJ, dpSTS and atSTS was cross-validated in an independent sample of 59 participants, using region-by-condition interactions. We also show that each region participates in non-overlapping networks of frontal, parietal and cerebellar regions. Our results challenge previous claims about functional specialisation in the left posterior superior temporal lobe and motivate future studies to determine the timing and directionality of information flow in the brain networks involved in speech perception and production.

1. Introduction

The goal of this study is to investigate functional subdivisions for speech processing within the left posterior superior temporal lobe. Many prior studies have already identified a range of functions that activate this part of the brain. However, integrating the results from a large number of independent neuroimaging studies is challenging. This is particularly true when (i) brain regions are labelled in different ways; (ii) the same anatomical labels and peak activation coordinates are associated with different functions; and (iii) when the same function is associated with different brain regions. The current fMRI study addresses these problems by using (i) anatomical regions of interest and (ii) a within-subjects, multi-factorial design to functionally segregate the response profiles in different left superior temporal lobe regions during auditory speech processing, short-term memory and speech production.

After considering several independent atlases, the most fine grained subdivision of the left posterior superior temporal lobe was provided by the Human Connectome Project multi-modal parcellation (HCP-MMP1.0; Glasser et al., 2016). The HCP atlas distinguishes four anatomical components of interest (illustrated in Fig. 1) described as (i) the dorsal surface of the horizontal stem of pSTS (dpSTS), (ii) the ventral surface of the horizontal stem of pSTS (vpSTS); (iii) the ascending terminal branch of the superior temporal sulcus (atSTS) that is referred to as the temporal-parietal-occipital junction (TPOJ1) in the HCP atlas; and (iv) a region on the temporo-parietal junction (TPJ), at the boundary of the posterior superior temporal lobe with the supramarginal gyrus,

* Corresponding author.
E-mail address: justyna.ekert.14@ucl.ac.uk (J.O. Ekert).

https://doi.org/10.1016/j.neuroimage.2021.118764.
Received 15 April 2021; Received in revised form 15 November 2021; Accepted 24 November 2021
Available online 27 November 2021.

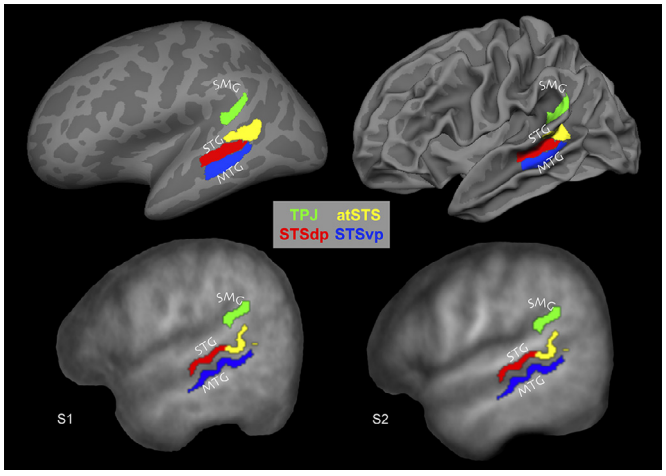

Fig. 1. Anatomical regions of interest from the Human Connectome Project. Regions of interest (ROIs) from the HCP-MMP1.0 atlas (Glasser et al., 2016; Horn, 2016). Top row: the ROIs are overlaid on two left hemisphere cortical surface reconstructions using csurf (Dale and Sereno, 1993). Bottom row: the ROIs are overlaid on the mean structural scans of the participants in the current study. Sample 1 (S1, left) includes 24 healthy participants. Sample 2 (S2, right) includes 59 healthy participants. As can be seen, the four regions of interest align well with the sulcal morphometry of both our samples. Green = the temporo-parietal junction (TPJ). Red = the dorsal surface of the horizontal section of pSTS (dpSTS). Blue = the ventral surface of the horizontal section of pSTS (vpSTS). Yellow = the anterior ascending terminal branch of the superior temporal sulcus (atSTS). SMG = supramarginal gyrus, STG = superior temporal gyrus, MTG = middle temporal gyrus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

that is referred to as the perisylvian language area in the HCP atlas. This is because it aligns with the functionally defined Sylvian parietal temporal region (Spt) that has been reported to respond to both speech perception and speech production in previous studies (Buchsbaum et al., 2001; Hickok et al., 2003; Hickok and Poeppel, 2004). In brief, we refer to these regions by the anatomical descriptions: dpSTS, vpSTS, atSTS and TPJ.

Almost all prior studies of the functional anatomy of speech and language processing did not make a distinction between the different parts of the left posterior superior temporal sulcus because these subdivisions started to emerge more recently with the advent of modern brain parcellation approaches such as that exemplified by the HCP-MMP1.0 atlas. Therefore, when highlighting the prior inconsistencies in structure-function relationships within the left posterior superior temporal lobe, we refer to the three different STS subdivisions collectively as the pSTS; and compare the conclusions drawn about left pSTS to those drawn about left TPJ before introducing how we attempted to reconcile prior inconsistencies.

1.1. The contribution of left pSTS to speech processing

Early functional imaging studies demonstrated that left pSTS activation is almost invariably observed during speech perception, even when acoustic processing is controlled by comparing speech to complex unintelligible sound stimuli, and irrespective of whether the speech is intelligible or not (Benson et al., 2006; Giraud et al., 2004; Hugdahl et al., 2003; Narain et al., 2003; Rimol et al., 2006; Scott et al., 2000). As part of left pSTS was also shown to be activated during verbal fluency (word generation without stimuli), Wise et al. (2001) proposed that the left pSTS is involved in transiently representing the temporally ordered sound structure of phonetic sequences (phonological short-term memory), whether heard or internally generated. The authors also highlighted the importance of phonological short-term memory for guiding speech production, and implicated left pSTS in both mimicry and language acquisition.

A role in phonological short-term memory does not, however, mean that the underlying function is specific to speech. Indeed, left pSTS activation increases with familiarity to non-verbal sounds (lacking any phonological content) even when auditory input is controlled (Dehaene-Lambertz et al., 2005; Dick et al., 2011; Leech et al., 2009; Liebenthal et al., 2003, 2010; Margulis et al., 2009; Meyer et al., 2005). A parsimonious explanation, that would explain the response to both verbal and non-verbal stimuli, is that left pSTS contributes to the short-term retention of auditory representations that underpins speech perception, speech production and other non-linguistic auditory tasks. Another possibility is that left pSTS might be a heterogeneous region with multiple functional subdivisions that have been conflated in prior studies. For example, Liebenthal et al. (2014) distinguished the posterior portion of the stem of STS (including dpSTS and vpSTS in Fig. 1), from the anterior terminal ascending branch of the STS (atSTS) which lies dorsal to pSTS and ventral to TPJ (Fig. 1). Specifically, in a large-scale meta-analysis of 253 studies, Liebenthal et al. (2014) found that pSTS activation was more frequently associated with non-linguistic than linguistic stimuli, whereas atSTS was most sensitive to linguistic material (although also activated by a range of executive and motor planning tasks).

The current within-subjects study tests whether different parts of left pSTS respond to the demands on: (1) auditory short-term memory (that is not specific to speech sounds); (2) phonological short-term memory (greater for speech than non-speech) and/or (3) retrieval of phonological representations that can be integrated with the articulatory system.

1.2. The contribution of left TPJ to speech processing

As reported for pSTS, part of left TPJ is independently activated by speech perception and production and during auditory-motor tasks on non-verbal sounds (Hickok et al., 2003, 2004; Buchsbaum et al., 2001). Hickok et al. (2003) therefore proposed that this region plays a role in auditory-motor integration. The same part of left TPJ has also been associated with short-term memory of verbal (Buchsbaum and D'Esposito, 2019, 2009; Koelsch et al., 2009; Kraemer et al., 2005; McGettigan et al., 2011) and non-verbal sounds (Koelsch et al., 2009; Kraemer et al., 2005). Drawing this work together, Buchsbaum and D'Esposito (2019) have described how this left TPJ region could support the temporary maintenance of auditory speech representations via feed-forward and feedback pathways that connect the auditory- and motor-speech systems.

On the other hand, activation in the same part of left TPJ has also been observed in the absence of auditory stimuli, speech production or a motor task. For example, left TPJ activation has been reported for imagining music, tones or environmental sounds (Aleman, 2004; Bunzeck et al., 2005; Zatorre and Halpern, 2005; Xu et al., 2006) or viewing visual stimuli that had previously been paired with sounds, music or rhythms (Jäncke and Shah, 2004; Pekkola et al., 2006; Wheeler et al., 2000; Hasegawa et al., 2004).

As with pSTS, we tested whether TPJ was sensitive to the demands on: (1) auditory short-term memory (that is not specific to speech sounds); (2) phonological short-term memory (greater for speech than non-speech) and/or (3) retrieval of phonological representations that can be integrated with the articulatory system.

1.3. Experimental rationale

Using a multi-factorial within-subjects design, we aimed to dissociate different parts of the left posterior superior temporal lobe on the basis of their response profiles. Our choice of conditions is founded on the type of processing that we expect to be engaged by the regions, stimuli and tasks. Once different response profiles are segregated, we and others can generate and test hypotheses about the function of each region for a given task, in the acknowledgement that the function of a region may

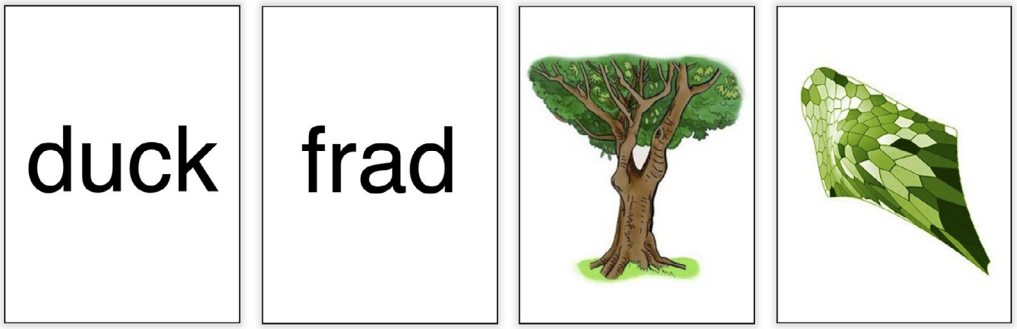

Fig. 2. Examples of visual stimuli.
Verbal (words/pseudowords) and non-verbal (pictures of objects and non-objects) visual stimuli.

Table 1
Task analysis.

Sensory / perceptual processing	Visual Conditions								Auditory Conditions							
Task set	Speech production				1-back matching				Speech production				1-back matching			
Type of stimulus	W	P	O	C	W	P	O	C	W	P	O	H	W	P	O	H
Speech acoustics from stimulus									**A**	**A**			**A**	**A**		
Hearing own speech	**A**	**A**	**A**	**A**					**A**	**A**	**A**	**A**				
Visual short-term memory	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**								
Auditory short-term memory									**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**
Semantic retrieval/memory	**A**		**A**		**A**		**A**		**A**		**A**		**A**		**A**	
Phonological retrieval	**A**	**A**	**A**	**A**	**A**	**A**	i	i			**A**	**A**			i	i
Phonological short-term memory	**A**	**A**	**A**	**A**	**A**	**A**	i	i	**A**	**A**	**A**	**A**	**A**	**A**	i	i
Articulatory recoding	**A**	**A**	**A**	**A**	**A**	**A**	i	i	**A**	**A**	**A**	**A**	i	i	i	i
Motor control of speech	**A**	**A**	**A**	**A**					**A**	**A**	**A**	**A**				

For each of the 16 different conditions (16 columns), A = the types of processing (rows) that are expected to be engaged. i = not required but may occur implicitly; W = words; P = pseudowords; O = objects; H = humming; C = coloured non-objects.

vary depending on the network of regions it contributes to during any given task (Price and Friston, 2005).

To dissociate the response profile of different regions, our experimental design presented 8 different stimulus types with two different tasks: overt speech production (e.g. repeating aloud, reading aloud, naming aloud) or silent 1-back matching, with a finger press response. Half the stimuli were verbal and half the stimuli were non-verbal. The auditory verbal stimuli were (1) spoken words or (2) spoken pseudowords. The visual verbal stimuli were (3) written words or (4) written pseudowords. We describe these stimuli as verbal rather than phonological because (i) phonological processing will be activated by non-verbal stimuli during speech production tasks (e.g., picture naming) and (ii) the term "verbal" is broader than phonological, incorporating the specific experiences we have with speech sounds and written material compared to other types of stimuli.

The auditory non-verbal stimuli were (5) sounds of animals and objects or (6) meaningless vocal humming sounds. The visual non-verbal stimuli were (7) pictures of animals and objects or (8) meaningless coloured non-objects (see Fig. 2). A task analysis of the types of processing we hypothesise to be tapped by each of the 16 conditions is provided in Table 1 and detailed below.

In all the auditory conditions, we expected that (i) auditory short-term memory would be required until a speech production or 1-back matching response had been made, (ii) the demands on auditory short-term memory would be greater when auditory stimuli lacked semantic content (pseudowords and humming) compared to stimuli that have rich semantic content (words and object sounds) that can be used to support task performance, (iii) speech processing would be engaged by verbal more than non-verbal stimuli in the auditory modality, and (iv) the demands on phonological retrieval would be greater for non-verbal than verbal stimuli because heard speech is available to guide the production of speech from verbal but not non-verbal stimuli.

The visual 1-back matching conditions were used to limit the auditory effects described above to higher order processing areas that are not specific to the auditory modality because they are also activated during phonological processing of visual stimuli. We expected that 1-back matching of visual words and pseudowords would involve phonological processing (encoding and short-term memory) because (1) even when the task does not require speech production, skilled readers are highly trained to rapidly link phonologically legal written text (words or pseudowords) to higher level representations of speech sounds (phonological encoding/retrieval) and articulation (articulatory recoding), and (2) these phonological representations can be held in short-term memory to support 1-back matching. We also expected that the demands on phonological retrieval would be higher for visual words and pseudowords than auditory words and pseudowords (because speech sounds are provided by auditory verbal stimuli but need to be retrieved from visual stimuli).

To examine whether the left posterior superior temporal lobe regions of interest differed in the task-relevant functional networks in which they participate, we used a covariance analysis (Seghier and Price, 2009). Finally, to confirm that condition-dependent responses were significantly different in functionally distinct regions, we report region-by-condition interactions. This necessitates data from a second sample of participants (Sample 2) that is completely independent of the sample of participants used to define the functionally distinct regions (Sample 1). Significant region-by-condition interactions in Sample 2 also serve to cross validate the findings of Sample 1.

2. Methods

2.1. Participants

There were two independent samples of participants (n = 24 and n = 59) with no overlap (total n = 83). All participants were healthy,

Table 2
Experimental design (Sample 1).

		Stimulus	Abb	Verbal	Semantic	Production
Visual modality	Speech production	Words	**W**	+	+	+
		Pseudowords	**P**	+	-	+
		Objects	**O**	-	+	+
		Colours	**C**	-	-	+
	1-back matching	Words	**W**	+	+	-
		Pseudowords	**P**	+	-	-
		Objects	**O**	-	+	-
		Colours	**C**	-	-	-
Auditory modality	Speech production	Words	**W**	+	+	+
		Pseudowords	**P**	+	-	+
		Objects	**O**	-	+	+
		Hums	**H**	-	-	+
	1-back matching	Words	**W**	+	+	-
		Pseudowords	**P**	+	-	-
		Objects	**O**	-	+	-
		Hums	**H**	-	-	-

Abb = abbreviations used in all Tables and Figures: W = words; P = pseudowords; O = objects; C = coloured non-objects, H = humming sounds. The plus signs were the activation conditions (weighted +1 in the statistical contrast) and the negative signs were the baseline (weighted -1 in the statistical contrast).

right handed (assessed with the Edinburgh Handedness Inventory; Oldfield, 1971) native English speakers, with normal or corrected-to-normal vision and hearing. Sample 1 included 12 females and 12 males (mean age = 31.4 years, standard deviation (SD) = 5.9 years). Sample 2 included 34 females and 25 males (mean age = 44.5 years, SD = 17.7 years). It was not necessary to match the samples for age because we were testing for consistency across samples, rather than searching for group differences. Written informed consent was obtained from each participant prior to scanning with ethical approval from the London Queen Square Research Ethics Committee.

2.2. Paradigm for sample 1

Sample 1 were scanned during 16 different conditions with 8 types of stimuli, in a 2 × 2 × 2 × 2 factorial design (see Tables 1 and 2). Factor 1 was stimulus modality (auditory versus visual). Factor 2 was verbal (words or pseudowords) versus non-verbal stimuli (pictures and sounds). Factor 3 was semantic (words and objects) versus non-semantic stimuli (pseudowords and meaningless baselines). Factor 4 was response modality with two tasks: overtly producing speech (i.e. speech production, SP), or 1-back matching with a finger press response.

For the speech production tasks, participants produced a single overt spoken response. In the visual modality, they: (1) named objects or animals in pictures; (2) read written object names; (3) read pseudowords; and (4) named the colour of meaningless non-objects (see Fig. 2). In the auditory modality, participants (1) named objects/animals after hearing environmental sounds associated with those objects; (2) repeated heard object names; (3) repeated pseudowords; and (4) named the gender of the voice ('male' or 'female') after hearing male or female humming sounds.

For the 1-back matching task, participants placed two fingers of the same hand over an fMRI compatible button box to indicate whether or not the stimulus was the same as the one preceding it (left button for 'same', right button for 'different'). Hand was counterbalanced evenly across participants. There was no overt speech production involved in any 1-back matching condition.

Data from Sample 1 have previously been reported in: Oberhuber et al. (2013) to demonstrate a functional posterior-anterior subdivision in the putamen; Hope et al. (2014) to dissect the functional anatomy of auditory word repetition; Oberhuber et al. (2016) to investigate functional subdivisions within the supramarginal gyrus; and Yamamoto et al. (2019) to highlight a special role for the right posterior superior temporal sulcus during speech production. The current focus on auditory short-term memory in the left posterior superior temporal lobe yields novel findings that were outside the scope of all previous analyses of the same dataset.

2.3. Stimulus presentation

The same set of 96 objects were presented as pictures, written words and heard words, with items rotated, across participants such that all 96 objects were presented in each of these three conditions across all participants (each participant experienced 32 items per condition). This ensured that the speech production responses (i.e. object names) were identical for reading, repetition and object naming when averaging across participants. A different set of 32 objects were presented in the object sound conditions because only a limited number of objects are recognisable from their sounds (e.g. motorbike and telephone but not suitcase or banana). All participants were therefore presented with the same set of 32 object sounds. Stimulus characteristics for each condition are provided in Table 3.

Condition order was fully counterbalanced across participants. Half the participants performed the 8 speech production tasks first, followed by the 1-back matching tasks. The others performed the 1-back matching tasks first, followed by the speech production tasks. Within task, we counterbalanced the order of other variables (stimulus modality, semantics and phonology). Stimuli for the speech production conditions were identical to those for the 1-back matching conditions.

The auditory words and pseudowords were recordings of a male, native, English speaker (with a Southern British accent approximating Received Pronunciation) reading aloud the written versions of the same stimuli. The auditory semantic non-verbal stimuli (sounds of animals and objects) were taken from the NESSTI sound library (Hocking et al., 2013). The auditory non-semantic, non-verbal stimuli were created by male and female voices humming with no phonological or semantic content. Critically, stimulus duration was longer for non-verbal sounds (objects and humming) than verbal sounds (words and pseudowords) because when the stimulus duration was shortened, participants were unable to name the source of animal and object sounds. On average, the non-verbal hums were shorter than the object sounds and longer than the words and pseudowords (see Table 3). We expected that as stimulus duration increased, so would the demands on acoustic processing, auditory attention and auditory short-term memory.

Written pseudowords were created using a non-word generator (Duyck et al., 2004). To ensure that the pseudoword stimuli were balanced with the word stimuli, we generated 128 written pseudowords

Table 3
Stimulus characteristics for each condition.

Stimulus	Syllables (SD)	Letters	Duration (seconds)	log10 word frequency (Zipf value)
Visual words	1.53 (0.68)	5.24 (1.68)	1.5	4.3 (0.6; 1.6-5.9)
Visual pseudowords	1.94 (0.92)	5.28 (1.94)	1.5	0
Visual objects	1.55 (0.69)	5.30 (1.75)	1.5	4.3 (0.6; 1.6-5.9)
Visual colours	1.36 (0.49)	4.89 (1.04)	1.5	4.8 (0.4; 4.3-5.4)
Auditory words	1.53 (0.68)	5.24 (1.68)	0.64 (0.10)	4.3 (0.6; 1.6-5.9)
Auditory pseudowords	1.90 (0.84)	5.35 (1.72)	0.68 (0.12)	0
Auditory objects	1.81 (0.92)	5.64 (2.21)	1.47 (0.12)	4.4 (0.7; 1.7-5.4)
Auditory humming	1.50 (0.51)	5.00 (1.01)	1.04 (0.43)	4.7 (0.01; 4.66-4.67)

The average number of syllables and letters (standard deviation in brackets) for each word, pseudoword, object name, colour name or gender name. The average duration of these stimuli is in seconds. Average log word frequency is from SUBTLEX-UK (van Heuven et al., 2014) with standard deviation, minimum and maximum in brackets. A Zipf value of 1 corresponds to very-low-frequency words (1 per 100 million words) and a value of 6 to very-high-frequency content words (10,000 per million words).

that were matched to the 128 objects names for bigram frequency, number of orthographic neighbours and word length. The visual non-semantic, non-verbal stimuli were coloured non-objects. They were created from the object pictures by scrambling the global and local features to render them unrecognisable and then manually editing the images to accentuate one of eight colours (brown, blue, orange, red, yellow, pink, purple and green). Illustrations of the visual stimuli are presented in Fig. 2.

Details of the stimulus properties (average number of syllables, average number of letters, average stimulus duration and word frequency) can be found in Table 3. There were separate runs for each of the 16 conditions. Within each 3.2 min run, there were 4 blocks of stimuli, alternating with rest. Each block presented 9 stimuli including 1 repeat, with an inter-stimulus interval of 2.52 s. The repeat was present for speech production and 1-back matching conditions and was only used to assess accuracy in the 1-back matching condition. All conditions were fully counterbalanced across participants.

2.4. Procedure

Prior to scanning, each participant was trained on all tasks using a separate set of stimuli, except for animal and object sounds which remained the same. During both visual and auditory conditions, participants were instructed to respond as fast as possible, keeping their body and head as still as possible, and their eyes open and fixated on a cross in the middle of the display screen.

Scanning started with the instructions 'Get Ready' written on the in-scanner screen while five dummy scans were acquired (15.4 s in total). This was followed by a written instruction (e.g. 'Repeat'), lasting 3.08s, which indicated the forthcoming start of a new block and reminded participants of the task that needed to be performed. Auditory stimuli were presented via MRI-compatible headphones (MR Confon, Magdeburg, Germany), which also attenuated the noise of the magnetic gradients and the helium pump via active gradient noise suppression. The initial headphone volume level was set to 89dB and adjusted for each participant before scanning. Spoken responses were recorded via a noise-cancelling MRI microphone (FOMRI IIITM Optoacoustics, Or-Yehuda, Israel) at a sampling rate of 44,100 Hz, for off-line analysis. The pictures subtended an angle of 7.4° (10 cm on screen, 78 cm viewing distance) with a pixel size of 350 × 350, and a screen resolution of 1024 × 768. The visual angle for the written words ranged from 1.47 to 4.41°, with the majority of words (with five letters) extending 1.84–2.2°. Visual verbal stimuli (words and pseudowords) were presented in lower case Helvetica.

In-scanner behaviour was measured for each of the 16 conditions. Correct responses were those that matched the target without delay or self-correction. All other responses were categorised as incorrect. For 1-back matching, accuracy and response times (from stimulus onset to button press) were computed automatically, according to the button pressed in response to each trial. For speech production, spoken responses were recorded via a microphone and monitored by the experimenter who either (i) ticked a check list to confirm that the expected response had been made or (ii) recorded an alternative (or null) response. For some stimuli, more than one response was considered correct. For example, a picture of a mug could be named "cup" or "mug". The same criteria were used for all participants.

Due to technical failure, response times were only available in the 1-back matching task. We conducted a repeated measures 2 × 2 × 2 ANOVA in SPSS (IBM SPSS 22, NY, USA) to test for main effects and interactions. Factor 1 was stimulus modality (visual vs. auditory), factor 2 was semantic versus non-semantic stimuli (words and objects versus pseudowords and baseline) and factor 3 was verbal versus non-verbal stimuli (words and pseudowords versus objects and baseline).

2.5. Data acquisition

Functional and anatomical data were collected on a 3T scanner (Trio, Siemens, Erlangen, Germany) using a 12-channel head coil. To minimise movement during acquisition, a careful head fixation procedure was used when positioning each participant's head. This ensured that none of the speech sessions were excluded after checking the realignment parameters. Functional images consisted of a gradient-echo planar imaging (EPI) sequence and 3 × 3 mm^2 in-plane resolution (TR/TE/flip angle = 3080 ms/30 ms/90°), field of view (FOV) = 192 mm, matrix size = 64 × 64, 44 slices, slice thickness = 2 mm, interslice gap = 1 mm, 62 image volumes per time series, including five "dummies" to allow for magnetisation to reach equilibrium. The TR was chosen to maximize whole brain coverage (44 slices) and to ensure that slice acquisition onset was offset-asynchronised with stimulus onset, which allowed for distributed sampling of slice acquisition across the study (Veltman et al., 2002). For anatomical reference, a high-resolution T1 weighted (w) structural image was acquired after completing the tasks using a three-dimensional Modified Driven Equilibrium Fourier transform (MDEFT) sequence (TR/TE/TI = 7.92 ms/2.48 ms/910 ms), flip angle = 16°, 176 slices, voxel size = 1 × 1 × 1 mm). The total scanning time was approximately 1 h and 20 min per participant, including set-up and the acquisition of the anatomical scan.

2.6. fMRI data preprocessing

Data preprocessing and statistical analysis were performed in SPM12 (Wellcome Centre for Human Neuroimaging, London, UK), running on

MATLAB 2012a. Functional volumes were spatially realigned to the first EPI volume and unwarped to compensate for non-linear distortions caused by head movement or magnetic field inhomogeneity. The unwarping procedure was used in preference to including the realignment parameters as linear regressors in the first-level analysis because unwarping accounts for non-linear movement effects by modelling the interaction between movement and any inhomogeneity in the $T2^*$ signal. After realignment and unwarping, the realignment parameters were checked to ensure that participants moved less than one voxel (3 mm^3) within each scanning run. The anatomical T1w image was co-registered to the mean EPI image generated during the realignment step and then spatially normalised to the MNI space using the unified normalisation-segmentation routine of SPM12. To spatially normalise all EPI scans to MNI space, the deformation field parameters that were obtained during the normalisation of the anatomical T1w image were applied to the functional volumes. The original resolution of the different images was maintained during normalisation (voxel size $1 \times 1 \times 1$ mm for structural T1w and $3 \times 3 \times 3$ mm for EPI images). After normalisation, functional images were spatially smoothed with a 6 mm full-width-half-maximum isotropic Gaussian kernel to compensate for residual anatomical variability and to permit application of Gaussian random-field theory for statistical inference (Friston et al., 1995).

2.7. First level statistical analyses of fMRI data

Each pre-processed functional volume was entered into a subject-specific fixed effect analysis using the general linear model. Stimulus onset times were modelled as single events with two regressors per run, one modelling the instructions and one modelling all stimuli of interest. Stimulus functions were convolved with the SPM canonical haemodynamic response function and high pass filtered with a cut-off period of 128 s.

For each scanning session/run (that alternated one condition of interest with fixation), we generated a single contrast that compared activation in response to the stimuli and task of interest to resting with fixation. This resulted in 16 different contrasts (one per condition) for each participant. Each contrast for each individual was inspected to ensure that there were no visible artefacts (e.g. edge effects, activation in ventricles) that might have been caused by within-scan head movements.

2.8. Second level statistical analyses of fMRI data

At the second level, the 16 contrasts for each participant were entered into a within-subjects one-way ANOVA in SPM12, with factorial analysis conducted at the contrast level.

Within the 8 auditory conditions, our statistical contrasts reflected the conventional analysis of a factorial design - main effect of (1) verbal versus non-verbal stimuli; (2) semantic versus non-semantic stimuli and (3) speech production versus 1-back matching (see Table 2). Each effect was tested in both directions (i.e. verbal > non-verbal; non-verbal > verbal; semantic > non-semantic; non-semantic > semantic; speech production > 1-back matching; and 1-back matching > speech production).

Activation for each auditory effect of interest is reported where there was also greater activation during 1-back matching in the visual modality for verbal than non-verbal stimuli (words and pseudowords > objects and coloured patterns). This ensured that the auditory effects we report are in higher level amodal areas, rather than areas that are specific to auditory processing. We did not include the visual speech production conditions in the conjunction because they involve auditory processing of the spoken response which is avoided in silent 1-back matching. The visual speech production conditions were however, used to test the main effects of task (speech production more than 1-back matching) and stimulus modality (visual versus auditory).

2.9. Statistical contrasts

Our rationale for the statistical contrasts is based on our task analysis (see Introduction and Table 1). In the visual modality, we compared 1-back matching on verbal stimuli to rest and non-verbal stimuli. In the auditory modality, we compared (1) verbal to non-verbal stimuli, and vice versa, (2) semantic to non-semantic stimuli and vice versa, and (3) speech production to 1-back matching and vice versa.

By reporting activation that was common for (1) visual 1-back matching on verbal compared to non-verbal stimuli and (2) each of the auditory contrasts (see above), we identified responses that were not specific to the auditory modality (i.e. involved in higher level processing).

2.10. Statistical thresholds

Voxel-wise correction for multiple comparisons was either (i) across the whole brain or (ii) within a single anatomical region (the four ROIs illustrated in Fig. 1 combined into a single binary mask). This is more conservative than individually correcting for multiple comparisons within each ROI for each statistical contrast. The individual ROIs were then used to anatomically determine which ROI were involved in each effect of interest. The whole brain analysis allowed us to identify other areas, outside our regions of interest, that co-activated with each ROI.

For the conjunctions of visual and auditory contrasts, we used the global conjunction in SPM with a statistical threshold of $p < 0.05$ after family-wise error correction for multiple comparisons across the whole brain (in height). The auditory contrast that entered the conjunction was computed across tasks and for each task separately.

2.11. Dissociating the whole brain neural systems associated with different left posterior superior temporal lobe regions

We dissociated the brain networks associated with different ROIs by using a second level covariance analysis. This is purely based on variance between subjects (not within subjects). The rationale is that brain regions that are part of the same functional network will show similar increases and decreases in activation across conditions; and these changes in activation will co-vary across subjects engaging these networks. In contrast, when regions are part of different functional networks, covariance will be out of sync across conditions and across subjects.

The advantage of this analysis is that it does not assume any a priori knowledge of the functional processing involved in each condition. This avoids the pitfalls associated with cognitive subtractions. It can also detect areas where the significance of the effect is low in the cognitive subtraction approach because of high within- or between-subject variance (see Seghier and Price, 2009).

Procedurally, we repeated our whole-brain second level analysis (with 16 different conditions). This time, the parameter estimates (activation compared to rest for each subject in each condition) at the coordinates for the peak voxels were entered into the analysis as separate covariates. The number of regressors was equal to the number of left posterior superior temporal lobe regions of interest. Variance associated with each regressor is therefore a combination of condition effects and inter-subject variability effects. By comparing each regressor to all others, within the same analysis, we identified the sets of distributed regions (across the whole brain) that covaried with one region of interest more than the others. We report these effects, after family-wise error correction for multiple comparisons across the whole brain, in height.

2.12. Region by condition interactions in sample 2

The 59 participants in Sample 2 performed the 8 speech production conditions used with Sample 1, but did not perform any of the 1-back matching tasks. Experimental details can be found in Oberhuber et al. (2016). After running the same pre-processing and first

Table 4
Response times for 1-back matching in seconds (standard deviation).

Stimulus	RT 1-back
Visual words	0.655 (0.11)
Visual pseudowords	0.648 (0.09)
Visual objects	0.683 (0.12)
Visual colours	0.762 (0.11)
Auditory words	0.880 (0.11)
Auditory pseudowords	0.959 (0.14)
Auditory objects	1.111 (0.33)
Auditory humming	1.125 (0.23)

The response time (RT), in seconds, from stimulus onset to finger press response during the 1-back matching task. Response times were not available for the speech production conditions.

level analysis steps as used for Sample 1, the second level analysis for Sample 2 included 8 (speech production) conditions. Using the F-map only (unbiased by condition), we extracted the subject specific data for all conditions from the voxel closest to each of the coordinates identified in Sample 1. We then conducted region by condition interactions (in IBM SPSS Statistics, version 25.0, using 1-tailed p values) to (i) confirm that the regions dissociated in Sample 1 showed the same response profiles in Sample 2; and (ii) demonstrate that the response profile differed significantly between regions. The condition specific effects of interest were subsequently investigated in SPM to illustrate their extent.

3. Results

3.1. In-scanner behavioural data

As reported in detail for the same dataset in Yamamoto et al. (2019), in-scanner accuracy was above 90% for each of the 16 conditions, except for 1-back matching on auditory humming (89%), repeating heard pseudowords (88%) and reading written pseudowords (86%). Reaction times during 1-back matching were slower for: (i) auditory > visual stimuli, because auditory features were delivered sequentially whereas visual features were delivered simultaneously; (ii) non-verbal > verbal auditory stimuli, because non-verbal stimuli had longer delivery duration than auditory speech stimuli; and (iii) non-semantic > semantic visual stimuli (written pseudowords > words, and colour > object stimuli) plausibly because on-line retention of non-semantic stimuli, until the matching decision, is not facilitated by semantic memory. The mean response times for each 1-back matching condition can be found in Table 4.

3.2. fMRI data

3.2.1. Visual 1-back matching on words and pseudowords (W&P)

When correcting for multiple comparisons within the binary mask (composed of the four anatomical ROIs), there were three spatially distinct activation peaks for 1-back matching on verbal stimuli more than rest. These were located in left pSTS (dorsal surface), left atSTS and left TPJ (see Fig. 3 and Table 5A). All three regions were also activated by (i) visual 1-back matching on words and pseudowords more than objects and coloured non-objects ($p < 0.001$ uncorrected) and (ii) speech production on all visual stimuli more than 1-back matching on all visual stimuli ($p < 0.05$ after correction for multiple comparisons within the binary mask comprising the four regions of interest), see plots in Fig. 3.

Across the whole brain, the only other areas to show higher activation for visual 1-back matching on words and pseudowords compared to objects and coloured non-objects were the left frontal operculum and left pars opercularis (Table 5A). This was significant after correction for multiple comparisons across the whole brain in extent, when the height-level statistical threshold was set at $p < 0.001$ uncorrected.

According to our task analysis (Table 1), the processing associated with visual 1-back matching of verbal (words and pseudowords) > non-verbal (objects and coloured non-objects) stimuli could reflect phonological processing (encoding/retrieval and/or short-term memory) or orthographic processing. To determine which parts are associated with phonological processing, we consider the results of the auditory contrasts.

3.2.2. Verbal (words and pseudowords) versus non-verbal (object and humming) auditory conditions

The only temporal lobe region to show more activation for verbal than non-verbal auditory conditions was located in the middle part of the STS [peak at -57, -18, -6] consistent with previous studies of acoustic speech processing (Dick et al., 2011; Norman-Haignere et al., 2015; Specht et al., 2009; Liebenthal et al., 2014). Within our left posterior superior temporal regions of interest, left TPJ activation was higher for non-verbal than verbal auditory conditions (see Fig. 3). As the same left TPJ was also more activated by visual 1-back matching on verbal more than non-verbal stimuli (see Table 5B), there was a highly significant interaction between [verbal versus non-verbal] and [visual versus auditory]; $p < 0.05$ corrected in height for multiple comparisons across the whole brain (peak coordinates for the interaction at: -57, -36, +21, Z score = 5.1 and -54, -42, +18, Z score = 5.0).

To summarise, contrary to expectation, we did not find any left posterior superior temporal lobe region that could be associated with modality-independent phonological short-term memory. Instead, we discovered that the part of left TPJ that corresponds to the putative Perisylvian Language area (PSL) (see above) is more strongly activated by non-verbal than verbal auditory conditions. This cannot be explained in terms of the demands on phonological retrieval because activation was higher for auditory verbal stimuli than visual verbal stimuli (see Fig. 3).

3.2.3. Non-semantic (pseudowords and humming) versus semantic (words and objects) auditory conditions

Within our anatomical regions of interest, activation was higher for non-semantic than semantic auditory conditions in left dpSTS, with a corresponding effect in right dpSTS (RdpSTS). There were no effects of non-semantic > semantic (or semantic > non-semantic) in either left TPJ or left atSTS ($p > 0.05$ uncorrected). A conjunction analysis confirmed that 142 voxels within left dpSTS and 271 voxels within right dpSTS were activated by (i) non-semantic > semantic auditory conditions and (ii) verbal more than non-verbal visual 1-back matching (see Table 5B).

3.2.4. The effects of speech production and stimulus modality in the areas co-activated by visual and auditory contrasts

The conjunction of auditory and visual contrasts (i.e. the effects reported in Table 5B) ensures that the activation we report is in amodal areas that are activated by both visual and auditory stimuli, during 1-back matching and speech production. Nevertheless, we also note, for completeness, that all the regions activated in the conjunction were more activated during (A) speech production than 1-back matching conditions in both the visual and auditory modalities and (B) auditory conditions more than visual conditions, in both the speech production and 1-back matching conditions.

3.2.5. Dissociating the whole brain neural systems associated with dpSTS, TPJ and atSTS

To dissociate the networks of regions associated with TPJ, dpSTS and atSTS, we (i) extracted the parameter estimates (activation for each condition compared to rest for each subject) averaging over voxels within 3 mm of the peak coordinates for the conjunction of auditory and visual contrasts (see Table 5B). We then searched for brain regions where activation, across conditions, co-varied with one of the regions of interest more than the other two.

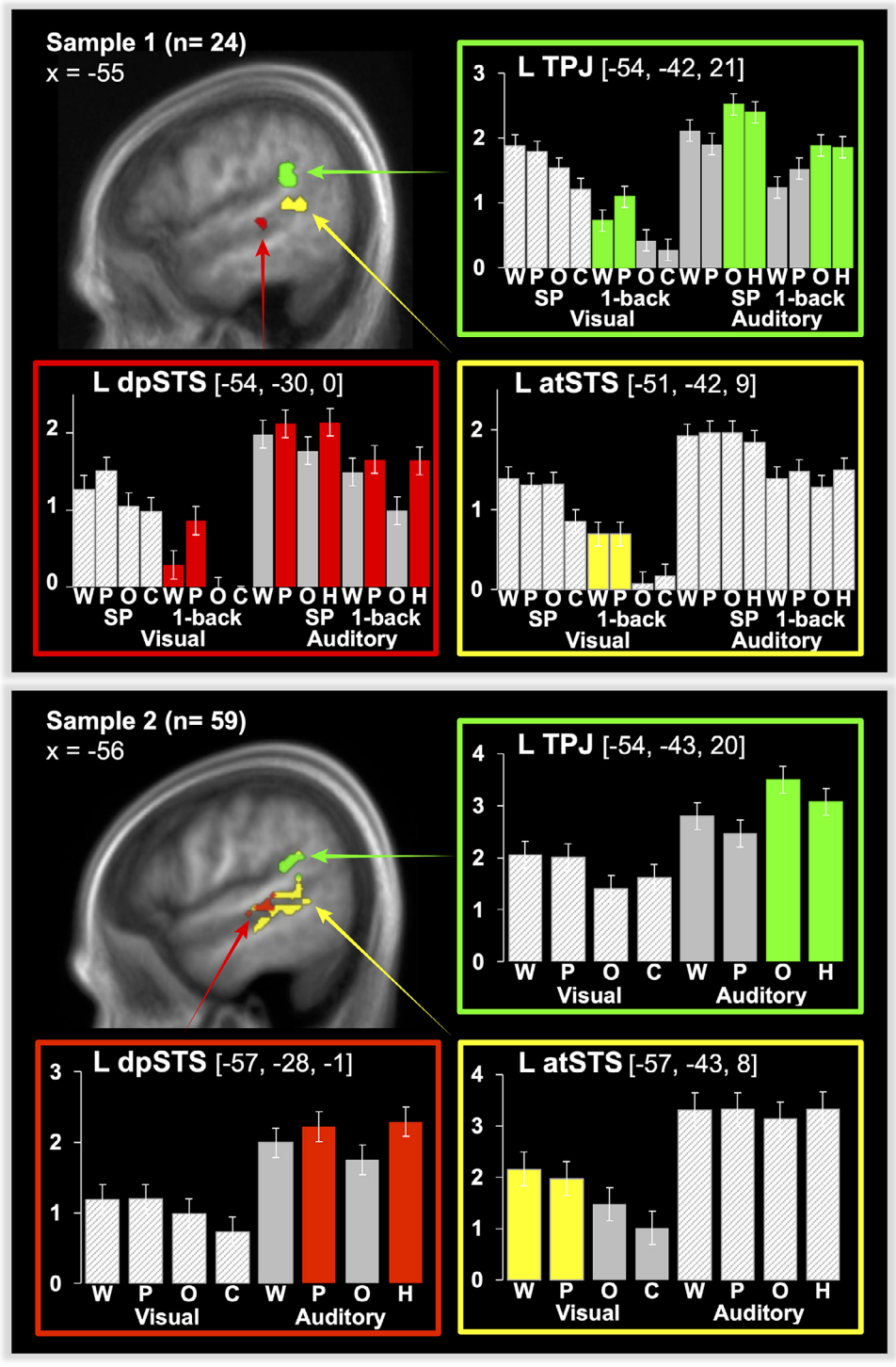

Fig. 3. Condition-specific responses in three left posterior superior temporal regions.
The upper section shows the results from Sample 1. The lower section shows the results from Sample 2. For each sample, activation within the anatomical regions of interest (Fig. 1) are shown on the mean structural image for that sample. The three plots for each sample show the mean relative activation per condition (16 for Samples 1, 8 for Sample 2), with standard error bars. The coloured bars show conditions of interest (weighted +1 in the statistical contrast), the grey bars show the corresponding baseline conditions (weighted -1 in the statistical contrast) and the hashed bars show conditions that were not included in the statistical contrasts. Different colours (in both the plots and the brain image) distinguish different statistical contrasts in the auditory modality: green activation is higher for non-verbal than verbal auditory stimuli; red activation is higher for non-semantic than semantic auditory stimuli and yellow activation was not significantly different for verbal, non-verbal, semantic or non-semantic auditory stimuli. For Sample 1, each statistical contrast in the auditory modality was in conjunction with the effect of verbal more than non-verbal stimuli during visual 1-back matching. The verbal stimuli are words (W) and pseudowords (P). The non-verbal stimuli are objects (O), coloured patterns (C) and humming (H). The semantic stimuli are W & O. The non-semantic stimuli are P, C and H. SP = speech production tasks; 1-back = 1-back matching tasks. For Sample 1, peak effects were all significant at $p < 0.05$ FWE-corrected for multiple comparisons within the anatomical regions of interest (see Table 5A), and the extent of the effect is illustrated at a height-level threshold of $p < 0.001$ uncorrected. For Sample 2, the significance threshold was set at $p < 0.05$ uncorrected (Table 5C) but the effects were all significant with small volume correction at the peak co-ordinate for Sample 1 (see text). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The results dissociated three different networks (see Fig. 4). The network associated with left TPJ primarily included bilateral postcentral gyri, SMA and bilateral superior cerebellum, with a smaller area in the left ventral premotor cortex [-57, 9, 6]. The network associated with left dpSTS primarily included the left anterior superior temporal cortex, the pre-SMA and two small regions in the dorsal and ventral premotor cortex ([-42, 3, 54] and [-54, 0, 12]). The network of regions associated with left atSTS primarily involved bilateral inferior frontal gyri and sulci (left hemisphere peak at [-54, 27, 18]) and bilateral anterior insulae/frontal opercula (left hemisphere peak at [-30, 24, 9] / [-42, 27, 3]). These results focus on how the covariance pattern of one ROI differs from that of the other ROIs. Common covariance across all three ROIs was, expectedly, extensive (e.g. bilateral striatum, thalami, cingulate, motor cortex, frontal opercula, cerebellum, and right temporal and parietal regions), arguably, because activation in each of these regions contributes to generic speech production and domain general attentional processes.

3.2.6. *Region by condition interactions in sample 2*

We extracted the subject specific parameter estimates from the voxel, in the F-map for Sample 2, that was nearest to the coordinates identified for the conjunction of auditory and visual contrasts in Sample 1 (see Table 5B). The proximity of coordinates for Samples 1 and 2 (no more than 1mm on each axis) is shown in Table 5C. The data from the four auditory conditions were entered into SPSS which confirmed the

Table 5
Statistical details for effects of interest.

(A) Visual 1-back matching on words (W) and pseudowords (P) compared to (i) rest and (ii) visual 1-back matching on objects (O) and coloured non-objects (C).

Anatomical Region (abbreviation)		MNI coordinates	*Z* scores W&P > rest	W&P > O&C
Left temporo-parietal junction	(L TPJ)	-54, -42, +21	5.2 *	3.1
Left anterior ascending terminal branch of the STS	(L atSTS)	-51, -45, +9	4.2 *	3.2
		-51, -45, +6	3.9	3.7
Left dorsal surface of the horizontal stem of posterior STS	(L dpSTS)	-54, -30, +0	3.5 *	4.0
		-57, -30, -3	3.1	4.3
Right posterior STS	(R dpSTS)	+57, -30, 0	3.3	3.2
Left frontal operculum	(L FO)	-42, +30, -3	3.8	4.3
Left pars opercularis	(L pOp)	-48, +15, +12	4.4	3.7

(B) Conjunctions of auditory effects and visual 1-back matching on verbal > non-verbal stimuli (W&P>O&C)

Auditory contrast	Region	*k* Voxels	MNI coordinates	*Z* scores Con.	Aud.	Vis.
Non-verbal > Verbal	L TPJ	182	-54, -42, +21	4.8	4.7	3.1
Non-semantic > Semantic	L dpSTS	142	-57, -27, 0	5.9	4.0	4.3
			-54, -42, +3	5.3	3.2	3.9
	R dpSTS	271	+57, -24, -3	5.3	6.2	3.5
			+54, -36, -3	4.8	5.1	3.2
Speech production > 1-back matching	L dpSTS	231	-60, -30, 0	6.0	3.9	4.2
	L atSTS		-57, -42, +9	4.9	3.4	3.4
	L TPJ		-54, -42, +21	4.8	3.3	3.1
	R dpSTS	40	+57, -30, 0	4.8	3.3	3.1

(C) Validating effects in Sample 2

Region	L atSTS	L TPJ	L dpSTS
Sample 1 conjunction	-57, -42, +9	-54, -42, +21	-57, -27, 0
Sample 2 F map	-57, -43, +8	-54, -43, +20	-57, -28, -1
Visual verbal > non-verbal (SP)	**4.96**	3.26	2.56
Auditory non-verbal > verbal (SP)	NS	**3.98**	NS
Auditory non-semantic > semantic (SP)	NS	NS	**3.03**

In Part A, *Z* scores that reached significance at $p < 0.05$ corrected for multiple comparisons within the anatomical region of interest (see Fig. 1) are masked with an asterisk (*). In part B, Con. = *Z* scores for conjunction of the auditory contrast (Aud) listed in column 1, and the visual contrast (Vis) which was always visual 1-back matching on words and pseudowords > objects and coloured non-objects (Vis). The *Z*-scores for the conjunctions were all significant at $p < 0.05$ FWE-corrected for multiple comparisons across the whole brain. *k* = number of voxels at $p < 0.001$ uncorrected. In part C, *Z* scores for Sample 2 are reported for the three different effects within 1 mm of the peak coordinates from the conjunction shown in Table 5B. The highest *Z* scores for each of the three effects is in bold. SP = speech production.

expected region by condition interactions: The effect of non-verbal > verbal auditory conditions (across tasks) was significantly higher in left TPJ than either left dpSTS ($F(1,58) = 19.248$; $p < 0.001$) or left atSTS ($F(1,58) = 11.533$; $p < 0.001$), and the effect of non-semantic > semantic auditory conditions (across tasks) was significantly higher in left dpSTS than either left atSTS ($F(1,58) = 4.710$; $p = 0.017$) or left TPJ ($F(1,58) = 16.907$; $p < 0.001$). The results of the SPM statistical comparisons between conditions are provided in Table 5C, and the extent of the effects within the anatomical regions of interest is illustrated in the lower section of Fig. 3.

3.2.7. The effect of stimulus duration in left TPJ in sample 2

Higher left TPJ activation for non-verbal than verbal auditory stimuli, in both Sample 1 and 2, might reflect longer stimulus duration for non-verbal than verbal stimuli. The non-verbal object/animal sounds were the longest (1.47 s) because our pilot study indicated that shorter sounds were not as well recognised, and word durations were shortest (0.64 s). The durations of half the humming sounds were matched to the object/animal sounds, the other half were matched to the word durations; with a mean duration of 1.04 s for the humming sounds. If left TPJ activation was sensitive to the duration of the auditory stimuli, then it should be higher for (A) object/animal sounds than humming; and (B) long humming sounds than short humming sounds. We did not find any evidence for either A in Sample 1 (see top right Fig. 3) or B in Sample 2 in a new analysis that modelled long and short humming sounds separately and found no significant difference (*Z* score = 0.97; $p = 0.165$ uncorrected) in left TPJ activation for longer (parameter estimate = 3.7) compared to shorter sounds (parameter estimate = 3.2).

4. Discussion

The main contribution of our study is to dissociate the response profiles of three different left posterior superior temporal lobe regions (dpSTS, TPJ and atSTS) when participants are engaged in speech production and 1-back matching tasks on different types of auditory stimuli. Critically, none of the three regions responds specifically to (a) auditory stimuli because they were all more activated during silent 1-back matching on visual words and pseudowords than the same task on objects and coloured non-objects or (b) speech and language processing because activation does not increase with semantic or phonological content in the auditory domain. To the contrary, left TPJ shows higher responses to non-phonological than phonological auditory stimuli and left dpSTS shows higher responses to non-semantic than semantic auditory stimuli.

The functional dissociation of dpSTS, TPJ and atSTS is demonstrated in two independent samples of participants. We also show that the three regions co-activate with different neural systems that include different frontal, parietal and cerebellar regions. Below, we consider the response profile of each region in detail and discuss how their functional role might be described and investigated further in future studies.

4.1. Left dpSTS

Here we distinguish left dpSTS from the adjacent anterior terminal ascending branch of the superior temporal sulcus (i.e. left atSTS). We found that the response in left dpSTS was higher, when the auditory stimuli were non-semantic compared to semantic, irrespective

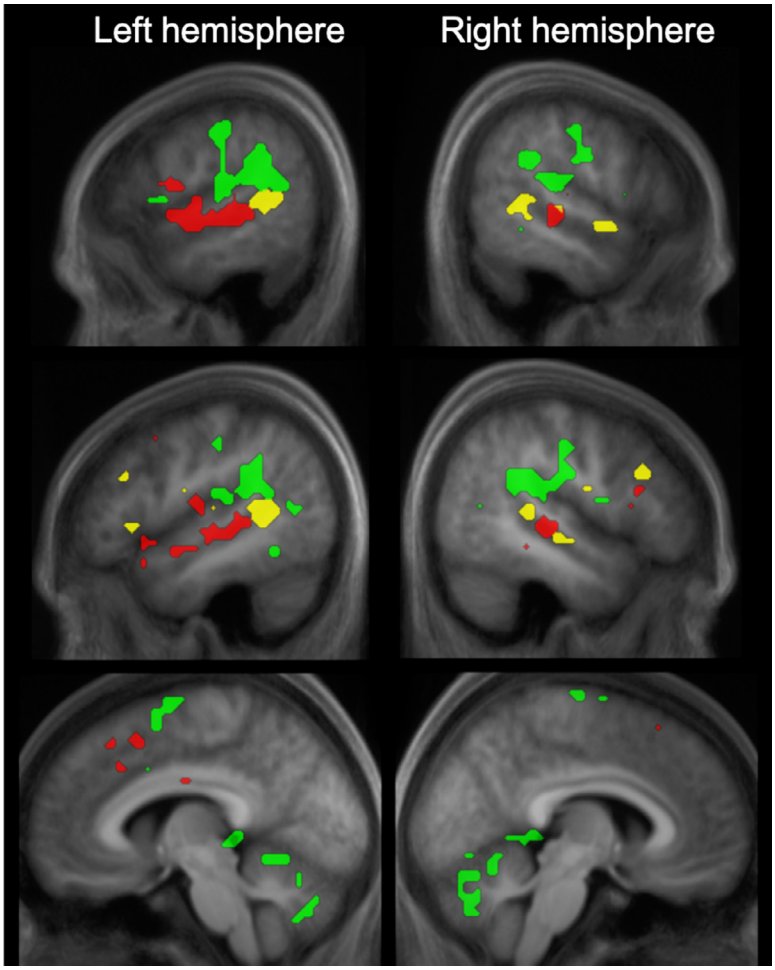

Fig. 4. Brain regions that covary with left dpSTS (red), left TPJ (green) or left atSTS (yellow)
Regions where activation co-varied with each of our three regions of interest illustrated in Fig. 3. Green = network associated with left TPJ, red = network associated with left dpSTS and yellow = network associated with left atSTS. Top row: x = -/+ 56; second row x = -/+ 48; bottom row x= -/+ 6. All coloured voxels (a) covaried positively and significantly with the corresponding region, (b) covaried significantly more with the corresponding region compared to the others and (c) were activated across all tasks compared to rest, with the threshold for both (a), (b) and (c) set at $p<0.05$ corrected for multiple comparisons across the whole brain. The threshold for the extent of the cluster was set to >20 voxels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of task. According to our task analysis (Table 1), the enhanced activation could be a consequence of an increased reliance on auditory short-term memory when facilitation from semantic processing is not available. Indeed, the peak coordinates of this effect [-57, -30, -3/-54, -30, 0] are in close proximity to the brain region [-56, -30, 1] where Richardson et al. (2011) showed, in neurotypical individuals, that higher digit span (a classic measure of verbal short-term memory) correlated with higher grey matter density.

Our results elucidate the functional contribution of left dpSTS in two ways. First, we show that left dpSTS activation is not sensitive to the phonological content of stimuli because activation was *not* higher for verbal compared to non-verbal auditory stimuli, or vice versa. Second, we show that, left dpSTS activation is not specific to auditory input because activation was also observed for silent 1-back matching on visual stimuli with verbal content (i.e. words and pseudowords). A parsimonious explanation is that left dpSTS may support the short-term retention of auditory representations that can be derived from either auditory or visual inputs.

If our design had not included the auditory non-verbal conditions, higher activation for verbal than non-verbal stimuli in the visual modality might have been interpreted as reflecting the demands on phonological processing. By showing that dpSTS activation is higher for non-semantic, non-verbal humming than spoken word processing, our findings are more consistent with a role for left dpSTS in short-term representation of sound features relevant to the task (Liebenthal et al., 2010), with demands on these short-term auditory representations increasing when the retention of auditory stimuli cannot rely on semantic memory.

This explanation can help interpret a range of prior findings. For example, reliance on auditory short-term memory may increase during audio-visual integration (Erickson et al., 2014; Szycik et al., 2012) and the attention, memory and executive tasks included in the meta-analysis conducted by Liebenthal et al. (2014) who reported greater left dpSTS activation for non-linguistic than linguistic stimuli. It is also possible that, in the absence of a behavioural task, the reliance on short-term representation of relevant sound features increases when passively listening to (1) non-verbal sounds as they become familiar (Dehaene-Lambertz et al., 2005; Dick et al., 2011; Leech et al., 2009; Liebenthal et al., 2010) and (2) non-semantic speech sounds compared to complex unintelligible sounds (Benson et al., 2006; Giraud et al., 2004; Narain et al., 2003; Rimol et al., 2006; Scott et al., 2000).

4.2. Left TPJ

We consider how the response we observed in left TPJ fits with two non-mutually exclusive perspectives reported in the prior literature: (A) this region plays a role in auditory-motor integration (Hickok et al., 2003, 2004; Buchsbaum et al., 2001; Buchsbaum and D'Esposito 2019); and/or (B) it plays a role in short-term memory of auditory representations (Buchsbaum and D'Esposito, 2019, 2009; Koelsch et al., 2009; Kraemer et al., 2005; McGettigan et al., 2011) that are not necessarily linked to phonology or auditory-motor integration (Aleman, 2004; Bunzeck et al., 2005; Zatorre and Halpern, 2005; Xu et al., 2006; Jäncke and Shah, 2004; Pekkola et al., 2006; Wheeler et al., 2000; Hasegawa et al., 2004).

The contributions of our study are as follows. First, we found that left TPJ activation was observed for silent 1-back matching on visual words and pseudowords. Its response is therefore not specific to auditory inputs or attention. Second, we found that activation was higher for non-verbal than verbal auditory stimuli. This is difficult to explain in terms of the demands on auditory-motor integration or stimulus duration (see Section 7 of the fMRI results). Third, we found that left TPJ was more activated by auditory repetition than reading of stimuli that were matched across conditions (i.e. the same phonology and semantics). This cannot be explained by the demands on phonological retrieval which are less for repetition than reading because the phonological representations are primed by the auditory stimulus. The role that left TPJ plays in auditory short-term memory, is therefore not specific to speech perception or speech production, as shown in prior studies (Koelsch et al., 2009; Kraemer et al., 2005). We are also cautious about defining the response in left TPJ as an auditory-motor integration area. Instead, like many other areas, it may contribute to auditory-motor integration, indirectly.

If left TPJ plays a role in auditory short-term memory, how does this differ from that in left dpSTS? Our study provides three distinguishing pieces of evidence. In TPJ, activation is higher for non-verbal than verbal auditory stimuli but did not significantly differ for non-semantic compared to semantic auditory stimuli. As non-verbal auditory stimuli had longer durations and took longer to process than verbal stimuli (see Table 4), enhanced left TPJ activation for non-verbal compared to verbal auditory stimuli may reflect either the load on memory encoding or the prolonged maintenance of information in auditory short-term memory until the task is completed. In contrast, left dpSTS activation was higher for non-semantic than semantic auditory stimuli but did not significantly differ for non-verbal compared to verbal auditory stimuli. This may reflect demands on auditory short-term memory when there is no support from semantic memory.

The whole brain activations associated with left TPJ and left dpSTS also suggest that these regions participate in partially non-overlapping neural systems (Fig. 3). Left dpSTS co-activated with extensive parts of the superior temporal gyrus (consistent with attention to auditory input), whereas left TPJ co-activated with extensive parts of the postcentral gyri that are associated with the sensory consequences of motor actions rather than motor planning. Although further studies are required to understand how these neural systems function, co-activation in left TPJ and the postcentral gyri raises an interesting hypothesis. Rather than driving motor responses (as implied from the auditory-motor integration hypothesis), left TPJ may contribute to speech production, at a post-articulatory stage, by holding auditory representations of expected speech on-line until the spoken output is matched to the intended speech. This hypothesis could be tested in future, using directional connectivity studies to determine whether left TPJ drives articulatory planning or is involved in sustaining auditory representations for post-articulatory processing.

In summary, the response we observe in left TPJ is most consistent with encoding and sustaining auditory representations on-line. This is required for both speech perception and speech production but is not limited to language tasks. Future studies are required to test whether left TPJ contributes: (i) directly to motor planning (e.g. driving premotor/motor regions), (ii) indirectly to motor planning (e.g. by sustaining activity in other regions that drive the motor response) and/or (iii) to post-articulatory processing of the spoken response.

4.3. Left atSTS

A distinction between the function of left atSTS and left dpSTS was previously reported in a large-scale meta-analysis of 253 studies by Liebenthal et al. (2014) who found that left dpSTS activation was more frequently associated with non-linguistic stimuli than linguistic stimuli whereas left atSTS was most sensitive to linguistic material (although also activated by a range of executive and motor planning tasks).

Our within-subjects study indicates that activation in left atSTS increased for (i) verbal more than non-verbal visual stimuli, (ii) speech production more than 1-back matching and (iii) auditory more than visual stimuli. In these ways, the response in atSTS was similar to those in dpSTS and TPJ. However, unlike left dpSTS, left atSTS did not respond differentially to semantic versus non-semantic auditory stimuli; and, unlike TPJ, atSTS did not respond differentially to verbal versus non-verbal auditory stimuli. We therefore found no evidence to suggest that left atSTS was sensitive to the demands on auditory or phonological short-term memory.

Based on connectivity patterns, Glasser et al. (2016) report that atSTS (corresponding to area TPOJ1 in HCP-MMP1.0) is one of three temporo-parieto-occipital regions that link higher auditory and higher visual areas. In addition, the same authors report that, relative to the dorsal surface of left dpSTS, left atSTS is more activated by motor tasks involving tongue movements, finger tapping and toe squeezing; and less activated for listening to stories compared to answering arithmetic questions ("LANGUAGE STORY-MATH" contrast).

In our covariance analysis, we found that activation in the left frontal operculum and left middle frontal gyrus/inferior frontal sulcus covaried more strongly with left atSTS than either left dpSTS or left TPJ. This provides regions of interest for future connectivity analyses to investigate whether left atSTS is driven bottom-up from the auditory cortex and/or top-down from left frontal regions. It will also be of interest to understand the direction of information flow between atSTS, dpSTS and TPJ.

5. Conclusions

The novel contribution of our study is the demonstration that dpSTS, atSTS and TPJ each have distinct response properties, with left TPJ responding to non-verbal more than verbal auditory stimuli and left dpSTS responding to non-semantic more than semantic stimuli; and left atSTS being significantly less sensitive, than TPJ and dpSTS, to the verbal and semantic content of the stimuli. Although none of these regions are specific to speech and language processing (as discussed in the Introduction), they each contribute to speech and language processing in different ways.

We have already provided some speculative hypotheses about how each region might contribute to language processing, but most importantly, our findings strongly motivate and guide future studies to probe the function of each region further and to use effective connectivity analyses (e.g. as in Parker Jones et al., 2013) to improve our understanding of how different parts of the speech and language network interact with one another to support speech comprehension and drive speech production. For example, how do our left posterior superior temporal lobe regions interact with each other and the rest of the brain during sensory, motor and higher-level cognitive/language processing? More specifically, is left temporo-parietal activation driven bottom-up from auditory inputs in the auditory cortex or top-down from the left posterior inferior frontal cortex; and how does this depend on stimulus modality, stimulus content and task? In addition, a greater understanding of inter-subject variability in the response of each region will also be essential for building maps of the functional anatomy of language that can be used to predict the behavioural consequences of brain damage or neurosurgery.

Data availability

The data that support the findings of this study are available upon request from the senior author [C.J.P.].

Credit authorship contribution statement

Justyna O. Ekert: Conceptualization, Writing – original draft, Writing – review & editing, Formal analysis, Visualization. **Andrea Gajardo-Vidal:** Investigation, Writing – review & editing. **Diego L. Lorca-Puls:**

Investigation, Writing – review & editing. **Thomas M.H. Hope:** Methodology, Software, Writing – review & editing. **Fred Dick:** Visualization, Writing – review & editing. **Jennifer T. Crinion:** Writing – review & editing. **David W. Green:** Conceptualization, Writing – review & editing. **Cathy J. Price:** Conceptualization, Methodology, Visualization, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

This work was funded by Wellcome (203147/Z/16/Z and 205103/Z/16/Z, C.J.P.) and the Middlesex Hospital Medical School General Charitable Trust. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We thank Eldad Druks for creating the picture stimuli.

References

Aleman, A., 2004. The functional neuroanatomy of metrical stress evaluation of perceived and imagined spoken words. Cereb. Cortex 15, 221–228. doi:10.1093/cercor/bhh124.

Benson, R.R., Richardson, M., Whalen, D.H., Lai, S., 2006. Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. Neuroimage 31, 342–353.

Buchsbaum, B.R., D'Esposito, M., 2019. A sensorimotor view of verbal working memory. Cortex 112, 134–148.

Buchsbaum, B.R., D'Esposito, M., 2009. Repetition suppression and reactivation in auditory–verbal short-term recognition memory. Cereb. Cortex 19, 1474–1485.

Buchsbaum, B.R., Hickok, G., Humphries, C., 2001. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cogn. Sci. 25, 663–678.

Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H.-J., Jancke, L., 2005. Scanning silence: mental imagery of complex sounds. Neuroimage 26, 1119–1127.

Dale, A.M., Sereno, M.I., 1993. Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. 5, 162–176.

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

Dick, F., Lee, H.L., Nusbaum, H., Price, C.J., 2011. Auditory-motor expertise alters "speech selectivity" in professional musicians and actors. Cereb. Cortex 21, 938–948.

Duyck, W., Desmet, T., Verbeke, L.P.C., Brysbaert, M., 2004. WordGen: a tool for word selection and nonword generation in Dutch, English, German, and French. Behav. Res. Methods Instrum. Comput. 36, 488–499. doi:10.3758/BF03195595.

Erickson, L.C., Zielinski, B.A., Zielinski, J.E.V, Liu, G., Turkeltaub, P.E., Leaver, A.M., Rauschecker, J.P., 2014. Distinct cortical locations for integration of audiovisual speech and the McGurk effect. Front. Psychol. 5, 534. doi:10.3389/fpsyg.2014.00534.

Friston, K.J., Frith, C.D., Turner, R., Frackowiak, R.S.J., 1995. Characterizing evoked hemodynamics with fMRI. Neuroimage 2, 157–165.

Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb. Cortex 14, 247–255.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. doi:10.1038/nature18933.

Hasegawa, T., Matsuki, K.I., Ueno, T., Maeda, Y., Matsue, Y., Konishi, Y., Sadato, N., 2004. Learned audio-visual cross-modal associations in observed piano playing activate the left planum temporale. An fMRI study. Cogn. Brain Res. 20, 510–518.

Hickok, G., Buchsbaum, B., Humphries, C., Muftuler, T., 2003. Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. J. Cogn. Neurosci. 15, 673–682.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99. doi:10.1016/j.cognition.2003.10.011.

Hocking, J., Dzafic, I., Kazovsky, M., Copland, D.A., 2013. NESSTI: norms for environmental sound stimuli. PLoS ONE 8. doi:10.1371/journal.pone.0073382.

Hope, T.M.H., Prejawa, S., Parker Jones, O., Oberhuber, M., Seghier, M.L., Green, D.W., Price, C.J., 2014. Dissecting the functional anatomy of auditory word repetition. Front. Hum. Neurosci. 8, 1–17. doi:10.3389/fnhum.2014.00246.

Horn, A., 2016. HCP-MMP1. 0 projected on MNI2009a GM (volumetric) in NIfTI format [WWW Document]. URL 10.6084/m9.figshare.3501911.v5 (accessed 2.15.21).

Hugdahl, K., Thomsen, T., Ersland, L., Rimol, L.M., Niemi, J., 2003. The effects of attention on speech perception: an fMRI study. Brain Lang. 85, 37–48.

Jäncke, L., Shah, N.J., 2004. 'Hearing'syllables by 'seeing'visual stimuli. Eur. J. Neurosci. 19, 2603–2608.

Koelsch, S., Schulze, K., Sammler, D., Fritz, T., Müller, K., Gruber, O., 2009. Functional architecture of verbal and tonal working memory: an FMRI study. Hum. Brain Mapp. 30, 859–873.

Kraemer, D.J.M., Macrae, C.N., Green, A.E., Kelley, W.M., 2005. Musical imagery: sound of silence activates auditory cortex. Nature 434, 158.

Leech, R., Holt, L.L., Devlin, J.T., Dick, F., 2009. Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. J. Neurosci. 29, 5234–5239.

Liebenthal, E., Binder, J.R., Piorkowski, R.L., Remez, R.E., 2003. Short-term reorganization of auditory analysis induced by phonetic experience. J. Cogn. Neurosci. 15, 549–558.

Liebenthal, E., Desai, R., Ellingson, M.M., Ramachandran, B., Desai, A., Binder, J.R., 2010. Specialization along the left superior temporal sulcus for auditory categorization. Cereb. Cortex 20, 2958–2970.

Liebenthal, E., Desai, R.H., Humphries, C., Sabri, M., Desai, A., 2014. The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. Front. Neurosci. 8, 289.

Margulis, E.H., Mlsna, L.M., Uppunda, A.K., Parrish, T.B., Wong, P.C.M., 2009. Selective neurophysiologic responses to music in instrumentalists with different listening biographies. Hum. Brain Mapp. 30, 267–275.

McGettigan, C., Warren, J.E., Eisner, F., Marshall, C.R., Shanmugalingam, P., Scott, S.K., 2011. Neural correlates of sublexical processing in phonological working memory. J. Cogn. Neurosci. 23, 961–977.

Meyer, M., Zysset, S., Von Cramon, D.Y., Alter, K., 2005. Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. Cogn. Brain Res. 24, 291–306.

Narain, C., Scott, S.K., Wise, R.J.S., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb. Cortex 13, 1362–1368.

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296.

Oberhuber, M., Hope, T.M.H., Seghier, M.L., Parker Jones, O., Prejawa, S., Green, D.W., Price, C.J., 2016. Four functionally distinct regions in the left supramarginal gyrus support word processing. Cereb. Cortex 26, 4212–4226. doi:10.1093/cercor/bhw251.

Oberhuber, M., Jones, O.P., Hope, T.M.H., Prejawa, S., Seghier, M.L., Green, D.W., Price, C.J., 2013. Functionally distinct contributions of the anterior and posterior putamen during sublexical and lexical reading. Front. Hum. Neurosci. 7, 1–10. doi:10.3389/fnhum.2013.00787.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113. doi:10.1016/0028-3932(71)90067-4.

Parker Jones, O., Seghier, M.L., Kawabata Duncan, K.J., Leff, A.P., Green, D.W., Pricel, C.J., 2013. The auditory-motor interactions for the production of native and non-native speech. J. Neurosci. 33, 2376–2387. doi:10.1523/JNEUROSCI.3289-12.2013.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Sams, M., 2006. Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. Hum. Brain Mapp. 27, 471–477.

Price, C.J., Friston, K.J., 2005. Functional ontologies for cognition: the systematic definition of structure and function. Cogn. Neuropsychol. 22, 262–275.

Richardson, F.M., Ramsden, S., Ellis, C., Burnett, S., Megnin, O., Catmur, C., Schofield, T.M., Leff, A.P., Price, C.J., 2011. Auditory short-term memory capacity correlates with gray matter density in the left posterior STS in cognitively normal and dyslexic adults. J. Cogn. Neurosci. 23, 3746–3756.

Rimol, L.M., Specht, K., Hugdahl, K., 2006. Controlling for individual differences in fMRI brain activation to tones, syllables, and words. Neuroimage 30, 554–562. doi:10.1016/j.neuroimage.2005.10.021.

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.

Seghier, M.L., Price, C.J., 2009. Dissociating functional brain networks by decoding the between-subject variability. Neuroimage 45, 349–359.

Specht, K., Osnes, B., Hugdahl, K., 2009. Detection of differential speech-specific processes in the temporal lobe using fMRI and a dynamic "sound morphing" technique. Hum. Brain Mapp. 30, 3436–3444.

Szycik, G.R., Stadler, J., Tempelmann, C., Münte, T.F., 2012. Examining the McGurk illusion using high-field 7 Tesla functional MRI. Front. Hum. Neurosci. 6, 95.

Van Heuven, W.J.B., Mandera, P., Keuleers, E., Brysbaert, M., 2014. SUBTLEX-UK: a new and improved word frequency database for British English. Q. J. Exp. Psychol. 67, 1176–1190.

Veltman, D.J., Mechelli, A., Friston, K.J., Price, C.J., 2002. The importance of distributed sampling in blocked functional magnetic resonance imaging designs. Neuroimage 17, 1203–1206. doi:10.1006/nimg.2002.1242.

Wheeler, M.E., Petersen, S.E., Buckner, R.L., 2000. Memory's echo: vivid remembering reactivates sensory-specific cortex. Proc. Natl. Acad. Sci. 97, 11125–11129.

Wise, R.J.S., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K., Warburton, E.A., 2001. Separate neural subsystems within Wernicke's area. Brain 124, 83–95.

Yamamoto, A.K., Parker Jones, O., Hope, T.M.H., Prejawa, S., Oberhuber, M., Ludersdorfer, P., Yousry, T.A., Green, D.W., Price, C.J., 2019. A special role for the right posterior superior temporal sulcus during speech production. Neuroimage 203, 116184. doi:10.1016/j.neuroimage.2019.116184.

Zatorre, R.J., Halpern, A.R., 2005. Mental concerts: musical imagery and auditory cortex. Neuron 47, 9–12.