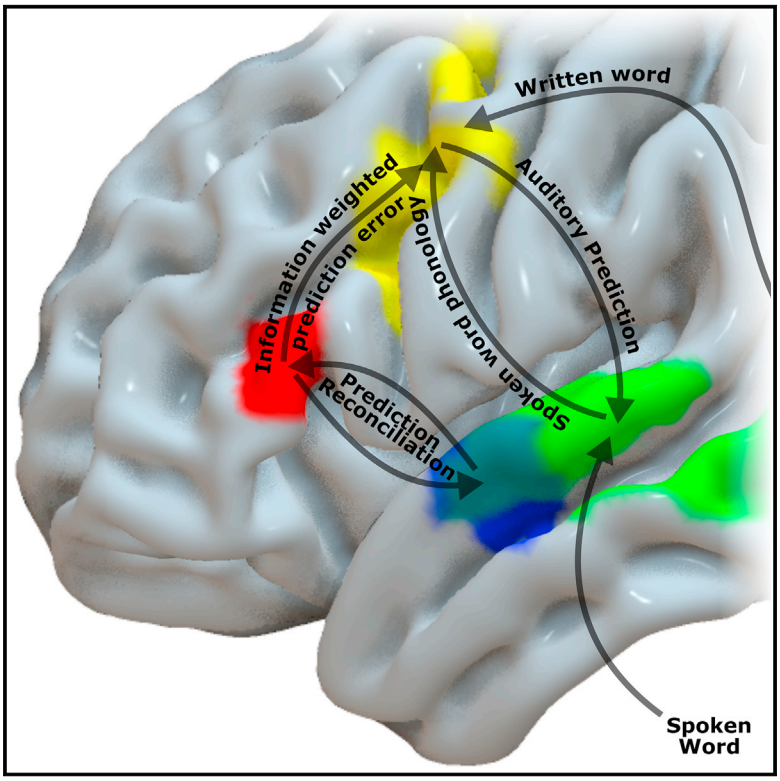
Cell Reports

Article

Temporal lobe perceptual predictions for speech are instantiated in motor cortex and reconciled by inferior frontal cortex

Graphical abstract



Authors

Thomas E. Cope, Ediz Sohoglu, Katie A. Peterson, ..., Karalyn Patterson, Matthew H. Davis, James B. Rowe

Correspondence

thomascope@gmail.com

In brief

Cope et al. use multivariate 7-T fMRI in patients with language impairment from selective frontal neurodegeneration (nfvPPA) to show that precentral gyrus instantiates and refines perceptual speech predictions, while inferior frontal gyrus contains distinct representations of verified and violated predictions that support their reconciliation with sensory input in anterior STG.

Highlights

- Left precentral gyrus represents spoken-word phonology and weighted prediction error
- Left inferior frontal gyrus represents verified and violated predictions independently
- Frontal neurodegeneration impairs integration of speech predictions in anterior STG
- A tripartite speech perception network using a predictive motor model is proposed

Cope et al., 2023, Cell Reports *42*, 112422
May 30, 2023
https://doi.org/10.1016/j.celrep.2023.112422

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Temporal lobe perceptual predictions for speech are instantiated in motor cortex and reconciled by inferior frontal cortex

Thomas E. Cope,[1,2,3,14,*] Ediz Sohoglu,[2,4] Katie A. Peterson,[1,5] P. Simon Jones,[1] Catarina Rua,[1] Luca Passamonti,[1] William Sedley,[6] Brechtje Post,[7] Jan Coebergh,[8,9] Christopher R. Butler,[10,11] Peter Garrard,[9,12] Khaled Abdel-Aziz,[8,9] Masud Husain,[10] Timothy D. Griffiths,[6] Karalyn Patterson,[1,2] Matthew H. Davis,[2,13] and James B. Rowe[1,2,3,13]

[1]Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 0SZ, UK
[2]Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 7EF, UK
[3]Cambridge University Hospitals NHS Trust, Cambridge CB2 0QQ, UK
[4]School of Psychology, University of Sussex, Brighton BN1 9RH, UK
[5]Department of Radiology, University of Cambridge, Cambridge CB2 0QQ, UK
[6]Biosciences Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[7]Theoretical and Applied Linguistics, Faculty of Modern & Medieval Languages & Linguistics, University of Cambridge, Cambridge CB3 9DA, UK
[8]Ashford and St Peter's Hospital, Ashford TW15 3AA, UK
[9]St George's Hospital, London SW17 0QT, UK
[10]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK
[11]Faculty of Medicine, Department of Brain Sciences, Imperial College London, London W12 0NN, UK
[12]Molecular and Clinical Sciences Research Institute, St. George's, University of London, London SW17 0RE, UK
[13]Senior author
[14]Lead contact
*Correspondence: thomascope@gmail.com
https://doi.org/10.1016/j.celrep.2023.112422

SUMMARY

Humans use predictions to improve speech perception, especially in noisy environments. Here we use 7-T functional MRI (fMRI) to decode brain representations of written phonological predictions and degraded speech signals in healthy humans and people with selective frontal neurodegeneration (non-fluent variant primary progressive aphasia [nfvPPA]). Multivariate analyses of item-specific patterns of neural activation indicate dissimilar representations of verified and violated predictions in left inferior frontal gyrus, suggestive of processing by distinct neural populations. In contrast, precentral gyrus represents a combination of phonological information and weighted prediction error. In the presence of intact temporal cortex, frontal neurodegeneration results in inflexible predictions. This manifests neurally as a failure to suppress incorrect predictions in anterior superior temporal gyrus and reduced stability of phonological representations in precentral gyrus. We propose a tripartite speech perception network in which inferior frontal gyrus supports prediction reconciliation in echoic memory, and precentral gyrus invokes a motor model to instantiate and refine perceptual predictions for speech.

INTRODUCTION

Perception is the result of the integration of sensory inputs with prior predictions.[1] Language comprehension is a natural domain in which to study such predictions, as speech perception relies on accurate inference to comprehend what has been said. The role of motor cortical regions in facilitating speech comprehension, via an articulatory model, is highly controversial.[2–5] Nonetheless, it is clear that humans use cross-modal cues such as lip reading and subtitles to improve speech perception, especially in noisy listening environments.[6] Such cross-modal cuing typically improves comprehension but can lead to false perception when cues mismatch.[7–9] Here, we address the issue of how and where the human brain encodes perceptual predictions for phonemes and how these predictions are reconciled with auditory inputs.

Independent manipulations of predictions and sensory inputs can create identical perceptual outcomes from differing input combinations.[10] Written text provides prior knowledge in a visual form that is strongly associated with, but sensorily separated from, the auditory speech signal that verifies or violates those predictions, avoiding confounding sensory neural activity by prior adaptation, habituation, or repetition suppression.[11] Providing an explicit cross-modal prime allows assessment of the neural mechanisms of perceptual prediction, independently of the generation of prediction identity, and with precise control of the phonemic overlap between written and spoken signals. Here we manipulated written text predictions of degraded

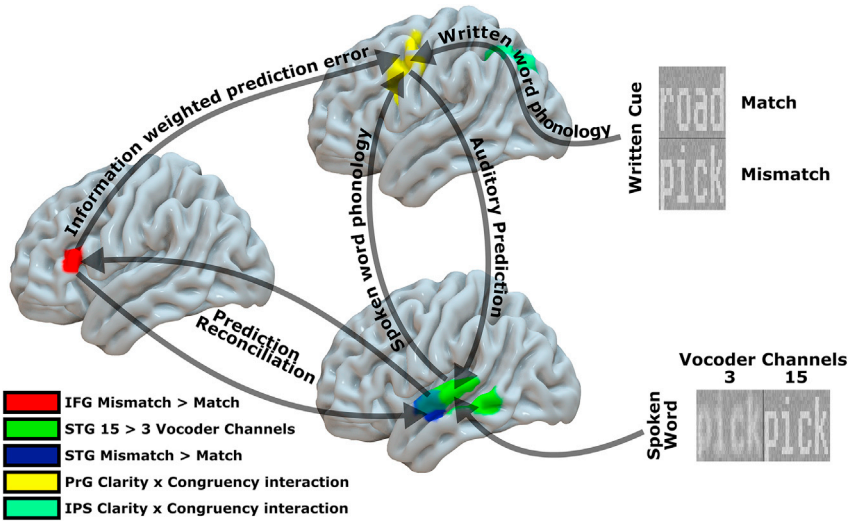


Figure 1. The proposed tripartite speech network involving STG, IFG, and PrG

The general experimental approach was a 2 × 2 manipulation of cue congruency (did the written and spoken words match or mismatch?) and sensory detail (high vs. low; 15 vs. 3 vocoder channels). Colored brain regions represent significant regions in our univariate contrasts, cluster thresholded at FDR $p < 0.05$. The red left IFG cluster and blue anterior STG cluster were defined by the match > mismatch contrast. The green cluster extending along the whole of STG and posterior middle temporal gyrus was defined by 15 > 3 channel vocoder speech. The yellow PrG and turquoise IPS clusters were defined by the sensory detail × congruency interaction, which was a crossover interaction such that activity was greater for match 3 and mismatch 15 conditions than for match 15 and mismatch 3. All colored regions were implicated in the overall contrast of written + spoken trials against written-only trials. Connectivity arrows are influenced by studies in nonhuman primates[19] and our physio-physiological interaction analysis.

speech during ultra-high-field (7-T) functional brain imaging, to test the hypothesis that motor cortex operates in combination with lateral prefrontal cortex, to first construct perceptual predictions for phonemes and then process sensory inputs by reconciling prediction errors, in order to make perceptual interpretations. We delineate the roles and interactions of functional sub-regions of the language network by combining multivariate analysis of the representations of verified and violated predictions with functional connectivity measures.

To test causal influences between regions in this network,[12] we study both healthy adults and people with non-fluent variant primary progressive aphasia (nfvPPA). This rare condition causes selective degeneration of frontal motor speech regions while preserving the structure and function of temporal cortex. We have previously shown with magnetoencephalography that nfvPPA causes delayed neuronal reconciliation of predictions as degraded fronto-temporal connections work harder to resolve mismatching sensory inputs and prior expectations, despite normal temporal lobe neural responses to bottom-up manipulations of sensory detail.[13] This results in inflexible, overly influential perceptual predictions that preserve speech-in-noise performance at the clinical cost of receptive agrammatism and failure to comprehend unexpected speech. In contrast to stroke aphasia, the frontal neurodegeneration in nfvPPA is stereotyped, partial, and subtle. This enables one to study a limited disruption of predictive mechanisms rather than a system reorganized following their complete absence.

Our hypothesis of multi-modal perception is as follows. Before sensory input, predictions are instantiated, based on a combination of immediate context and lifelong experience of the environment; here we simplify this step by providing the prediction identity in the form of written text. At the time of a sensory event and immediately afterward, the sensory input is reconciled with the prediction to infer its content, resulting in perception. This reconciliation is the result of two distinct prediction error assessments, one of higher-level expectations and another of lower-level sensory representations. In lay terms, the first assessment considers whether the predicted event occurred, and the second considers whether the sensory signals for that event were as expected. It was previously unknown whether these assessments were conducted by the same neuronal populations firing differentially[14] or by separate populations.[15] The difference between predicted and observed sensory input is then used to refine predictions for future sensory events. We propose that the refinement signal based on prediction errors is information weighted; a large error with low information content provides little basis for prediction refinement beyond an overall weakening of existing associations, while a small but precise error can be highly instructive.[16,17] Note that this information weighting is similar to previous conceptualizations of precision weighting,[16] but it is based on how informative the prediction error in a single trial, rather than long-run uncertainties.

Our results provide evidence for distinct roles in a tripartite speech perception network that includes superior temporal gyrus (STG), inferior frontal gyrus (IFG), and precentral gyrus (PrG) (Figure 1). This network is partially lateralized to the dominant hemisphere, and for simplicity our model focuses on left-sided structures, but we show evidence of bilateral representations in STG and PrG but not IFG. In line with previous work,[13] we show that the role of left IFG in speech perception is to reconcile perceptual predictions. We extend this by demonstrating that the IFG achieves this role by processing representations of verified and violated predictions in functionally segregated neural populations.[18]

In healthy individuals, but not patients with nfvPPA, these differential prediction outcome representations were also observed in anterior STG regions that are implicated in echoic memory.[20,21] This is consistent with IFG's role in restoring absent but inferred speech in auditory temporal regions,[22] modulated by linguistic knowledge.[23] In contrast, we show that only PrG contains the combined representations of sensory input and weighted prediction errors necessary to support the generation and refinement of future perceptual predictions. We demonstrate that a computational model based on our proposals can recapitulate both univariate (signal magnitude) and multivariate (representational information) functional MRI (fMRI) results in

PrG, supporting the hypothesis that we use a motor model to construct perceptual predictions for speech.

RESULTS

Structural MRI: nfvPPA atrophy of the frontal lobes only

We confirmed that patients with early nfvPPA have selective atrophy of frontal cortex. There was Bayesian evidence for the lack of atrophy in auditory temporal cortical regions (Figure 2A; Table S1). In our left-sided regions of interest (ROIs), patients with nfvPPA had very strong evidence for reduced cortical thickness in frontal operculum (BF10 = 82.3), pars triangularis (BF10 = 224.9), and precentral gyrus (BF10 = 161.1), but crucially there was evidence for no atrophy in temporal lobe primary auditory cortex (banks of superior temporal sulcus BF_null = 6.0) and planum temporale (transverse temporal BF_null = 2.9). Right-sided cortical thickness was reduced in homologous areas, but to a lesser degree, not meeting cluster-wise significance.

Behavior: nfvPPA causes inflexible perceptual predictions

In this new cohort with nfvPPA, we replicate the previous behavioral and Bayesian modeling findings[13] that patients have a bigger perceptual clarity benefit than controls from matching prior information, because they make overly precise and inflexible perceptual predictions (Figure S1). For all individuals, the perceptual clarity of vocoded words was significantly increased by matching text cues, compared with neutral or mismatching cues. This effect was significantly greater in patients with nfvPPA than controls, and persisted throughout the experimental session. If patients were simply making imprecise or incorrect predictions, the effect of prior knowledge would instead have been attenuated. This cannot be explained by noisier sensory input, as we show that people with nfvPPA are excellent at reporting unprimed vocoded speech and respond to manipulations of sensory detail and response difficulty in exactly the same way as healthy individuals (Figure S1). Finally, we explicitly model, and account for, single-subject sensory input noise in our Bayesian behavioral modeling. All of these findings are direct replications of the behavior shown in our previous, independent cohort with nfvPPA, in whom we additionally provided evidence of intact sensory cortical responses to bottom-up manipulations of sensory detail, and precisely quantified auditory perceptual performance.[13]

As in previous work,[13] the inflexibility of perceptual predictions (estimated as a parameter from the Bayesian perceptual modeling) correlated with the degree of frontal atrophy, but not temporal atrophy (with cortical thickness in IFG pars triangularis $r(31) = 0.522$, $p = 0.002$ [Figure S2]; pars opercularis $r(31) = 0.398$, $p = 0.022$; and PrG $r(31) = 0.429$, $p = 0.013$; but not Heschl's gyrus $r(31) = -0.026$, $p = 0.888$ or planum temporale $r(31) = 0.050$, $p = 0.783$).

The fMRI experimental manipulation was successful in modifying behavior in both groups, with better performance for reporting heard speech when it was presented with 15 rather than three vocoder channels, or matching rather than mismatching prior expectations (see supplementary results for more detail).

Patients and controls were above chance at reporting even the most degraded speech in the open set. With the closed set from the in-scanner experiment, patients correctly reported 55.8% of words even in the most difficult listening condition (chance level 25%; Figure S3).

Univariate fMRI: Response amplitude is modulated by spoken language and by context

Univariate voxel-wise analyses confirmed greater activation throughout the left-lateralized language network when a written and spoken word were presented together than when a written word was presented alone; there were peaks of differential activation in STG, PrG, IFG, and intraparietal sulcus (IPS) (Figure 2C; Table 1A). There were matching right-sided activations in STG and PrG, but not right IFG or IPS. There were no clusters in which controls had greater activation than patients with nfvPPA, but patients with nfvPPA had greater activation than controls in a small cluster in posterior right STG.

These left-sided language regions were each differentially sensitive to our experimental manipulations of cue congruency and sensory detail.

Cue congruency modulated activity in left IFG and anterior left STG; greater activation was observed in these regions during trials with mismatching compared with matching written cues (Figure 2C; Table 1A). No clusters showed the reverse contrast or significantly differed by group.

Sensory detail modulated activity along left and right STG and left IPS, as well as left posterior middle temporal gyrus; greater activation was observed in these clusters during trials with high (15 vocoder channels) compared with low (three vocoder channels) sensory detail (Figure 2C; Table 1A). No clusters showed the reverse contrast or significantly differed by group.

The interaction between cue congruency and sensory detail modulated activity in left and right PrG, left middle frontal gyrus, left and right IPS, and left posterior middle temporal gyrus (Figure 2C and Table 1A). These were all crossover interactions, such that there was greater activity for low sensory detail speech that matched the written cue, and high sensory detail speech that mismatched the written cue (Figures 3B and S4). No clusters differed significantly by group.

Multivariate fMRI

We approached the information present in brain response patterns systematically, proceeding step by step from global to local comparisons. We first combine all conditions and both groups in a whole-brain searchlight approach to ascertain where in the brain a given representation can be found. Only then do we assess the consistency of these representations between conditions and groups using an ROI approach, conservatively defined with orthogonal contrasts.[25] Despite the complexity of our study design, this conservative approach means that our results cannot be due to multiple comparisons: all of our reported searchlight analyses demonstrated significant clusters at $p < 0.001$ with both false discovery rate (FDR) and family wise error (FWE) whole-brain correction, we tested only hypothesis-driven representational similarity analysis (RSA) grids, and, despite a whole-brain approach, we see significant results only where they might have been expected *a priori*: in left PrG, left and right STG, and left IFG.

Our multivariate analyses do not include self-to-self decoding of word identity as this is a relatively non-specific measure of speech

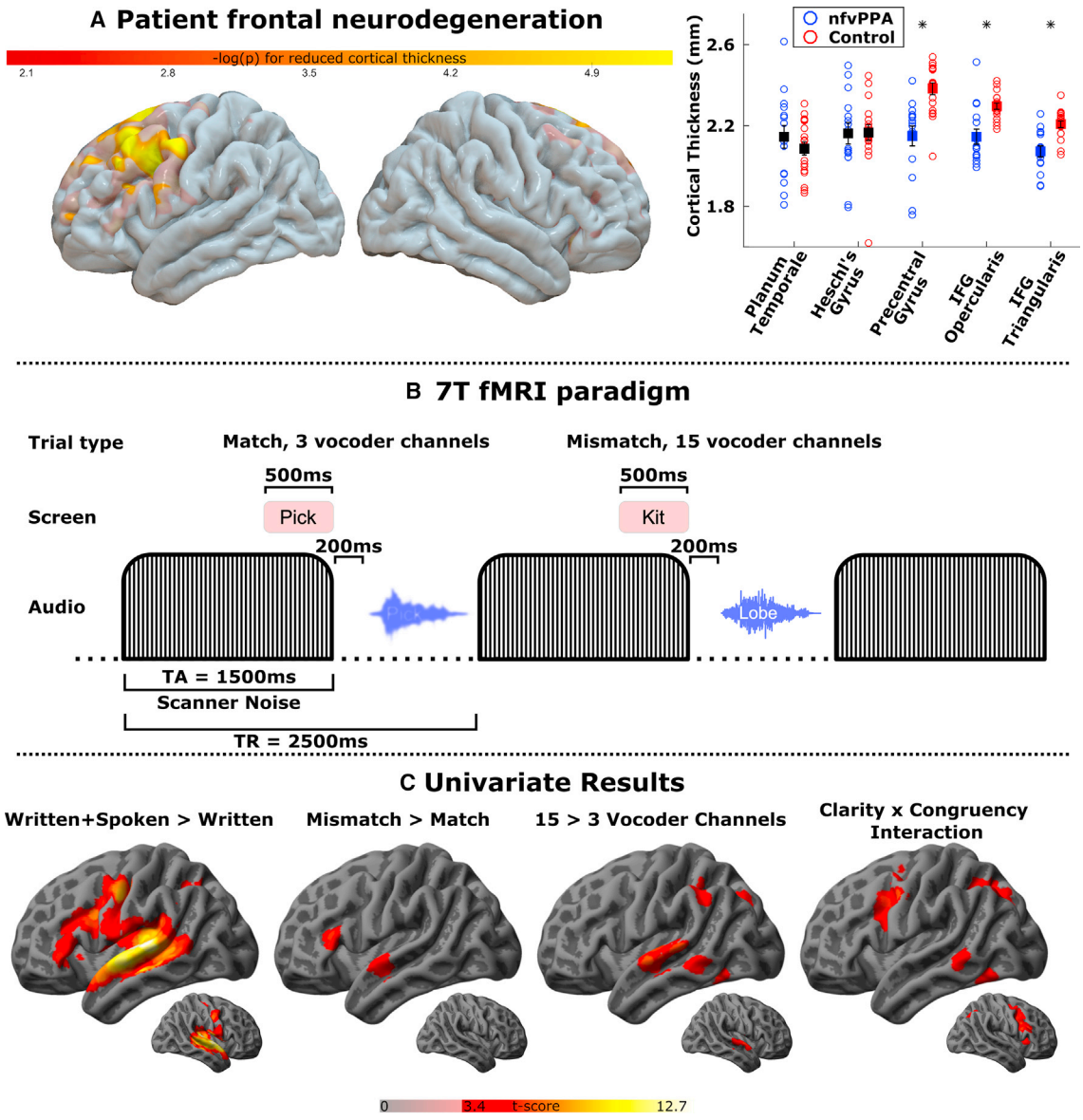


Figure 2. Structural MRI, fMRI paradigm, and univariate fMRI results

(A) Cortical thickness in 15 patients with nfvPPA compared with 19 controls. Rendered brains show cluster-corrected whole-brain results for the comparison of patients with nfvPPA against controls, with a 10-mm surface smoothing kernel. Negative log p maps are shown overlaid on a template cortical surface, thresholded above 2 for visualization (i.e., uncorrected $p < 0.01$). Areas meeting permutation-based cluster-wise significance after 10,000 iterations are shown as opaque, while other areas are shown with transparency. Also shown are single-subject average cortical thicknesses in atlas-based ROIs from the left hemisphere. Circles represent single subjects. Squares represent the mean, and error bars standard error of the mean. Squares are shown in black where they do not significantly differ, and colored by group where they do. Patients displayed significant atrophy in IFG and PrG (all $t(32) < -4.05$, $p \leq 0.0003$) but, crucially, normal volume in auditory temporal brain regions (all one tailed $t(32) \geq 0.06$). Bayesian analysis confirmed very strong evidence for a group difference in frontal ROIs (all $BF10 > 82$) but also moderate evidence for no atrophy in auditory ROIs ($BF01s$ 2.9 and 6.0).

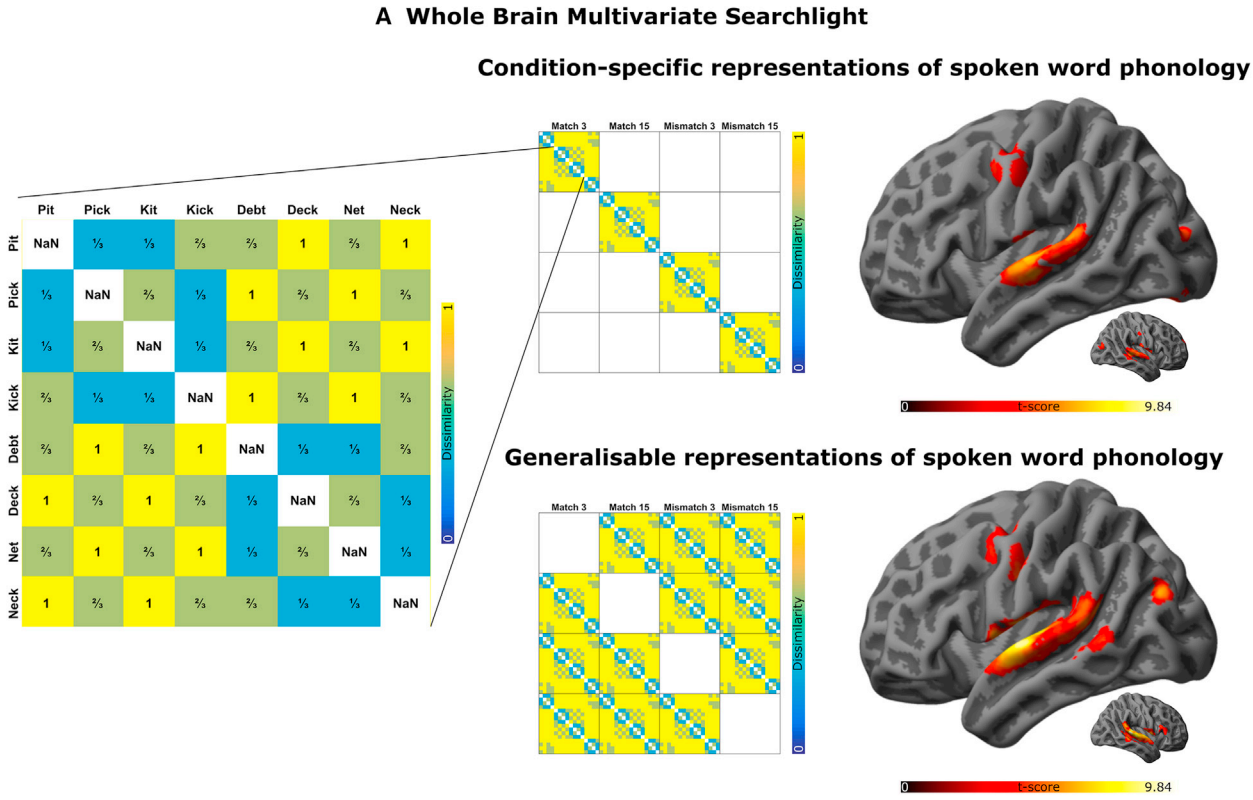
(B) A schematic of written + spoken trials from the in-scanner fMRI paradigm, which were the basis of the multivariate representational similarity analysis (RSA), and represented eight-elevenths of all trials. Participants also experienced written-only trials two-elevenths of the time, in which the spoken word was omitted. Finally, one-eleventh of trials were response trials, included to monitor attention and ensure in-scanner behavioral effects from our manipulation. Response trials followed the same pattern as written + spoken trials, but, 1,050 ms after the onset of the sound stimulus, participants were presented with a written response cue "What did you hear?" above a number of alternatives. They had 6 s to respond using a button box, during which no further written or auditory stimuli were presented.

(C) Univariate fMRI contrasts, cluster thresholded at FDR $p < 0.05$.

Table 1. Cluster-based statistics to accompany neuroimaging analyses

Contrast	Region	MNI co-ordinates	Degrees of freedom	Peak t score	Cluster sizes (voxels)	FDR p value	FWE p value
A: Univariate analysis							
Written + spoken > written	left STG	−52 −18 6	31	12.52	5474	<0.001	<0.001
	left PrG	−52 −8 42	31	8.49	1314	<0.001	<0.001
	left IFG	−51 34 −1	31	5.54	149	<0.001	<0.001
	left IPS	−30 −58 42	31	6.42	125	<0.001	<0.001
	Right STG	46 −25 11	31	12.72	4325	<0.001	<0.001
	Right PrG	51 −7 40	31	6.67	229	<0.001	<0.001
Written + spoken > written; nfvPPA > controls	Right STG	58 −44 6	31	5.13	60	0.004	0.018
Mismatch > match	left IFG	−52 30 16	31	5.74	78	0.001	0.002
	left anterior STG	−57 −7 −7	31	5.14	89	0.001	0.001
15 > 3 vocoder channels	left STG	−62 −8 −2	31	6.43	156, 89, 57	<0.001	<0.001
	left IPS	−50 −42 53	31	5.34	86	<0.001	<0.001
	left posterior MTG	−62 −43 −6	31	4.46	45	0.013	0.045
	Right STG	63 −4 0	31	5.85	122	<0.001	<0.001
Congruency by sensory detail interaction	left PrG	−33, −1, 50	31	5.21	91	<0.001	<0.001
	left MFG	−46 8 41	31	5.81	276	<0.001	<0.001
	left IPS	−28 −66 46	31	5.19	327	<0.001	<0.001
	left IPS	−26 −67 38	31	6.28	90	<0.001	<0.001
	left posterior MTG	−63 −42 4	31	4.61	59	0.007	0.012
B: multivariate analysis: spoken-word phonology							
Within conditions	left PrG	−44 0 47	31	4.74	507	<0.001	<0.001
	left STG	−59 −14 2	31	6.72	4187	<0.001	<0.001
	Right STG	53 −27 10	31	6.62	4825	<0.001	<0.001
Between conditions	left PrG	−52 −10 34	31	5.02	453, 345	<0.001	<0.001
	left STG	−60 −8 4	31	8.76	8170	<0.001	<0.001
	Right STG	66 −32 8	31	9.84	9333	<0.001	<0.001
C: multivariate analysis: prediction error phonology							
Model 1, partial correlations	left IFG	−51 14 7	31	5.33	717	<0.001	<0.001
Model 2, sparse matrices	left IFG	−51 14 7	31	4.96	504	<0.001	<0.001
Controls > nfvPPA, model 1	left anterior STG	−61 −1 −5	31	5.97	407	<0.001	<0.001
Controls > nfvPPA, model 2	left anterior STG	−61 −1 −7	31	5.50	462	<0.001	<0.001
D: physio-physiological connectivity analysis							
PrG > STG connectivity	left insula	−32 17 2	31	6.28	95	<0.001	0.001
	left middle frontal gyrus	−48 26 30	31	5.89	91	<0.001	0.001
	left frontal pole	−38 54 10	31	5.51	161	<0.001	<0.001
	left IPS	−46 −48 48	31	6.04	77	0.001	0.004
Physio-physiological interaction	left occipital fusiform	−34 −67 −16	31	6.77	60	0.006	0.006
	left occipital fusiform	−24 −72 −7	31	6.18	54	0.007	0.012
	left occipital fusiform	−16 −90 −13	31	4.88	56	0.007	0.009
	left inferior occipital gyrus	−27 −91 5	31	5.56	127	<0.001	<0.001
	left IPS	−26 −49 47	31	5.40	63	0.006	0.004
	Right occipital fusiform	21 −78 −7	31	5.91	66	0.006	0.003
	Right lateral occipital	34 −80 29	31	5.19	46	0.015	0.028

Phonological representations in PrG and STG, but not IFG or intraparietal sulcus



(legend on next page)

representations: words may differ along many different dimensions (e.g., semantic, phonological, acoustic, lexical). Instead, we focus on phonology by assessing the number of shared segments between word pairs (counting individual phonemes from the consonant-vowel-consonant [CVC] syllables).

Multivariate fMRI: Phonological information is represented in STG and PrG, not IFG or IPS

Representational dissimilarity matrices based on spoken-word phonological representations correlated with observed multivariate fMRI patterns in left and right STG and left PrG (Figure 3A; Table 1B). These matrices convey the number of shared segments between each spoken word and all other words; i.e., where in the brain was the representation of "pit" most similar to that of "pick" and "kit," with which it shares two segments, less similar to "kick," "debt," and "'net," with which it shares one segment, and most dissimilar to all other words? The first matrix assessed these representations within condition; i.e., where were there phonological representations keeping cue congruency and sensory detail constant? The second matrix assessed the ability of these representations to be generalized between conditions; i.e., where were there consistent phonological representations despite manipulations of cue congruency and sensory detail? There was strong agreement in the topography of the whole-brain searchlight maps created from within-condition and between-condition comparisons.

Multivariate fMRI: STG represents spoken-word phonology, while PrG contains dual representations

The STG represents spoken-word phonology across all conditions (Figure 3B, intercept $F(1,32) = 29.3$, $p = 5.99 \times 10^{-6}$; no significant main effects or interactions of sensory detail, congruency, and diagnosis). However, in PrG, there was a significant crossover interaction between vocoder detail and cue congruency (Figures 3B and 3$F(1,32) = 14.8$, $p = 5.37 \times 10^{-4}$), which did not significantly vary by group ($F(1,32) = 1.44$, $p = 0.239$). Crucially, this was in the opposite direction to the univariate interaction that had defined the ROI, such that phonological representations explained more of the multivoxel pattern in conditions where overall activity was lower. This is exactly the relationship that would be predicted of a brain region combining spoken-word phonology and information-weighted prediction errors for prediction refinement (cf. section "illustrative computational modeling of univariate and multivariate fMRI").

Multivariate fMRI: nfvPPA reduces the consistency of phonological representations in PrG but not STG

We examined the robustness of phonological representations to manipulations of cue congruency and sensory detail. Specifically, we assessed the consistency of phonological representations in six individual cross-condition pairs (Figure 3C) and compared this across groups with repeated-measures ANOVAs in the STG and PrG ROIs.

In STG, both groups represented spoken-word phonology consistently in all condition pairs ($F(1,32) = 40.4$, Greenhouse-Geisser $p = 3.91 \times 10^{-7}$). There was no significant main effect of diagnosis or condition, or interaction between diagnosis and condition. In PrG, again both groups represented spoken-word phonology consistently across all conditions ($F(1,32) = 22.1$, Greenhouse-Geisser $p = 4.66 \times 10^{-5}$). However, this representation was stronger in controls than in patients with nfvPPA (diagnosis $F(1,32) = 5.06$, Greenhouse-Geisser $p = 0.032$). Overall, therefore, both groups represent phonology in PrG, but this representation is less generalizable to other conditions in patients with nfvPPA.

Multivariate fMRI: IFG represents verified and violated predictions in distinct neural populations

Written-to-spoken word mismatches were consistent. This allowed us to examine representational dissimilarity in verified (match) and

Figure 3. Phonological representations of spoken words were observed in PrG and STG, but not IFG or IPS

(A) Whole-brain searchlight results for the RSA of spoken-word shared segments (i.e., number of CVC elements in common across spoken words) across all 34 participants, excluding self-to-self comparisons. Left: design matrices for RSA. There were 16 words presented in four different combinations of vocoder detail and congruency. In the shared-segments model, spoken words had a dissimilarity of 1 if they shared no consonants or vowels, two-thirds if they shared one consonant or vowel, and one-third if they shared two consonants or vowels. Right: whole-brain maps cluster thresholded for visualization at FDR $p < 0.05$, but note that, in both comparisons, left PrG and right and left STG clusters were FWE $p < 0.001$. The top map shows which brain regions contain consistent within-condition phonological similarity (i.e., when cue congruency and vocoder detail were kept constant, where is the brain representation of "pit" more similar to "pick" than it is to "road"?). The bottom map shows the same comparison between conditions (i.e., where was there shared similarity between "pit" and "pick" when cue congruency or vocoder detail were changed?). There was strong agreement between the maps, with definitive evidence for phonological representations in PrG and STG but not IFG or IPS. Note that, although the input data to these maps were only smoothed at 3 mm, the RSA searchlight had an 8-mm radius, effectively introducing a degree of smoothing into the output. The appearance of representations above the lateral fissure is therefore likely an artifact of the fact that searchlights centered on these locations will have included voxels below the fissure, contributing sufficient information for searchlight decoding.

(B) ROI analysis broken down by condition and group. For univariate bar charts, where statistics were done on the whole brain and these figures are illustrative of the effects, error bars represent between-subject standard error to show variability in response magnitude. For multivariate bar charts, where the ROI was independently determined so the results are not double dipped, error bars represent the standard error of the mean after removing between-subject variance, suitable for repeated-measures comparisons.[24] The STG ROI was defined by greater univariate responses for 15 > 3 channel vocoded speech, and showed above-chance multivariate representational similarity of shared segments in all conditions ($F(1,32) = 29.3$, $p = 5.99 \times 10^{-6}$), which did not significantly vary by group or condition. The PrG ROI was defined by the univariate interaction, which was a crossover interaction with greater fMRI signal magnitude in the match 3 and mismatch 15 conditions than in the mismatch 3 and match 15 conditions. Multivariate representational similarity in this region displayed the opposite crossover interaction between vocoder detail and cue congruency ($F(1,32) = 14.8$, $p = 5.37 \times 10^{-4}$), which did not significantly vary by group ($F(1,32) = 1.44$, $p = 0.239$).

(C) ROI analysis for between-condition representations of phonological shared segments. In the STG ROI, both groups represented spoken-word phonology consistently across all conditions ($F(1,32) = 40.4$, Greenhouse-Geisser $p = 3.91 \times 10^{-7}$). There was no significant main effect of diagnosis or condition, or interaction between diagnosis and condition. In the PrG ROI defined from the whole-brain between-condition, shared-segments analysis, again both groups represented spoken-word phonology consistently across all conditions ($F(1,32) = 22.1$, Greenhouse-Geisser $p = 4.66 \times 10^{-5}$). However, this representation was stronger in 19 controls than in 15 patients (diagnosis $F(1,32) = 5.06$, Greenhouse-Geisser $p = 0.032$).

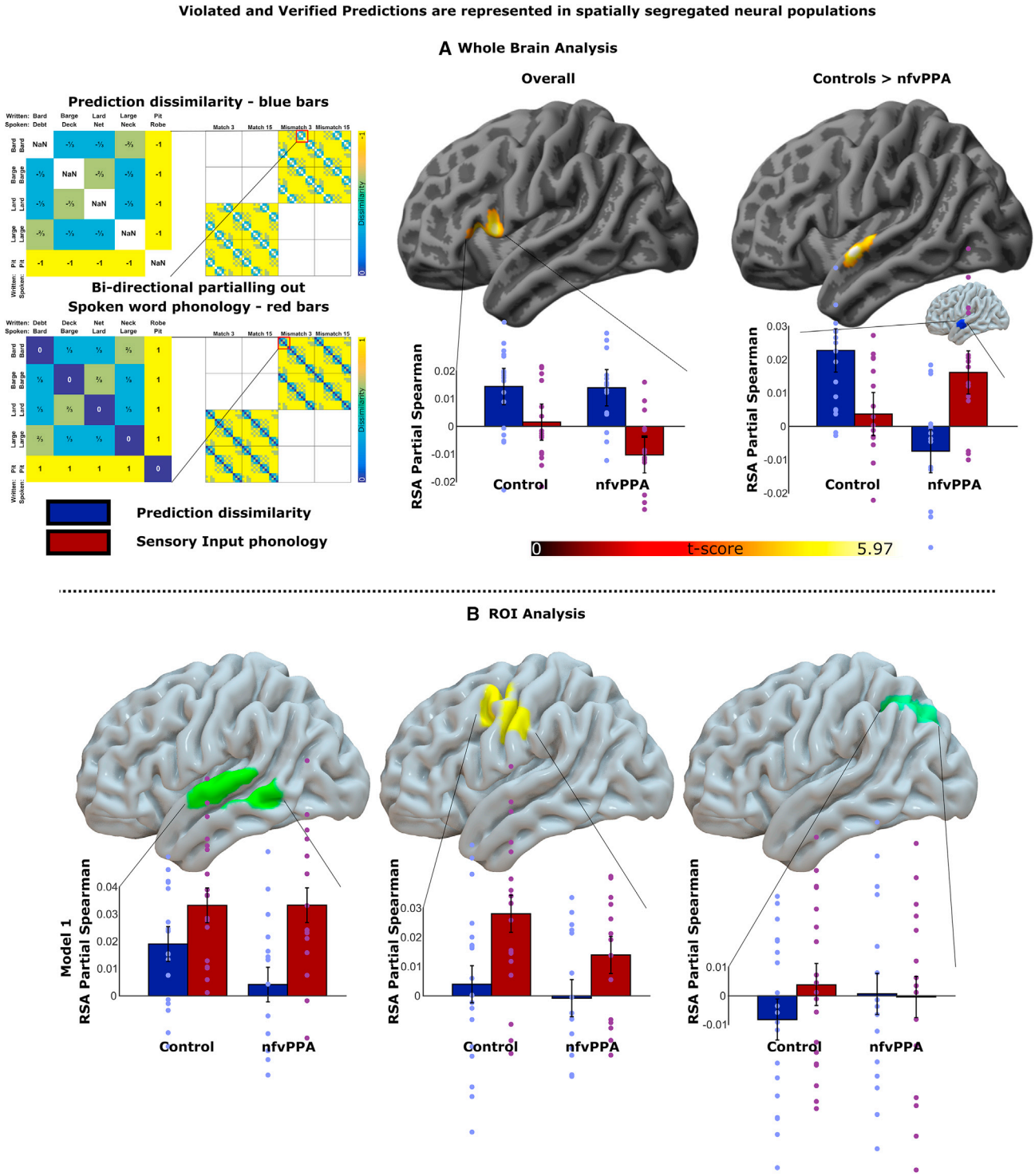


Figure 4. Multivariate assessment of consistent relationships between verified and violated predictions

All observed RSA partial correlations were such that verified and violated predictions were more dissimilar than would be observed by chance; we have inverted the prediction RSA matrix such that these correlations are displayed as positive dark blue bars. Error bars represent between-subject standard error.

(A) Whole-brain analysis for all 34 subjects, cluster corrected at FDR $p < 0.05$. Bar charts show separate ROI analyses of the model of prediction dissimilarity and its matching model of spoken-word phonology. Single-subject datapoints are shown. The IFG ROI is as shown, defined from all subjects. However, using the anterior STG cluster from the whole-brain controls > nfvPPA contrast would be double dipping. We therefore assessed the anterior STG ROI defined from our univariate mismatch > match contrast, displayed in blue on the inset illustrative brain. There was a group-by-condition interaction in this region ($F(1,32) = 5.51$,

(legend continued on next page)

violated (mismatch) predictions, controlling for the phonological information in the spoken word. Specifically, we set up a dissimilarity matrix based upon written-word shared segments across match and mismatch conditions, excluding self-to-self comparisons. In this way, we assessed whether the verified prediction for the word "pick" is more or less similar than chance to the violated predictions for the words "pit" and "kick" (Figure 4, upper left).

For a clearly verified prediction (the match 15 condition, see STAR Methods and Figure 2B), prediction errors are predominantly positive, because high precision sensory inputs exceed uncertain predictions. In contrast, for violated (mismatch) predictions there are negative prediction errors as expectations remain unfulfilled. If positive and negative prediction errors are encoded by differential firing patterns of the same neural units, the hemodynamic blood-oxygen-level-dependent (BOLD) response would be more similar than chance as spatial patterns of metabolic demand overlap. However, if they are encoded by distinct neural populations, albeit in the same brain region, the BOLD patterns would be more dissimilar than chance (i.e., a negative RSA correlation). In such a comparison, it is important to exclude shared phonology in the sensory input. We did this in two different ways. First, we performed partial correlations of prediction error representational similarity, removing the variance that could be explained by spoken-word shared segments (model 1). Second, we assessed prediction error relationships and spoken-word shared segments with separate sparse design matrices, in which any comparison that contained non-zero representational dissimilarity in one design matrix was excluded from the other (model 2). We examined both models and tested for prediction dissimilarity while controlling for sensory input and vice versa.

The results from both models were concordant (Figures 4 and S5; Table 1C). All significant correlations were in the direction of prediction *dissimilarity* between match and mismatch conditions; i.e., verified and violated predictions were more dissimilar than chance. Across all subjects, in both models, left IFG showed a significant correlation with the prediction dissimilarity model, confirming that verified and violated predictions are represented in distinct neural populations in this brain region. The robustness of this finding is emphasized by IFG being the only region consistently implicated across individuals using both RSA models, in two independent groups (controls and patients with nfvPPA), using a cross-validated multi-voxel pattern analysis (MVPA) distance metric.

Multivariate fMRI: Healthy adults but not patients show prediction reconciliation in anterior STG

We tested for the presence of a stronger correlation with the prediction dissimilarity model for control participants than patients with nfvPPA. At the whole-brain level, we found this differential effect only in left anterior STG (FWE $p < 0.001$; Figure 4A; Table 1C). Controls showed consistent representations of prediction dissimilarity in this region, while patients did not, and instead continued to strongly represent sensory input phonology. In other words, while both groups displayed representations of verified and violated predictions in IFG, only controls were able to integrate these with sensory input in anterior STG. It is not surprising that there is no main effect in anterior STG in the presence of this interaction, as Figure 4A shows this to be a crossover effect. This represents failed reconciliation of prediction error in the patient population, explaining their overly precise behavioral predictions, and low perceptual clarity in mismatch conditions (Figure S1, replicating Cope et al.[13]). There were no regions in which patients had a stronger correlation than controls.

Throughout these analyses, we have shown that representations are present in particular regions, not that they are absent in others.[26] For illustration, we therefore examined three further ROIs that had been implicated in our univariate contrasts but not in the whole-brain analyses of prediction representations. These data are shown qualitatively, as it would be statistically questionable, and in violation of our pre-specified analysis strategy, to show ROI statistics for regions that were not implicated at the whole-brain level. However, they are helpful in demonstrating that there is no trend toward a representation of prediction dissimilarity in other regions (Figure 4B, blue bars). Across the whole of STG, both patients and controls represent sensory input phonology, but only controls represented prediction dissimilarity consistently. In PrG, neither group represented prediction dissimilarity consistently across all conditions, but both represented phonology. In IPS, neither representation was found in either group.

Functional connectivity fMRI: PrG connectivity allows integration from multiple information sources

We propose that PrG, IFG, and STG form a network during speech perception. To define the connectivity patterns within this network that support the observed representations of prediction and sensory input, we performed a physio-physiological connectivity analysis (Figure 5). Specifically, we tested whether IFG was conveying the information-weighted prediction error signal to PrG directly, or via STG. Our analysis supports direct connectivity from IFG to PrG, with multiple left frontal clusters more strongly functionally connected to PrG than to STG (Table 1D). PrG was also more strongly connected than STG to IPS. This pattern of connectivity did not significantly vary by group.

We then examined the physio-physiological interaction, which identifies brain regions that are preferentially correlated with activity in one of our seeds when activity in the other seed was low. This is of interest, because the strongest modulator of univariate activity in our experiment is STG activation following the spoken word (Figure 2C). This analysis therefore elucidates PrG connectivity during periods of silence. The resultant clusters resemble a

Greenhouse-Geisser $p = 0.025$). An alternative model based on separate sparse design matrices produced almost identical results (Figure S5; $F(1,32) = 5.95$, Greenhouse-Geisser $p = 0.020$).

(B) Analysis within ROIs not implicated in the whole-brain analysis of prediction dissimilarity, showing shared representations of consistent prediction error and sensory input phonology in STG in controls but no consistent prediction dissimilarity representations in patients in STG, or either group in PrG. STG ROI defined from the 15 > 3 vocoder channel univariate contrast. PrG ROI defined from the multivariate between-condition shared-segment analysis. IPS ROI defined from the univariate interaction between cue congruency and sensory detail. Figure S5 provides a technical replicate of this figure using a different design matrix.

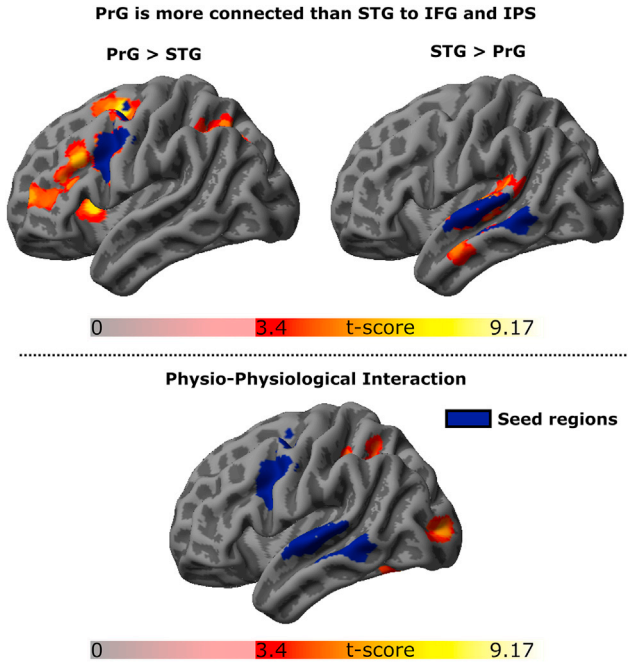


Figure 5. Physio-physiological connectivity analysis results

The physio-physiological interaction can be conceptualized as an extension of the seed-based connectivity approaches commonly employed in resting-state fMRI analyses. It relies on the correlation of regional time series after regressing out univariate behavioral effects at the first level, and as such is a non-directional assessment of the connectivity of two regions to the rest of the brain. Here we specified two seed regions in STG and PrG, shown in blue on the maps above and defined from our univariate contrasts of sensory detail and the interaction of congruency by sensory detail respectively, and assessed these across all 34 subjects. The upper panels show those brain regions that are more strongly connected to PrG than STG, and vice versa, cluster thresholded at FDR $p < 0.05$. PrG was more functionally connected to both IFG and IPS nodes than was STG. The lower panels show the negative physio-physiological interaction (i.e., which brain regions were preferentially correlated with activity in one seed when activity in the other seed was low). Note that there is involvement of a posterior reading network, including both IPS and the visual word form area.[27]

posterior reading network, with both dorsal and ventral components.[27,28] Significant clusters were observed in the visual word form area of left occipital fusiform, as well as left inferior occipital gyrus and left IPS (Table 1D). Right-sided clusters were limited to occipital visual regions. This interaction did not significantly vary by group.

DISCUSSION

This study demonstrates distinct neural representations of phonemic predictions and sensory inputs in a tripartite speech perception network comprising STG, IFG, and PrG (Figure 1). The important findings are that (1) IFG contains distinct representations of verified and violated predictions, which are maximally dissimilar, indicating their processing by distinct neural populations despite their anatomical co-localization; (2) PrG displays responses, representations, and connectivity that are consistent with the integration of sensory input and prediction refinement signals; (3) neurodegeneration of frontal cortex results in inflexible prior expectations, causing a failure to reconcile prediction errors with sensory signals in regions of anterior STG that support echoic memory, and reducing the stability of phonological representations between conditions in PrG. Overall, the data support the hypothesis that humans use a motor model[29] to make perceptual predictions for speech, in which precentral cortex instantiates and refines predictions, and prefrontal cortex supports their top-down reconciliation with sensory inputs in echoic memory.

A tripartite network for speech prediction and perception

Most hierarchical generative models for perception[30–33] and predictive coding[31,34,35] propose a linear arrangement in which the instantiation, reconciliation, and refinement of predictions in a lower-order node is supported by bidirectional connectivity with the same higher-order node. We have previously shown that the reconciliation of perceptual predictions for speech in STG is causally dependent on left frontal cortex,[13] and this study demonstrates that IFG contains distinct representations of verified and violated predictions to support this role. However, it does not necessarily follow that left IFG is solely responsible for the instantiation of predictions, or their refinement based on perceptual events. In fact, the current study shows that PrG contains the necessary combination of representations and connections to perform these roles. This dual support from PrG and IFG allows STG to hold combined representations of the phonological information in prediction outcomes[36,37] and in sensory input (Figure 4).

Such a tripartite arrangement is well suited to flexible and adaptive speech perception. IFG can support speech prediction reconciliation regardless of the nature, strength, or source of the prior expectation. When the prior expectation is of a particular word, from cross-modal information or semantic context, it can reconcile predictions created in precentral gyrus from a motor model. However, IFG also supports the reconciliation of grammatical predictions. Agrammatism is a core diagnostic symptom in nfvPPA,[38] with particular difficulties understanding complex grammatical structures containing hierarchical structures or the passive voice. In such sentences, the listener must re-orient from an expected linear structure to parse a sentence or sub-clause.[39] This linguistic processing is a specialized function of a more general cognitive computational system for complex and flexible thought. It is based on dynamic functional interactions between inferior frontal and superior temporal cortex,[40–42] explaining the close integration between language-selective and domain-general regions in IFG.[43] In the presence of IFG atrophy, there is greater activation of non-dominant STG in response to auditory speech. This may reflect contralateral attempts at compensation, mirroring those seen during auditory change detection.[40]

Illustrative computational modeling of univariate and multivariate fMRI

To illustrate how the roles we propose for each node of the tripartite speech perception network relate to the neural data, we modeled signal magnitude (cf. univariate fMRI power) and representational information (cf. multivariate fMRI). This modeling was undertaken at different stages of hierarchical predictive

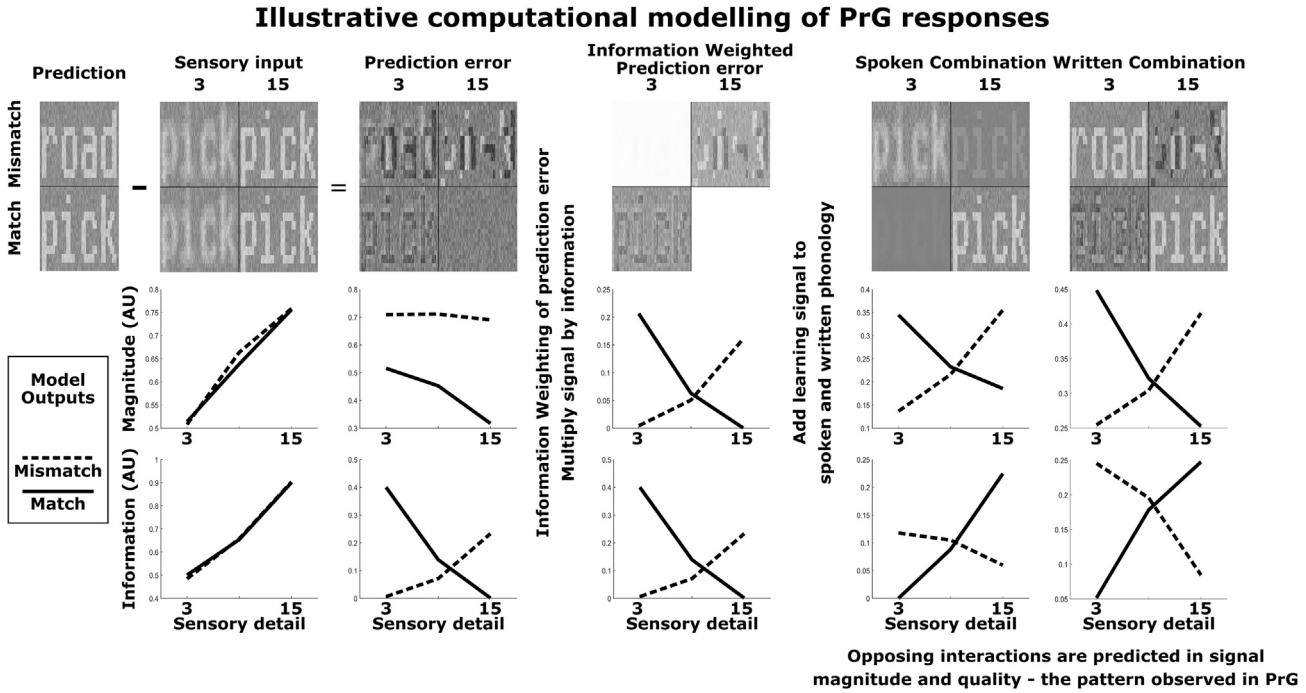


Figure 6. Illustrative computational modeling

Illustrative computational modeling of the magnitude and information content of representations at different stages of hierarchical predictive processing, based on Sohoglu and Davis.[37] In our experimental context, the prediction is the written word and the sensory input is the spoken word, which was manipulated in sensory detail by the application of a 3- or 15-channel noise vocoder. Sensory input increased in magnitude and information as the number of vocoder channels increased, and did not differ between match and mismatch conditions. Prediction error magnitude was greater for mismatch than match conditions, where word identity was unexpected, and also for match 3 than for match 15, where the sensory consequences of the prediction were only partially fulfilled. We propose that a refinement signal based on prediction errors would be information weighted; to construct this representation, we simply multiplied the prediction error in each condition by its information. Match 15 has a minimal refinement signal, because the prediction was completely fulfilled. Mismatch 3 also has a minimal refinement signal, because the prediction error representation contains little information; we know that our prediction was incorrect, but we have little information from which to improve it next time. mismatch 15 and match 3 have large refinement signals, as they represent clear errors of identity and sensory expectations respectively. Finally, we can model the univariate and multivariate responses in a region responsible for the refinement of predictions by combining our spoken and written inputs with this information-weighted prediction error. In both cases, opposite crossover interactions were observed for signal magnitude and information, consistent with those observed in PrG (Figure 3B). Mismatch 15 and match 3 contain combined representations of phonology and information-weighted prediction error that destructively interact, such that they have higher signal magnitude but lower signal information. Match 15 and mismatch 3 contain only the phonology, resulting in lower signal magnitude but higher signal information.

processing (Figure 6), extending the Sohoglu and Davis hypothesis.[37]

Basic modeling of sensory detail and prediction error recapitulated our results in STG and IFG. Increasing the amount of sensory detail in turn increased the modeled magnitude of response to sensory input, paralleling the univariate finding in STG of greater activity for 15 compared with three vocoder channels (Figure 2C). Representational information also increased, but it was high even for noisy inputs (cf. Figures 3B and S3). Prediction error magnitude was greater for mismatch than match conditions, where word identity was unexpected, paralleling this univariate contrast in IFG and anterior STG (cf. Figures 2 and S4). Prediction error representational information displayed an interaction pattern that we did not observe in our experimental data but that has previously been demonstrated in posterior STG in paradigms designed to capture lower-level prediction errors in isolation.[36,37] In our experiment, the consistent relationship between written and spoken words in mismatch trials meant that STG representations were a mixture of sensory input, lower-level, and consistent higher-level prediction errors, at least one of which was available in each trial type.

We extended the basic model by proposing that a prediction-based refinement signal would be information weighted (within trial). This means that, when the written and spoken word match and the spoken word is presented with high sensory detail (cf. match 15 condition), the prediction is fully verified and there is very little error to be used for refinement. Similarly, mismatch 3 also has a minimal refinement signal, because the prediction error representation contains little information. In lay terms, this is because one knows that the prediction was incorrect but has little information from which to improve it next time. In contrast, mismatch 15 and match 3 have large refinement signals, as they represent clear and consistent errors of identity (mismatch15) and sensory expectations (match 3) respectively. In other words, when a *mismatching* written word is presented with *high* fidelity (mismatch 15), there is a large prediction error that contains a large amount of information about the incorrect higher-level expectation; the heard word was clearly not that

which was expected. When there is a *matching* written word presented with *low* fidelity (match 3), there is a prediction error of moderate size, because the prediction of lower-level sensory representation is incompletely fulfilled. Participants are able to ascertain that their higher-level expectations were fulfilled, accounting for the increased perceptual clarity compared with *mismatching low-fidelity* speech (Figure S1, replicating other studies[10,13,44]). This means that there is significant information present in the prediction error about the range of lower-level sensory stimuli that can result from the same higher-level prediction.

We propose that, to refine future predictions, the information-weighted prediction error is combined with representations of phonology in the spoken and written inputs. Adding information-weighted prediction errors to either spoken or written inputs predicts the same opposing interactions for signal magnitude and information. Match 15 and mismatch 3 contain only consistent phonology from the spoken- and written-word inputs, resulting in low signal magnitude but high signal information. In contrast, mismatch 15 and match 3 contain combined representations of phonology and prediction error that destructively interact, resulting in higher signal magnitude but lower signal information. In the experimental data, this pattern was observed in PrG (Figure 3B).

Precentral gyrus is an integrative motor speech hub

PrG contained a representation of the phonology of spoken words that was consistent across listening conditions (Figure 4). This is in keeping with previous demonstrations that precentral gyrus represents phonological (rather than acoustic) information during speech perception.[45] However, we show that this representation is obscured by prediction errors within those conditions that allow prediction refinement (Figures 3B and 6). PrG showed strong functional connectivity with IFG and IPS (Figure 5), and there was a physio-physiological interaction with a posterior reading network,[27,28] consistent with PrG receiving written information before speech onset.

We propose a role for PrG that reconciles seemingly contradictory views of articulatory coding[4,5] and extends and generalizes theories of sensorimotor integration.[29] Recall that our experiment used written words as primes with independent and precise manipulation of predictions and sensory input. Our design therefore avoided confounding effects of prior adaptation, habituation, or repetition suppression, but we propose that these mechanisms are likely to be employed regardless of the source of word identity expectations. We suggest that humans generate auditory predictions of upcoming words on the basis of an internal model of the sounds they would themselves make if reading the word aloud.[46] PrG is already known to play this role during lip reading, when there is an explicit motor component to the cross-modal cue.[47,48] Here we provide evidence to generalize this role to information containing no intrinsic acoustic or motor signal (written text).

This generalization extends the classical motor theory of speech,[2] which held that we perceive speech by inferring the intended motor gestures of other talkers.[3,49] Rather than a process of analysis by synthesis, in which the speech production process is inverted to infer the nature of an incoming phoneme,[50] we propose that an internal model of the motor program for speech is generated pro-actively in PrG, to feedforward perceptual predictions for expected speech sounds before speech is heard. These predictions are then passed to both STG and IFG, where they are later reconciled with sensory input by an iterative settling process.[36] After speech input, this model is refined in PrG by combining sensory input and information-weighted prediction error. This integrative function explains PrG's widespread connectivity with language network nodes.[19]

Reconciliation of predictions into a single percept in echoic memory

Our experimental design, in which mismatching written and spoken words were presented in consistent pairs, allowed us to undertake representational similarity analyses that examined the consistency of the neural pattern of a violated prediction and assess how it related to the neural pattern of a verified prediction for the same word. We demonstrated that there were stable and dissimilar representations of verified and violated predictions in IFG in both groups. This is consistent with distinct neural populations processing the perceptual outcomes of higher-level expectations and lower-level representations,[15] rather than the same populations firing differentially.[14]

Controls but not patients demonstrated the same consistent stable and dissimilar representations of verified and violated predictions in anterior STG. This same brain region also displayed more univariate activity in the mismatch condition in both groups, and has been implicated in previous work as the seat of echoic memory for speech.[20,21,51]

Cross-modal cues reflect predictive processes because the perceptual effect depends on written text being presented before spoken words, and they are effective in improving the perceptual quality of degraded speech only for the duration of echoic memory.[44] In our experiment, we demonstrated that STG contains dual representations of sensory input and prediction error phonology. However, the subjective experience of an ecological cross-modal conflict such as the McGurk effect is not one of dual perception; when a healthy observer is presented with video of a speaker uttering /ga/synchronized to audio of a speaker uttering /ba/, they hear only a combined representation /da/, and are unaware of the conflict until instructed to close their eyes. This perceptual experience reflects the successful reconciliation of cross-modal predictions in echoic memory. Here we demonstrate that, while spoken-word phonology is represented in anterior STG (Figure 3A), controls fully segregate the representations of verified and violated predictions, accounting for more representational information than was present in sensory input alone (Figure 4A). This confirms top-down reconciliation resulting in the restoration of absent but inferred speech signals in STG,[22] modulated by linguistic knowledge,[23] and also a suppression of these neural populations when they represent incorrect predictions. It also accounts for the earlier modulation of univariate activity by predictive context in posterior compared with anterior STG,[52] as only anterior regions rely on IFG-mediated iterative settling to reconcile violated predictions in echoic memory.[36,46]

The effects of frontal lobe neurodegeneration

To test causality of the proposed role of IFG in the language network, we compared healthy individuals with people with focal neurodegeneration of frontal language regions, causing progressive non-fluent aphasia. The neurophysiological consequence of

this frontal neurodegeneration is the delayed reconciliation of perceptual predictions for speech in intact STG.[13] Here we replicate the finding that this has the behavioral consequence of making perceptual predictions inflexible, increasing the effect of cue congruency (Figure S1). This patient population allows us to make a stronger inference about the causal role of frontal and motor parts of the network for speech prediction and perception. An important starting point is that we show a consistent general pattern of results in controls and patients for univariate and multivariate control analyses, providing an internal replication of the primary findings. However, two crucial between-group differences relate to frontal neurodegeneration and inflexible predictive processing.

First, while both patients and controls represent verified and violated predictions in distinct neural populations in IFG, only controls also showed this pattern in anterior STG (Figure 4). This anterior STG locus spatially overlapped with the univariate contrast for greater activity during mismatch compared with match trials, which was displayed by both groups. There was a group-by-condition interaction in the multivariate analysis, replicated across two models (Figures 4 and S5), such that patients continued to represent only the phonology of sensory input along the whole of STG, while, in controls, the phonological representation was replaced by one of prediction outcome in anterior STG. This represents a failure to completely reconcile predictions in patients with nfvPPA, due to impaired iterative settling processes between STG and IFG.[13,36] While controls represent a combined perceptual outcome in echoic memory, patients continue to represent sensory input similarly in match and mismatch conditions, consistent with persisting and competing representations of sensory input and prediction error. This accounts for the patients' experience of high perceptual clarity when sensory input matches expectations, but confusion and low perceptual clarity for unexpected sensory events (Figure S1).

Second, while both patients and controls represented the phonology of spoken words in PrG, this representation was stronger overall (Figure 4B) and more consistent between conditions (Figure 3C) in controls compared with patients. This was not a non-specific effect of signal quality, as there was no group difference in the strength of within-condition phonological representations in PrG (Figure 3B). We speculate that this reduced consistency between conditions may underpin the impaired perceptual learning in nfvPPA, as patients are less able to generalize new knowledge to other perceptual and predictive circumstances despite retaining the ability to learn specific grammatical and auditory associations.[53–55]

LIMITATIONS

Our experiment is designed to assess the neural mechanisms of perceptual prediction for phonemes. This is necessary but not sufficient for word comprehension. Lexical processing requires additional neural processing, which is likely to involve additional brain regions outside of our tripartite network. For example, we have previously shown that anterior temporal lobe is necessary for the efficient, lateralized processing of spoken-word identity.[56] We argue that these anterior temporal responses are amodal, semantic, and account for the phenomenon of surface dyslexia in patients with semantic dementia.[57] In the current study, we included only regularly pronounced CVC words to avoid this confound. Intriguingly, there is emerging evidence that disconnection of anterior temporal lobe leads to an increase in the reliance of auditory cortex on frontal and motor connectivity, perhaps in compensation.[58]

In order to precisely control the experimental and neural context, we provide participants with an explicit prediction in the form of a written prime that correctly predicted the heard word 50% of the time. This provides a situation in which auditory targets are predicted by cues that draw on participants' lifelong experience of the arbitrary associations between visual symbols (letter strings) and auditory signals (spoken words) more powerfully than could be achieved by a newly learned arbitrary cross-modal association,[59] which may be confounded by differential probabilistic learning in healthy individuals and people with nfvPPA.[53] However, ecological perception of running speech involves the ongoing generation of predictions by rapid integration of novel information with the linguistic context, enabling predictions of what will be said next, taking into account both semantics[52,56,60] and grammatical structure.[53,61] Accordingly, our experiment assesses only perceptual prediction and predictive coding, not the process of prediction identity generation, which would likely invoke wider-scale probabilistic networks such as the multiple demand system, which would overlap with the language networks studied here.[18,43] This overlap would make it difficult to dissociate activations, especially with the poor temporal resolution of MRI; running speech prediction paradigms may be more suitable to M/EEG, where the time course of prediction identity generation and perceptual resolution could be more easily separated, but at the cost of spatial resolution.

There was a mid-study change in the number of response options (from two to four) provided to participants during the in-scanner behavioral task, and the coronavirus pandemic prevented this being fully counter-balanced across groups. While undesirable, we are confident that this could not have affected the data presented here. The aim of the behavioral in-scanner task was solely to maintain participants' attention to the stimuli, by intermittently asking for a response. The neural data from these response trials were not analyzed. All of the reported neural results come from the standard trials, in which a response was not requested. All of the behavioral data in the main manuscript were collected out of scanner, and those tasks were identical across all individuals.

The effect sizes we report here may appear small, but in fact they are equal or greater to those in comparable studies. Multivariate fMRI studies of language representations commonly demonstrate MVPA Spearman correlations around 0.02–0.04.[36,45,62] These correlation values are low for two reasons: the complexity of the neural signal being evaluated compared with the sparse RSA matrix being evaluated, and the signal-to-noise ratio of the measurement technique. This does not undermine the use of MVPA. What one evaluates with RSA is whether the theoretical matrix is consistently represented in the neural data across subjects, not whether it is a holistic explanation of all the variance that is represented in those data.

The current study is a theoretically motivated, task-based functional imaging study of people with precisely phenotyped nfvPPA. Our group sizes are small compared with some recent structural imaging and neuropsychological descriptive studies of primary progressive aphasia (PPA). However, because testing

our hypothesis required a dissociation between frontal lobe atrophy and temporal lobe preservation (Figure 2A), we prioritized the recruitment of this extremely rare uniform cohort over a larger group of patients with unselected PPA who would not have allowed such precise mechanistic conclusions. The reliability of our data is emphasized by (1) exact replication of the key behavioral and neuroanatomical findings from Cope et al.[13] in a new cohort, explained within the same computational framework; and (2) replication within the new study of the primary multivariate fMRI findings in both the patient and control groups. Our study was ambitious, and only made possible by the use of a 7-T scanner, which has significantly superior signal to noise compared with 3-T, allowing the acquisition of data either at higher resolution, higher signal to noise, or more quickly. Here, the priority was acquisition speed, because we know that patients often cannot tolerate the scanner environment for more than an hour; using 7-T allowed us to collect our data with 40 min of echo planar imaging (EPI) scanning, compared with 2 h for a comparable 3-T acquisition. The higher spatial resolution imaging brought additional benefits. There is controversy in the literature about the value of high resolution for multivariate analyses,[63] as it depends on the spatial scale of the signal, compared with the noise and the smoothing filter.[64] We assessed the importance of high-resolution acquisition in our data by repeating our analyses with 8-mm spatial smoothing, simluating lower resolution data, and this made little difference in most regions. However, it was crucial in left IFG, where we argue there are two competing codes at fine-grained spatial scale, and the results were consequently lost with 8-mm smoothing. These results are not artifactual, because they survive independently in both patients and controls and are present across multiple conditions and comparisons at stringent statistical thresholds but are at a spatial scale where they may have been obscured at 3-T.

Conclusions

We provide univariate activation, multivariate representation, and causal lesion evidence for a motor model of predictions in speech perception. There is a tripartite speech perception network in the dominant hemisphere in which (1) STG simultaneously represents sensory information and prediction errors, (2) IFG contains distinct neural representations of verified and violated predictions, and (3) precentral gyrus contains representations of spoken-word phonology in combination with information-weighted prediction errors. We propose that the IFG plays a primary role in supporting the top-down reconciliation of predictions with sensory input, while precentral gyrus plays a primary role in instantiating and refining those predictions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Ethics
  - Participants
- METHOD DETAILS
  - Scan protocol
  - In-scanner paradigm
  - Out-of-scanner experiments
  - Structural MRI preprocessing
  - Functional MRI pre-processing
  - fMRI Physio-physiological interaction
  - Illustrative computational modelling
  - Data visualisation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Structural MRI analysis
  - Functional MRI univariate analysis
  - Functional MRI multivariate analysis
  - fMRI physio-physiological interaction analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112422.

ACKNOWLEDGMENTS

This work was primarily funded by grants to author T.E.C. from the Association of British Neurologists, Patrick Berthoud Charitable Trust, and Academy of Medical Sciences (SGL024\1069). Additional support was provided by the National Institute for Health Research both directly and through the Cambridge Biomedical Research Centre (BRC-1215-20014; NIHR203312), Alzheimer's Research UK (ARUK-RS2019-002), the Medical Research Council (MR/M008983/1; MR/M009041/1; SUAG092/MC_UU_0030/14; MC_UU_00005/5; MC_UU_00030/6), and the Wellcome Trust (220258). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

We thank Prof. Matt Lambon Ralph for encouragement and intellectual discussion of the paper. We thank Prof. Jonathan Peelle for sharing his mesh projection code, on which our data visualization technique is based.

AUTHOR CONTRIBUTIONS

Conceptualization, T.E.C., E.S., M.H.D., and J.B.R.; data curation, T.E.C. and P.S.J.; formal analysis, T.E.C.; funding acquisition, T.E.C.; investigation, T.E.C. and K.A.P.; methodology, T.E.C., E.S., P.S.J., C.R., L.P., K.P., M.H.D., and J.B.R.; project administration, T.E.C., K.A.P., and J.B.R.; resources (patient identification and recruitment), T.E.C., J.C., C.R.B., P.G., K.A.-A., M.H., and J.B.R.; software, T.E.C., E.S., and P.S.J.; supervision, J.B.R.; visualization, T.E.C.; writing – original draft, T.E.C.; writing – review and editing, T.E.C., E.S., L.P., W.S., B.P., T.D.G., K.P., M.H.D., and J.B.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure gender balance in the recruitment of human subjects. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority

in their field of research. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community.

Received: August 4, 2022
Revised: December 23, 2022
Accepted: April 5, 2023
Published: April 24, 2023

REFERENCES

1. von Helmholtz, H. (1925). Helmholtz's Treatise on Physiological Optics (Optical Society of America).
2. Lane, H. (1965). The motor theory of speech perception: a critical review. Psychol. Rev. *72*, 275–309.
3. Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition *21*, 1–36.
4. Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. Nat. Rev. Neurosci. *11*, 351–360.
5. Scott, S.K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. Nat. Rev. Neurosci. *10*, 295–302.
6. Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215.
7. Van Wassenhove, V., Grant, K.W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci. USA *102*, 1181–1186.
8. McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. Nature *264*, 746–748.
9. Blank, H., Spangenberg, M., and Davis, M.H. (2018). Neural prediction errors distinguish perception and misperception of speech. J. Neurosci. *38*, 6076–6089.
10. Sohoglu, E., Peelle, J.E., Carlyon, R.P., and Davis, M.H. (2012). Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. *32*, 8443–8453. https://doi.org/10.1523/JNEUROSCI.5069-11.2012.
11. Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. Trends Cognit. Sci. *10*, 14–23.
12. Wolff, S.B., and Ölveczky, B.P. (2018). The promise and perils of causal circuit manipulations. Curr. Opin. Neurobiol. *49*, 84–94.
13. Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R.P., Griffiths, T.D., et al. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat. Commun. *8*, 2154.
14. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.
15. Keller, G.B., and Mrsic-Flogel, T.D. (2018). Predictive processing: a canonical cortical computation. Neuron *100*, 424–435.
16. Friston, K. (2018). Does predictive coding have a future? Nat. Neurosci. *21*, 1019–1021.
17. Haarsma, J., Fletcher, P.C., Griffin, J.D., Taverne, H.J., Ziauddeen, H., Spencer, T.J., Miller, C., Katthagen, T., Goodyer, I., Diederen, K.M.J., and Murray, G.K. (2020). Precision weighting of cortical unsigned prediction error signals benefits learning, is mediated by dopamine, and is impaired in psychosis. Mol. Psychiatr. *26*, 5320–5333.
18. Fedorenko, E., and Blank, I.A. (2020). Broca's area is not a natural kind. Trends Cognit. Sci. *24*, 270–284.
19. Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. *12*, 718–724.
20. Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr. Biol. *28*, 1860–1871.e4.
21. Buchsbaum, B.R., Olsen, R.K., Koch, P., and Berman, K.F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. Neuron *48*, 687–697.
22. Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual restoration of masked speech in human cortex. Nat. Commun. *7*. 13619–9.
23. Kim, S.-G., De Martino, F., and Overath, T. (2021). Linguistic modulation of the neural encoding of phonemes. Preprint at bioRxiv. https://doi.org/10.1101/2021.07.05.451175.
24. Loftus, G.R., and Masson, M.E. (1994). Using confidence intervals in within-subject designs. Psychon. Bull. Rev. *1*, 476–490.
25. Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., and Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. *12*, 535–540.
26. Gelman, A., and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. Am. Statistician *60*, 328–331.
27. Dehaene, S., and Cohen, L. (2011). The unique role of the visual word form area in reading. Trends Cognit. Sci. *15*, 254–262.
28. Cohen, L., Dehaene, S., Vinckier, F., Jobert, A., and Montavont, A. (2008). Reading normal and degraded words: contribution of the dorsal and ventral visual pathways. Neuroimage *40*, 353–366.
29. Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. Neuron *69*, 407–422.
30. Hinton, G.E. (2007). Learning multiple layers of representation. Trends Cognit. Sci. *11*, 428–434.
31. Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836.
32. Heeger, D.J. (2017). Theory of cortical function. Proc. Natl. Acad. Sci. USA *114*, 1773–1782.
33. Friston, K. (2008). Hierarchical models in the brain. PLoS Comput. Biol. *4*, e1000211. https://doi.org/10.1371/journal.pcbi.1000211.
34. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.
35. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711. https://doi.org/10.1016/j.neuron.2012.10.038.
36. Blank, H., and Davis, M.H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biol. *14*, e1002577.
37. Sohoglu, E., and Davis, M.H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. Elife *9*, e58077.
38. Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., et al. (2011). Classification of primary progressive aphasia and its variants. Neurology *76*, 1006–1014.
39. Traxler, M.J., Morris, R.K., and Seely, R.E. (2002). Processing subject and object relative clauses: evidence from eye movements. J. Mem. Lang. *47*, 69–90.
40. Cope, T.E., Hughes, L.E., Phillips, H.N., Adams, N.E., Jafarian, A., Nesbitt, D., Assem, M., Woolgar, A., Duncan, J., and Rowe, J.B. (2022). Causal evidence for the multiple demand network in change detection: auditory mismatch magnetoencephalography across focal neurodegenerative diseases. J. Neurosci. *42*, 3197–3215. https://doi.org/10.1523/JNEUROSCI.1622-21.2022.
41. Friederici, A.D., Chomsky, N., Berwick, R.C., Moro, A., and Bolhuis, J.J. (2017). Language, mind and brain. Nat. Human Behav. *1*, 713–722. https://doi.org/10.1038/s41562-017-0184-4.

42. Milne, A., Wilson, B., and Christiansen, M. (2018). Structured sequence learning across sensory modalities in humans and nonhuman primates. Current Opinion in Behavioral Sciences *21*, 39–48.

43. Fedorenko, E., Duncan, J., and Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Broca's area. Curr. Biol. *22*, 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011.

44. Sohoglu, E., Peelle, J.E., Carlyon, R.P., and Davis, M.H. (2014). Top-down influences of written text on perceived clarity of degraded speech. J. Exp. Psychol. Hum. Percept. Perform. *40*, 186–199. https://doi.org/10.1037/a0033206.

45. Evans, S., and Davis, M.H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cerebr. Cortex *25*, 4772–4788.

46. Davis, M.H., and Johnsrude, I.S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear. Res. *229*, 132–147.

47. Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. *25*, 1649–1653. https://doi.org/10.1016/j.cub.2015.04.049.

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

49. Galantucci, B., Fowler, C.A., and Turvey, M.T. (2006). The motor theory of speech perception reviewed. Psychon. Bull. Rev. *13*, 361–377.

50. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychol. Rev. *74*, 431–461.

51. Davis, M.H., and Gaskell, M.G. (2009). A complementary systems account of word learning: neural and behavioural evidence. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 3773–3800.

52. Davis, M.H., Ford, M.A., Kherif, F., and Johnsrude, I.S. (2011). Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. J. Cognit. Neurosci. *23*, 3914–3932.

53. Cope, T.E., Wilson, B., Robson, H., Drinkall, R., Dean, L., Grube, M., Jones, P.S., Patterson, K., Griffiths, T.D., Rowe, J.B., and Petkov, C.I. (2017). Artificial grammar learning in vascular and progressive non-fluent aphasias. Neuropsychologia *104*, 201–213. https://doi.org/10.1016/j.neuropsychologia.2017.08.022.

54. Hardy, C.J.D., Marshall, C.R., Bond, R.L., Russell, L.L., Dick, K., Ariti, C., Thomas, D.L., Ross, S.J., Agustus, J.L., Crutch, S.J., et al. (2018). Retained capacity for perceptual learning of degraded speech in primary progressive aphasia and Alzheimer's disease. Alzheimer's Res. Ther. *10*, 70.

55. Henry, M.L., Meese, M.V., Truong, S., Babiak, M.C., Miller, B.L., and Gorno-Tempini, M.L. (2013). Treatment for apraxia of speech in nonfluent variant primary progressive aphasia. Behav. Neurol. *26*, 77–88. https://doi.org/10.3233/BEN-2012-120260.

56. Cope, T.E., Shtyrov, Y., MacGregor, L.J., Holland, R., Pulvermüller, F., Rowe, J.B., and Patterson, K. (2020). Anterior temporal lobe is necessary for efficient lateralised processing of spoken word identity. Cortex *126*, 107–118.

57. Woollams, A.M., Ralph, M.A.L., Plaut, D.C., and Patterson, K. (2007). SD-squared: on the association between semantic dementia and surface dyslexia. Psychol. Rev. *114*, 316–339.

58. Kocsis, Z., Jenison, R.L., Cope, T.E., Taylor, P.N., Calmus, R.M., McMurray, B., Rhone, A.E., Sarrett, M.E., Kikuchi, Y., and Gander, P.E. (2022). Immediate neural network impact after the loss of a semantic hub. Preprint at bioRxiv. https://doi.org/10.1101/2022.04.15.488388.

59. Kok, P., Jehee, J.F.M., and De Lange, F.P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. Neuron *75*, 265–270.

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

61. Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical learning by 8-month-old infants. Science *274*, 1926–1928.

62. Du, Y., Buchsbaum, B.R., Grady, C.L., and Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. USA *111*, 7126–7131.

63. Op de Beeck, H.P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage *49*, 1943–1948.

64. Kriegeskorte, N., Cusack, R., and Bandettini, P. (2010). How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage *49*, 1965–1976.

65. Zeinali-Rafsanjani, B., Faghihi, R., Mosleh-Shirazi, M.A., Moghadam, S.-M., Lotfi, M., Jalli, R., Sina, S., and Mina, L. (2018). MRS shimming: an important point which should not be ignored. J. Biomed. Phys. Eng. *8*, 261–270.

66. Andersson, J.L.R., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage *20*, 870–888.

67. Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn. Reson. Med. *63*, 1144–1153.

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

69. Sohoglu, E., and Davis, M.H. (2016). Perceptual learning of degraded speech by minimizing prediction error. Proc. Natl. Acad. Sci. USA *113*, E1747–E1756.

70. Gaser, C., and Dahnke, R. (2016). CAT-a computational anatomy toolbox for the analysis of structural MRI data. Hbm *2016*, 336–348.

71. McLaren, D.G., Ries, M.L., Xu, G., and Johnson, S.C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage *61*, 1277–1286.

72. Hebart, M.N., Görgen, K., and Haynes, J.-D. (2014). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinf. *8*, 88.

73. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.

74. Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage *137*, 188–200.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Organisms/strains		
Human: patients with nfvPPA	1) Cambridge University Hospitals NHS Foundation Trust 2) Oxford University Hospitals NHS Foundation Trust 3) St George's Healthcare NHS trust	N/A
Human: control	Join Dementia Research	https://www.joindementiaresearch.nihr.ac.uk/
Software and algorithms		
Matlab 2014a	Mathworks	https://www.mathworks.com/
Psychtoolbox	Brainard, 1997	http://psychtoolbox.org/
SPM12	Wellcome Department of Cognitive Neurology, London, United Kingdom	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
Freesurfer 7.1.0	Laboratory for Computational Neuroimaging, Athinoula A. Martinos Center for Biomedical Imaging.	https://surfer.nmr.mgh.harvard.edu/
Surf Ice	Neuroimaging Tools and Resources Collaboratory	https://www.nitrc.org/projects/surfice/
Other		
7T Terra MRI Scanner	Siemens	https://www.siemens-healthineers.com/en-uk/magnetic-resonance-imaging/7t-mri-scanner/magnetom-terra
Single-channel transmit, 32-channel receive head coil	Nova Medical, Wilmington MA, USA	https://www.siemens-healthineers.com/en-uk/magnetic-resonance-imaging/options-and-upgrades/coils/nova-medical-head-coil
S15 insert earphones	Sensimetrics	https://www.sens.com/products/model-s15/
PROPixx projector	VPixx	https://vpixx.com/products/propixx/
Rigid rear projection screen	Comar Optics	
PCI 6503 card	National Instruments	https://www.ni.com/en-gb/support/model.pci-6503.html
HD250 linear 2 headphones	Sennheiser	
UCA 202 external sound card	Behringer	https://www.behringer.com/product.html?modelCode=P0484

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact Thomas Cope (thomascope@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw neuroimaging data are not publicly available due to consent, ethical, and governance approval restrictions. The data that support the findings of this study are available on request from the corresponding author.
- All original code has been deposited at the relevant github repositories linked below, and is publicly available as of the date of publication.

- Any additional information required to reanalyse the data reported in this work is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ethics

Study procedures were approved by the UK Health Research Authority after review by the Cambridge Central Ethics Committee (16/EE/0084, 16/EE/0351). Participants had mental capacity and gave written informed consent to participation.

Participants

Seventeen patients with early nfvPPA were identified according to consensus diagnostic criteria,[38] of whom fifteen were able to complete the neuroimaging protocol and are included in the final analysis (one patient did not tolerate the scanner environment, and another did not understand the behavioural task). As nfvPPA is a rare diagnosis we took a multi-centre approach. Potential patient participants were identified in the specialist cognitive clinics of authors TEC, JC, CB, PG, KA, and JBR, then screened for 7T MRI suitability, and asked if they would be willing to travel to the main study site (Cambridge). There the diagnosis was verified by authors TEC and KP, and all study procedures including scanning were performed. Twenty-two age and gender matched controls were recruited, primarily from the National Institute for Health Research 'Join Dementia Research' volunteer database. Nineteen were able to complete the neuroimaging protocol and are included in the final analysis (two did not tolerate the scanner environment, and one had scanner technical problems). Participant demographics are shown in the table below, and average pure-tone audiograms are shown in Figure S6. There were no statistically significant group differences in auditory threshold at any frequency in either ear.

Group Mean (sd)	Number and gender	Age	Years of Education	ACE-R (/100)	ACE-R excluding fluency (/86)	MMSE (/30)	Raven's matrices (/60)	Boston naming (/15)
Control	19 (11M, 8F)	69 (6)	16.1 (2.2)	95 (4)	83 (3)	29 (2)	48 (5)	15 (1)
nfvPPA	15 (9M, 6F)	71 (8)	14.4 (3.4)	82 (9)	75 (5)	27 (2)	33 (12)	14 (2)
Difference	$X^2(1) = 0.153$ $p = 0.910$	$t(25) = 0.592$ $p = 0.559$	$t(21) = 1.66$ $p = 0.112$	**$t(16) = 4.68$ $p < 0.001$**	**$t(17) = 3.52$ $p = 0.002$**	$t(20) = 1.72$ $p = 0.101$	**$t(16) = 4.31$ $p < 0.001$**	**$t(19) = 2.23$ $p = 0.038$**

Demographics table: Included participant demographics. ACE-R = Addenbrooke's cognitive examination, with the total score and the score excluding verbal fluency reported separately. MMSE = Mini-Mental State Examination. Boston naming tests were scored ignoring phonemic and phonetic errors, as we were interested in excluding anomia. Difference tests employed chi-squared for categorical data, and unpaired t-tests with unequal variance for continuous data.

METHOD DETAILS

Scan protocol

All participants underwent a single session of standardised magnetic resonance imaging at the Wolfson Brain Imaging Center, University of Cambridge, using a 7T Siemens Terra scanner, with a Siemens single-channel transmit, 32-channel receive head coil (Nova Medical, Wilmington MA, USA). Following scout images and field maps, the vendor's automatic shimming (at least three iterations) and interactive manual shimming were performed to generate an homogeneous magnetic field across the brain, targeting an FWHM of the water peak measured in-line at the scanner console of <40Hz. Note that the FWHM scales linearly with magnetic field, so this is similar to usual targets of less than 20 Hz at 3T, or 10 Hz at 1.5T.[65]

After shimming and acquisition of final B0 field maps, participants performed four blocks of the fMRI task described below. After two of these blocks (i.e. half way through), four volumes with the same scan parameters were acquired with the phase-encode direction reversed to enable *topup* correction of susceptibility distortions.[66] fMRI was acquired with a gradient echo echo planar imaging (EPI) sequence with a GRAPPA acceleration factor of 2 in single-band mode implemented by CMRR.[67] If movement occurred during the reference image, resulting in ghosting in the first acquired volumes, the block was immediately aborted and restarted after a reminder to the participant to stay very still. 1.5mm isotropic voxels were acquired in 72 interleaved slices, each with a 150x150 voxel matrix. The acquisition plane was tilted from axial so as to avoid the eyes and maximise coverage. The resulting 108x225x225mm oblique volume was sufficient for whole-brain coverage in most participants, but dorsal parietal lobe was sacrificed for those with exceptionally large brain size. Sparse acquisition allowed for an acquisition time (TA) of 1500ms, followed by a silent gap of 1000ms during which the auditory stimulus was presented, giving a repetition time (TR) of 2500ms. Echo time (TE) was 23.4ms and flip angle 50 degrees. The task paradigm began after four volumes, to allow settling and time to restart the block in the case of movement during the reference image. 238 volumes were acquired, meaning that each block was slightly more than ten minutes in duration, including the acquisition of the acceleration reference volumes, resulting in 40 minutes of total task-based fMRI per session.

Next, a structural MP2RAGE image was acquired with TR 4300ms, TE 1.99ms, flip angles 5/6 degrees, inversion times 840/2370ms, resulting in 224 slices of 0.75mm isotropic voxels in a 300x320 matrix. Finally, a "fast" T2-weighted image was acquired

for clinical reporting and study governance with 34 slices of 3.9mm thickness acquired at an in-plane resolution of 0.225x0.225mm with matrix size 1024x768. Total MRI session duration was approximately 75 minutes.

In-scanner paradigm

Each participant performed four blocks of an audio-visual fMRI paradigm designed to evaluate the influences of prior knowledge and sensory degradation on the perception of spoken language by presenting a written word before a vocoded spoken word (Figure 2B). Participants were explicitly informed, and implicitly learned through practice, that there was a 50% chance that the trial would be congruent (written word matching the spoken word), and a 50% chance that it would be incongruent (the spoken word being one of the other words). This strongly incentivises prediction, because in half of trials the prediction would be fulfilled, while also necessitating perceptual flexibility, because in half of trials participants would need to correctly perceive a largely unexpected word. This has a very consistent perceptual effect, increasing the perceptual clarity of the heard word when written and spoken text match.

All of our multivariate analyses are based on 'written+spoken' trials, in which participants were presented with a written word, followed 700ms later by a consonant-vowel-consonant (CVC) spoken word, which was acoustically degraded using a noise vocoder with either 3 or 15 channels, creating low and high sensory detail respectively.[68] No response was requested from participants on these trials. This allowed for a factorial manipulation of predictions, by presenting written text that either matched or mismatched with the speech, and sensory detail, by varying the number of channels in the noise vocoder. The spoken word was synchronised with the fMRI sparse acquisition, beginning 200ms into the one-second silent gap. These combined written+spoken pairs comprised 8/11 of trials (128 trials per block, 512 trials in total).

For modelling purposes, in 2/11 of trials (32 trials per block, 128 trials per session), participants were presented with a written word, but no spoken word was presented; a 'written-only trial'. Each written+spoken or written-only trial lasted one scanner TR, i.e. 2.5 seconds in total.

One eleventh of the trials were 'response trials' (16 trials per block, 64 trials per session), with the purpose of maintaining participant attention to the screen and auditory stimuli. These 'response trials' did not contribute to any fMRI contrasts, and do not contribute to any of the neural data presented in the study. In these trials, 1050ms after the onset of the sound stimulus, participants were presented with a written response cue "What did you hear?" above written-word alternatives. Participants had six seconds to select an alternative using a button box in their left hand, after which time the text was removed and a null response was recorded. Once the participant responded, the cue disappeared to avoid repeated responses, but the next trial did not begin until four scanner TRs (10 seconds in total) had elapsed. For the first twenty participants (11 patients, 9 controls), two alternatives were presented, one being the heard word and one being another word from the experimental set that shared no phonemes with that word (for example 'pit' and 'robe'). Interim analysis showed that this was helpful in confirming attention and perception, but was so easy that performance was close to ceiling even for mismatching text with low sensory detail. Therefore, to confirm the perceptual salience of our experimental manipulation in the scanner environment, for the remaining subjects (4 patients, 10 controls), four alternatives were presented, one being the heard word, and the other three being close neighbours that shared a vowel (for example 'bard', 'barge', 'lard', and 'large'). It was not possible to scan more patients with the four-alternative forced-choice because of a national lockdown during the coronavirus pandemic, resulting in a group imbalance on the in-scanner response trials, which differed only in the number of response options provided. Sufficient behavioural data were acquired to confirm the salience of the in-scanner manipulations in both groups (see supplementary results). All of the behavioural data in the main manuscript were collected out-of-scanner, and that task was identical across all individuals.

The experimental set was restricted to sixteen words, split into four sets across four vowels, designed to facilitate representational similarity analysis. Within each set, every word had two close neighbours that shared two phonemes (consonant plus vowel), and one neighbour that shared only the vowel. Some consonants were shared across vowel sets. The sixteen words were: bard, barge, lard, large; pit, pick, kit, kick; debt, deck, net, neck; robe, road, lobe, load. That is to say, after reading the word 'pit', the participants performing optimally would create a perceptual prediction that they have a 50% chance of hearing the word 'pit', and a 3.33% chance of hearing any one of the other words in the experiment.

Crucially, trials where written and spoken text mismatched were consistent, such that if a written word did not predict itself, it instead predicted a word eight items later on the list above. Therefore, reading the word 'bard' meant that the subject would always subsequently hear 'bard' (a matching written+spoken trial), 'debt' (a mismatching written+spoken trial), or nothing at all (a written-only trial). Similarly reading 'debt' predicted hearing either 'debt', 'bard', or nothing at all. This consistent relationship allowed representational similarity analyses of the consistency of the neural pattern of a prediction error, and how it related to the neural pattern of a confirmed prediction for the same word.

In each block, every word was presented in written form eleven times, and in spoken form nine times, such that there were two identical exemplars of every 'written+spoken' and 'written-only' trial type, and one exemplar of a response trial. To ensure consistency, stimuli and trial types were presented in fixed random order, such that their order differed between blocks, but was the same for every participant. The first trial was always 'written+spoken'. There were between 1 and 8 other trial types between 'written-only' trials (i.e. they were never consecutive), and between 6 and 15 other trial types between response trials. A response trial never followed immediately after a 'written-only' trial, but 'written-only' and 'written+spoken' trials had a probability of occurring immediately after a response trial proportional to their overall frequency.

In-scanner auditory stimuli were presented binaurally through Sensimetrics S15 insert earphones, and visual stimuli were presented with a VPixx PROPixx projector onto a Comar Optics rigid rear projection screen. Experiment scripts are available at https://github.com/thomascope/PINFA_paradigm_scripts/tree/main/fMRI%20task (https://doi.org/10.5281/zenodo.7777386). The task was delivered with Psychtoolbox, running in Matlab 2014a, synchronised with the MRI scanner pulses through a National Instruments PCI 6503 card. Participants indicated responses with their left hand using a bespoke four-button box, interfacing with the same card. After the participants had been positioned in the scanner, we confirmed that they could see the whole screen clearly, hear stimuli in both ears, and respond appropriately using the button box.

Out-of-scanner experiments

Participants performed all study procedures in a stereotyped order. After informed consent, patient participants provided an audio recording of their speech for scoring and assessment. This comprised a description of the 'cookie theft' picture from the Boston Diagnostic Aphasia Examination, followed by a free speech description of their hobbies and interests. All participants then completed handedness, earedness, and general health questionnaires, digit span, revised Addenbrooke's cognitive examination, Boston Diagnostic Aphasia Examination (BDAE) short form naming, and, for controls only, a Wechsler test of adult reading (WTAR).

Approximately forty minutes before the MRI scan, participants undertook the same out-of-scanner primed clarity rating task as described in,[13] based on[10,44] In this task, participants are presented with a written word, followed 1050 (±50) ms later by a spoken word, which is acoustically degraded using a noise vocoder.[68] After a further 1050 (±50) ms, participants are asked to rate the perceptual clarity of the vocoded word. In this way, the perceptual effects of cue congruency and sensory detail can be independently manipulated and assessed.

Next, participants undertook an unprimed vocoded word identification task as described in,[13] except that the number of channels in the noise vocoder was reduced from 4/8/16 to 3/6/15 to match the fMRI experiment. In this task, no prior written text was provided. Participants simply heard a noise vocoded word and, 1050 (±50) ms later, were presented with four written alternatives, from which they selected the word that they had heard. The closeness of the three distractor items to the correct response was manipulated by controlling the number of shared segments between the spoken word and the alternatives. None of the words presented in either of these out-of-scanner tasks was part of the word set presented during fMRI scanning.

Participants then underwent MRI scanning as described above, followed by a lunch break.

Then, participants repeated the unprimed vocoded word identification task, but this time using only the words presented inside the scanner environment. Because all the in-scanner words were in sets of four, there was no manipulation of distractor difficulty – target words shared a vowel with all three distractor words, plus an onset consonant with one and an offset consonant with another.

Next, participants repeated the unprimed vocoded word identification task from Cope *et al.*,[13] and then repeated the clarity rating task from Cope *et al.*,[13] in order to assess the consistency of effects across the experimental session.

After this, participants had a short break, then completed Raven's progressive matrices, followed by a pure-tone audiogram.

Finally, for controls only, there was one more behavioural test. To enable MVPA analysis of prediction errors, there was a consistent relationship between written and spoken words in the mismatch case. For example, reading the word 'bard' in a written+spoken trial meant that the subject would always subsequently hear 'bard' (a matching trial), or 'debt' (a mismatching trial). It simplifies the interpretation of our results if participants did not learn this relationship. To assess this, participants were presented with a written word, then asked what they might hear next, if they were not to hear the same word they had read. Four alternatives were presented, none of which was the same as the written word, and one of which was the word representing the consistent mismatch from the fMRI environment. Universally, control participants reported that they did not know which answer was correct. We gave them the instruction to: "Guess, according to your 'gut feeling' of which word was most likely, because in the scanner the words were not presented randomly." Performance for this task was at chance across the control cohort (mean 23.9%, standard error of the mean 1.59%, chance performance 25%).

All tasks were administered on a Dell XPS 15 laptop in a quiet clinic room, with sounds presented through Sennheiser HD250 linear 2 headphones, driven by a Behringer UCA 202 external sound card. Participants indicated responses either by pressing a number on a keyboard (clarity rating task) or a button on a custom made response box (all other tasks).

Behavioural clarity rating and word report data from before and after 7T imaging were modelled using hierarchical Bayesian inference simulations previously described,[13] based on,[69] code available at: https://github.com/thomascope/7T_pilot_analysis/blob/master/module_bayesian_behaviour.m (https://doi.org/10.5281/zenodo.7777380). The only modification to this procedure was a scaling to account for the difference in the ratio of vocoder channel numbers between the out-of-scanner word identification experiments presented here (3/6/15) and those described previously (4/8/16). Based on unprimed word identification performance across all participants, we applied a scaling of 1.09x to the modelled sensory detail of the intermediate, 6-channel, condition to better model expected clarity ratings.

Structural MRI preprocessing

Code for the MRI analysis pipeline is available at: https://github.com/thomascope/7T_pilot_analysis/blob/master/batch_7T_preprocess.m (https://doi.org/10.5281/zenodo.7777380)

First, we created a mask for skull stripping by performing segmentation in SPM12, based on both the 'unified' and 'second inversion' from the MP2RAGE sequence, with custom bias egularization of '0.00001' and bias FWHM of '30'. Providing both of these

images improves segmentation, as the unified image has superior grey/white matter contrast, while the second inversion has better grey/CSF contrast. The mask was created by adding together the white matter, grey matter and CSF compartments, followed by the application of the Matlab 'imfill' function to ensure a contiguous brain mask. This mask was then applied to produce a skull stripped brain, which was AC-PC aligned to an elderly template brain in SPM12.

For quantitative structural data analyses we used Freesurfer 7.1.0 to assess cortical thickness. The aligned and skullstripped brain images were provided to the 'autorecon1' module of Freesurfer 'recon-all' with the '-noskullstrip', '-hires', 'notal-check', '-cw256', and '-bigventricles' flags. Next, we copied the T1.mgz image to brainmask.auto.mgz and brainmask.mgz, before submitting the 'autorecon2' and 'autorecon3' stages with the same flags. Every cortical segmentation was manually visually checked for quality, and confirmed to be of good quality in all individuals in our regions of interest across superior temporal, frontal, parietal, and occipital language regions.

Functional MRI pre-processing

First, we realigned the EPI from all runs in SPM12, including those with reversed phase encoding, to match the first in the time series using 5th degree B-spline interpolation. Next, we applied distortion correction to all runs using '*topup*' from FSL version 5.0.3. This is a two-stage process. In the first stage, distortions are calculated by comparing EPI with opposite phase encoding directions; here we used the reference image taken at the start of the third run (i.e. half way through the fMRI experiment) and a reverse-phase reference image taken immediately beforehand. In the second stage, corrections against these distortions are applied to all images. The resulting images were then co-registered to the native-space structural MP2RAGE.

To create the input images for our first-level (single subject) native-space multivariate (MVPA) analyses, these re-aligned, undistorted, functional time-series were then resliced into the same space in SPM12, before smoothing with a 3mm FWHM kernel (twice the voxel size).

Next we obtained deformation fields to IXI template space for every individual by performing normalisation of the structural MP2RAGE using the CAT12 toolbox.[70] As our multivariate analyses were performed in native space but regions of interest were defined from group-level contrasts in template space, this was a particularly crucial step to allow the accurate transformation of data between spaces.

To create the input images for our first-level template-space univariate and connectivity analyses, we applied these deformation fields to the re-aligned, undistorted, unsmoothed functional images using the SPM12 normalise write function with 4th degree B-spline interpolation and an output voxel size of 1.5mm isotropic. These images were then smoothed with a 3mm FWHM kernel.

Every participant performed four blocks of the task, however for three individuals (one patient, two controls), one block failed visual data quality checks due to motion-related 'ghosting' artefact and was discarded. Two of the discarded blocks were the final block, and one was the penultimate block.

fMRI Physio-physiological interaction

To assess how connectivity between brain regions supported the construction of neuronal representations we assessed the physio-physiological interaction from the first-level univariate design matrix described above, using the gPPI toolbox.[71] Note that this approach is related to, but differs from, the similarly-named psycho-physiological interaction analysis, which can be used to assess the effect of task on connectivity in block-design fMRI paradigms, but is not suitable for randomly presented short events as in our experiment here. The physio-physiological interaction can be conceptualised as an extension of the seed-based connectivity approaches commonly employed in resting-state fMRI analyses. It relies on the correlation of regional time-series after regressing out univariate task-related effects, and as such is a non-directional assessment of the connectivity of two regions to the rest of the brain.

We were specifically interested in the relative connectivity of PrG and STG during the instantiation, reconciliation, and refinement of predictions. This tests which brain regions were preferentially connected to PrG, which were preferentially connected to STG, and which were engaged by the negative interaction of the activity in these regions. The negative interaction is of interest, because the strongest manipulation of the magnitude of the time-series is the auditory presentation of the spoken word, meaning that this interaction is most sensitive to connectivity during the instantiation of prediction while reading, when STG activity is low.

Illustrative computational modelling

We constructed a computational model of hierarchical predictive processing, based on that published in.[37] In our experimental context, the prediction is the written word and the sensory input is the spoken word, which was manipulated in sensory detail by the application of a 3 or 15 channel noise vocoder. The model uses pixel-based synthetic representations of words, with the inputs filtered according to sensory detail, added to uniformly distributed noise, and scaled to sum to one. Prediction error was calculated from a subtraction of the prediction from the input. Signal magnitudes for sensory input and prediction error were quantified as the absolute difference between the observed pixel value and the mean pixel value, summed across all voxels, and thus quantified the combined magnitude of both 'positive' and 'negative' prediction errors. Signal information was quantified as the squared Pearson correlation (i.e. r^2) between the overall pattern and an undistorted representation of the speech input. We propose that a refinement signal based on prediction errors would be information-weighted; a large error with low information would provide little basis for specific learning beyond an overall weakening of existing associations, while a small but precise error may be highly instructive.[16,17]

While this information weighting is conceptually similar to precision weighting.[16] it quantifies the informativeness of the prediction error in a single trial, rather than the long-run uncertainty in the environment. The model creates the information-weighted prediction error by multiplying the prediction error in each condition by its representational information. More complex mathematical relationships beyond simple multiplication may underlie real neuronal processes, but would not change the overall pattern for low vs high sensory detail contrasts in matching vs fully mismatching predictive contexts. Finally, we can model the univariate and multivariate representations in a region responsible for the refinement of predictions by adding the information-weighted prediction error to the sensory input from the spoken word. Thus, we assume that BOLD signals reflect the combination of two distinct speech representations (information-weighted prediction error and sensory input), possibly encoded by different neural populations. Model sensory input signal magnitudes were up to 0.75 arbitrary units, while information-weighted prediction error representations had magnitudes of <=0.2 arbitrary units. We propose these signals to be roughly equally scaled for prediction refinement, and therefore performed this combination by adding the information-weighted prediction error to the sensory input divided by four. Changing this multiplier would not change the qualitative nature of the modelled interactions, merely their scaling.

Data visualisation

In this paper we use *Surf Ice* (https://www.nitrc.org/projects/surfice/) to represent non-quantitative volumetric regions of interest on the BrainNet BrainMesh_ICBM152.lh surface template. We also use it to represent our surface-based structural analysis on the Freesurfer lh.pial template, where it can display quantitative data accurately.

To display the quantitative fMRI results in a visually comparable way, we developed a Matlab script to display volumetric data on the same cortical mesh; available at: https://github.com/thomascope/7T_pilot_analysis/blob/master/Thresholded_multivariate/jp_spm8_surfacerender2_version_tc.m (https://doi.org/10.5281/zenodo.7777380). Because the resolution of 7T fMRI exceeds the vertex spacing of cortical meshes, the code iterates through every vertex of the mesh, and finds the highest value voxel within a 3mm sampling radius, for projection to that location. This method resembles *Surf Ice* when rendering the same data, but crucially preserves data range (Figure S7).

QUANTIFICATION AND STATISTICAL ANALYSIS

Structural MRI analysis

Between-group whole-brain comparisons were implemented in Freesurfer FSGLM using a 'DOSS' approach with Age as a covariate of no-interest and a 10mm cortical smoothing kernel. The GLM model was evaluated first for raw significance, and then with cluster-wise statistics based on 10,000 simulations per-hemisphere. Clusters were retained at corrected $p<0.05$. Our analyses focus on the left hemisphere - the right hemisphere results are, however, shown for illustrative purposes to demonstrate that nfvPPA is a relatively asymmetric but not strictly unilateral disease.

We then extracted the single subject cortical thickness in regions of interest defined by the Desikan-Killiany Atlas. Our regions of interest were IFG pars opercularis and triangularis, PrG, and STG planum temporale ('banks of superior temporal sulcus') and primary auditory cortex ('transverse temporal gyrus'). In each region, we tested for between-group differences both with a traditional, frequentist two-sample t-test, and with a Bayesian t-test using the bayesFactor Matlab toolbox (https://klabhub.github.io/bayesFactor/). This allowed us to test the strength of evidence both for and against atrophy, by reporting the Bayes Factor, BF10, and its inverse, BF_null.

Results, subject numbers and definitions are found in Figure 2A legend and supplementary table. Data met the assumptions of the statistical analysis approaches.

Functional MRI univariate analysis

A first-level design matrix was created in SPM12 from the behavioural task. A single event was specified at the onset of the spoken word (i.e. 700ms after the onset of the written word) in every trial, as this is the timepoint at which it is possible to begin to reconcile predictions and sensory evidence. For written-only trials there was no spoken word, but the event was still specified at the time it would have occurred to model the absence of spoken input. Events were categorised into six types – 'Match 3', 'Match 15', 'Mismatch 3', 'Mismatch 15', 'Written-only' and 'Response', with the latter used only for regressing out effects of no interest.

The design matrix included each event type as a separate column, along with six head motion regressors of no interest, replicated across all blocks. Because of the increased signal homogeneity in 7T data compared to 3T, the implicit masking threshold was reduced from the default of 0.8 to a more inclusive threshold of 0.3. To ensure analyses were restricted to brain, an explicit grey-matter mask was applied at the second level, constructed from an 80% majority consensus of the control participants' SPM c1 segmentation at a 5% threshold.

Contrasts were evaluated for written+spoken trials at the single subject first-level and then group second-level for the main effects of congruency and sensory detail, and the interaction between the two. An additional contrast was specified for all written+spoken trials against written-only trials (i.e. to assess the neural effect of the presence versus absence of the spoken word). Response trials were not analysed in these contrasts. Age was included as a covariate of no interest in all analyses.

Results, subject numbers and definitions are found in Figure 2 legend and Table 1A. Data met the assumptions of the statistical analysis approaches.

Functional MRI multivariate analysis

For the multivariate analysis, a more complex first-level design matrix was created in SPM12 from the behavioural task. As before, single events were specified at the onset of the spoken word (i.e. 700ms after the onset of the written word), but now each of the six trial types was broken down into 16 separate spoken words, resulting in 96 event types, each occurring twice per block (eight times in total), except the response trial words, which only occurred once per block (four times in total). An additional event was specified for the button press. Again, the design matrix included each event type as a separate column, along with six head motion regressors of no interest, replicated across all blocks.

The native-space output of this first-level model was then subjected to whole-brain searchlight and ROI-based multivariate representational similarity analysis using the decoding toolbox.[72] Distance measures were calculated using the cross-validated Mahalanobis distance,[73,74] which can be conceptualised as a cross-validated Euclidean distance accounting for noise covariance, and is an unbiased metric with interpretable zero. Distance was evaluated from every written+spoken event to every other, resulting in a 64x64 dissimilarity matrix (16 words across 4 conditions) in each ROI or searchlight location. Searchlight analysis employed an 8mm-radius sphere, evaluated with a cent at every voxel in the native-space image.

Each dissimilarity matrix was then compared with Spearman correlations against several candidate representational dissimilarity matrices designed to assess specific hypotheses, as described in the results section. For searchlight analysis, this resulted in whole-brain native space images of representational similarity, which were then deformed into template space and resliced to 1mm isotropic voxels. These standardised maps were evaluated at the second level, with age as a covariate of no interest.

Region of interest analysis was also performed, with ROIs determined from orthogonal contrasts with patients and controls weighted equally to avoid double-dipping.

Results, subject numbers and definitions are found in Figures 3 and 4 legends, Tables 1B and 1C, and Figures S4 and S5. Data met the assumptions of the statistical analysis approaches.

fMRI physio-physiological interaction analysis

We applied the gPPI toolbox to the first-level univariate design matrix, which convolves task events with the haemodynamic response function. It then adds regressor columns of the time-course of the activity in two seed regions of interest, and the interaction between them. Here, seeds were based on the second level univariate effects across all subjects, cluster-thresholded at $p<0.05$: the contrast for sensory detail that preferentially engaged STG (greater activity for 15 channel compared to 3 channel vocoded speech) and the interaction between sensory detail and congruency that delineated a cluster in PrG.

The first-level single-subject contrasts for STG > PrG connectivity, PrG > STG connectivity, and the negative physio-physiological interaction were evaluated at the second (group) level, with age as a covariate of no interest.

Results, subject numbers and definitions are found in Figure 5 legend and Table 1D. Data met the assumptions of the statistical analysis approaches.