PLOS | BIOLOGY

RESEARCH ARTICLE

Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception

Helen Blank*, **Matthew H. Davis**

MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom

* Helen.Blank@mrc-cbu.cam.ac.uk

OPEN ACCESS

Citation: Blank H, Davis MH (2016) Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS Biol 14(11): e1002577. doi:10.1371/journal.pbio.1002577

Academic Editor: Robert Zatorre, McGill University, CANADA

Received: May 10, 2016

Accepted: October 19, 2016

Published: November 15, 2016

Data Availability Statement: The data is available at https://osf.io/2ze9n/ doi: 10.17605/OSF.IO/2ZE9N

Funding: This research was supported by UK Medical Research Council (MRC) funding of the Cognition and Brain Sciences Unit MC-A060-5PQ80 (to MHD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Successful perception depends on combining sensory input with prior knowledge. However, the underlying mechanism by which these two sources of information are combined is unknown. In speech perception, as in other domains, two functionally distinct coding schemes have been proposed for how expectations influence representation of sensory evidence. Traditional models suggest that expected features of the speech input are enhanced or sharpened via interactive activation (Sharpened Signals). Conversely, Predictive Coding suggests that expected features are suppressed so that unexpected features of the speech input (Prediction Errors) are processed further. The present work is aimed at distinguishing between these two accounts of how prior knowledge influences speech perception. By combining behavioural, univariate, and multivariate fMRI measures of how sensory detail and prior expectations influence speech perception with computational modelling, we provide evidence in favour of Prediction Error computations. Increased sensory detail and informative expectations have additive behavioural and univariate neural effects because they both improve the accuracy of word report and reduce the BOLD signal in lateral temporal lobe regions. However, sensory detail and informative expectations have interacting effects on speech representations shown by multivariate fMRI in the posterior superior temporal sulcus. When prior knowledge was absent, increased sensory detail enhanced the amount of speech information measured in superior temporal multivoxel patterns, but with informative expectations, increased sensory detail reduced the amount of measured information. Computational simulations of Sharpened Signals and Prediction Errors during speech perception could both explain these behavioural and univariate fMRI observations. However, the multivariate fMRI observations were uniquely simulated by a Prediction Error and not a Sharpened Signal model. The interaction between prior expectation and sensory detail provides evidence for a Predictive Coding account of speech perception. Our work establishes methods that can be used to distinguish representations of Prediction Error and Sharpened Signals in other perceptual domains.

Abbreviations: fMRI, functional magnetic resonance imaging; IFG, inferior frontal gyrus; RDM, representational dissimilarity matrix; ROI, region of interest; RSA, representational similarity analysis; STG, superior temporal gyrus; STS, superior temporal sulcus.

Author Summary

Perception inevitably depends on combining sensory input with prior expectations. This is particularly critical for identifying degraded input. However, the underlying neural mechanism by which expectations influence sensory processing is unclear. Predictive Coding theories suggest that the brain passes forward the unexpected part of the sensory input while expected properties are suppressed (i.e., Prediction Error). However, evidence to rule out the opposite mechanism in which the expected part of the sensory input is enhanced or sharpened (i.e., Sharpening) has been lacking. In this study, we investigate the neural mechanisms by which sensory clarity and prior knowledge influence the perception of degraded speech. A univariate measure of brain activity obtained from functional magnetic resonance imaging (fMRI) is in line with both neural mechanisms (Prediction Error and Sharpening). However, combining multivariate fMRI measures with computational simulations allows us to determine the underlying mechanism. Our key finding was an interaction between sensory input and prior expectations: for unexpected speech, increasing speech clarity increases the amount of information represented in sensory brain areas. In contrast, for speech that matches prior expectations, increasing speech clarity reduces the amount of this information. Our observations are uniquely simulated by a model of speech perception that includes Prediction Errors.

Introduction

The observation that our perception of the world not only depends on sensory input but also on our prior knowledge has been of longstanding interest in psychology [1] and neuroscience [2–5]. There is widespread agreement that sensory input and prior knowledge are combined in neural representations; by which we mean the specific patterns of neural activity that are associated with the content of our sensory experiences. However, despite extensive experimental work in many sensory modalities [6–16], the neural and computational mechanisms by which prior knowledge guides perception are unclear [17,18].

One proposal is that neural representations of expected sensory signals are enhanced or tuned [19,20]. Critically, in this account, perceptual representations are sharpened by relevant prior expectations in much the same way as if the quality of the sensory input was increased [17,18]. Alternatively, Predictive Coding schemes suggest that expected sensory input is explained away and unexpected information is represented in the form of prediction errors (cf. in engineering [21,22] and neuroscience [3,23,24]). One intuitively attractive aspect of Predictive Coding, both for engineering and neuroscience, is its assumption that minimal effort should be invested in further processing of sensory information that is already known or expected.

Our goal in this work is to distinguish these two fundamental coding schemes for how prior expectations influence perception. Do neural representations of sensory signals contain only the unexpected parts of the sensory evidence (from now on we will refer to these as "Prediction Errors")? Or do they contain an enhanced version of the expected sensory evidence (from now on "Sharpened Signals")? Our approach allows us to test each of these coding schemes against behavioural and fMRI data to determine how expected sensory signals are neurally coded.

Sharpening and Predictive Coding schemes have proved hard to distinguish in neuroscience [2,5,25]. Predictive Coding theories have proposed that each level of a cortical hierarchy contains two functionally distinct subpopulations (i.e., prediction and prediction error units [3,20,24,26]). In these accounts, the signals that are passed forward from one level of the hierarchy to the next (i.e., the feedforward signals) represent Prediction Error. This Prediction Error

signal is also used to update prediction units within the same level of the cortical hierarchy (through lateral interactions), such that prediction units represent a sharpened version of the sensory signal [3]. Therefore, evidence for Sharpened Signal representations has been used to support both Predictive Coding theories [20] as well as pure Sharpening theories without computation of Prediction Errors [27]. However, evidence for Prediction Error representations would be uniquely consistent with Predictive Coding and challenge pure Sharpening accounts.

Speech perception provides a biologically significant domain in which prior knowledge has been clearly shown to guide perception (for review, see [28]). Behavioural experiments show that numerous sources of proximal and distal prior knowledge (including subtitles, lip-reading, lexical constraint, or semantic predictability) can enhance subjective and objective perceptual outcomes for degraded speech [29–33]. The dominant computational theories of speech perception have included interactive-activation mechanisms that lead to enhanced representations of expected signals (i.e., Sharpened Signals), most notably in the TRACE model [34] but also in other influential models of speech perception [35–38]. More recent work has proposed Predictive Coding schemes, which use Prediction Error signals [4,7,39] to explain how prior expectations improve sensory processing. However, evidence to overturn Sharpening accounts has been lacking.

One challenge for existing research is that both suggested computational schemes predict reduced neural activity during perception of expected speech signals, either due to suppression of unexpected noise (in Sharpened Signals) or suppression of expected signals (in Prediction Errors). Brain regions in and around the left posterior superior temporal sulcus (STS) are proposed to support perceptual processing of speech [40,41] and integrate expectations from different modalities with speech input [8,39,42–46], and activity in this region is proposed to show effects of prior training on speech responses [47–49]. While these studies provide abundant evidence that prior knowledge can influence the magnitude of activity in the posterior STS during speech perception, they do not determine the computational mechanism by which relevant prior knowledge enhances perception of speech.

However, multivariate analyses of the representational content of brain responses can differentiate these two accounts by testing whether representations of speech signals are enhanced (in line with Sharpened Signals) or suppressed (Prediction Errors) when they match prior expectations. Therefore, we used representational similarity analysis [50] on multivoxel response patterns in the posterior STS. This approach is "information based" because it measures how much information about the phonetic form of speech is contained in spatial fMRI activation patterns in each of the experimental conditions that we tested [51,52]. We focus on the posterior STS because this is both a region in which effects of prior knowledge on speech processing have been repeatedly shown and also a region in which syllable identity can be decoded from multivariate BOLD signals [53–57].

To guide our interpretation of this data, we constructed two computational simulations based on either Sharpened Signals or Prediction Errors. Both these simulations can explain observations of perceptual enhancement and reduced fMRI responses in the left posterior STS for degraded speech that matches prior expectations. Crucially, however, these simulations make distinct predictions for the results of multivariate representational similarity analysis. In our Sharpened Signal model, simulated neural representations are enhanced for degraded speech that matches prior expectations in the same way as for speech that is presented with more sensory detail (Fig 1A). However, in our Prediction Error model (Fig 1B), these two manipulations have an interactive effect on simulated neural representations: the effect of increasing sensory detail depends on whether or not speech matches prior expectations. Increased sensory detail for expected speech leads to reduced information about the phonetic form of speech in simulated Prediction Errors. In contrast, increased sensory detail for unexpected speech leads to more Prediction Error and, hence, more information in simulated neural

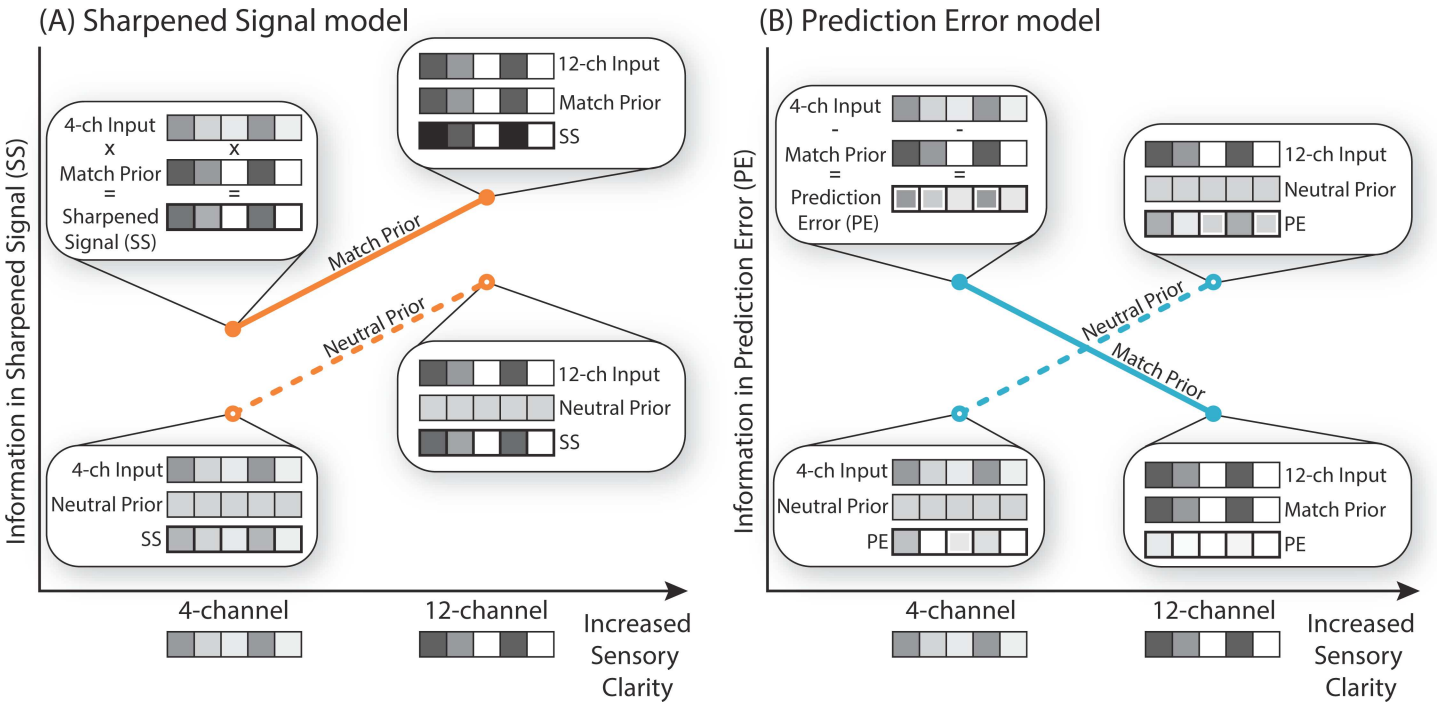


Fig 1. Two computational models for how matching or neutral prior expectations influence processing of sensory signals at different levels of clarity: **(A)** Sharpened Signal model and **(B)** Prediction Error model. For both accounts, neural representations are derived by combining the sensory input with prior expectation. However, the underlying computations and information content in neural representations differ. **(A)** Sharpened Signal model: Prior expectation is used to multiply sensory input, leading to more specific representations for expected compared to unexpected sensory input (Sharpened Signals, SS). This leads to additive effects of sensory detail and matching prior expectation on the information content of neural representations. **(B)** Prediction Error model: Prior expectation is subtracted from the sensory input such that neural representations encode the difference between expected and actual input (Prediction Error, PE). This leads to an interaction between sensory detail and prior expectations, with most informative neural representations found when clearer signals follow neutral expectations, or when degraded signals match informative prior expectations. Critically, when clear signals match informative prior expectations, this produces a small and uninformative Prediction Error (Match 12-channel condition). The information content of neural representations (*y*-axis) contained in SS (A) and in PE (B) refers to the signal that is passed forward after the input and prior have been combined (bottom bars). This allows us to test which of these neural representations best describes measured fMRI pattern information. In each model, neural activity patterns are represented by greyscale values over sets of units. Negative Prediction Error values are shown with a white outline.

doi:10.1371/journal.pbio.1002577.g001

representations. In our experimental work, we test both these proposals using representational similarity analysis (RSA) fMRI applied to BOLD responses time-locked to the onset of a degraded spoken word.

To obtain experimental evidence to differentiate these two computational accounts, we therefore simultaneously manipulated (1) prior knowledge of speech content by having participants read matching/mismatching written words or neutral text ("XXXX") before spoken words [8,33,58] and (2) sensory detail in speech by presenting vocoded spoken words at one of two different levels of acoustic degradation (Fig 2) [59,60]. In this way, we could test whether representations of the phonetic form of speech in the posterior STS [55,57,61] are enhanced similarly by changes in prior knowledge as by changes to sensory detail (in line with Sharpened Signals) or whether these two factors interact (in line with Prediction Errors).

Results

Behavioural Results

First, we confirmed that, consistent with both Predictive Coding and Sharpening, providing informative prior expectations improves perception of degraded speech. Participants' report of

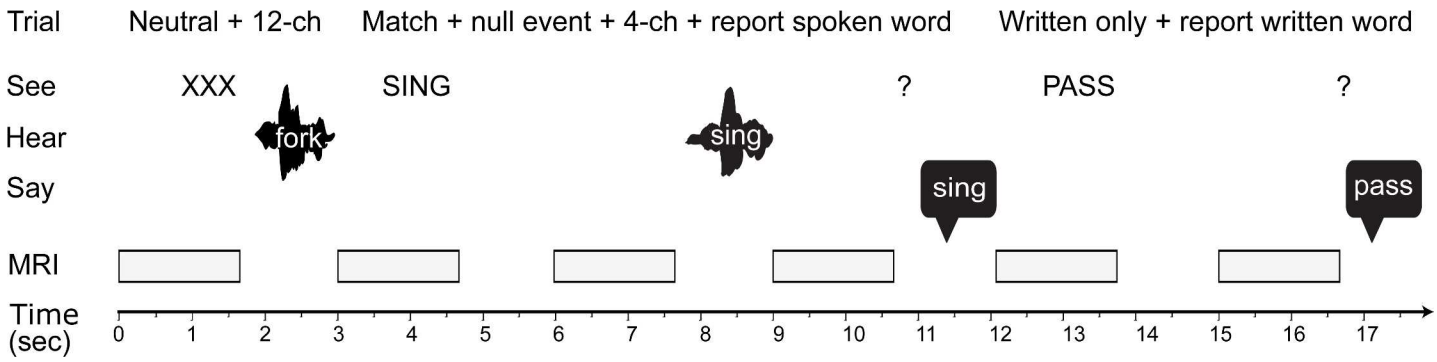


Fig 2. Design and experimental conditions. We used sparse imaging to record fMRI responses while participants see written words, hear subsequently presented degraded spoken words, and say what word they heard or read previously. We used two levels of sensory detail (4- and 12-channel) for presentation of the spoken words and conditions containing different pairings of written and spoken words: (1) matching written text + spoken words ("SING" + sing); (2) neutral written text ("XXXX") + spoken words (e.g., fork); and (3) written-only text ("PASS"). Following 1/6 of all trials, participants were cued with a question mark to say aloud the previous written or spoken word. In addition, we inserted fixation crosses, null events, and trials in which written text partially or totally mismatched with spoken words (see Materials and Methods for details).

doi:10.1371/journal.pbio.1002577.g002

the degraded spoken words was improved by both increased sensory detail and matching prior information from a preceding written word (Fig 3). A two-way repeated measures ANOVA with the factors sensory detail (4- versus 12-channel) and prior knowledge (Match versus Neutral) revealed significant main effects of sensory detail on word report (12-channel: 85.39% > 4-channel: 57.83% correct; $F(1, 20) = 133.419$, $p < 0.001$, eta squared = 86.96) and prior knowledge (Match: 84.42% > Neutral: 63.49% correct; $F(1, 20) = 89.582$, $p < 0.001$, eta squared = 81.75), and a significant interaction ($F(1, 20) = 74.997$, $p < 0.001$, Fig 3A).

These effects of sensory detail and prior knowledge combined such that 4-channel vocoded speech in the Match condition was reported with equivalent accuracy as 12-channel vocoded speech in the Neutral condition (79.17% versus 83.53% correct, $t(20) = -1.427$, $p = 0.169$). Nonetheless, word report was further enhanced in the Match 12-channel condition compared to the Neutral 12-channel condition (89.68% versus 83.53% correct, $t(20) = 3.267$, $p = 0.004$) and the Match 4-channel condition (89.68% versus 79.17% correct, $t(20) = -4.460$, $p < 0.001$). Word report in the Match 12-channel condition was also more accurate than in a condition in which the spoken word was omitted and participants were prompted to report the preceding written word (89.68 versus 82.14% correct in the written only condition, $t(20) = 2.348$, $p = 0.029$). These findings confirmed that participants used prior knowledge to enhance perception of degraded speech even when relatively clear 12-channel speech was presented. Behavioural responses in the Mismatch conditions resemble the pattern of results in the Neutral condition (see S1 Fig).

Univariate fMRI Results

Second, we sought to localise the univariate BOLD activity decrease for degraded spoken words that follow matching written words relative to words following neutral cues. These observations replicate previous findings but do not distinguish between accounts in which this effect is due to suppression of unexpected noise (Sharpened Signals) or suppression of expected signals (Prediction Errors). Univariate BOLD responses were influenced by both increased sensory detail and matching written text. A two-way repeated measures ANOVA with the factors sensory detail (4- versus 12-channel) and prior knowledge (Match versus Neutral) revealed a main effect of matching versus neutral prior knowledge on responses in the left posterior STS, as predicted, and in other regions of the speech processing network (Fig 3B and 3C, S1 Table:

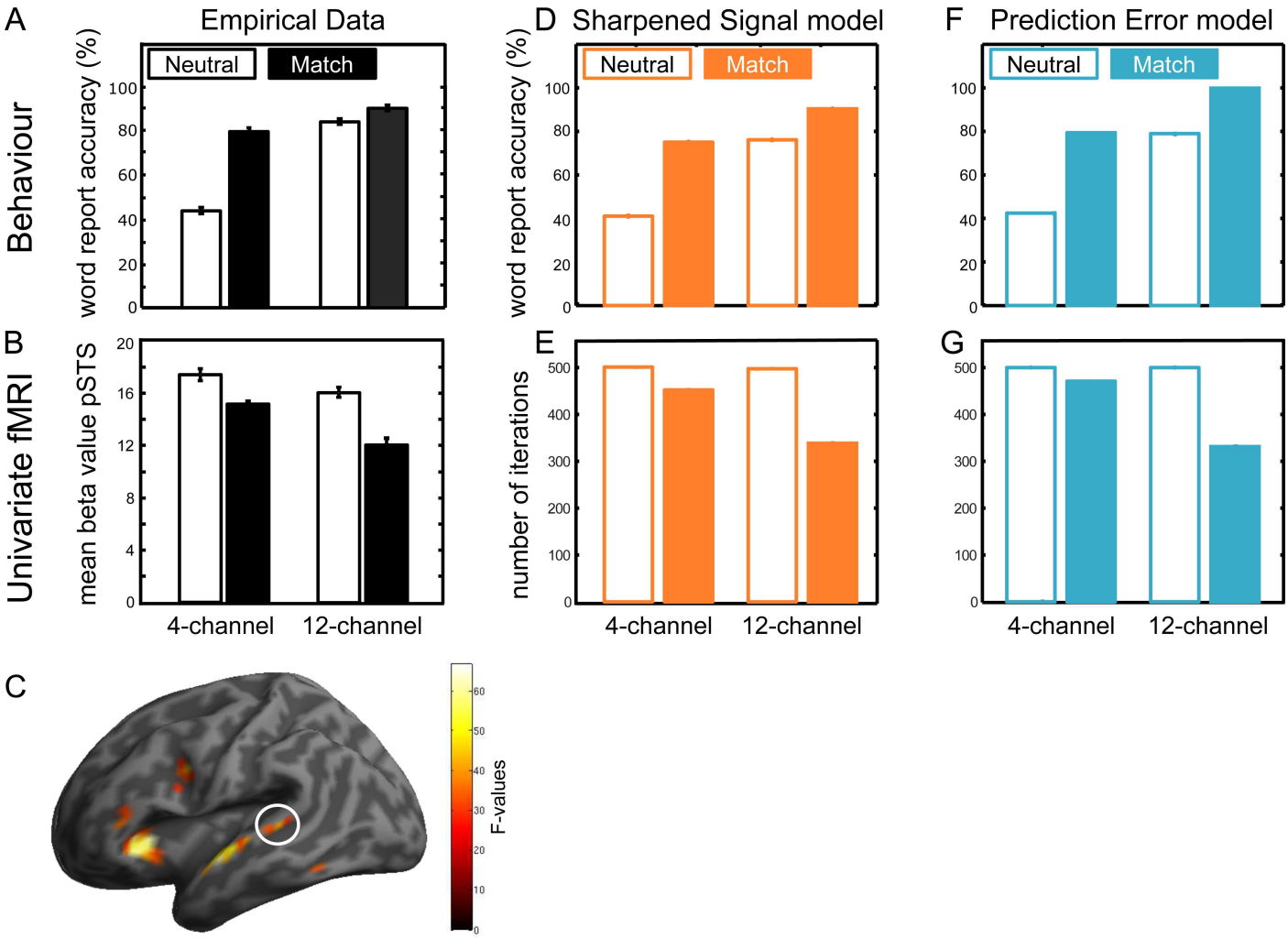


Fig 3. Comparison of behavioural and univariate fMRI results with model output. (A) Behavioural results. Matching expectations and increased sensory detail improved perception of degraded spoken words. **(B)** Univariate results. Mean beta values extracted from the posterior STS (pSTS, MNI: x = -52, y = -38, z = 6) show reduced BOLD signal during Match conditions (solid) in contrast to Neutral conditions (open). Error bars for the empirical data indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons [62]. **(C)** Main effect of prior expectations rendered on a canonical brain ($p < 0.05$ voxelwise FWE, n = 21). White circle marks the region of interest in the posterior STS. **(D/E)** Sharpened Signal model (orange) and **(F/G)** Prediction Error model (blue). For comparison with the behavioural results (**D**/**F**) we assessed word recognition accuracy in the model based on the final lexical representation (i.e., which word the model selected as presented), and for comparison with the univariate results (**E/G**) we assessed the number of activation updates required to reach the stopping criterion. Error bars for both simulations indicate the standard error of the mean over 1,000 replications. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

doi:10.1371/journal.pbio.1002577.g003

main effect of Match/Neutral, $p < 0.05$ FWE voxel correction). Mean beta values extracted from the left posterior STS showed a reduction during Match in contrast to Neutral conditions (Fig 3; inspection of contrast estimates from all other clusters also revealed less activity for Match than Neutral). In addition, there was a main effect of sensory detail in bilateral insula, SMA, left premotor, and orbitofrontal cortex (S2 Table; main effect of 4/12-channel, $p < 0.05$ FWE). Inspection of contrast estimates revealed less activity for 12- than 4-channel in most clusters; the reverse pattern was only observed in the right middle orbitofrontal gyrus). The interaction of prior knowledge and sensory detail did not reach corrected significance (S3 Table).

Increased BOLD activity for Mismatch > Match resembles the difference in BOLD activity found for Neutral > Match (see S1 Fig and S4 Table). This confirms that our observed effects are not due to differences in attention, anticipation of more difficult trials, or baseline differences between the Match and Neutral conditions (see S1 Text), but rather due to the influence of matching prior knowledge on speech perception.

Model Simulation of Behavioural and Univariate fMRI Results

The behavioural and univariate results appear to be in line with both Sharpening and Predictive Coding theories. Although the underlying coding schemes differ, both accounts suggest that increased sensory detail and matching prior information should improve recognition performance and that prior matching knowledge should reduce univariate fMRI responses. To confirm this, we constructed two computational models of spoken word recognition, which only differed by using representations of Sharpened Signals or Prediction Errors to simulate how sensory information and prior knowledge are combined (see S2 Fig for details). In both these models, behavioural performance (i.e., word recognition) was simulated by the model's ability to identify the correct word presented in degraded speech, and univariate fMRI results (i.e., the magnitude of hemodynamic activity in the left posterior STS) were simulated by the number of processing iterations required for the model to settle. By simulating the univariate fMRI signal with the number of model iterations, we assume that the hemodynamic signal as measured by fMRI integrates over several seconds of neural activity and that a longer duration of neural processing should result in an increased amplitude of the fMRI signal [63]. Six parameters were optimised for each model: the amount of sensory degradation used to simulate 4- and 12-channel vocoded speech (which influences word report and processing time), variability and confidence in behavioural responses (which influences word report), and the rate and duration of model updating (which primarily influences processing time; see S3 Fig for sensitivity analysis of the optimized parameters).

We used Akaike weights to compare goodness of fit to word report and univariate hemodynamic responses in the left posterior STS (see Materials and Methods for details). Based on 1,000 replications using the best-fitting set of parameters, a probability density function for the predicted outcome of behavioural and univariate results was generated for both model simulations. We then used the evidence ratio of Akaike weights to compare the relative likelihood of the two models given the observed data. The ratio of the Akaike weights revealed a slightly higher likelihood of Sharpened Signal model than of the Prediction Error model for both the behavioural results ($w_{PE}/w_{Sharp} = 0.9307$) and the univariate results ($w_{PE}/w_{Sharp} = 0.8149$). Both of these values are close to 1, indicating that there is a negligible difference between the two models [64]. The good fit observed between these models and behavioural and univariate hemodynamic data from the current experiment suggests that computation of Sharpened Signals and Prediction Errors can explain the effect of increased sensory detail and matching prior information during perception of degraded words (model simulations and experimental results shown in Fig 3).

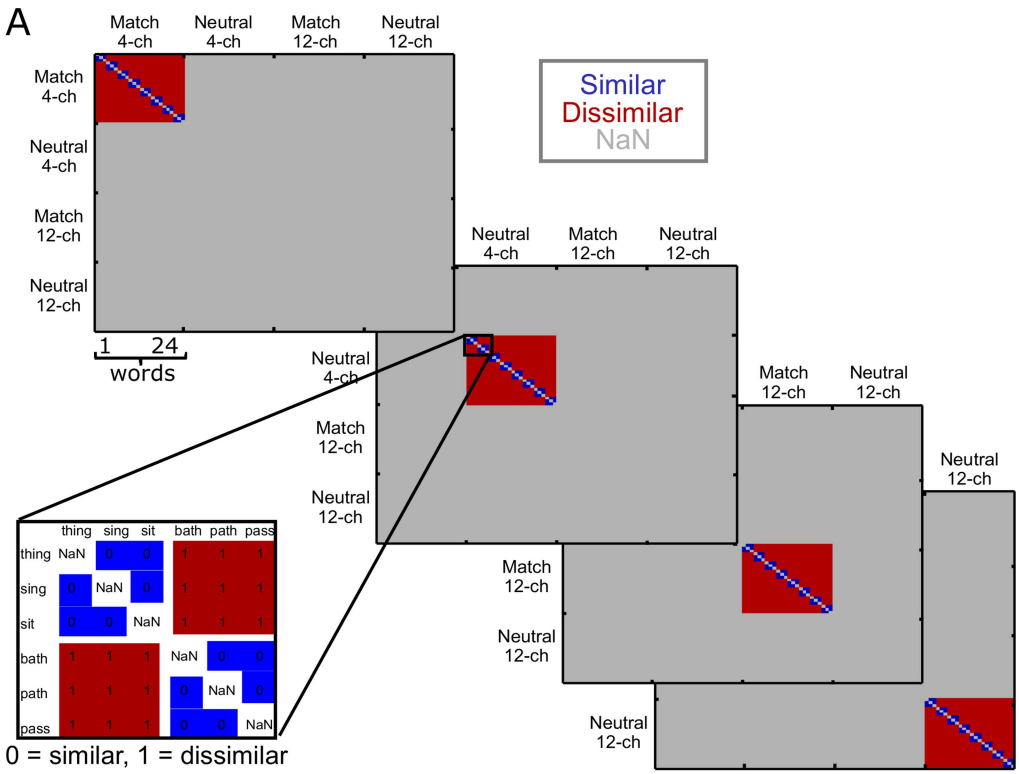
Multivariate fMRI Results

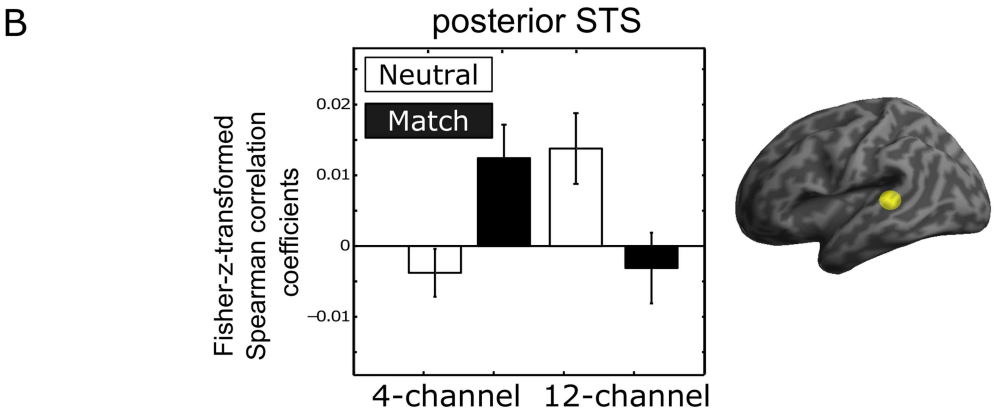
Although both models can accurately simulate behavioural and univariate fMRI results, they perform different underlying computations and make different assumptions about the effect of matching prior knowledge on neural representations of speech signals. The Sharpened Signal model predicts that degraded speech is better represented in the STS when it matches prior knowledge, because expected sensory features of the speech input are enhanced and unexpected sensory features are suppressed. In contrast, the Prediction Error model assumes that

the expected part of the speech input is explained away (i.e., reduced) and only Prediction Errors (i.e., the difference between heard and expected speech) are represented in the STS. To test these two simulations, we assessed the neural representation of speech information by means of RSA [50]. This approach allowed us to quantify the amount of information about the phonetic form of speech that is carried by the spatial pattern of fMRI activity in each of our four critical conditions.

We designed our experiment to test for categorical representations of syllable similarity, because previous studies (in fMRI [55,57] and intracranial recordings [61]) showed that categorical representations of speech, such as vowels and syllables rather than acoustic cues, are decodable from the STS. Neural representational similarity was first measured by computing a representational dissimilarity matrix (RDM) for multivoxel fMRI responses for each item and condition (see Materials and Methods for details). To quantify the amount of speech information, we computed the Fisher-z-transformed Spearman correlation between the observed RDM and a hypothesised RDM of interest that tested for increased similarity between pairs of syllables that shared the same vowel and had other segments in common (e.g., "sing" and "thing") compared to pairs of unrelated words (e.g., "sing" and "bath", see Fig 4A). This similarity measure was computed separately for each condition. This analysis targets speech representations in the posterior STS by testing for similarity of words that have similar phonetic forms but different lexical or semantic representations. We did not compare identical words presented in different scanning sessions.

Regions of interest (ROIs) analysis. Fisher-z-transformed correlation coefficients for searchlight locations were computed for two left posterior STS ROIs. The first of these was based on a 6-mm sphere centred on a coordinate defined by multivariate syllable coding in independent data in the left posterior STS (MNI: x = -57, y = -39, z = 8, [57]). The second ROI was defined by the univariate analysis of the present data (centre of mass MNI: x = -56, y = -35, z = 6). Mean correlation coefficients for these ROIs were entered into a repeated measures ANOVA with factors sensory detail (4- versus 12-channel) and prior knowledge (Match versus Neutral). This showed a significant cross-over interaction of sensory detail and prior knowledge (independent ROI: $F(1,20) = 9.306$, $p = 0.006$; and univariate ROI: $F(1,20) = 5.449$, $p = 0.030$) and no main effects of sensory detail (independent ROI: $F(1,20) = 0.037$; and univariate ROI: $F(1,20) = 0.675$) and prior knowledge (independent ROI: $F(1,20) = 0.005$; and univariate ROI: $F(1,20) = 0.043$, Fig 4B). For the Neutral condition, greater sensory detail leads to an increase in representational similarity (12- versus 4-channel speech, independent ROI: $t(20) = 2.551$, $p = 0.0095$; univariate ROI: $t(20) = 2.542$, $p = 0.0097$), whereas for the Match condition, increased sensory detail led to reduced representational similarity (comparison of 12- versus 4-channel speech, independent ROI: $t(20) = -1.884$, $p = 0.037$), though this was not significant in the univariate ROI ($t(20) = -1.082$, $p = 0.146$). Post-hoc one-sample t tests revealed that representational similarity was significantly greater than zero for the Match 4-channel and Neutral 12-channel conditions (independent ROI: $t(20) = 2.263$, $p = 0.018$; univariate ROI: $t(20) = 1.792$, $p = 0.044$ and independent ROI: $t(20) = 1.913$, $p = 0.035$; univariate ROI: $t(20) = 2.179$, $p = 0.021$, respectively), but not for the Match 12-channel and Neutral 4-channel conditions (independent ROI: $t(20) = -0.559$, $p = 0.709$; univariate ROI: $t(20) = 0.018$, $p = 0.493$ and independent ROI: $t(20) = -0.880$, $p = 0.805$; univariate ROI: $t(20) = -0.725$, $p = 0.762$, respectively). For completeness, we also tested other STS ROIs using clusters observed in the univariate analysis. There were no significant effects of sensory detail, prior knowledge, or interaction in either the left anterior STS (Sensory detail: $F(1,20) = 1.96$, $p = 0.177$) or the right STS (all other effects $F < 1$). In addition, we used the two regions in the inferior frontal gyrus (IFG) identified by the univariate analysis on prior knowledge, but Fisher-z-transformed correlation coefficients


A
Match 4-ch
Neutral 4-ch
Match 12-ch
Neutral 12-ch
Similar
Dissimilar
NaN
1
24
words
thing sing sit bath path pass
0 = similar, 1 = dissimilar


B
posterior STS
Neutral
Match
Fisher-z-transformed Spearman correlation coefficients
4-channel
12-channel

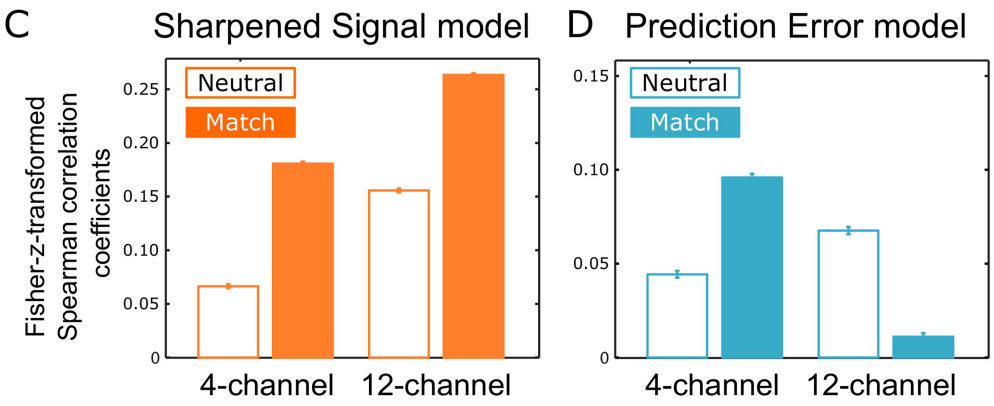

C
Sharpened Signal model
D
Prediction Error model
Neutral
Match
Fisher-z-transformed Spearman correlation coefficients
4-channel
12-channel

Fig 4. Multivariate fMRI results and simulation. (A) Hypothesized representational dissimilarity matrices. These four matrices were used to test similarity between words that share vowels within each of the four critical conditions (Match 4-channel, Neutral 4-channel, Match 12-channel, and Neutral 12-channel). Similarity between responses to identical items (on the main diagonal) was excluded, as was similarity between items in different conditions ("Not a Number" [NaN] values depicted in grey). Similarity between items containing the same vowel was predicted (zeroes in blue), whereas items containing different vowels were predicted to have more dissimilar representations (ones in red). These matrices are correlated with observed and simulated representational similarity. **(B)** RSA results. Fisher-z-transformed Spearman correlation coefficients for each of the four conditions in the left posterior STS (extracted from an independent ROI, [57]) show a significant interaction between sensory detail and prior expectation. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons [62]. **(C,D)** Model comparison. Fisher-z-transformed Spearman correlation coefficients for each of the four conditions in the two models. **(C)** Sharpened Signal model (in orange) shows that both prior knowledge and sensory detail increase similarity for words that share the same vowel. **(D)** Prediction Error model (in blue) shows opposite effects of sensory detail in neutral and matching prior knowledge conditions, consistent with the RSA results (B). Error bars in (C) and (D) indicate standard error of the mean over 1,000 replications of these simulations. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

doi:10.1371/journal.pbio.1002577.g004

extracted from either of the regions of interest in the IFG did not reveal any significant main effect or interaction (see S1 Text, S4 Fig, and S1 Table).

To illustrate how our results depend on the assumptions made about the similarity of specific syllable pairs, we also explored other ways of testing representational similarity for speech at different levels of abstraction. We therefore compared representational similarity in the independent STS ROI to hypothesised dissimilarity on the basis of early acoustic, feature, and segmental properties (see S5 Fig). Both of the more abstract RDMs (Syllable and Segment) showed a significant interaction between prior knowledge and sensory detail (see S1 Text). This is consistent with the proposal that representations of phonetic form in the STS/superior temporal gyrus (STG) reflect the abstract, categorical similarity of syllables independent of their acoustic realisation (see S1 Text) [55,57]. The low correlation values observed in these analyses are comparable with those observed in similar studies with speech stimuli [53,57]. Analysis of cross-subject consistency of observed RDMs suggests some potential for alternative hypothesis RDMs to provide higher correlation values with the observed RDMs, but confirms the crossover interaction between sensory detail and prior knowledge (see S1 Text and S6A Fig).

Whole brain analysis. In order to further test for differences in representational similarity between conditions, we conducted a repeated measures ANOVA with factors sensory detail (4- versus 12-channel) and prior knowledge (Match versus Neutral) using searchlight similarity values for the whole brain. This revealed a significant interaction in the left middle occipital gyrus ($p < 0.05$, FWE voxelwise corrected) and an interaction in the left posterior STS and the left precentral gyrus at a more lenient threshold ($p < 0.001$ uncorrected, $k > 10$ voxels, S7 Fig, S5 Table). The interaction in the posterior STS showed the same pattern as the ROI analysis for the posterior STS (as depicted in Fig 4B, S5H Fig). This cluster in the left posterior STS was significant, with small volume correction (MNI: x = -57, y = -40, z = 10, $p = 0.003$) based on an independent coordinate (defined by multivariate syllable identity coding in the left posterior STS MNI: x = -57, y = -39, z = 8, [57]). Even at this lenient threshold there was no main effect of prior information on multivoxel fMRI pattern similarity and only an effect of sensory detail in the right postcentral gyrus that failed to reach corrected significance (MNI: x = 54, y = -13, z = 40, $p < 0.001$ uncorrected, k = 14).

Model Simulation of Multivariate fMRI Results

To test our two computational simulations of spoken word recognition, we applied the same multivariate analysis to representations of the sensory input in the Sharpened Signal and

Prediction Error models for each of our four conditions (for details, see Materials and Methods). As for the multivoxel fMRI RSA, we quantified the difference in pattern similarity between pairs of similar and dissimilar syllables (e.g., "sing" and "thing" versus "sing" and "bath;" see Fig 4A). The simulation for the Sharpened Signal model showed increased similarity for both increased sensory detail and matching prior information (Fig 4C). In contrast, the simulation for the Prediction Error model showed an interaction between sensory detail and prior information (Fig 4D). Specifically, there was greater pattern similarity for similar syllable pairs in the Neutral 12-channel than in the Neutral 4-channel condition, whereas in the Match 12-channel there was less pattern similarity than in the Match 4-channel condition. This outcome resembles the interaction of sensory detail and prior knowledge shown for multivariate fMRI results in the posterior STS ROI (Fig 4B). In addition, we repeated the cross-subject consistency analysis on representations generated by individual simulated participants. For the Prediction Error but not for the Sharpened Signal model, this showed the same crossover interaction of sensory detail and prior knowledge as in the equivalent fMRI analysis, suggesting a common underlying explanation (see S1 Text and S6A–S6C Fig).

Again, we used the evidence ratio of Akaike weights to compare the evidence for both models given the pattern similarity results in the left posterior STS (see Materials and Methods). Importantly, both models used parameters optimised to simulate the behavioural and univariate fMRI results, and no modifications or parameter optimisation were performed when simulating similarity in spatial patterns of fMRI activity. For the multivariate fMRI results, the evidence ratio of the Akaike weights revealed that the multivariate fMRI patterns very strongly supported the Prediction Error model over the Sharpened Signal model (w_{PE}/w_{Sharp} = 1.898 x 10^{11}, tested based on the independent ROI in the posterior STS [57]). Hence, computational simulations provided compelling evidence that multivariate fMRI results are more consistent with computation of Prediction Errors than of Sharpened Signals in the posterior STS during the perception of degraded speech.

Discussion

We used multiple approaches (behavioural, computational, univariate, and multivariate fMRI) to investigate how prior expectations improve perception of degraded speech in order to distinguish Sharpened Signal and Prediction Error computations. Our experimental findings, first of all, replicate the existing literature [31–33,65] by showing that behavioural report of degraded words was improved both by matching expectations and by increased sensory detail (Fig 3A). Second, we show that matching expectations reduced BOLD activity during speech processing in left posterior STS (Fig 3B and 3C). Like other previous observations in the literature [8,39,43,45,46,66], these findings are in line with either Sharpened Signal or Prediction Error computations for combining prior knowledge and sensory input. This is confirmed by our computational simulations, which show that a good fit to behavioural and univariate fMRI data is achieved by models that include either of these two coding schemes (Fig 3D–3G). These model simulations are also consistent with the proposal that BOLD responses in the Match condition are lower because word identification is easier (as suggested by the behavioural improvements we observed). More informative results come from fMRI multivoxel pattern similarity, which revealed an interaction between prior knowledge and sensory detail in the posterior STS (Fig 4B). Specifically, for degraded speech that follows neutral expectations, increased sensory detail improved the amount of sensory information contained in fMRI multivoxel patterns. However, for speech that matched expectations, increased sensory detail led to a reduction in the amount of information represented in the posterior STS as measured by similarity analysis. This interaction is uniquely consistent with a Prediction Error model in which

expected sensory input is explained away, and deviations from expectation are represented as Prediction Errors (Fig 1B). Our results, therefore, provide evidence for computation of Prediction Errors but not of Sharpened Signals (see simulations in Fig 4C and 4D).

Why is this interaction between sensory detail and prior knowledge shown in multivariate representations of speech so diagnostic of Prediction Error computations? In explaining this interaction, we will first consider the situation in which listeners have uninformative prior expectations. In the absence of specific expectations (as in the Neutral condition), both Sharpening and Prediction Error accounts propose that the amount of sensory information represented in neural patterns should increase with the amount of sensory detail in the input. In Prediction Error schemes, the brain does not pass forward the sensory input directly, but rather the discrepancy between expectations and sensory input. These Prediction Errors will provide an informative representation of the sensory input if these expectations are uninformative and the sensory input is sufficiently clear. Thus, our observation of enhanced coding for Neutral 12-channel compared to Neutral 4-channel stimuli is equally consistent with Prediction Error as with the traditional view that the brain directly represents the sensory input. The true test of Prediction Error schemes is provided by conditions in which specific and accurate expectations guide perceptual processing.

The hallmark of Prediction Error in our data is that for speech that matches prior expectations increasing the sensory detail reduces the informativeness of multivariate representations (Fig 1B). This is a counterintuitive finding, because clear speech that matches a previously presented written word (our Match 12-channel condition) is most accurately perceived, whereas multivariate representations are more informative in the less intelligible Match 4-channel condition. This is to be expected because Prediction Errors will be substantially reduced for conditions in which sensory input matches prior knowledge. Hence, increases in sensory detail lead to a better correspondence between sensory input and listeners' prior expectations of clear speech. Our observation of reduced representation of speech content for Match 12-channel compared to Match 4-channel stimuli is entirely consistent with Prediction Errors but stands in marked contrast to the outcome expected for Sharpened Signals—or, indeed, any account in which sensory representations directly encode perceptual outcomes. Low pattern similarity for the condition with the clearest perceptual outcome (Match 12-channel) might appear surprising given previous findings that perceptual representations can be decoded from low-level response patterns [67–69]. However, these findings can be reconciled with Prediction Error schemes by recalling that these previous experiments used presentation conditions similar to the Neutral condition in our experiment (i.e., an uninformative prior).

Prediction Error can also explain the apparent increase in the informativeness of speech representations in the Match 4-channel condition compared to the Neutral 4-channel condition. Our simulations reveal that when sensory signals are severely degraded (such as for 4-channel vocoded speech), informative Prediction Errors are derived from the residual of matching expectations (in the Match condition). A specific expectation, as provided by our written word cue, when combined with a less informative stimulus, remains "unfulfilled" and is therefore represented as a negative but informative Prediction Error. Informative Prediction Errors (either positive or negative) are absent when prior expectations are uninformative (in the Neutral condition). Hence, both Prediction Error and Sharpened Signal models can explain our observation of increased representation of 4-channel speech that matches prior expectations. Other similar studies in the literature have explored whether visual representations of expected stimuli are sharpened or reduced [20,26] but have yielded contradictory findings. While Prediction Error was supported by univariate hemodynamic responses to unexpected classes of visual stimuli (faces versus houses, [26]), multivariate responses supported sharpening of expected visual gratings [20]. Two other differences between our work and this previous

multivariate study are noteworthy. First, we separated the neural response to the cue (written word) and stimulus (spoken word). Second, we tested the interaction between sensory information and prior knowledge. Only Prediction Error can explain the full interaction of sensory detail and prior knowledge described above.

By the Prediction Error scheme, there should be a negative correlation between neural representations in the Neutral 12-channel condition (a positive Prediction Error) and the Match 4-channel condition (a negative Prediction Error, apparent in positive and negative Prediction Error in Fig 1B). However, an additional analysis of the present data showed that there was neither a negative nor a positive correlation between these conditions (we tested for a positive correlation because a negative Prediction Error could evoke a positive hemodynamic response due to metabolic costs of neural inhibition). These null findings cannot rule out the possibility that both conditions do indeed contain complementary information based on positive and negative Prediction Errors. Direct neural data (e.g., from intracranial recordings) might provide a more sensitive test of this proposal. Taken to an extreme (i.e., without any sensory input), computation of negative Prediction Errors could also explain previous results showing that the omission of an expected stimulus causes an increased signal [70–72] from which stimulus identity can be decoded [72–74].

Implications for Predictive Coding Theories

Current Predictive Coding theories suggest that cortical regions involved in sensory processing contain two subpopulations of neurons: (1) prediction error units that represent the unexpected part of the incoming sensory information and (2) prediction units that represent the expected part of the incoming sensory information (and can be sharpened by matching prior expectations) [3,24,75]. These models have thus drawn support from empirical evidence showing either Prediction Errors [26,39,49,76] or Sharpened Signals [20] by attributing neural responses to prediction error and prediction units, respectively. Our goal in this study was to test two functionally distinct coding schemes in isolation by building computational models in which a simulated cortical area passes only one type of information forward (only Prediction Errors or Sharpened Signals). In the context of these simulations, our results provide clear evidence for representations of Prediction Errors. However, our multivariate fMRI findings do not oppose theories of Predictive Coding that propose Sharpened Signals coded by prediction units in addition to Prediction Errors in prediction error units [3,23,24]. The absence of evidence for Sharpened Signals in our data from the STS could be explained by previous proposals that fMRI measurements are dominated by responses from prediction error units (as [26,77] have argued for visual cortex). It could be that other neural measures, such as neurophysiological recordings with depth electrodes [78] or laminar-specific ultra-high field strength fMRI [79,80] are better able to detect responses from prediction units and could provide evidence of laminar-specific representations of Prediction Errors and Sharpened Signals.

Nonetheless, the interaction observed in the present study favours Predictive Coding theories (with representations of Prediction Error) over the traditional view that the brain directly passes forward the sensory input, as hypothesised in a Sharpening scheme without representations of Prediction Error. Our simulations show that in Sharpening schemes, the Match 12-channel condition should contain the clearest representation of speech content. This was not observed in the present data (compare Fig 4B and 4C). Our work not only provides evidence to support the hypothesis that integration of prior expectation and perceptual input for speech is achieved through computation of Prediction Errors or Sharpened Signals, but also introduces a new and critical diagnostic finding for Prediction Error responses: For unexpected stimuli, increased sensory detail should improve the amount of sensory information contained

in neural patterns. However, for stimuli that match expectations, increased sensory detail should lead to a reduction in the amount of information represented. Future studies in other sensory modalities and domains might benefit from adopting similar methods.

Implications for the Perception of Speech and Other Sensory Signals

Our work joins a number of recent fMRI and MEG/EEG studies in proposing an important role for Prediction Error computations in speech perception [4,7,8,39,81]. In these earlier studies, the observation of decreased activation for expected stimuli in the STG has been interpreted as a neural correlate of reduced Prediction Error and, hence, as evidence for Predictive Coding theories. However, almost all established computational theories of speech perception can also explain this observation. For example, TRACE [34] implements a form of neural sharpening in which prior knowledge enhances the representation of expected sensory signals and suppresses sensory noise, producing a reduced neural response overall. Similar, interactive activation models [35–38] might predict exactly the same decrease in STG activity for expected stimuli, as observed in these previous neuroimaging studies. Thus, existing empirical evidence proposed for Predictive Coding is also largely consistent with Sharpening theories. Even our previous comparison of Predictive Coding and Lexical Competition accounts of spoken word recognition [39] challenged the competitive lexical selection mechanism implemented in TRACE, but did not test the Sharpening mechanism traditionally described as Interactive Activation.

In this context, then, the results of our study have important implications for understanding speech perception, a domain in which the presence and function of top-down processes has been much debated [82,83]. By directly quantifying the information represented in multivariate signals during perception of correctly expected and unexpected speech, we provided evidence that the neural mechanisms underlying speech perception are in line with Prediction Error simulations. Prior knowledge of speech content is used to explain away sensory evidence such that speech representations encode Prediction Error.

The present multivariate interaction of sensory detail and prior information supports a Predictive Coding theory for how matching expectations improve perception of degraded speech. In contrast, enhanced representation of attended compared to unattended speech supports Sharpening mechanisms [84–87]. These findings could be reconciled by theories proposing that expectation (Prediction Error) and attention (Sharpening) operate in parallel, as suggested in some Predictive Coding theories [3,88]. However, more detailed computational specification of attentional mechanisms will be required to test these theories with experimental data. Comparing neural representations of attended and unattended speech signals at varying levels of expectation and degradation may be informative.

There are three reasons why our results are of general interest for the study of speech and other domains of perception. One key aspect of our approach is that we assessed the perception of speech presented at varying levels of signal degradation. As in accounts proposing Bayesian perceptual inference [89], this provides the best opportunity to observe influences of prior knowledge on perception. In doing so, we also test the perception of speech in listening conditions similar to the way that speech is most often heard in the real world [90]. A second form of generality is that prior expectations for speech were derived from written text. Our results may therefore also inform other situations in which prior knowledge and sensory information are combined across different modalities for speech [91–93] and other cross-modal stimuli [94–96]. Third and perhaps most important, however, is that the representations of Prediction Error that we have observed during speech perception might apply to many other sensory domains in which prior knowledge has been shown to influence perception (such as audition [6,7,76,97], vision [9–12,20,98,99], touch [13], gustation [14,100], olfaction [15], and pain

[16]). The interactive effect of prior knowledge and sensory input on neural representation of degraded stimuli provides a stronger test of Predictive Coding theories of perception than has been provided by existing methods, as it offers the potential to challenge alternative views based purely on Sharpening mechanisms.

Conclusions

In summary, the present results show that both increased sensory detail and matching prior expectations improved accuracy of word report for degraded speech but had opposite effects on speech coding in the posterior STS. Following neutral text, increased sensory detail enhanced the amount of speech information, whereas matching prior expectations reduced the amount of measured information during presentation of clearer speech. These findings support the view that the brain reduces the expected and, therefore, redundant part of the sensory input during perception, in line with representations of Prediction Error proposed in Predictive Coding theories.

Materials and Methods

Ethics Statement

Ethical approval was provided by Cambridge Psychology Research Ethics committee (CPREC) under approval number 2009.46. All participants provided their written informed consent.

Participants

Twenty-five healthy native-English speakers (aged 18–40, with self-reported normal hearing and language function) participated in the experiment. Three participants had to be excluded because they were insufficiently attentive to the written text during the scanning runs (they reported less than 50% of the written words correctly when prompted). One additional participant had to be excluded due to technical problems. The reported analyses are therefore based on 21 participants (mean age 25 y [range 19 to 38 y], 9 females).

Stimuli

Word stimuli consisted of 24 different monosyllabic words, each with a consonant-vowel-consonant structure. The words were selected as eight triples of three similar words, each sharing the same vowel and with offset and onset changes between items (eight triples: thing/sing/sit, bath/path/pass, deep/peep/peak, pork/fork/fort, doom/tomb/tooth, take/shake/shape, kite/tight/type, zone/moan/mode). These stimuli were recorded by a male native speaker of Southern British English and noise-vocoded (4- and 12-channel) using custom scripts written in Matlab [59]. The syllables were filtered into 4 or 12 approximately logarithmically spaced frequency bands from 70 to 5,000 Hz [101], with each pass band 3 dB down with a 16 dB/octave roll off. In each band, envelopes were extracted using half wave rectification, and pitch synchronous oscillations above 30 Hz were removed with a second-order Butterworth filter. The resulting envelopes were multiplied with a broadband noise and then band pass filtered in the same frequency ranges as the source and recombined. To ensure that acoustic intensity was matched across all stimuli, the RMS amplitude of each sound file was equalised. Finally, we applied an additional filter to ensure a flat frequency response when the spoken words were presented via Sensimetrics insert headphones in the scanner (http://www.sens.com).

fMRI Procedure

Participants read written words and listened to subsequently presented degraded spoken words (see Fig 2). There were four conditions containing different pairings of written and spoken words: (1) matching written text + spoken words ("SING" + *sing*); (2) neutral written text ("XXXX") + spoken words (*sing*); (3) partially mismatching written text + spoken words ("SIT" + *sing*); (4) totally mismatching written text + spoken words ("SING" + *doom*). In addition, we included a fifth condition in which only written text ("SING") was presented to test whether participants attended to the written words. Only the match and neutral conditions (condition 1 and 2) were repeated sufficiently (six presentations per item per condition) to permit multivariate RSA (see below for details). In occasional catch trials, a response cue, which consisted of a visual display of a question mark, was presented 1,000 ms after trial onset. This cued participants to say aloud the written or spoken word that they saw or heard previously. This design does not allow the analysis of response times, because participants were cued to respond after a delay. A previous behavioural study in our lab showed that response times for reporting vocoded spoken words are uninformative even when collected in such a way as to permit response time analyses [102]. The partial and total mismatch conditions (condition 3 and 4) were included to make sure that participants paid attention to both the written and the spoken word; these conditions ensured that they could not simply report the preceding written word. Due to the small number of trials, RSA analysis was not possible for neural responses measured in the Mismatch condition. We can, however, report behavioural and univariate fMRI results for the Mismatch condition; this confirms that behavioural and neural enhancement following matching written text is not due to prestimulus attention or anticipation (because prestimulus processes will be identical following mismatching text but enhanced perception is not typically observed) [8,33].

Trials commenced with presentation of a fixation cross (1,000 ms), followed by presentation of a written word (500 ms), again followed by a fixation cross (500 ms), and finally the presentation of a spoken word. Written cues (i.e., written words, neutral "XXXX", and fixation cross) were presented in grey in the centre of the black screen. Trials were 3 to 9 s long, depending on the number of inserted null events to decorrelate the events within each run (76 trials of 3 s without null event, 45 trials of 6 s with a null event of 3 s, and 15 trials of 9 s with a null event of 6 sec, resulting in 211 TRs per run with null events).

Spoken words were presented after 4- or 12-channel noise-vocoding to produce two different levels of sensory detail in the speech input. Altogether, this resulted in 816 trials, including 1/6 catch trials (136 trials) in which participants had to give their verbal response (24 neutral and 24 match words x 6 repetitions x 2 levels of sensory detail = 576 trials, 24 written-only words x 6 repetitions = 144 trials, 24 partial mismatch and 24 total mismatch words x 2 levels of sensory detail without repetition on the word level = 96 trials; i.e., 11.8% of the trials contained mismatching information). These trials were split into 6 runs of 136 trials each, ensuring that each word in each condition occurred once in each scanning run. With additional catch trials, each run took 11.7 min, and the overall experiment lasted approximately 70 min for all 6 runs. Stimulus delivery was controlled and behavioural responses were recorded with E-Prime 2.0 software (Psychology Software Tools, Inc.).

Scanning Parameters

Structural scanning. MRI data were acquired on a 3-Tesla Siemens Tim Trio scanner using a 32-channel head coil. A T1-weighted structural scan was acquired for each subject using a three-dimensional MPRAGE sequence (TR 2,250 ms, TE: 2.99 ms, flip angle: 98 deg, field of view: 256 x 240 x 160 mm, matrix size: 256 x 240 x 160 mm, spatial resolution, 1 x 1 x 1 mm).

Functional scanning. The fMRI session was split into 6 runs of 11.7 min. We used sparse imaging to acquire fMRI data. For each participant and scanning run, 239 echo planar imaging (EPI) volumes comprising 26 slices of 3 mm thickness were acquired using a continuous, descending acquisition sequence (TR 3,000 ms, TA 1,600 ms, TE 30 ms, FA 78 deg, matrix size: 64 x 64, in plane resolution: 3 x 3 mm, inter-slice gap 25%). We used transverse-oblique acquisition, with slices angled away from the eyes to avoid artefacts from eye movements. Visual stimuli were projected on a screen at the head-end of the scanner table and reflected onto a mirror attached to the head coil above the participants' eyes. We used Sensimetrics headphones (Sensimetrics Corporation, Malden, MA, USA, model S14) to deliver the sound stimulation and a MR-compatible microphone (FOMRI II, Optoacoustics) to record verbal response.

Behavioural Analysis

Verbal responses recorded in the scanner were transcribed by two independent raters (the first author and a native English speaker with a PhD in phonetics who was naïve to the stimulus set) and disagreements adjudicated by a third rater (the senior author). All raters were blind to which word and stimulus condition was presented in each trial. Responses were scored for whole-word accuracy and analysed using Matlab. Because the percent correct performance scores were bound to [0;1], we applied an arcsine transformation [103] before we computed a two-way repeated measures ANOVA and the corresponding post-hoc pared *t* tests.

Univariate fMRI Analysis

Data were analysed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) applying automatic analysis (aa) pipelines [104]. The first three volumes of each run were removed to allow for T1 equilibrium effects. Scans were realigned to the first EPI image. The structural image was coregistered to the mean functional image and the parameters from the segmentation of the structural image were used to normalise the functional images, which were resampled to 2 mm isotropic voxels. The realigned normalised images were then smoothed with a Gaussian kernel of 8 mm full width half maximum. Data were analysed using the general linear model with a 128 s high pass filter. We included the onset of 11 event types in the GLM, each convolved with the canonical SPM haemodynamic response: eight conditions come from specifying the onset of spoken words paired with four types of written text (matching, neutral, partially mismatching, and totally mismatching) crossed with two types of vocoding (4- and 12-channel). We also specified onsets for written words and neutral strings ("XXXX") as well as the onset of the visual task cue that instructed participants to say the spoken word. Following parameter estimation of the first level model, we conducted a repeated measures ANOVA with two factors: prior knowledge (matching versus neutral text) and level of sensory detail (4- versus 12-channel) to assess the main effects and interaction of these factors.

We were interested in the effect of hearing speech that matches prior expectations on BOLD responses in the left posterior STS. To locate these ROIs for the multivoxel RSA (see below), we tested for a main effect of prior knowledge (F-contrast "Neutral versus Match") and identified a cluster at $p < 0.05$ FWE voxel-corrected in the left posterior STS.

Multivariate RSA fMRI Analysis

Multivariate analyses were conducted on realigned data within each participant's native space without normalisation or spatial smoothing. An additional first-level model was constructed for each participant that contained the same set of regressors as the first level model used for the univariate analysis, except that regressors for individual spoken words were used in each of the four conditions for which there were sufficient numbers of repetitions for item-specific

modelling (4- and 12-channel vocoded words following neutral or matching text). This resulted in 103 conditions per participant per run: 24 words for each of these four conditions and the remaining seven conditions from the univariate model. For each of the 96 item-specific regressors in these four conditions, we estimated single-subject T-statistic images for the contrast of speech onset compared to the unmodelled resting period, averaged over the six scanning runs.

We used the resulting single condition and item T-images for RSA [50] using the RSA toolbox [52]. We used T-images so that effect sizes were weighted by their error variance, which reduces the influence of large but variable response estimates for multivariate analyses [105]. RSA involves testing whether the observed similarity of brain responses in specific conditions (a neural RDM) corresponds to a hypothetical pattern of similarity between these conditions (hypothesis RDM). We constructed four hypothesis RDMs to test for greater similarity between syllable pairs within the same stimulus triple (i.e., syllables that shared the same vowel and had similar onset or offset segments like "sing" and "thing," as compared to dissimilar syllables like "sing" and "bath") within each of four critical conditions: Match 4-channel, Neutral 4-channel, Match 12-channel, and Neutral 12-channel. The design of our experiment was motivated by previous work that showed that STS encodes vowel and syllable similarity [55,61], rather than spectrotemporal acoustic cues [61]. The comparisons used in our ROI analysis test for global similarity in representations of the phonetic form of similar-sounding spoken words because multiple consonantal features as well as the vowel are preserved within each syllable triple (e.g., bath/path/pass). We chose to analyse similarity of neural representations for phonetically similar but non-identical words for two reasons: (1) this approach allowed us to merge all six runs into a single analysis, which reduced the noise in the estimation of the T-images relative to a split-half method, and (2) comparing similar but non-identical word pairs makes our method insensitive to other forms of lexical or semantic similarity that could lead to similar neural representations for identical word pairs (e.g., in regions that code for word meaning [106]). Similarity between items in different conditions and between identical items (i.e., the main diagonal) was therefore not included in our hypothesis RDMs (see Fig 4A).

We measured multivoxel RDMs by computing the dissimilarity (1–Pearson correlation across voxels) of T-statistics for a specific item and condition. In a searchlight analysis, the sets of voxels were extracted by specifying grey-matter voxels (voxels with a value > 0.33 in a probabilistic grey-matter map) within an 8-mm radius sphere of each grey matter voxel (with a voxel size of 3 x 3 x 3.75 mm, i.e., a maximum of 65 voxels per sphere). This was repeated for all searchlight locations in the brain. The similarity between the observed RDM and each of the hypothetical RDMs was computed using a Spearman correlation for each searchlight location, and the resulting correlation coefficient returned to the voxel at the centre of the searchlight. This resulted in a Spearman correlation map for each participant in each grey matter voxel. To assess searchlight similarity values across participants at the second level, the Spearman correlation maps for each participant were Fisher-z-transformed to conform to Gaussian assumptions, normalized to MNI space, and spatially smoothed with a 10-mm FWHM Gaussian kernel for group analysis. These second-level analyses used a within-subject analysis of variance similar to those used for the univariate fMRI analysis.

Region of interest (ROI) analysis. In a region of interest (ROI) analysis using MarsBaR (http://marsbar.sourceforge.net/), we extracted similarity values from searchlights within ROIs defined on the basis (1) of an independent coordinate (defined by multivariate syllable identity coding in the left posterior STS MNI: x = -57, y = -39, z = 8, [57]) and (2) of the univariate fMRI analysis. We used the independent ROI in the left posterior STS to make sure that the results were not caused by any potential dependencies of univariate and multivariate analyses.

In addition, the univariate ROIs allowed us to test for differences in observed multivoxel similarity in each of the four conditions within STS regions defined on the basis of showing hemodynamic response reductions for degraded words following matching written words. To locate this region, we tested for a main effect of prior knowledge (F-contrast "Neutral versus Match") and identified a cluster at $p < 0.05$ FWE voxel-corrected in the left posterior STS (centre of mass MNI: x = -56, y = -35, z = 6, k = 99 voxels). For completeness, we also considered two other STS clusters from this univariate analysis: left anterior STS (centre of mass MNI: x = -57, y = -10, z = -5, k = 229 voxels) and right STS (centre of mass MNI: x = 56, y = -13, z = -4, k = 92 voxels). For each ROI, we obtained one Fisher-z-transformed Spearman correlation value for each of our four conditions. We then tested for differences between these conditions in a repeated measures ANOVA with factors sensory detail (4- versus 12-channel) and prior knowledge (Neutral versus Match). We conducted post-hoc one-sided paired *t* tests on the data extracted from the independent ROI in the left posterior STS (MNI: x = -57, y = -39, z = 8, [57], sphere 6 mm, 896 mm volume) and based on the ROI defined by the univariate analysis (centre of mass MNI: x = -56, y = -35, z = 6, k = 99 voxels, 782 mm volume) to test for the Neutral condition whether sensory detail led to an increase in representational similarity and for the Match condition whether sensory detail led to a decrease in representational similarity. In addition, we conducted post-hoc one-sample *t* tests on the data extracted from the independent ROI in the left posterior STS [57] and the ROI defined by the univariate analysis to test whether the correlation was significantly greater than zero for the four conditions, individually.

Computational Simulations of Spoken Word Recognition using Sharpened Signals or Prediction Errors

We used two computational implementations of Sharpened Signal and Prediction Error models of spoken word recognition (using update mechanisms based on [75]), to simulate observed behavioural performance (i.e., word recognition), univariate fMRI results (the magnitude of hemodynamic activity in the STS), and RSA fMRI results (the similarity of representations for word pairs in the left posterior STS) in each of our four experimental conditions. The sensory representations supplied at the input, the output lexical representations, and the specification of matching or neutral prior knowledge was identical for both simulations. We used a localist lexical representation (i.e., a set of 24 units, each of which was activated to represent a single word), as in previous models of spoken word recognition such as TRACE [34] or Shortlist [107]. The input to the model was provided as a distributed set of phonetic features (derived from [108]). These are similar to the acoustic/phonetic features supplied as the input to TRACE or in recurrent network simulations such as the Distributed Cohort Model [109]. However, to avoid the complexity of representing temporal information (and given the slow haemodynamic responses measured by fMRI), we assumed that speech information is provided in parallel over three groups of units for the initial consonant, medial vowel, and final consonant of our CVC words.

The key difference between the Sharpened Signal and Prediction Error models concerns the computations by which prior knowledge is combined with degraded sensory representations of expected spoken words. In the Sharpened Signal simulation, expected sensory features receive additional activation through increased sensory gain [19,20], whereas in the Prediction Error model, prior expectations contribute to perception by subtracting expected input from sensory representations (i.e., computation of Prediction Error [3,23,24]). In both simulations, an iterative settling procedure was used such that feature representations of the input are combined with prior knowledge to generate feature representations that convey Sharpened Signals or Prediction Errors respectively (hereafter "sharpened features" and "prediction error features").

These representations were used to update lexical activations, and updated lexical activations in turn led to modified top-down expectations. This settling procedure continued until a settling criterion was reached or a maximum number of iterations had been performed.

Representations of speech input, lexical knowledge, and perceptual expectations. The representations of the speech input, lexical knowledge, and perceptual expectations were the same for both the Sharpened Signal and the Prediction Error model (see S2 Fig). The sensory input for each degraded spoken word was determined by a feature matrix that transformed a phonological transcription of each of the 24 words into an articulatory feature representation based on phonetic descriptions of each segment [108]. We used articulatory representations because they appropriately model the similarity of different spoken words, and there is considerable evidence from intracranial recordings [110] and multivariate fMRI to support the presence of articulatory representations in superior temporal regions [106,111]. Representing the segments of the 24 words in our stimulus set required 13 consonantal features and 11 vowel features, concatenated into a set of 37 binary features for the CVC syllables used in the experiment. The 13 consonantal features were divided into four groups: (1) place of articulation (six features: bilabial, labiodental, dental, alveolar, palato-alveolar, velar), (2) manner of articulation (three features: stop, sibilant, non-sibilant), (3) nasality (three features: nasal, oral), and (4) voicing (two features: voiceless, voiced). The 11 vowel features were divided into four groups: (1) height (five features: high, mid-high, mid, mid-low, low), (2) backness (two features: front, back), (3) rounding (two features: rounded, unrounded), and (4) length/diphthong (two features: long, short). Based on these position-specific features, we constructed a feature-to-word transformation matrix that included positive binary values in each row to indicate which phonetic features were relevant for each word (see S2 Fig). Each row contained 12 active features (four features for each consonant and vowel). This matrix served as a set of connection weights to link phonetic features to words in both models and thereby encoded long-term knowledge of the form of each spoken word.

To generate different levels of degradation of the sensory input (equivalent to 4- or 12-channel vocoded speech), we set noise parameters (for low- and high-sensory detail) that determined the degree to which the appropriate input features remained active and inappropriate features inactive following degradation. Noise was added to each group of features (place, manner, etc.) individually, such that the sum of all active features within each group remained 1 and, hence, the pattern of activation within each of the feature groups could be interpreted as a probability distribution. For example, if the current "place" feature was 1 for bilabial (as in the initial segment of "bath"), this group of features would be [1 0 0 0 0] for clear speech, but with a noise parameter of 0.5 the input representation would be set to [0.5 0 0 0 0] and a uniform random amount (that sums to 0.5) assigned to all five features. Thus, with a noise parameter of 1, no information would remain concerning the place of articulation of the speech input. The noise parameter for low and high sensory detail conditions was fitted separately for each model based on the aggregate behavioural and univariate results (i.e., 4- and 12-channel vocoded speech, **low sensory noise** and **high sensory noise**; names of fitted parameters are highlighted as shown). This is sufficient to allow our model to simulate the overall accuracy of perception (though not the fine-grained pattern of perceptual confusions, which is beyond the scope of the present simulation). We note that the similarity of these simulated degraded feature representations resembles the similarity of the acoustic forms of the vocoded spoken words.

The two prior knowledge conditions (Neutral and Match) were differentiated by prior lexical expectations, i.e., the prior probability of each of the 24 words in the models vocabulary. The prior expectation for the Neutral condition was defined as a uniform distribution over all the words in the set (i.e., each word was assigned a prior probability of 0.042 equivalent to 1/24). For the Match condition, the prior probability was determined by the probability of

hearing matching speech after a written word was presented. In the experiment overall, there were 288 match trials and 48 mismatch trials; hence, a prior of 0.857 for the specific written word that was presented. However, because the written word was not always remembered correctly by participants, we multiplied this probability by behavioural performance in the "written only" condition (82.14% correct on average) to estimate the prior probability of the matching word and made all other words equally probable, such that the summed activation of lexical units was 1. Lexical expectations in the Neutral and Match conditions were transformed from the lexical level into phonetic feature expectations by multiplication of lexical probabilities by the word-to-feature transformation matrix.

A simulated word recognition trial in both models began by specifying the prior lexical knowledge for a Match or a Neutral trial at the output and presenting a degraded speech representation for one of the 24 words to the input (both as described above). Based on these initial activation values, an iterative updating process operated to combine prior knowledge and sensory input until a **stopping criterion** (defined on the basis of changes in lexical activation) or until a maximum number of iterations was performed. For both the Sharpened Signal and Prediction Error models, the maximum number of iterations was set to 500.

Sharpened signal model. In the Sharpened Signal model, sensory input that corresponds to expected words is enhanced and therefore plays a greater role in updating lexical activation values. This was achieved by generating a sharpened feature representation by multiplying the observed sensory input (over a set of features, i = 1:37) by a representation of the expected sensory input (derived from the set of expected words, w = 1:24).

First, the expected word was transformed from a lexical representation into a feature representation:

$$\text{expected features (i)} = \text{prior word (w)} * \text{feature}-\text{to}-\text{word matrix (w, i)}^{\text{T}}$$

Then, the expected features were used to enhance the expected part of the sensory input:

$$\text{sharpened features (i)} = \text{sensory input (i)} * (1 + \text{expected features (i)})$$

This sharpened set of phonetic features was then normalized and combined with the sensory input to form the input for the next iteration. An **update weight** parameter was fitted for the Sharpened Signal model (the same for all words and noise levels) to determine how much the sensory representation changed in each iteration.

$$\text{updated sharpened features (i)} = \text{sensory input (i)} + (\textbf{update weight} * \text{sharpened features (i)})$$

These sharpened features were then transformed to generate an updated word representation:

$$\text{updated word (w)} = \text{updated sharpened feature(i)} * \text{feature}-\text{to}-\text{word matrix (w, i)}$$

Iterations continued until a single lexical item became more strongly activated than any other item at the output based on a stopping criterion parameter, based on the difference between the maximum word value and the mean plus one standard deviation of all word activation values:

$$\text{if } \max(\text{updated word(w)}) - (\text{mean(updated word(w))} + \text{standard deviation(updated word(w))}) > \textbf{stopping criterion}$$

This **stopping criterion** parameter was fitted for the Sharpened Signal model and was the same for all words and noise levels.

Prediction error model. In the Prediction Error model, Prediction Errors were computed by comparing the heard sensory features (i = 1:37) with sensory features derived from the expected word (w = 1:24).

First, the expected word was transformed from a lexical representation to a feature representation:

$$\text{expected features (i)} = \text{prior word (w)} * \text{feature}-\text{to}-\text{word matrix (w, i)}^{T}$$

Then, the expected features were used to explain away the expected part of the sensory input:

$$\text{prediction error features (i)} = \text{sensory input (i)} - \text{expected features (i)}$$

Based on this, the feature prediction error was transformed into a word prediction error:

$$\text{prediction error word (w)} = \text{prediction error features (i)} * \text{feature}-\text{to}-\text{word matrix (w, i)}$$

From this, an updated word representation can be computed by adding the word prior to the word prediction error multiplied by an **update weight** (equivalent to that used in the Sharpened Signal model), and a precision value:

$$\text{updated word (w)} = \text{word prior (w)} + (\textbf{update weight} * \text{precision} * \text{prediction error word (w)})$$

The "update weight" parameter was fitted for the Prediction Error model and was the same for all words and noise levels. The precision of the Prediction Error was determined for each word and noise level by combining the precisions of its constituents

$$\begin{aligned}\text{precision} = \quad & \text{standard deviation(word prior (w))} \,/\, \text{sum(word prior (w))} + \\ & \text{standard deviation(sensory input (i))} \,/\, \text{sum(sensory input (i))).}\end{aligned}$$

Iterations continued until the prediction error was smaller than a **stopping criterion**:

$$\text{if sum(abs(prediction error word (w)))} < \textbf{stopping criterion}$$

This **stopping criterion** parameter was fitted for the Prediction Error model and was the same for all words and noise levels.

Relating model output to behavioural and fMRI measures. Several different measures can be derived from the operation of these computational models, which we used to simulate the behavioural, univariate, and multivariate fMRI results.

To simulate the behavioural performance, we tested whether each word presented was correctly identified by the model based on the state of the lexical representations at the end of the iterative update process (i.e., the posterior word representation). These output representations were transformed into probabilities using a softmax transfer function with a **temperature** parameter, fitted independently for Sharpened Signal and Prediction Error simulations to determine the degree of competition between active words. To simulate inconsistent or uncertain behavioural responses, we added Gaussian random noise to the word probabilities with the amount of noise determined by a **behavioural noise** parameter (again, fitted independently for each simulation Sharpened Signal and Prediction Error) and selected the word with the highest value as the response. The addition of random noise simulates word reports as resulting from additional "noisy" processes that follow computation of the likely word candidates (e.g., memory, attention, motor mapping that in turn influence how precepts lead responses). Based on whether the word chosen matches the word presented, we can calculate the word recognition performance of the model.

To simulate the univariate fMRI results, we counted the number of iterations the models needed to satisfy the **stopping criterion** (as described for each simulation). Our reasoning was that the number of processing iterations in the model serves as a proxy for the duration of the word recognition process and that, all other things being equal, a longer period of neural processing should lead to an increased BOLD signal during identification of a spoken word (see [63,112] for further discussion). This is not to say that other differences between conditions equated for processing time would not also give rise to differences in the BOLD response; only that, all other things being equal, longer processing time will lead to an increased BOLD response. Furthermore, this outcome measure (unlike, for instance, Prediction Error) is common to both sets of simulations.

To simulate the multivariate fMRI results, we tested the similarity of the sharpened feature and prediction error feature representations after the first model iteration (for the Sharpened Signal and Prediction Error model, respectively). We decided to use representations from the first iteration because we did not want to make further assumptions for how these signals are integrated over time that might favour one model or other (because Sharpened Signal and Prediction Error models show different settling dynamics) or that differentially impact one or more experimental condition (since settling dynamics may also differ between conditions). We leave it to later work using temporally sensitive neural measures (such as MEG or eCog) to explore how settling dynamics impact on neural representations of speech content. Similarly, to ensure that Sharpened Signal and Prediction Error models are more comparable, we removed the sign of the Prediction Error signal such that multivariate analyses are always performed on positive feature representations in both models. We assumed that these feature representations (or, equivalently, these representations multiplied by the feature-to-word matrix) can serve as a surrogate for multi-voxel patterns of searchlights in our posterior STS ROI. To simulate the influence of measurement noise on measured fMRI responses and, hence, multivariate similarity measures, we added a noise pattern to each of the activation patterns prior to computing correlations between feature representations. Specifically, we added Gaussian noise (with a standard deviation of 2 for both simulations) to the sharpened feature and prediction error feature representations before we conducted RSA. For each computational model, we then computed a dissimilarity matrix (based on a 1–Pearson correlation) for the feature representations for all word pairs in the model simulation of all four conditions. We then used this observed RDM and applied the same hypothetical model RDMs used in the multivariate analysis of the fMRI data (i.e., greater similarity for word-pairs within each triple that share the same vowel compared to words in different triples with different vowels). This comparison was conducted separately for each of the four key experimental conditions (Neutral/Matching priors, 4- and 12-channel speech), and Fisher-z-transformed similarity values were computed as for the fMRI data.

Model fitting procedure. We used a standard non-linear optimisation procedure implemented in Matlab (fminsearch, Matlab, The MathWorks, Inc.) to separately fit the following six parameters for the Sharpened Signal and Prediction Error models: (1) **low sensory detail**: the level of noise added to simulate 4-channel speech; (2) **high sensory detail**: the level of noise added to simulate 12-channel speech; (3) **update weight**: the amount by which prior representations are updated during a single processing iteration; (4) **stopping criterion**: the measure computed to determine when the iterative model process converged; (5) **temperature** parameter: this determined the degree of winner-take-all competition during response selection; (6) **behavioural noise**: simulating the degree of uncertainty and guessing in model responses. Sensitivity analyses (S3 Fig) show that parameters (1) and (2) influence both behavioural outcomes and univariate responses, parameters (3) and (4) largely influence settling time (and, hence, univariate responses), and parameters (5) and (6) influence behavioural outcomes.

The models were fitted to minimise the sum-squared error difference between model outcomes and observed behavioural and univariate results averaged over participants (see Fig 3). The analyses reported here constitute fixed-effects analyses, because we fitted the models to the group means. Specifically, we computed the sum of two error terms to quantify the difference between the model prediction and observed data. First, we computed the difference between the behavioural performance predicted by the model and the actual behaviour in the four conditions. Second, we computed the difference between the univariate results predicted by the model and the actual univariate results in the posterior STS. To relate model iterations to BOLD signal estimates, we normalized the univariate fMRI results (by dividing the extracted beta values in each condition by the maximum beta value) and the model outcome (by dividing the number of iterations by the maximum possible number of iterations, i.e., 500). Because simulations of both behavioural and fMRI responses were prone to chance variation (due to the influence of the various sources of added noise described above) we used the average results of 10 replications for each condition and item when computing and optimising model fit. The resulting six model parameters for the Sharpened Signal model were [0.2456, 0.4094, 0.002022, 2.198, 2.556, 0.01057], and the model parameters for the Prediction Error model were [0.3559, 0.5825, 0.03414, 0.407, 1.327, 0.00281].

Model fit evaluation. Due to the presence of several sources of random noise in the simulation model, we used Monte-Carlo methods to evaluate the goodness of fit of the two models to the data. We used the optimal parameters listed above to compute the distribution of model outcomes for 1,000 replications of each condition and item from the experiment. From these distributions, we could observe the likelihoods of the data given each model simulating (1) the behavioural results, (2) the mean parameter estimates in the left posterior STS from the univariate fMRI analysis, and (3) multivariate results for the left posterior STS.

These likelihoods were fit with a 1-dimensional kernel estimation function (kdensity function in Matlab) with a width specified based on visual inspection of the individual empirical probability density functions (kernel width set to 0.1 for all simulations and conditions). We fitted these kernel density estimates for the observed data for each model, condition, and data type. We then combined the density estimates over the four conditions for each model and data type by computing joint probabilities (i.e., the products of the four kernel density estimates of each condition). Essentially the same results were obtained by fitting four-dimensional kernel densities over all four conditions simultaneously. We computed the evidence ratio of Akaike weights to estimate how much support the data provides in favour of the Prediction Error over the Sharpened Signal model. Because both models have the same number of free parameters, the ratio of the Akaike weights can be directly calculated by $\frac{Likelihood\ \ Predictive\ \ Coding\ \ model}{Likelihood\ \ Sharpening\ \ model}$ [64].

Supporting Information

S1 Fig. Effect of mismatching prior expectations. (A) Behavioural results. **(B)** Univariate results: Main effect of prior knowledge (Matching versus Mismatching Prior) depicted on a rendered brain ($p < 0.05$ voxelwise FWE, n = 21). **(C)** Mean beta values extracted from the independent region of interest in the posterior STS [57] illustrate reduced BOLD signal during Match conditions (solid black) in contrast to Neutral (white) and Mismatch (green) conditions. Error bars indicate standard error of the mean after between-subject variability has been removed suitable for repeated measures comparisons [62]. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures. (TIF)

S2 Fig. Network architecture and example representations for (A) Sharpened Signal and (B) Prediction Error models. Common components of both models are outlined in black. Differences between the two models are coloured in orange (Sharpened Signal) and blue (Prediction Error). Both models map from a feature-based representation of consonant-vowel-consonant symbols that have been degraded by the addition of random, probabilistic noise within the different groups of units representing specific feature types (place, manner, voicing, etc.). Input for the word "thing" is shown for both models, using representations degraded to simulate 4-channel and 12-channel noise vocoded speech (based on clarity parameters fit for each of the simulations). A clear speech (un-degraded) representation of the word "thing" is shown for comparison, though this wasn't presented to either model. Hinton diagrams show the activation of each individual unit with the area of the squares proportional to activation values or probabilities, supplemented by colour scales as shown. In both models, lexical representations are specified over a bank of 24 localist units (one for each word in the models' vocabulary and experimental item set). These lexical representations are initialised to express the prior probability of each word being presented based on prior written text ("THING," Match condition) or a neutral string ("XXXX," Neutral condition). In both models, a word-to-feature matrix links words to their constituent phonetic features and a feature-to-word matrix links phonetic features to words (these two matrices are the transpose of each other). There are some key differences between the two models. In the Sharpened Signal model (A), prior knowledge is used to increase the gain of expected sensory features, such that expected features are preferentially activated in Sharpened Feature representations at the intermediate level of the model. These Sharpened Features are then used to update lexical representations. Thus, Match trials lead to Sharpened Feature representations that resemble those from speech signals with greater sensory detail. In contrast, in the Prediction Error model (B), expected sensory features are subtracted from the observed sensory input, and Prediction Error feature representations at the intermediate level are used to update lexical representations. These Prediction Error representations contain negative values (blue colours) for expected features that are presented in a degraded form; these negative prediction errors carry information concerning the identity of the speech signal in Match 4 trials that is absent for Match 12 trials in which speech is less degraded.
(TIF)

S3 Fig. Sensitivity analysis. (A) Prediction Error model. (B) Sharpened Signal model. The blue curves illustrate how the sum squared error (SSE, *y*-axis) for model fit to the behavioural (left column), univariate fMRI (middle columns), and multivariate fMRI (right columns) data changes for a range of parameters (along the *x*-axis). Each graph therefore shows the influence of each of the six parameters: (1) low clarity, (2) high clarity, (3) prior update weight, (4) stopping criterion, (5) temperature, and (6) behavioural noise on model fit. The red dot on each graph indicates the final parameters chosen by nonlinear optimisation. Univariate and multivariate fMRI data come from ROI coordinates based on univariate analysis (Fig 3C). Please refer to S2 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.
(TIF)

S4 Fig. Representation of phonetic form in Inferior Frontal regions (A) Univariate results: Main effect of prior knowledge (Matching versus Neutral Prior) depicted on a rendered brain ($p < 0.05$ voxelwise FWE, n = 21). White circle marks post-hoc defined clusters of interest in the left Inferior Frontal Gyrus (IFG). **(B,C)** Fisher-z-transformed Spearman correlation coefficients for each of the four conditions in two left IFG clusters (defined by the univariate analysis) show a significant correlation in the Match 4-channel condition and a significant reduction

in correlation with increased sensory detail Match 4-channel compared to Match 12-channel. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons [62]. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.
(TIF)

S5 Fig. Comparison of four different, hierarchically organised hypothesis RDMs of speech perception. Left Panel: **(A)** dissimilarity of the acoustic properties of the speech stimuli used in our study (see Supplementary Methods for details), **(B)** dissimilarity of feature representation for the canonical forms of the speech provided as the input to our computational simulations, **(C)** dissimilarity of the segment representations of the word stimuli used in the experiment, scored based on the number of position-specific phonemes shared between words pairs, and **(D)** main hypothesis RDM assuming increased similarity between pairs of syllables that shared the same vowel (e.g., "sing" and "thing" should have more similar patterns than "sing" and "bath"). These RDMs can be considered to describe a hierarchy of speech representations from the fine-grained acoustic RDM to the most abstract syllable RDM used in our main analysis. These hypothesis RDMs are positively correlated with each other and hence can be considered as testing related proposals concerning neural representations of spoken words. Right panel **(E–H)** shows the results for the Kendall's Tau A correlation coefficients (suitable for comparisons between binary and fine-grained RDMs; see Supplementary Methods for details) as extracted from the independent region of interest in the left posterior STS (pSTS, Fig 4B). Only the segment **(G)** and the syllable RDM **(H)** revealed a significant interaction of sensory detail and prior knowledge, similar to that shown in Fig 4B. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.
(TIF)

S6 Fig. Cross-subject consistency based on empirical and simulated RDMs. (A) Empirical RDMs were extracted from the independent ROI in the left posterior STS (pSTS, Fig 4B), and the Simulated RDMs based on either **(B)** the Sharpened Signal or **(C)** the Prediction Error model were computed for 21 simulated participants. The cross-subject consistencies from the empirical RDMs and simulated RDMs from the Prediction Error model show the same cross-over interaction of sensory detail and prior knowledge shown before (Fig 4B–4D). Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.
(TIF)

S7 Fig. Representational similarity searchlight analysis in the whole brain. Interaction of Prior information (Match/Neutral) x Sensory detail (4- versus 12-channel) depicted on rendered brain (F-contrast, $p < 0.001$ uncorrected, $k > 10$ voxels; searchlight analysis with a voxel size of 3 x 3 x 3.75 mm; see S4 Table for coordinates). https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N).
(TIF)

S1 Table. Univariate Analysis—F-contrast: Main effect Match/Neutral, $p < 0.05$ FWE (voxelwise correction)
(XLS)

S2 Table. Univariate Analysis—F-contrast: Main effect sensory detail, $p < 0.05$ FWE (voxelwise correction)
(XLS)

S3 Table. Univariate Analysis—F-contrast: Prior information (Match/Neutral) x Sensory detail full interaction, $p < 0.001$ uncorrected, k > 10 voxels
(XLS)

S4 Table. Univariate Analysis—F-contrast: Main effect Match/Mismatch, $p < 0.05$ FWE (voxelwise correction)
(XLS)

S5 Table. RSA—F-contrast: Prior information (Match/Neutral) x Sensory detail full interaction, $p < 0.001$ uncorrected, k > 10 voxels (searchlight analysis with a voxel size of 3 x 3 x 3.75 mm)
(XLS)

S1 Text. Text file describing the supplementary methods and supplementary results and discussion.
(DOCX)

Acknowledgments

We would like to thank Helen Lloyd and Steve Eldridge for their assistance in radiography. Thanks to Tibor Auer for support with automatic analysis (aa), to Ed Sohoglu for providing the speech recordings, and to Ed Sohoglu and Sam Evans for comments on the manuscript. Special thanks to Arnold Ziesche for stimulating discussion and support with computational modelling.

Author Contributions

Conceptualization: HB MHD.

Funding acquisition: MHD.

Investigation: HB.

Methodology: HB MHD.

Writing – review & editing: HB MHD.

References

1. von Helmholtz H, Nagel WA. Handbuch der physiologischen Optik: L. Voss; 1909.
2. Clark A. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. Behav Brain Sci. 2012; 36(3):181–204.
3. Friston K. A theory of cortical responses. Philos Trans R Soc London [Biol]. 2005; 360(1456):815–36.
4. Arnal LH, Giraud AL. Cortical oscillations and sensory predictions. Trends Cogn Sci. 2012; 16 (7):390–8. doi: 10.1016/j.tics.2012.05.003 PMID: 22682813
5. Summerfield C, de Lange FP. Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci. 2014; 15:745–56. doi: 10.1038/nrn3838 PMID: 25315388
6. Chennu S, Noreika V, Gueorguiev D, Blenkmann A, Kochen S, Ibanez A, et al. Expectation and Attention in Hierarchical Auditory Prediction. J Neurosci. 2013; 33(27):11194–U983. doi: 10.1523/JNEUROSCI.0114-13.2013 PMID: 23825422
7. Giraud AL, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience. 2012; 15(4):511–7. doi: 10.1038/nn.3063 PMID: 22426255
8. Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive Top-Down Integration of Prior Knowledge during Speech Perception. J Neurosci. 2012; 32(25):8443–53. doi: 10.1523/JNEUROSCI.5069-11.2012 PMID: 22723684

9. Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron. 1999; 22(4):751–61. PMID: 10230795
10. Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J. Predictive codes for forthcoming perception in the frontal cortex. Science. 2006; 314(5803):1311–4. doi: 10.1126/science.1132028 PMID: 17124325
11. Kok P, Brouwer GJ, van Gerven MAJ, de Lange FP. Prior Expectations Bias Sensory Representations in Visual Cortex. The Journal of neuroscience. 2013; 33(41):16275–84. doi: 10.1523/JNEUROSCI.0742-13.2013 PMID: 24107959
12. Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Dale AM, et al. Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A. 2006; 103(2):449–54. doi: 10.1073/pnas.0507062103 PMID: 16407167
13. van Ede F, Jensen O, Maris E. Tactile expectation modulates pre-stimulus beta-band oscillations in human sensorimotor cortex. Neuroimage. 2010; 51(2):867–76. doi: 10.1016/j.neuroimage.2010.02.053 PMID: 20188186
14. Gardner MP, Fontanini A. Encoding and tracking of outcome-specific expectancy in the gustatory cortex of alert rats. J Neurosci. 2014; 34(39):13000–17. doi: 10.1523/JNEUROSCI.1820-14.2014 PMID: 25253848
15. Zelano C, Mohanty A, Gottfried JA. Olfactory Predictive Codes and Stimulus Templates in Piriform Cortex. Neuron. 2011; 72(1):178–87. doi: 10.1016/j.neuron.2011.08.010 PMID: 21982378
16. Buchel C, Geuter S, Sprenger C, Eippert F. Placebo Analgesia: A Predictive Coding Perspective. Neuron. 2014; 81(6):1223–39. doi: 10.1016/j.neuron.2014.02.042 PMID: 24656247
17. De-Wit L, Machilsen B, Putzeys T. Predictive Coding and the Neural Response to Predictable Stimuli. J Neurosci. 2010; 30(26):8702–3. doi: 10.1523/JNEUROSCI.2248-10.2010 PMID: 20592191
18. Murray SO, Schrater P, Kersten D. Perceptual grouping and the interactions between visual cortical areas. Neural Networks. 2004; 17(5–6):695–705. doi: 10.1016/j.neunet.2004.03.010 PMID: 15288893
19. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A. 2003; 20 (7):1434–48.
20. Kok P, Jehee JFM, de Lange FP. Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron. 2012; 75(2):265–70. doi: 10.1016/j.neuron.2012.04.034 PMID: 22841311
21. Shi Y, Sun H. Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards: Boca Raton CRC Press.; 1999.
22. Schroeder M. Computer speech: recognition, compression and synthesis. Berlin: Springer-Verlag; 1999.
23. Mumford D. On the Computational Architecture of the Neocortex .2. The Role of Corticocortical Loops. Biol Cybern. 1992; 66(3):241–51. PMID: 1540675
24. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999; 2(1):79–87. doi: 10.1038/4580 PMID: 10195184
25. Koster-Hale J, Saxe R. Theory of Mind: A Neural Prediction Problem. Neuron. 2013; 79(5):836–48. doi: 10.1016/j.neuron.2013.08.020 PMID: 24012000
26. Egner T, Monti JM, Summerfield C. Expectation and surprise determine neural population responses in the ventral visual stream. The Journal of neuroscience. 2010; 30(49):16601–8 doi: 10.1523/JNEUROSCI.2770-10.2010 PMID: 21147999
27. Hsieh PJ, Vul E, Kanwisher N. Recognition Alters the Spatial Pattern of fMRI Activation in Early Retinotopic Cortex. J Neurophysiol. 2010; 103(3):1501–7. doi: 10.1152/jn.00812.2009 PMID: 20071627
28. Obleser J. Putting the Listening Brain in Context. Language and Linguistics Compass. 2014; 8 (12):646–58.
29. Sumby WH, Pollack I. Visual Contribution to Speech Intelligibility in Noise. Journal of the Acoustical Society of America. 1954; 26(2):212–5.
30. Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech-Perception without Traditional Speech Cues. Science. 1981; 212(4497):947–50. PMID: 7233191
31. Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. J Exp Psychol. 1951; 41(5):329–35. PMID: 14861384
32. MacLeod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. British journal of audiology. 1987; 21(2):131–41. PMID: 3594015
33. Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Top-Down Influences of Written Text on Perceived Clarity of Degraded Speech. J Exp Psychol Human. 2014; 40(1):186–99.

34. McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive psychology. 1986; 18 (1):1–86. PMID: 3753912

35. Mirman D, McClelland JL, Holt LL. An interactive Hebbian account of lexically guided tuning of speech perception. Psychon Bull Rev. 2006; 13(6):958–65. PMID: 17484419

36. Grossberg S, Stone G. Neural Dynamics of Word Recognition and Recall—Attentional Priming, Learning, and Resonance. Psychol Rev. 1986; 93(1):46–74. PMID: 3961051

37. Grossberg S, Boardman I, Cohen M. Neural dynamics of variable-rate speech categorization. J Exp Psychol Hum Percept Perform. 1997; 23(2):481–503. PMID: 9104006

38. Vitevitch MS, Luce PA, Pisoni DB, Auer ET. Phonotactics, neighborhood activation, and lexical access for spoken words. Brain Lang. 1999; 68(1–2):306–11. doi: 10.1006/brln.1999.2116 PMID: 10433774

39. Gagnepain P, Henson RN, Davis MH. Temporal predictive codes for spoken words in auditory cortex. Curr Biol. 2012; 22(7):615–21. doi: 10.1016/j.cub.2012.02.015 PMID: 22425155

40. Hickok G, Poeppel D. The cortical organization of speech processing. Nature reviews Neuroscience. 2007; 8(5):393–402. doi: 10.1038/nrn2113 PMID: 17431404

41. Scott SK, Johnsrude IS. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences. 2003; 26(2):100–7. doi: 10.1016/S0166-2236(02)00037-1 PMID: 12536133

42. Davis MH, Ford MA, Kherif F, Johnsrude IS. Does Semantic Context Benefit Speech Understanding through "Top–Down" Processes? Evidence from Time-resolved Sparse fMRI. J Cognitive Neurosci. 2011; 23(12):3914–32.

43. Blank H, von Kriegstein K. Mechanisms of enhancing visual–speech recognition by prior auditory information. NeuroImage. 2013; 65(0):109–18.

44. Nath AR, Beauchamp MS. Dynamic Changes in Superior Temporal Sulcus Connectivity during Perception of Noisy Audiovisual Speech. The Journal of neuroscience. 2011; 31(5):1704–14. doi: 10.1523/JNEUROSCI.4853-10.2011 PMID: 21289179

45. Lee H, Noppeney U. Temporal prediction errors in visual and auditory cortices. Curr Biol. 2014; 24(8): R309–10. doi: 10.1016/j.cub.2014.02.007 PMID: 24735850

46. Noppeney U, Josephs O, Hocking J, Price CJ, Friston KJ. The effect of prior visual information on recognition of speech and sounds. Cerebral Cortex. 2008; 18(3):598–609. doi: 10.1093/cercor/bhm091 PMID: 17617658

47. Mottonen R, Calvert GA, Jaaskelainen IP, Matthews PM, Thesen T, Tuomainen J, et al. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage. 2006; 30(2):563–9. doi: 10.1016/j.neuroimage.2005.10.002 PMID: 16275021

48. Dehaene-Lambertz G, Pallier C, Serniclaes W, Sprenger-Charolles L, Jobert A, Dehaene S. Neural correlates of switching from auditory to speech perception. Neuroimage. 2005; 24(1):21–33. doi: 10.1016/j.neuroimage.2004.09.039 PMID: 15588593

49. Sohoglu E, Davis MH. Perceptual learning of degraded speech by minimizing prediction error. Proc Natl Acad Sci U S A. 2016; 113(12):E1747–56. doi: 10.1073/pnas.1523266113 PMID: 26957596

50. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2008; 2:4. doi: 10.3389/neuro.06.004.2008 PMID: 19104670

51. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. P Natl Acad Sci USA. 2006; 103(10):3863–8.

52. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014; 10(4):e1003553. doi: 10.1371/journal.pcbi.1003553 PMID: 24743308

53. Du Y, Buchsbaum BR, Grady CL, Alain C. Noise differentially impacts phoneme representations in the auditory and speech motor systems. P Natl Acad Sci USA. 2014; 111(19):7126–31.

54. Lee YS, Turkeltaub P, Granger R, Raizada RDS. Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based Analysis. J Neurosci. 2012; 32(11):3942–8. doi: 10.1523/JNEUROSCI.3814-11.2012 PMID: 22423114

55. Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 2008; 322(5903):970–3. doi: 10.1126/science.1164318 PMID: 18988858

56. Boets B, Op de Beeck HP, Vandermosten M, Scott SK, Gillebert CR, Mantini D, et al. Intact But Less Accessible Phonetic Representations in Adults with Dyslexia. Science. 2013; 342(6163):1251–4. doi: 10.1126/science.1244333 PMID: 24311693

57. Evans S, Davis MH. Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. Cerebral cortex. 2015; 25(12):4772–88. doi: 10.1093/cercor/bhv136 PMID: 26157026

58. Wild CJ, Davis MH, Johnsrude IS. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage. 2012; 60(2):1490–502. doi: 10.1016/j.neuroimage.2012.01.035 PMID: 22248574

59. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270(5234):303–4. PMID: 7569981

60. Davis MH, Johnsrude IS. Hierarchical processing in spoken language comprehension. J Neurosci. 2003; 23(8):3423–31. PMID: 12716950

61. Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT. Categorical speech representation in human superior temporal gyrus. Nature Neuroscience. 2010; 13(11):1428–U169. doi: 10.1038/nn.2641 PMID: 20890293

62. Loftus GR, Masson MEJ. Using Confidence-Intervals in within-Subject Designs. Psychon B Rev. 1994; 1(4):476–90.

63. Grill-Spector K, Henson R, Martin A. Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences. 2006; 10(1):14–23. doi: 10.1016/j.tics.2005.11.006 PMID: 16321563

64. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. Psychon Bull Rev. 2004; 11 (1):192–6. PMID: 15117008

65. Obleser J, Wise RJS, Dresner MA, Scott SK. Functional integration across brain regions improves speech perception under adverse listening conditions. J Neurosci. 2007; 27(9):2283–9. doi: 10.1523/JNEUROSCI.4663-06.2007 PMID: 17329425

66. Arnal LH, Morillon B, Kell CA, Giraud AL. Dual neural routing of visual facilitation in speech processing. J Neurosci. 2009; 29(43):13445–53. doi: 10.1523/JNEUROSCI.3194-09.2009 PMID: 19864557

67. Nienborg H, Cumming BG. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature. 2009; 459(7243):89–92. doi: 10.1038/nature07821 PMID: 19270683

68. Haynes JD, Rees G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci. 2005; 8(5):686–91. doi: 10.1038/nn1445 PMID: 15852013

69. Kilian-Hutten N, Valente G, Vroomen J, Formisano E. Auditory Cortex Encodes the Perceptual Interpretation of Ambiguous Sound. J Neurosci. 2011; 31(5):1715–20. doi: 10.1523/JNEUROSCI.4572-10.2011 PMID: 21289180

70. den Ouden HEM, Friston KJ, Daw ND, McIntosh AR, Stephan KE. A Dual Role for Prediction Error in Associative Learning. Cerebral cortex. 2009; 19(5):1175–85. doi: 10.1093/cercor/bhn161 PMID: 18820290

71. Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. P Natl Acad Sci USA. 2011; 108 (51):20754–9.

72. Sanmiguel I, Widmann A, Bendixen A, Trujillo-Barreto N, Schroger E. Hearing silences: human auditory processing relies on preactivation of sound-specific brain activity patterns. J Neurosci. 2013; 33 (20):8633–9. doi: 10.1523/JNEUROSCI.5821-12.2013 PMID: 23678108

73. Kok P, Failing MF, de Lange FP. Prior expectations evoke stimulus templates in the primary visual cortex. J Cogn Neurosci. 2014; 26(7):1546–54. doi: 10.1162/jocn_a_00562 PMID: 24392894

74. Hsu YF, Le Bars S, Hamalainen JA, Waszak F. Distinctive Representation of Mispredicted and Unpredicted Prediction Errors in Human Electroencephalography. J Neurosci. 2015; 35(43):14653–60. doi: 10.1523/JNEUROSCI.2204-15.2015 PMID: 26511253

75. Spratling MW. Reconciling predictive coding and biased competition models of cortical function. Frontiers in computational neuroscience. 2008; 2:4. doi: 10.3389/neuro.10.004.2008 PMID: 18978957

76. Todorovic A, van Ede F, Maris E, de Lange FP. Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. J Neurosci. 2011; 31(25):9118–23. doi: 10.1523/JNEUROSCI.1425-11.2011 PMID: 21697363

77. de Gardelle V, Stokes M, Johnen VM, Wyart V, Summerfield C. Overlapping multivoxel patterns for two levels of visual expectation. Front Hum Neurosci. 2013; 7.

78. Lakatos P, O'Connell MN, Barczak A, Mills A, Javitt DC, Schroeder CE. The Leading Sense: Supramodal Control of Neurophysiological Context by Attention. Neuron. 2009; 64(3):419–30. doi: 10.1016/j.neuron.2009.10.014 PMID: 19914189

79. Muckli L, De Martino F, Vizioli L, Petro LS, Smith FW, Ugurbil K, et al. Contextual Feedback to Superficial Layers of V1. Current Biology. 2015; 25(20):2690–5. doi: 10.1016/j.cub.2015.08.057 PMID: 26441356

80. Kok P, Bains LJ, van Mourik T, Norris DG, de Lange FP. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. Current Biology. 2016; 26(3):371–6. doi: 10.1016/j.cub.2015.12.038 PMID: 26832438

81. Arnal LH, Wyart V, Giraud A-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature Neuroscience. 2011; 14(6):797–801. doi: 10.1038/nn.2810 PMID: 21552273

82. Norris D, McQueen JM, Cutler A. Merging information in speech recognition: Feedback is never necessary. Behav Brain Sci. 2000; 23(3):299-+. PMID: 11301575

83. McClelland JL, Mirman D, Holt LL. Are there interactive processes in speech perception? Trends in Cognitive Sciences. 2006; 10(8):363–9. doi: 10.1016/j.tics.2006.06.007 PMID: 16843037

84. Crosse MJ, Butler JS, Lalor EC. Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. J Neurosci. 2015; 35(42):14195–204. doi: 10.1523/JNEUROSCI.1829-15.2015 PMID: 26490860

85. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485:233–6. doi: 10.1038/nature11020 PMID: 22522927

86. Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC. At what time is the cocktail party? A late locus of selective attention to natural speech. European Journal of Neuroscience. 2012; 35(9):1497–503. doi: 10.1111/j.1460-9568.2012.08060.x PMID: 22462504

87. Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, et al. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron. 2013; 77 (5):980–91. doi: 10.1016/j.neuron.2012.12.037 PMID: 23473326

88. Summerfield C, Egner T. Expectation (and attention) in visual cognition. Trends in cognitive sciences. 2009; 13(9):403–9. doi: 10.1016/j.tics.2009.06.003 PMID: 19716752

89. Norris D, McQueen JM. Shortlist B: A Bayesian model of continuous speech recognition. Psychol Rev. 2008; 115(2):357–95. doi: 10.1037/0033-295X.115.2.357 PMID: 18426294

90. Mattys SL, Liss JM. On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. Perception & Psychophysics. 2008; 70(7):1235–42.

91. McGurk H, MacDonald J. Hearing Lips and Seeing Voices. Nature. 1976; 264(5588):746–8. PMID: 1012311

92. Blank H, Kiebel SJ, von Kriegstein K. How the Human Brain Exchanges Information Across Sensory Modalities to Recognize Other People. Hum Brain Mapp. 2015; 36(1):324–39. doi: 10.1002/hbm.22631 PMID: 25220190

93. von Kriegstein K, Dogan O, Gruter M, Giraud AL, Kell CA, Gruter T, et al. Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A. 2008; 105 (18):6747–52. doi: 10.1073/pnas.0710826105 PMID: 18436648

94. Botvinick M, Cohen J. Rubber hands 'feel' touch that eyes see. Nature. 1998; 391(6669):756-. doi: 10.1038/35784 PMID: 9486643

95. Hairston WD, Wallace MT, Vaughan JW, Stein BE, Norris JL, Schirillo JA. Visual localization ability influences cross-modal bias. J Cognitive Neurosci. 2003; 15(1):20–9.

96. Shams L, Kamitani Y, Shimojo S. Illusions—What you see is what you hear. Nature. 2000; 408 (6814):788. doi: 10.1038/35048669 PMID: 11130706

97. Todorovic A, de Lange FP. Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. J Neurosci. 2012; 32(39):13389–95. doi: 10.1523/JNEUROSCI.2227-12.2012 PMID: 23015429

98. Krol ME, El-Deredy W. When believing is seeing: The role of predictions in shaping visual perception. Q J Exp Psychol. 2011; 64(9):1743–71.

99. Rauss K, Schwartz S, Pourtois G. Top-down effects on early visual processing in humans: A predictive coding framework. Neurosci Biobehav R. 2011; 35(5):1237–53.

100. Siegrist M, Cousin ME. Expectations influence sensory experience in a wine tasting. Appetite. 2009; 52(3):762–5. doi: 10.1016/j.appet.2009.02.002 PMID: 19501777

101. Greenwood DD. A Cochlear Frequency-Position Function for Several Species—29 Years Later. Journal of the Acoustical Society of America. 1990; 87(6):2592–605. PMID: 2373794

102. Hervais-Adelman A, Davis MH, Johnsrude IS, Carlyon RP. Perceptual learning of noise vocoded words: Effects of feedback and lexicality. J Exp Psychol Human. 2008; 34(2):460–74.

103. Studebaker GA. A Rationalized Arcsine Transform. J Speech Hear Res. 1985; 28(3):455–62. PMID: 4046587

104. Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke A, et al. Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Frontiers in Neuroinformatics. 2015; 8.

105. Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage. 2010; 53(1):103–18. doi: 10.1016/j.neuroimage.2010.05.051 PMID: 20580933

106. Correia JM, Jansma BMB, Bonte M. Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. J Neurosci. 2015; 35(44):15015–25.

107. Norris D. Shortlist—a Connectionist Model of Continuous Speech Recognition. Cognition. 1994; 52 (3):189–234.

108. Ladefoged P, Johnson K. A Course in Phonetics. 6th ed. Michael Rosenberg; 2010.

109. Gaskell MG, Marslen-Wilson WD. Integrating form and meaning: A distributed model of speech perception. Lang Cognitive Proc. 1997; 12(5–6):613–56.

110. Mesgarani N, Cheung C, Johnson K, Chang EF. Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science. 2014; 343(6174):1006–10. doi: 10.1126/science.1245994 PMID: 24482117

111. Arsenault JS, Buchsbaum BR. Distributed Neural Representations of Phonological Features during Speech Perception. J Neurosci. 2015; 35(2):634–42. doi: 10.1523/JNEUROSCI.2454-14.2015 PMID: 25589757

112. Taylor JSH, Rastle K, Davis MH. Interpreting response time effects in functional imaging studies. Neuroimage. 2014; 99:419–33. doi: 10.1016/j.neuroimage.2014.05.073 PMID: 24904992

Supplementary Material

Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception - H Blank, H & M H Davis

Supplementary Methods

Comparison of four different, hierarchically organised hypothesis RDMs of speech perception.

We tested three additional hypothesis Representational Dissimilarity Matrices (RDMs) relating to (1) the acoustic properties of the speech stimuli used in our study, (2) the feature representation as used as input to our computational simulations, and (3) the segmental representation of the word stimuli, as scored based on the number of shared phonemes (SI Fig 5 left column). We will describe how these additional similarity matrices were generated in turn.

For the acoustic properties we computed acoustic (dis)similarity between pairs of speech tokens using methods described by Billig and colleagues (1). Specifically, we generated a Gammatone-based spectro-temporal representation for each speech token. A spectral dissimilarity matrix was then generated between pairs of spectro-temporal representations tokens by computing 1 minus the sample linear correlation between log-scaled spectra at all time slices. Next, the maximum-dissimilarity path through this spectral-dissimilarity matrix was found using dynamic time warping. Summed dissimilarity values along this path were computed and rescaled (dissimilarity / maximum dissimilarity) such that two identical sound files were assigned a score of 0 and the two most dissimilar sound files given a score of 1. Note that as reported by (1) greatest similarity is seen for pairs of syllables that contain the same vowel. The gammatone representation and dynamic time warping were performed using Matlab implementations of standard algorithms written by Dan Ellis (downloaded from http://www.ee.columbia.edu/ln/rosa/matlab/).

For the feature properties we used the input representation of speech used in our computational simulations (for details see Materials and Methods in the Main Text). A feature dissimilarity matrix was then generated by computing 1 minus the sample linear correlation between pairs of feature representations for all 24 words used in our experiment and simulations.

For the segmental properties we counted the common segments of all word pairs based on the phonemic transcription of each word (CELEX Database). This score based on the number of shared phonemes ranged from 0 to 3, because each word consisted of three segments. The common segments were transformed to a dissimilarity value by [(3 – number of common segments) / 3].

The three additional RDMs included more fine-grained similarity values (e.g., Acoustic RDM) compared to the binary RDM used in the main analysis (i.e., Syllable RDM). To not favour a simplified RDM with tied ranks (such as Syllable RDM) we repeated the RSA searchlight analysis with Kendall's Tau A (instead of Spearman correlation). Kendall's Tau A is more likely than Spearman correlation coefficient to prefer the true RDM over a simplified RDM containing tied ranks (2).

Estimation of cross-subject consistency and maximum possible correlation of the observed RDM in left posterior STS

To provide an estimate of the maximum possible correlation value between the observed RDM and the hypothesized RDMs, we used the procedure described in (2) for computing the upper bound of the noise ceiling of the observed RDMs for the fMRI data. Specifically, the rank-transformed single-subject RDMs were averaged and we used in an iterative procedure to find the RDM with the maximum average correlation to the single subject RDMs (using published code from (2)).

In addition, to provide an estimate of the expected correlation value between the observed RDM and the hypothesized RDMs, given the degree of inter-subject variation in the fMRI data, we computed the cross-subject consistency of the observed RDMs (using the procedure described for computing the lower bound of the noise ceiling in (2) and the corresponding published code). Specifically, we used a leave-one-subject-out procedure in which we correlated (using Kendall's Tau A coefficient) each subject's empirically observed RDM with the mean observed RDM of the remaining 20 subjects, separately for the four conditions. Then we computed the mean over these correlation values for each condition to estimate an empirically-derived hypothesis RDM for similarity between the word stimuli used in our experiment. The empirical RDMs were computed within an 8 mm sphere (corresponding to the 8 mm sphere used in the whole-brain searchlight analysis) centred on the voxel specified for the independent ROI in the left posterior STS (3). To compare this empirical cross-subject consistency with the expected cross-subject consistency based on the Prediction Error and Sharpened Signal models, we performed the same leave-one-subject-out correlation analysis on single subject RDMs for 21 simulated participants (i.e., treating individual simulation runs as individual participants). We increased the amount of Gaussian noise added to the prediction error and sharpened signal representations (5 standard derivations) so that overall similarity was comparable for empirical and simulated RDMs. Importantly, the same amount of noise was added to all four conditions and to both models.

Supplementary Results and Discussion

Comparison of responses following mismatching written text

Behavioural Analysis.

We confirmed that providing informative prior expectations improves perception of degraded speech in comparison to both providing neutral or mismatching prior information **(**SI Fig 1 A**).** A two-way repeated measures ANOVA with the factors sensory detail (4- vs. 12-channel) and prior knowledge (Match vs. Neutral vs. Mismatch) revealed significant main effects of sensory detail on word report ($F(1, 20) = 139.988$, $p < 0.001$, eta squared = 87.50) and prior knowledge ($F(1, 20) = 80.652$, $p < 0.001$, eta squared = 80.13), and a significant interaction ($F(1, 20) = 14.617$, $p < 0.001$). Post-hoc paired t-tests revealed that word report for degraded speech that mismatched with prior text was less accurate than for speech that matched prior text for both the 4-channel ($t(20) = 8.343$, $p < 0.001$) and the 12-channel conditions ($t(20) = 4.590$, $p < 0.001$). However, word report did not differ between Mismatch and Neutral condition at either level of sensory detail (4-channel: $t(20) = 1.71$, $p = 0.102$; 12-channel: $t(20) = 1.531$, $p = 0.141$). This suggests that differences between Match and Neutral trials reflect the facilitatory perceptual effect of matching prior knowledge, rather than any non-specific effect of hearing degraded words after reading a written text cue.

Univariate Results.

We sought to localise the univariate BOLD activity decrease for degraded spoken words that follow matching written words relative to words following mismatching cues (SI Fig 1 B/C). We conducted a repeated measures ANOVA with two factors: prior knowledge (Match vs. Mismatch) and level of sensory detail (4- vs. 12-channel) to assess the main effect of prior information a whole brain analysis. We collapsed across both types of mismatching conditions (partial and total mismatch; e.g., ‘shape’ - ‘shake’ and ‘shape’ - ‘zone’, respectively) between written and spoken words to increase the number of trials that could be included in this analysis. The magnitude of the BOLD responses in the left posterior STS in the Mismatch condition resembles the magnitude of the BOLD responses in the Neutral condition (SI Fig 1 C). This confirms that reduced activity observed for degraded speech that matches previously written words (compared to speech following Neutral text “XXXX”) is due to the facilitatory effect of hearing degraded speech that matches prior knowledge rather than a generic modulation of auditory responses following written text. Increased activity for speech that follows mismatching in comparison to matching written words also confirms that the difference found for Neutral > Match is not due to changes in “attention”, or baseline activation following written text since decreased activity for Match trials is not specific to a comparison with responses following uninformative cues in the Neutral condition.

Representations of phonetic form in Inferior Frontal Regions

To provide a more complete picture of our data, we used the two regions in the Inferior Frontal Gyrus as identified by the univariate analysis on prior expectation (responses greater following Neutral than following Matching text, SI Fig 4 + SI Table 1). These regions are potentially of interest because these regions have been proposed to contribute to (predictive) processing, in particular for speech heard in adverse listening conditions (4-6). Multivariate pattern analysis has further shown that inferior frontal and adjacent precentral gyrus regions represent the identity, but not the acoustic form of heard syllables (3, 7, 8), particularly if speech is degraded.

For these two regions, we conducted ROI analyses of multivariate information content in each of our four experimental conditions. Specifically, we conducted a Repeated Measures ANOVA with factors sensory detail (4- vs. 12-channel) and prior knowledge (Match vs. Neutral). Fisher-z-transformed correlation coefficients extracted from either of the regions of interest in the IFG (Orbitalis: 623 voxels, peak MNI: x = -32, y = 38, z = 0 and Opercularis: 164 voxels peak at MNI: x = -42, y = 4, z = 26; SI Fig 4) did not reveal any significant main effect or interaction (Main effect Prior: $F < 1$ for both ROIs; Main effect sensory detail: $F(20) = 3.965$, $p = 0.060$; $F(20) = 2.768$, $p = 0.112$; Interaction: $F(20) = 1.255$, $p = 0.276$; $F(20) = 3.728$, $p = 0.069$; for left IFG Orbitalis; left IFG Opercularis, respectively). Post-hoc t-tests in both regions revealed a significant correlation only in the Match 4-channel condition ($t(20) = 2.709$, $p = 0.007$; $t(20) = 3.682$, $p < 0.001$) and for the paired t-test of Match 4-channel vs. Match 12 ($t(20) = 2.268$, $p = 0.017$; $t(20) = 2.726$, $p = 0.007$; for left IFG Orbitalis; left IFG Opercularis, respectively). This result for the Match 4-channel in the IFG is particularly interesting because the IFG has previously been suggested as the source of top-down predictions when written text informs the perception of degraded speech (4, 6). Furthermore, these top-down mechanisms seem to be especially important for perceptual learning observed when matching prior expectations can be used to guide perception of highly degraded speech (5, 9, 10).

Comparison of four different, hierarchically organised hypothesis RDMs of speech perception.

The similarity values computed in the three additional RDMs are positively correlated (Acoustic to Feature RDM: $r = 0.4653$, $p < 0.0001$; Feature to Segment RDM: $r = 0.5872$; $p < 0.0001$). Importantly, the most abstract segment level description is also highly correlated with the similarity matrix constructed on the basis of the syllable triples used in the experiment (Segment to Syllable RDM = 0.6440, $p < 0.0001$, see also correlations for Acoustic to Syllable RDM: $r = 0.3549$, $p < 0.0001$; Feature to Syllable RDM: $r = 0.4331$, $p < 0.0001$). This indicates that acoustic, feature, segment, and syllable characteristics of the word stimuli used in our experiment are related to each other. Further evidence for these being a hierarchy of representations comes from comparisons between these correlations tested using one-sided t-tests for dependent correlations (11). We see

significantly higher correlations between Segment to Syllable than Feature to Syllable RDMs ($t(273) = 4.912$, $p < 0.001$), a trend for higher correlations between Feature to Syllable than Acoustic to Syllable RDMs ($t(273) = 1.379$, $p = 0.085$) and significantly higher correlations between Feature to Segment than Acoustic to Segment RDMs ($t(273) = 4.189$, $p < 0.001$). This suggests a hierarchy from acoustic representations to syllable representations in the order shown in SI Fig 5.

However, results confirm that only the Segment and the Syllable hypothesis RDMs show the interaction of sensory detail and prior knowledge in the STS (Syllable RDM: $F(1,20) = 9.302$, $p = 0.006$; Segment RDM: $F(1,20) = 6.237$, $p = 0.021$; main effects were not significant for either RDM: $F(1,20) < 0.1$). There were no significant main effects or interactions of sensory detail and prior knowledge for either the Acoustic or the Feature RDM (all effects: $p > 0.05$). This result is in line with previous findings that categorical, segmental representations are an important organizing principle in STG/STS regions (12, 13). Recordings from fMRI (13) and intracranial high-density cortical surface arrays showed that the posterior STG represents the underlying identity of spoken syllables rather than producing a linear response to changes in spectrotemporal acoustic or phonetic cues (12).

Cross-subject consistency and maximum possible correlation of the observed RDM in left posterior STS

The upper bound (that is, the maximum possible correlation value that could be observed in our fMRI data from the posterior STS) is very similar across the four conditions (Neutral 4-channel: 0.1488; Match 4-channel: 0.1715; Neutral 12-channel: 0.1573; Match 12-channel: 0.1481) and there is neither a significant interaction of sensory detail and prior knowledge ($F(1,20) = 1.785$), nor a main effect (Sensory detail: $F(1,20) = 0.344$; Prior knowledge: $F(1,20) = 0.323$). In all four conditions, the upper bound of the maximal possible correlation is substantially smaller than 1. This indicates limitations of our experimental data (e.g., low spatial resolution, high measurement noise and/or limited amounts of data). Nonetheless, none of these limitations differentially affect our four critical conditions and hence measurement noise or other extraneous factors cannot explain the significant interaction of sensory detail and prior knowledge seen in multivariate analyses.

The relatively small effect sizes that we have observed are common for multivariate fMRI analyses of speech stimuli. For RSA of speech perception, similar values of (Fisher-z transformed) correlation between fMRI-response based- and hypothesized similarity matrices have been observed previously (3, 8). Similarly, low classification accuracies are also common in decoding task events using Multivariate Classification of fMRI data (14, 15). Despite these small effect sizes, condition-specific differences in the observed correlations (i.e., the interaction of sensory detail and prior knowledge) provide compelling statistical support for neural representations of Prediction Error.

The obtained correlation values for the cross-subject consistency of the observed RDMs in the STS are higher than the correlation values obtained for the main RSA analysis (i.e., the correlation with Syllable RDM, compare SI Fig 5H and SI Fig 6A). This indicates that the observed correlation values in our fMRI RSA analysis are smaller than expected due to limitations of the hypothesis RDMs (2) and that there is potential for alternative hypothesis RDMs to provide higher correlation values with the observed RDMs. However, the cross-subject consistency of the observed RDMs also showed a significant cross-over interaction of sensory detail and prior knowledge ($F(1,20) = 6.443$, $p = 0.0196$) and no main effects (Sensory detail: $F(1,20) = 0.968$; Prior knowledge: $F(1,20) = 1.298$, SI Fig 6A). This suggests that the information present in multivoxel fMRI patterns differs among our four conditions even when this is tested without assuming a hypothesis RDM. Simulations show that this is in line with the Prediction Error model. Simulated cross-subject consistency from the Sharpened Signal model (SI Fig 6B) showed two significant main effects (Sensory detail: $F(1,20) = 153.023$, $p < 0.001$; Prior knowledge: $F(1,20) = 111.232$, $p < 0.001$), but no interaction ($F(1,20) = 0.340$). This is the same pattern as observed in the main simulation using the Syllable RDM as the hypothesis RDM (Fig 4C) which does not resemble the empirical data (i.e., Fig 4B, SI Fig 6A). In contrast, the cross-subject consistency simulated with the Prediction Error model (SI Fig 6C) showed a significant cross-over interaction of sensory detail and prior knowledge ($F(1,20) = 15.217$, $p = 0.0009$) and no main effects (Sensory detail: $F(1,20) = 0.059$; Prior knowledge: $F(1,20) = 0.906$). The reduction of simulated cross-subject consistency in both the Neutral 4-channel and the Match 12-channel conditions is explained by uninformative Prediction Errors in these conditions. This is due to either uninformative sensory information (Neutral 4-channel) or informative sensory information explained away by matching prior expectations (Match 12-channel). Again, this is the same pattern as observed in the main simulation using the Syllable RDM as the hypothesis RDM (Fig 4D) which resembles the empirical data (Fig 4B, SI Fig 6A). These cross-subject consistency measures suggest that with an appropriate hypothesis RDM we could have improved the correlation values obtained with our theoretically motivated hypothesis RDM (Syllable RDM). However, since these correlation values still differed across the four conditions our conclusions that neural representations of sensory detail and prior knowledge are in line with our Prediction Error simulation would still hold. Indeed, the good correspondence seen between simulated and observed multivariate analyses of cross-subject consistency further strengthens this conclusion.

References

1. Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP. Lexical influences on auditory streaming. Curr Biol. 2013;23(16):1585-9.

2. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Comput Biol. 2014;10(4):e1003553.

3. Evans S, Davis MH. Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. Cerebral cortex. 2015;25(12):4772-88.

4. Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive Top-Down Integration of Prior Knowledge during Speech Perception. J Neurosci. 2012;32(25):8443-53.

5. Eisner F, McGettigan C, Faulkner A, Rosen S, Scott SK. Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. J Neurosci. 2010;30(21):7179-86.

6. Obleser J, Kotz SA. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb Cortex. 2010;20(3):633-40.

7. Lee YS, Turkeltaub P, Granger R, Raizada RDS. Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based Analysis. J Neurosci. 2012;32(11):3942-8.

8. Du Y, Buchsbaum BR, Grady CL, Alain C. Noise differentially impacts phoneme representations in the auditory and speech motor systems. P Natl Acad Sci USA. 2014;111(19):7126-31.

9. Sohoglu E, Davis MH. Perceptual learning of degraded speech by minimizing prediction error. Proc Natl Acad Sci U S A. 2016;113(12):E1747-56.

10. Sehm B, Schnitzler T, Obleser J, Groba A, Ragert P, Villringer A, et al. Facilitation of Inferior Frontal Cortex by Transcranial Direct Current Stimulation Induces Perceptual Learning of Severely Degraded Speech. J Neurosci. 2013;33(40):15868-78.

11. Steiger JH. Tests for Comparing Elements of a Correlation Matrix. Psychol Bull. 1980;87(2):245-51.

12. Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT. Categorical speech representation in human superior temporal gyrus. Nature Neuroscience. 2010;13(11):1428-U169.

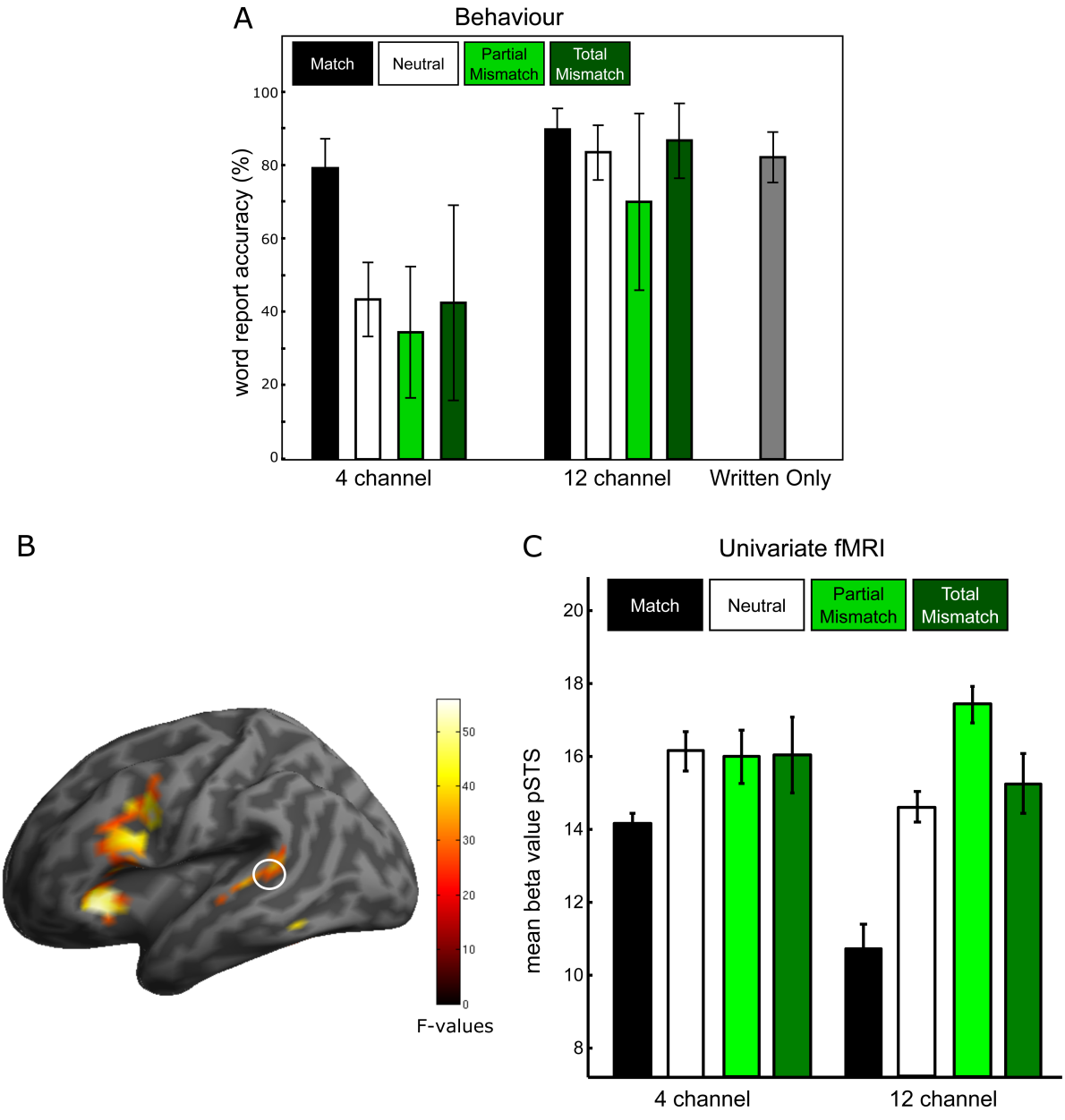
13. Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 2008;322(5903):970-3.

14. Erez Y, Duncan J. Discrimination of Visual Categories Based on Behavioral Relevance in Widespread Regions of Frontoparietal Cortex. J Neurosci. 2015;35(36):12383-93.

15. Correia JM, Jansma BMB, Bonte M. Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. J Neurosci. 2015;35(44):15015-25.

S1 Fig. Effect of mismatching prior expectations.

(A) Behavioural results. (B) Univariate results: Main effect of prior knowledge (Matching versus Mismatching Prior) depicted on a rendered brain ($p < 0.05$ voxelwise FWE, $n = 21$). (C) Mean beta values extracted from the independent region of interest in the posterior STS [57] illustrate reduced BOLD signal during Match conditions (solid black) in contrast to Neutral (white) and Mismatch (green) conditions. Error bars indicate standard error of the mean after between-subject variability has been removed suitable for repeated measures comparisons [62]. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

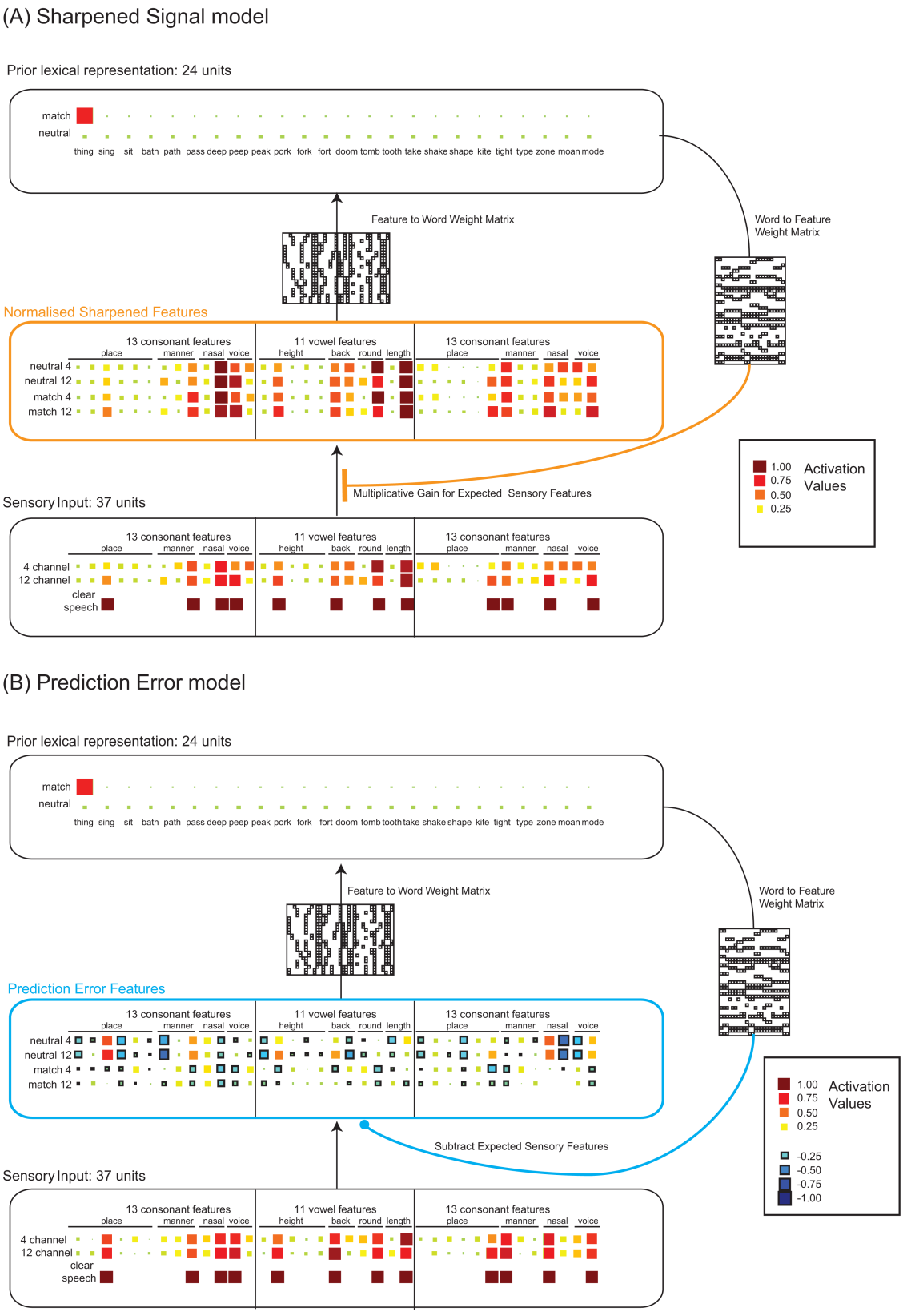


S2 Fig.

Network architecture and example representations for (A) Sharpened Signal and (B) Prediction Error models. Common components of both models are outlined in black. Differences between the two models are coloured in orange (Sharpened Signal) and blue (Prediction Error). Both models map from a feature-based representation of consonant-vowel-consonant symbols that have been degraded by the addition of random, probabilistic noise within the different groups of units representing specific feature types (place, manner, voicing, etc.). Input for the word "thing" is shown for both models, using representations degraded to simulate 4-channel and 12-channel noise vocoded speech (based on clarity parameters fit for each of the simulations). A clear speech (un-degraded) representation of the word "thing" is shown for comparison, though this wasn't presented to either model. Hinton diagrams show the activation of each individual unit with the area of the squares proportional to activation values or probabilities, supplemented by colour scales as shown. In both models, lexical representations are specified over a bank of 24 localist units (one for each word in the models' vocabulary and experimental item set). These lexical representations are initialised to express the prior probability of each word being presented based on prior written text ("THING," Match condition) or a neutral string ("XXXX," Neutral condition). In both models, a word-to-feature matrix links words to their constituent phonetic features and a feature-to-word matrix links phonetic features to words (these two matrices are the transpose of each other). There are some key differences between the two models. In the Sharpened Signal model (A), prior knowledge is used to increase the gain of expected sensory features, such that expected features are preferentially activated in Sharpened Feature representations at the intermediate level of the model. These Sharpened Features are then used to update lexical representations. Thus, Match trials lead to Sharpened Feature representations that resemble those from speech signals with greater sensory detail. In contrast, in the Prediction Error model (B), expected sensory features are subtracted from the observed sensory input, and Prediction Error feature representations at the intermediate level are used to update lexical representations. These Prediction Error representations contain negative values (blue colours) for expected features that are presented in a degraded form; these negative prediction errors carry information concerning the identity of the speech signal in Match 4 trials that is absent for Match 12 trials in which speech is less degraded.

doi:10.1371/journal.pbio.1002577.s002

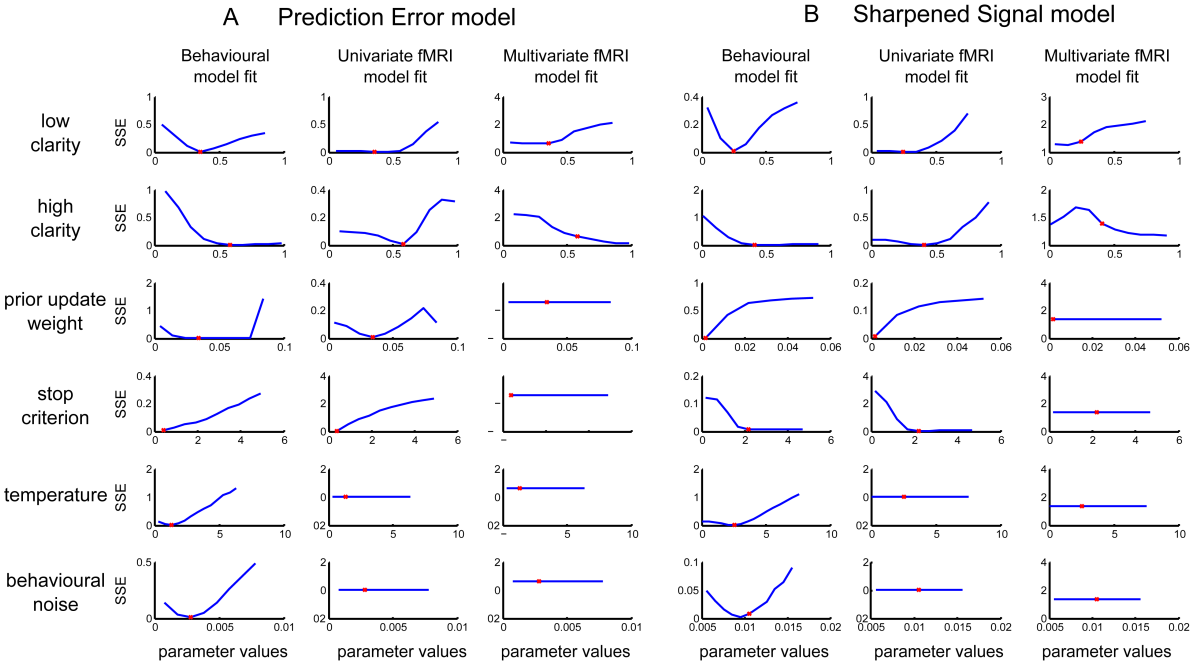
S2 Fig.


(A) Sharpened Signal model
Prior lexical representation: 24 units
match
neutral
thing sing sit bath path pass deep peep peak pork fork fort doom tomb tooth take shake shape kite tight type zone moan mode
Feature to Word Weight Matrix
Word to Feature Weight Matrix
Normalised Sharpened Features
13 consonant features
place manner nasal voice
11 vowel features
height back round length
13 consonant features
place manner nasal voice
neutral 4
neutral 12
match 4
match 12
Multiplicative Gain for Expected Sensory Features
Sensory Input: 37 units
4 channel
12 channel
clear speech
1.00
0.75
0.50
0.25
Activation Values
(B) Prediction Error model
Prior lexical representation: 24 units
Feature to Word Weight Matrix
Word to Feature Weight Matrix
Prediction Error Features
Subtract Expected Sensory Features
-0.25
-0.50
-0.75
-1.00

S3 Fig. Sensitivity analysis.

(A) Prediction Error model. (B) Sharpened Signal model. The blue curves illustrate how the sum squared error (SSE, y-axis) for model fit to the behavioural (left column), univariate fMRI (middle columns), and multivariate fMRI (right columns) data changes for a range of parameters (along the x-axis). Each graph therefore shows the influence of each of the six parameters: (1) low clarity, (2) high clarity, (3) prior update weight, (4) stopping criterion, (5) temperature, and (6) behavioural noise on model fit. The red dot on each graph indicates the final parameters chosen by nonlinear optimisation. Univariate and multivariate fMRI data come from ROI coordinates based on univariate analysis (Fig 3C). Please refer to S2 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

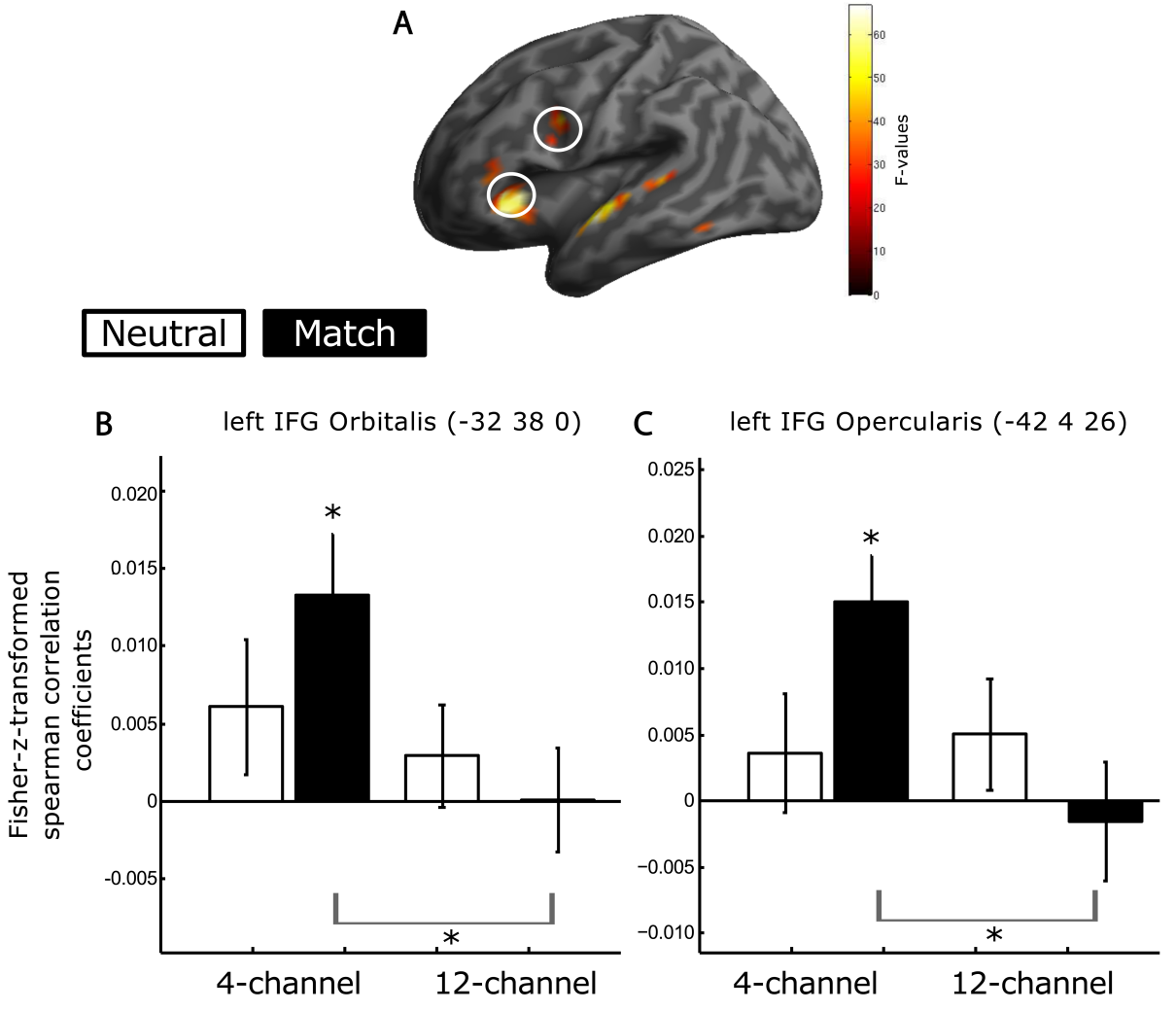
doi:10.1371/journal.pbio.1002577.s003



S4 Fig. Representation of phonetic form in Inferior Frontal regions

(A) Univariate results: Main effect of prior knowledge (Matching versus Neutral Prior) depicted on a rendered brain ($p < 0.05$ voxelwise FWE, $n = 21$). White circle marks post-hoc defined clusters of interest in the left Inferior Frontal Gyrus (IFG). (B,C) Fisher-z-transformed Spearman correlation coefficients for each of the four conditions in two left IFG clusters (defined by the univariate analysis) show a significant correlation in the Match 4-channel condition and a significant reduction in correlation with increased sensory detail Match 4-channel compared to Match 12-channel. Error bars indicate standard error of the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons [62]. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

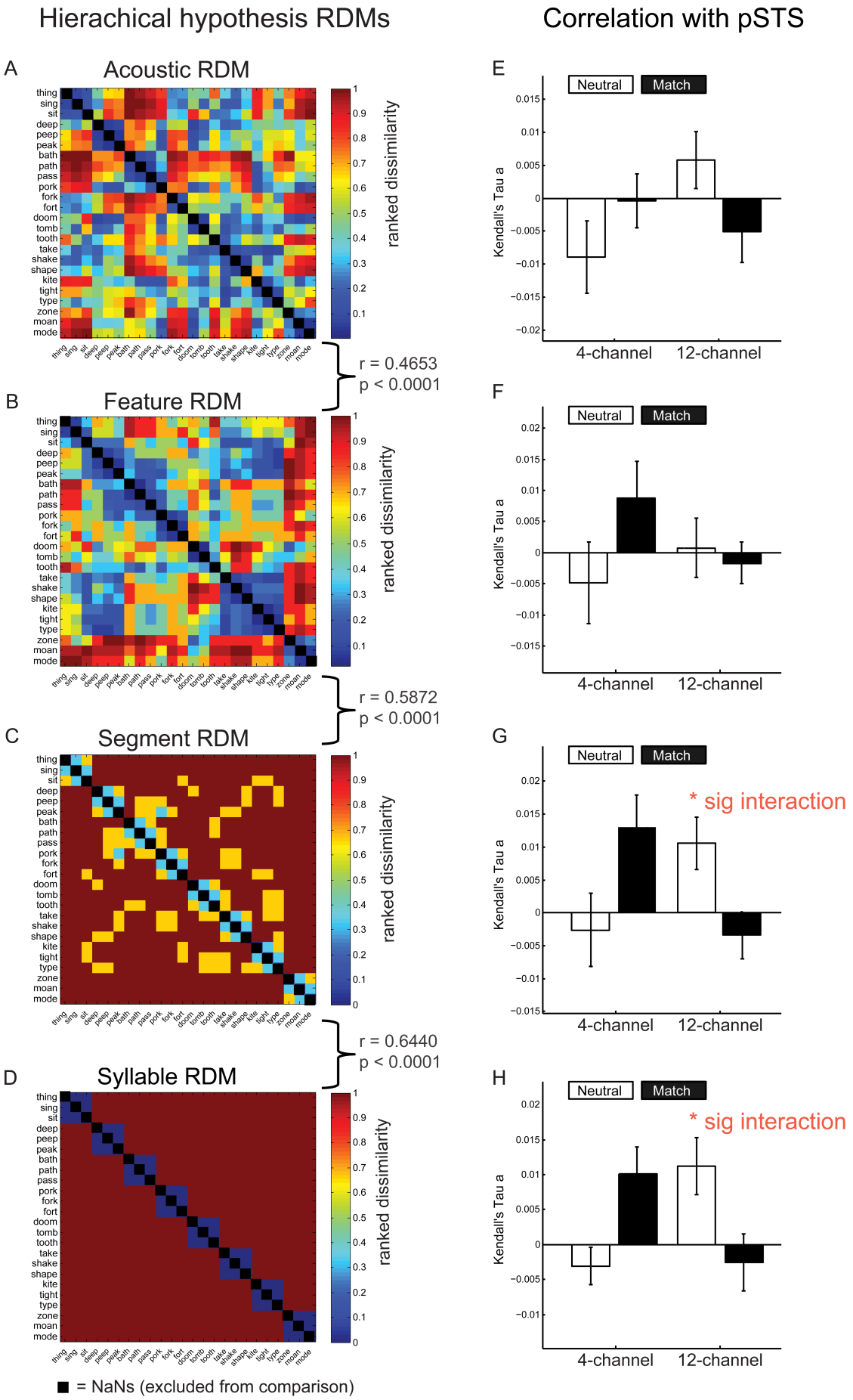
doi:10.1371/journal.pbio.1002577.s004



S5 Fig. Comparison of four different, hierarchically organised hypothesis RDMs of speech perception.

Left Panel: (A) dissimilarity of the acoustic properties of the speech stimuli used in our study (see Supplementary Methods for details), (B) dissimilarity of feature representation for the canonical forms of the speech provided as the input to our computational simulations, (C) dissimilarity of the segment representations of the word stimuli used in the experiment, scored based on the number of position-specific phonemes shared between words pairs, and (D) main hypothesis RDM assuming increased similarity between pairs of syllables that shared the same vowel (e.g., "sing" and "thing" should have more similar patterns than "sing" and "bath"). These RDMs can be considered to describe a hierarchy of speech representations from the fine-grained acoustic RDM to the most abstract syllable RDM used in our main analysis. These hypothesis RDMs are positively correlated with each other and hence can be considered as testing related proposals concerning neural representations of spoken words. Right panel (E–H) shows the results for the Kendall's Tau A correlation coefficients (suitable for comparisons between binary and fine-grained RDMs; see Supplementary Methods for details) as extracted from the independent region of interest in the left posterior STS (pSTS, Fig 4B). Only the segment (G) and the syllable RDM (H) revealed a significant interaction of sensory detail and prior knowledge, similar to that shown in Fig 4B. Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

doi:10.1371/journal.pbio.1002577.s005

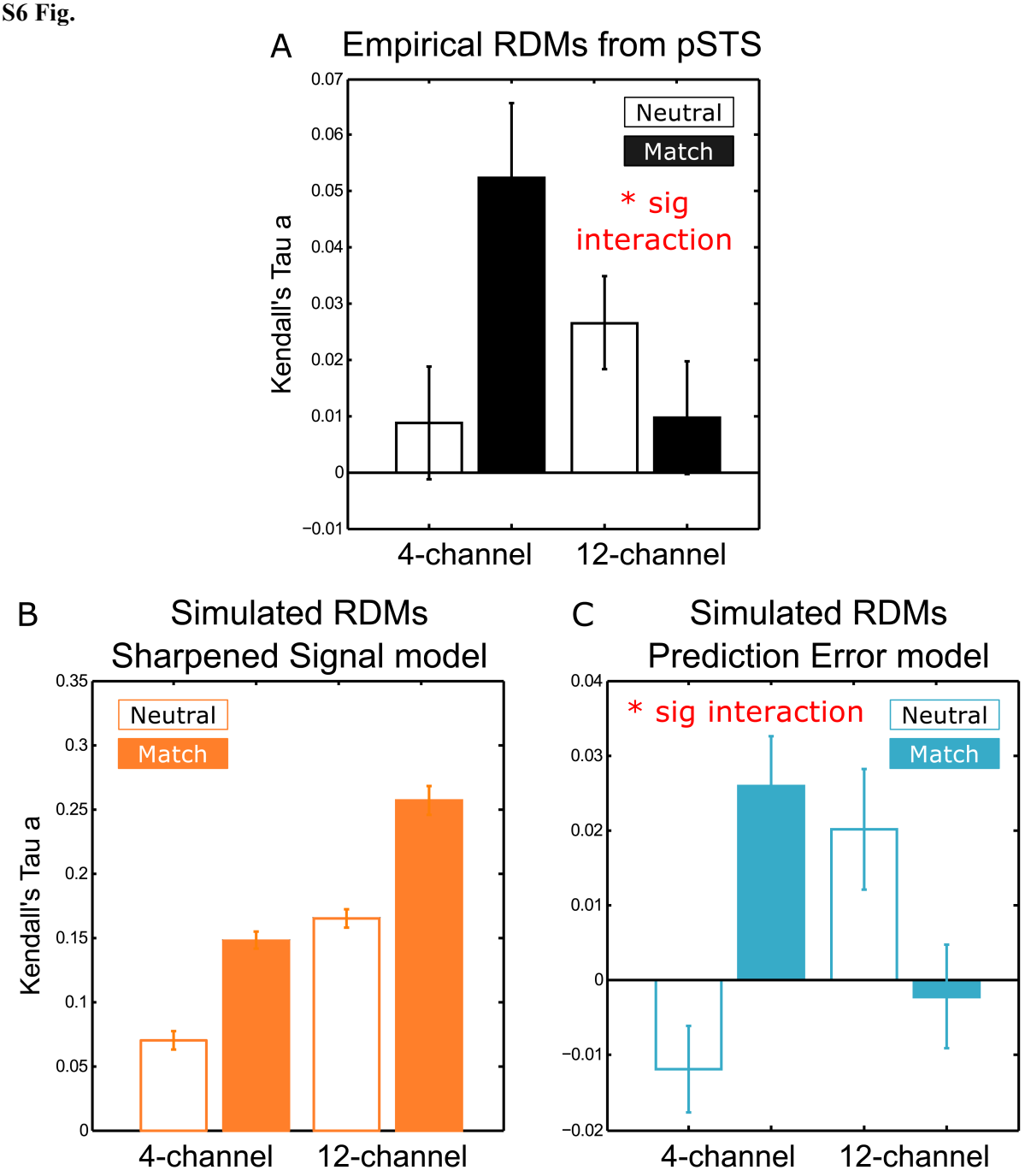

Hierachical hypothesis RDMs
Correlation with pSTS
A
Acoustic RDM
B
Feature RDM
C
Segment RDM
D
Syllable RDM
thing
sing
sit
deep
peep
peak
bath
path
pass
pork
fork
fort
doom
tomb
tooth
take
shake
shape
kite
tight
type
zone
moan
mode
ranked dissimilarity
r = 0.4653
p < 0.0001
r = 0.5872
p < 0.0001
r = 0.6440
p < 0.0001
■ = NaNs (excluded from comparison)
E
F
G
H
Neutral
Match
Kendall's Tau a
4-channel
12-channel
* sig interaction

S6 Fig. Cross-subject consistency based on empirical and simulated RDMs.

(A) Empirical RDMs were extracted from the independent ROI in the left posterior STS (pSTS, Fig 4B), and the Simulated RDMs based on either (B) the Sharpened Signal or (C) the Prediction Error model were computed for 21 simulated participants. The cross-subject consistencies from the empirical RDMs and simulated RDMs from the Prediction Error model show the same crossover interaction of sensory detail and prior knowledge shown before (Fig 4B–4D). Please refer to S1 Data at https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N) for the numerical values underlying these figures.

doi:10.1371/journal.pbio.1002577.s006

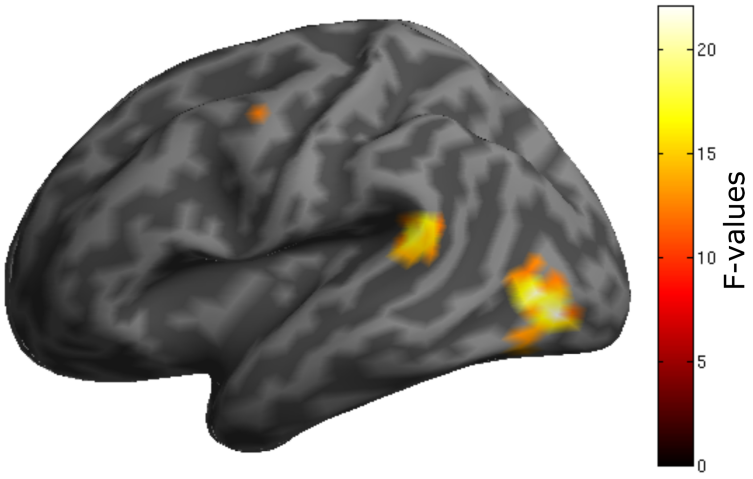
S6 Fig.


A
Empirical RDMs from pSTS
Kendall's Tau a
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
−0.01
Neutral
Match
* sig
interaction
4-channel
12-channel
B
Simulated RDMs
Sharpened Signal model
Kendall's Tau a
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Neutral
Match
4-channel
12-channel
C
Simulated RDMs
Prediction Error model
0.04
0.03
0.02
0.01
0
−0.01
−0.02
* sig interaction
Neutral
Match
4-channel
12-channel

S7 Fig. Representational similarity searchlight analysis in the whole brain.

Interaction of Prior information (Match/Neutral) x Sensory detail (4- versus 12-channel) depicted on rendered brain (F-contrast, $p < 0.001$ uncorrected, $k > 10$ voxels; searchlight analysis with a voxel size of 3 x 3 x 3.75 mm; see S4 Table for coordinates). https://osf.io/2ze9n/ (doi: 10.17605/OSF.IO/2ZE9N).

doi:10.1371/journal.pbio.1002577.s007



S1 Table. Univariate Analysis—F-contrast: Main effect Match/Neutral, p < 0.05 FWE (voxelwise correction)

doi:10.1371/journal.pbio.1002577.s008

Cluster	Peak p(FWE-corr)	peak F	x,y,z (mm)	Anatomy label
229	0	80.28	-58 -10 -6	Left middle temporal Gyrus (anterior)
623	0	76.14	-32 28 0	Left inferior frontal gyrus
	0.001	42.18	-52 32 6	Left middle temporal gyrus
	0.002	39.13	-44 28 10	
410	0	71.96	-8 20 42	Left superior frontal gyrus
	0.001	44.28	-10 26 30	
184	0	52.31	-42 4 26	Left inferior frontal gyrus
99	0	49.53	-52 -38 6	Left middle temporal gyrus (posterior)
	0	47.78	-60 -32 6	
161	0.001	43.35	28 22 -4	Right insula
92	0.002	40.41	52 -12 -6	Right superior temporal gyrus
	0.012	33.24	62 -14 0	
26	0.002	40.31	24 -74 -50	Right cerebellum
55	0.002	38.86	-48 -48 -14	Left inferior temporal gyrus
	0.008	34.36	-44 -40 -12	
24	0.004	36.79	-4 -14 2	Left thalamus
31	0.008	34.65	12 22 36	Right middle cingulate cortex
2	0.041	28.84	14 26 24	Right anterior cingulate cortex
1	0.047	28.4	10 18 40	Right middle cingulate cortex

S2 Table. Univariate Analysis—F-contrast: Main effect sensory detail, p < 0.05 FWE (voxelwise correction)

doi:10.1371/journal.pbio.1002577.s009

Cluster	peak p(FWE-corr)	peak F	x,y,z (mm)	Anatomy label
932	0	97.67	-32 28 0	Left inferior frontal gyrus
	0	64.91	-40 18 6	
376	0	82.32	-6 20 48	Superior medial gyrus
	0	45.93	8 20 44	
284	0	57.97	32 26 -2	Right insula lobe
292	0.001	44.31	-42 4 28	Left inferior frontal gyrus
32	0.001	43.41	6 58 -8	Right middle orbitofrontal gyrus
1	0.043	28.67	-52 -38 6	Left middle temporal gyrus

S3 Table. Univariate Analysis—F-contrast: Prior information (Match/Neutral) x Sensory detail full interaction, p < 0.001 uncorrected, k > 10 voxels

doi:10.1371/journal.pbio.1002577.s010

Cluster	peak p(FWE-corr)	peak F	peak equivZ	x,y,z (mm)	Anatomy label
127	0.142	24.59	4.37	16 -62 -22	Right cerebellum
	0.995	13.55	3.29	6 -62 -20	
	0.999	12.86	3.20	4 -58 -10	
257	0.400	20.76	4.05	-26 -52 -28	Left cerebellum
	0.872	16.46	3.62	-16 -60 -22	
	0.925	15.74	3.55	-2 -48 -30	
41	0.632	18.67	3.85	24 -6 -2	Right amygdala/pallidum
	0.999	12.86	3.20	26 -6 -12	
42	0.662	18.41	3.82	-58 -8 -8	Left middle temporal gyrus
16	0.751	17.66	3.75	-32 -12 16	Left insula lobe
21	0.799	17.21	3.70	-50 -10 -14	Left middle temporal gyrus
11	0.946	15.40	3.51	-4 -56 -8	Left cerebellum
16	0.977	14.60	3.42	22 -52 -50	Right cerebellum
10	0.990	14.01	3.35	22 -44 -42	Right cerebellum

S4 Table. Univariate Analysis—F-contrast: Main effect Match/Mismatch, p < 0.05 FWE (voxelwise correction)

doi:10.1371/journal.pbio.1002577.s011

Cluster	peak p(FWE-corr)	peak F	peak equivZ	x,y,z (mm)	Anatomy Label
211	0	61.96	6.83	-8 22 46	Left SMA
436	0	61.76	6.82	-30 26 -2	Left insula lobe
1133	0	60.55	6.76	-44 6 26	Left inferior Frontal gyrus
	0	48.38	6.16	-42 2 34	
	0	40.98	5.73	-46 18 20	
240	0	55.71	6.54	-54 -36 6	Left middle temporal gyrus
	0.003	34.37	5.3	-54 -22 -4	
80	0	46.24	6.04	-48 -48 -14	Left inferior temporal gyrus
151	0.001	38.29	5.56	32 28 -2	Right insula lobe
16	0.001	37.72	5.52	26 -72 -50	Right cerebellum

S5 Table. RSA—F-contrast: Prior information (Match/Neutral) x Sensory detail full interaction, p < 0.001 uncorrected, k > 10 voxels (searchlight analysis with a voxel size of 3 x 3 x 3.75 mm)

doi:10.1371/journal.pbio.1002577.s012

Cluster	peak p(FWE-corr)	peak F	peak equivZ	x,y,z (mm)	Anatomical Label
160	0.028	25.01	4.41	-39 -76 2	Left middle occipital gyrus
97	0.197	18.57	3.84	-60 -46 14	Left superior temporal gyrus
36	0.372	16.29	3.6	-36 -7 40	Left precentral gyrus