# Convergent neural signatures of speech prediction error are a biological marker for spoken word recognition

**Ediz Sohoglu[1], Loes Beckers[2,3,4] and Matthew H. Davis[2]**

1) School of Psychology, University of Sussex, Brighton, UK
2) MRC Cognition and Brain Sciences Unit, Cambridge, UK
3) Department of Otorhinolaryngology, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, Netherlands
4) Cochlear Ltd., Mechelen, Belgium

Keywords: Predictive coding, Speech perception, MEG, fMRI

**Acknowledgements:** This research was supported by intermural funding from the Medical Research Council to the Cognition and Brain Sciences Unit (MC_UU_0005/5 and MC_UU_00030/6 to M.H.D.). We thank Sander Van Bree and Clare Cook for assistance with MEG data collection and Steve Eldridge and Karen Kabakulu for assistance with radiography.

**Address correspondence to:**

Ediz Sohoglu (E.Sohoglu@sussex.ac.uk) or Matthew Davis (matt.davis@mrc-cbu.cam.ac.uk)

# Abstract

We used MEG and fMRI to determine how predictions are combined with speech input in superior temporal cortex. We compared neural responses to bisyllabic spoken words in which the first syllable strongly or weakly predicts the form of the second syllable (e.g. “bingo” versus “tango”). We further compared neural responses to the same second syllables when heard in an unfamiliar pseudoword and therefore in a situation in which predictions mismatch with sensory input (e.g. “snigo” and “meago”). Across multiple imaging modalities and analysis approaches (including information-based pattern analysis methods), we show that neural representations of second syllables are suppressed by strong predictions when predictions match sensory input. However, neural representations of the same second syllables show the opposite effect (i.e. enhanced representations following strongly than weakly-predicting syllables) when predictions mismatch with sensory input. Computational simulations show that this interaction between prediction strength and (mis)match is consistent with prediction error but not alternative (sharpened signal) computations. Neural signatures of prediction error are observed early (beginning 200 ms after the onset of the second syllable), localise to early auditory regions (in fMRI, bilateral Heschl’s gyrus and STG) and are expressed as changes in low-frequency (theta and alpha) power. Our study therefore provides convergent neural evidence that speech perception is supported by the computation of prediction errors in auditory brain regions. These prediction error computations play a central role in the identification of familiar spoken words and perception of unfamiliar pseudowords.

# Introduction

The information reaching our senses is incomplete and ambiguous such that perception requires active computations for successful interpretation. Nowhere is this more evident than with speech perception. To paraphrase Heraclitus, “No listener ever hears the same speech twice, for it’s not the same sounds and they’re not the same listener”. No two productions of the same word are ever acoustically identical; each will be differently realised due to the ubiquitous variability of natural speech produced by different talkers, and in different contexts. No listener is ever the same as they will have different expectations or prior knowledge for the speech they will hear at different times. Consider for example, hearing a word starting with the syllable “tan”. It can be unclear whether this is the start of the word “tangle”, “tango” or another word that starts with the same initial sounds. This ambiguity can be present even if speech is clearly heard in a quiet background. Yet when listening during everyday conversation, our subjective impression is that spoken word recognition is immediate, clear and accurate. How do listeners achieve such reliable perception despite transient sensory ambiguity?

In answering this question, the notion of Bayesian inference is commonly invoked. That is, listeners combine sensory evidence with prior knowledge about which speech sounds they will hear [1]. This proposal is appealing because it can explain how listeners process speech both rapidly and accurately [2,3]. However, despite widely held agreement that listeners use prior knowledge during speech processing, it remains unclear how this is implemented neurally. One possibility is that neural representations of speech are enhanced by prior knowledge [4]. This has been termed the sharpened signal account [5–7] because in this proposal, neural responses to expected stimuli come to resemble a sharpened (i.e. less ambiguous or less noisy) version of the sensory input. However, it is also the case that expected speech sounds carry relatively little information. Therefore, an alternative and potentially more efficient strategy would be to compute prediction errors i.e. the difference between heard and predicted sounds [8–11]. In this prediction error account, neural representations are suppressed by prior knowledge and unexpected (and therefore novel and unexplained) sensory information is prioritised for processing.

Sharpened signal and prediction error accounts have been difficult to distinguish experimentally. This is because in both accounts, prior knowledge has a similar effect on the overall strength of neural responses [5,6]. In the case of the prediction error account, brain responses tuned to the heard stimulus features are suppressed when those stimulus features are expected, resulting in an overall reduction in brain activity. With the sharpened signal account, although brain responses tuned to the heard stimulus features are enhanced, this is accompanied by the suppression of responses tuned away from the stimulus input features i.e. competing interpretations. This suppression of brain responses tuned away from the input features also leads to overall weaker activity. Therefore, previous findings that neural responses correlate with moment-by-moment phoneme and word predictability during story comprehension [12–16] do not provide unambiguous support for prediction error over sharpened signal accounts. Some accounts have proposed that sharpened signal and prediction error computations operate in parallel [8], and localise to different cortical layers [17]. Nonetheless, we focus on contrasting sharpened signal and prediction error mechanisms since single mechanism accounts have been proposed in speech perception [4,18,19] and remain to be distinguished [20].

Multivariate pattern analysis methods that permit measurement of the information conveyed by neural responses can distinguish between sharpened signal and prediction error computations. Indeed, previous work using fMRI [18] and MEG [21] has demonstrated a distinctive neural marker of prediction errors which is that neural representations of sensory features of speech show an interaction between signal quality and prior knowledge. This interaction arises because sensory signals that match strong prior expectations are explained away more effectively as signal quality increases and hence neural representations are suppressed even as perceptual outcomes improve. Whereas for sensory signals that follow less informative prior knowledge, increased signal quality leads to a corresponding increase in sensory information that remains unexplained. By contrast, under a sharpened signal account, computational simulations show that neural representations (like perceptual outcomes) are similarly enhanced by increased signal quality and prior knowledge [18,21]. Therefore, by experimentally manipulating both signal quality and prior knowledge, and observing the consequences on pattern-based neural measures, it is possible to distinguish between accounts.

However, in the aforementioned studies [18,21], listeners heard highly distorted (and unnatural sounding) spoken words while prior knowledge was manipulated by presenting matching or nonmatching written words before speech was heard. This is far removed from real-world conditions in which speech is clearly heard (non-distorted) and predictions are not derived from external written cues but from linguistic knowledge e.g. word probabilities vary based on usage frequency and preceding sentence context. Indeed, some investigators have argued that previously documented effects of prior knowledge or prediction on language processing are the consequence of “prediction encouraging” paradigms and fail to generalise to settings more representative of real-world perception [22].

In the current study, we return to this issue by presenting listeners with clearly spoken words. We experimentally manipulated the strength and accuracy of listeners’ prior knowledge with the goal of adjudicating between sharpened signal and prediction error accounts. Here our manipulation of prior knowledge is based on information intrinsic to the speech signal and arises from listeners’ long-term linguistic knowledge of the sounds of familiar spoken words. We compare neural responses to bisyllabic spoken words (e.g. “bingo”, “tango”) in which the first syllable strongly (in “bingo”) or weakly (in “tango”) predicts the form of the second syllable (Figure 1A). In addition, we compare neural responses to the same second syllables when heard in a pseudoword context (e.g. “snigo”, “meago”). Pseudowords, by definition, are unfamiliar to participants and therefore the second syllable of these items mismatches with listeners’ predictions. This creates a listening situation in which strong predictions entirely mismatch with sensory input. This cannot be observed in other experimental situations that forego manipulation of speech content (e.g. story comprehension) and creates a further opportunity for adjudicating between sharpened signal and prediction error computations.

We first performed computational simulations (using model architectures and representations schematised in Figure 1B, C and Figure S6) to show that, as intended, our experimental manipulations of prediction strength and match/mismatch result in dissociable effects for sharpened signal and prediction error computations (Figure 1E). While prediction strength and prediction congruency act to enhance sharpened signal representations, prediction error computations show an interaction between prediction strength and congruency. This interaction is driven by a reduction of prediction errors when strong predictions match the sensory input compared with

weak predictions. When predictions mismatch with sensory input, however, the opposite occurs with prediction errors slightly increasing with prediction strength. We then measured neural responses using 204-channel MEG and 3T fMRI in separate groups of listeners while they performed an incidental (pause detection) listening task to maintain attention (Figure S1A). Across multiple imaging modalities (MEG, fMRI), analysis approaches (univariate signal magnitude and multivariate pattern), and signal domains (time and time-frequency), we provide convergent neural evidence that speech perception is supported by the computation of prediction errors in auditory brain regions.

# Figure 1

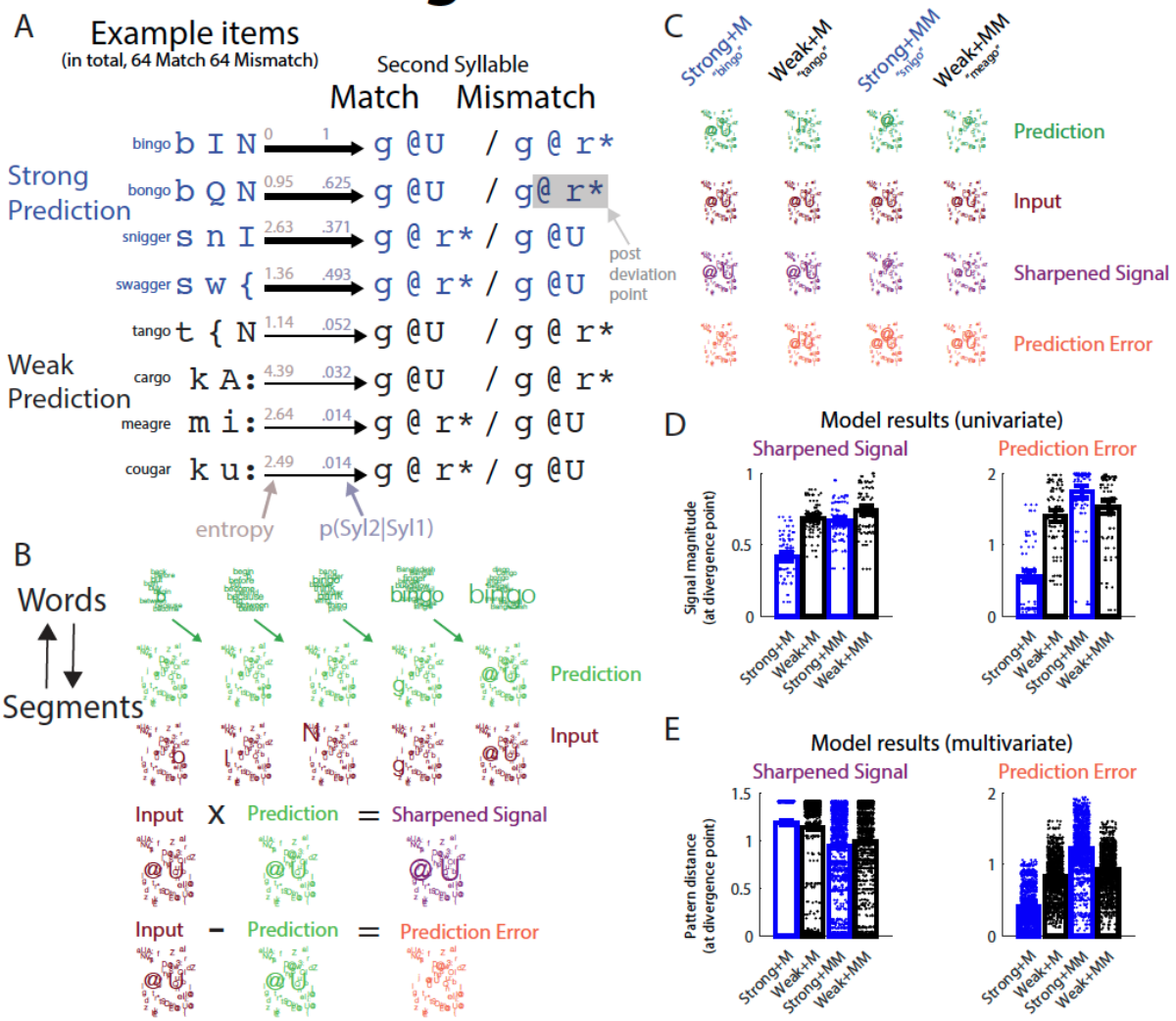



*Figure 1- (A) Our stimuli consisted of multiple sets of bi-syllabic spoken words ("bingo", "bongo", etc. transcribed phonetically as /b I N – g @U/, /b Q N – g @U/) and pseudowords ("bingger", "bongger", etc. transcribed /b I N – g @ r*/, /b Q N – g @ r*/). Purple numbers indicate the strength of predictions for the second syllable of each item, operationalized as p(Syl2|Syl1): the probability of hearing the second syllable, conditioned on the first*

*syllable (the negative log of this probability is equivalent to syllable surprisal; see [23]). Brown numbers indicate the entropy of predictions over all possible syllables (equivalent to the weighted average of syllable surprisal over syllables [23]). Thickness of arrows depicts the overall higher p(Syl2|Syl1) and lower entropy for Strong vs Weak Prediction items. By cross-splicing the syllables in our word items, we created a set of pseudowords in which predictions mismatched with Syl2. The earliest speech segment when the words diverged from the pseudowords (the divergence point) is indicated in bold and always occurred at the second segment of Syl2. (B) Model representations for the example word "bingo", visualized using word and segment clouds with larger text indicating higher probabilities (for a more complete illustration of these computations, see Figure S6). Both sharpened signal and prediction error computations are derived from conditional (posterior) word probabilities, as estimated using Bayes theorem by combining prior word probabilities (frequency of occurrence) and the sensory evidence. Here the sensory evidence is derived from the acoustic similarity between individual speech segments and expresses the degree to which the sensory input matches internal representations of spoken words. Word level representations show posterior probabilities for each word as each segment in the input word is heard prior to the divergence point (only the ten words with the highest probabilities at any one timepoint are shown for illustration purposes). Word posteriors provide predictions for which segment will be heard next (segment predictions shown as green colored phonetic transcriptions). Segment predictions are then combined with the sensory evidence for individual segments ('Input'; shown as brown colored phonetic transcriptions) through sharpened signal (multiplication) or prediction error (subtraction) computations with outcomes depicted on the right as purple and orange segment clouds, respectively (shown only for the divergence point). Note that the segment probabilities in these simulations are not directly related to the stimulus probabilities in panel A as the simulations are based on segment representations and assume some uncertainty in the sensory input (using acoustic similarity). In contrast the stimulus probabilities in Figure 1A, which were used to select the stimuli, are based on syllable representations and assume ideal observer behavior (i.e. no sensory uncertainty). (C) Example model representations for all four experimental conditions at the divergence point for Strong + Match ("bingo"), Weak + Match ("tango"), Strong + Mismatch ("snigo"), Weak + Mismatch ("meago") items. Due to differential predictions (top-line, green), identical speech input (/@U/, second line, brown), elicits different neural representations for sharpened signal (third line, purple) and prediction error (fourth line, orange) representations. (D) Results from sharpened signals and prediction error simulations at the divergence point, used to predict the magnitude of neural responses (univariate signal magnitude analysis). Sharpened signals were summarised as the sum of log activity over segment units while prediction errors were summarised as the sum of absolute activity. Dots indicate the model signals for each individual item (32 items in each condition) with bars and error bars indicating the mean and standard error across the full item set). (E) Results from simulations used to predict neural pattern distances, shown in the same way as panel D except the individual datapoints and bars depict the dissimilarity between model representations for all pairs of items (496 item pairs, i.e. 32x31/2) within each condition (multivariate pattern analysis).*

# Results

## Stimulus properties and behavioural responses

To manipulate listeners' predictions, we chose a set of 64 bisyllabic words in which the first syllable (Syl1) either strongly or weakly predicts the same second

syllable (Syl2; see Figure 1A which depicts some example items). Here we operationalise prediction strength as the probability of hearing Syl2, conditioned on Syl1 (p(Syl2|Syl1); see Methods section for details and refs. [2,11,12,23]). For example, after hearing the first syllable /b I N/, there is only one potential word candidate (*bingo*). Therefore, based on their experience of spoken English, listeners will predict that the second syllable /g @U/ is most likely after hearing /b I N/ (p(Syl2|Syl1) = 1). In contrast, after hearing the first syllable “tan”, there are more potential word candidates (e.g. *tango, tangle, tanker, tankard, tang*) and of these words, *tango* is less frequently used than the competitors *tangle* and *tang*. As a result, listeners can only make a weak prediction for /g @U/ after hearing /t { N/ (p(Syl2|Syl1) = .052). Note that p(Syl2|Syl1) can also be expressed as syllable surprisal [see 23], which is equivalent to the negative log of p(Syl2|Syl1).

We validated our experimental manipulation of prediction strength by asking a separate group of listeners to perform a free report gating task [see ref 24 and SI Methods and Results]. This confirmed that listeners were more likely to predict the second syllables of our Strong items than for Weak items. Further validation of our experimental manipulation is provided when assessing listeners’ memory for the pseudowords in a cued recall task (see SI Results).

In addition to manipulating the strength of listeners’ predictions, we manipulated whether Syl2 matched or mismatched with listeners’ predictions. This was achieved by cross-splicing Syl1 and Syl2 between words to create a set of corresponding pseudowords sharing the same set of syllables as the word items. For real word items, Syl2 always matches predictions, at least to some degree. For pseudowords, p(Syl2|Syl1) approaches zero. This is by definition, since these Syl1 and Syl2 combinations will never previously have been heard by participants as English spoken words prior to the experiment. Therefore, for these items, Syl2 entirely mismatches with listeners’ predictions.

It is important to note that even though p(Syl2|Syl1) approaches zero for all Mismatch items, we can still expect a difference in prediction strength between Strong+Mismatch and Weak+Mismatch items since Strong and Weak items differ not only in terms of p(Syl2|Syl1), but also in entropy over all possible second syllables i.e. the overall uncertainty of predictions (see SI Results and Figure S4A). Indeed in natural speech, conditional probabilities for upcoming speech sounds are negatively correlated with entropy [12,14,23]. Going back to the example above, after hearing

the first syllable /b I N/, entropy is relatively low because only one word (*bingo*) can be predicted. In this case, the unexpected second syllable (/g @ r*/) follows a state of precise predictions (low uncertainty/entropy). For *tango* on the other hand, entropy is relatively high after hearing the first syllable because predictions are made for multiple second syllables each arising from different word candidates. Here the unexpected second syllable follows a state of imprecise predictions (high uncertainty/entropy). Thus, while we constructed our stimuli based on p(Syl2|Syl1), our manipulation of prediction strength reflects differences in both p(Syl2|Syl1) and syllable entropy. Both aspects of predictability may contribute to perceptual and neural responses in the Match (real word) condition. In the Mismatch condition however, effects of prediction strength can only reflect differences in entropy (as explained above).

To assess neural responses to these stimuli, we recorded MEG (N=19) and fMRI (N=21) data in two separate groups of listeners (Figure S1A). To maintain listeners' attention, participants were asked to detect brief (200 ms) pauses inserted between Syl1 and Syl2 in occasional target items. As shown in Figure S1B and S1C, participants could perform this task quickly and accurately (d-prime higher than 3 and with response times on target-present trials less than 1100 ms).

**Computational simulations**

We performed simulations of sharpened signal and prediction error computations supporting spoken word and pseudoword recognition (model architectures and representations schematised in Figure 1B, C with further illustration of model computations provided in Figure S6). The underlying mechanisms were derived from established Bayesian models of spoken word recognition [e.g. 2] guided by observations of functionally equivalent but neurally distinct implementations of these. We focus on the timepoint at which the words and pseudowords in our stimuli diverge (the divergence point; highlighted as bold speech segments in Figure 1A). This is the time at which our experimental manipulation of prediction congruency can affect brain responses: in phonemes, this is at the onset of the second speech segment during Syl2 (98 ms after Syl2 onset on average, with a standard deviation of 46 ms).

Both sharpened signal and prediction error computations are derived from conditional (posterior) word probabilities, as estimated using Bayes theorem (for details, see Methods). As with previous work [e.g. 2,11,12], predictions for the next speech segment are estimated from the relative frequencies of the words matching the preceding speech signal, which is appropriate for spoken words heard in isolation (see Methods for details). Our estimates of the sensory evidence are derived from the acoustic similarity between individual speech segments and expresses the degree to which the sensory input matches internal representations of spoken words [2,25]. Because our stimuli consisted of clear speech recordings, we simulated conditions in which the sensory evidence was high (i.e. there was minimal sensory uncertainty; see Methods) but dissociable sharpened signal and prediction error computations are observed when simulating a range of sensory uncertainty levels (Figure S3).

To relate these simulations to observed neural responses, we apply signal magnitude (univariate) and pattern (multivariate) analysis procedures used in our MEG and fMRI analyses (described in a subsequent section) to simulated representations of speech segments at the divergence point (results shown in Figure 1D, E). This is performed for each of the four experimental conditions (the 2-by-2 crossing of prediction strength and congruency).

The signal magnitude analysis (Figure 1D) focusses on the overall strength (rather than pattern) of neural responses. For this magnitude analysis of prediction error representations, we summed model signals (absolute prediction error) over all possible segment representations (48 in total). However, this approach is ill-suited for the sharpened signal simulation since here the model representations correspond to posterior probability distributions and hence all simulated signals are normalized to sum to one (and will hence never differ between conditions). Therefore, to relate sharpened signals to univariate neural responses, we summarised model responses as the normalized sum of log-transformed probabilities over the 48 segment representations (see Methods). This choice of linking function between the model and neural responses was motivated by the proposal that neural responses encode log rather than linear probabilities [26]. We also considered an alternative linking function based on entropy, which was motivated by the observation that increased uncertainty in neural responses can be accompanied by an increase in the mean and therefore overall strength of neural responses [9]. The two linking functions produced equivalent

results (this is unsurprising given that entropy is defined as a weighted sum of log probabilities). In the following section we report only the results of using the first linking function (the normalised sum of log-transformed sharpened signals).

This univariate signal magnitude analysis shows a broadly similar pattern for sharpened signal and prediction error computations. For familiar words (Match items), both sharpened signal and prediction error simulations show a reduced univariate response to strongly versus weakly predicted second syllables (Figure 1D). In addition, univariate responses are overall larger for pseudowords (Mismatch items) versus words (Match items). Within the Mismatch condition, sharpened signal and prediction error simulations diverge somewhat with the former showing a reduced response to strongly versus weakly predicted second syllables and the latter showing a larger response. For both models, prediction strength and congruency have interactive effects on univariate responses as the effect of prediction strength is largest for familiar words (Match items) and reduced for pseudowords (Mismatch items).

The multivariate pattern analysis (Figure 1E), on the other hand, shows clearer differences between sharpened signal and prediction error computations. This multivariate analysis quantifies how strongly phonetic content in the speech input is represented in simulated patterns at the segment level. This is achieved by quantifying the dissimilarity between model representations (pattern distances) for all pairs of items within each condition.

The results of this multivariate pattern analysis show that while sharpened signal representations are enhanced for Match items, prediction strength and congruency show a cross-over interaction which only occurs for prediction error representations. Specifically, for signals that match prior predictions (during word processing), prediction errors decrease in magnitude and pattern distance for stronger predictions. However, the opposite effect occurs when predictions mismatch with predictions during pseudoword processing; that is prediction error is increased, and more informative for syllables that mismatch with stronger predictions. In addition, prediction errors are overall larger for Mismatch compared with Match items, reflecting the absence of accurate predictions following the divergence point of the pseudoword (Mismatch) stimuli.

These simulations, particularly when analyzed using multivariate pattern analysis, confirm that our experimental manipulations can distinguish between sharpened signal and prediction error computations. We can therefore evaluate the

extent to which observed brain responses are more consistent with sharpened signal or prediction error computations.

**MEG signal magnitude**

As shown in Figure 2A, we assessed MEG responses timelocked to the onset of Syl2, with respect to a baseline period prior to Syl1 (i.e. speech onset; see Methods). Note that as a result, the MEG responses at negative latencies shown in Figure 2A are non-zero (coinciding with sensory input from the Syl1 portion of our spoken stimuli).

Before assessing the neural timecourse of between-condition differences, we selected 20 (planar gradiometer) sensors with the strongest responses (signal power averaged over conditions, separately for each hemisphere and participant). Sensor selections (i.e. those sensors with the strongest responses) were distributed over temporal and frontal sites bilaterally (shown inset within Figure 2A), consistent with neural generators in superior temporal cortex. We then summarised MEG responses in these “speech responsive” sensors in a univariate fashion by taking the RMS amplitude across sensor selections for each timepoint. Reported effects are all FWE (family-wise error) rate corrected for multiple comparisons across timepoints using cluster permutation methods and a clusterwise threshold of $p < 0.05$.

Following Syl2 onset, prediction strength (Strong vs Weak) and congruency (Match vs Mismatch) had interactive influences on MEG responses from around 200 to 600 ms (results for left hemisphere sensors are shown in Figure 2A; similar results are seen for right hemisphere sensors, depicted in Figure S2A). Stronger predictions resulted in weaker MEG responses but only when speech matched with predictions (significant timepoints for the interaction between prediction strength and congruency are indicated as horizontal cyan bars; the simple effect of Strong vs Weak for Matching items is shown as the grey shaded area between MEG traces). MEG responses also showed a main effect of prediction congruency with overall larger responses for Mismatch vs Match items (significant timepoints shown as red horizontal line). These two patterns – interaction between prediction strength and congruency, together with a main effect of Mismatch > Match –are broadly consistent with both sharpened signal and prediction error computations (compare with Figure 1D). Note that although MEG responses do not increase with prediction strength for

mismatching items in this particular analysis (a pattern present only in the prediction error simulations; see Figure 1D), they do so for other analyses that will be reported below.

Prior to Syl2 onset, a different pattern of results was observed. Here MEG responses increased with prediction strength (indicated as horizontal purple bars in Figure 2A). While the timing of this effect (before Syl2 onset) suggests a neural signal reflecting predictions rather than prediction errors, we cannot rule out that this pre-Syl2 effect reflects acoustic differences since the Syl1 portion of Strong items is longer in duration than for Weak items (see SI Results and Figure S4A). This duration difference could potentially lead to larger MEG signal when timelocking to Syl2 onset as neural responses would be shifted later in time for Strong versus Weak items. Indeed, when timelocking to Syl1 onset to control for this difference in timing, the increased MEG response for Strong versus Weak items is no longer apparent (Figure S2C). Note that these or other acoustic confounds can only affect pre-Syl2 neural responses since our cross-splicing procedure ensures that post-Syl2, there are no acoustic differences between our four conditions (see Figure 1A and Figure S4A). Since representations of Syl2 are critical for distinguishing sharpened signal from prediction error computations (Figure 1E), we subsequently focus on post-Syl2 effects.

To test whether Syl2 responses are modulated by prediction strength for single items, we correlated p(Syl2|Syl1) with MEG responses across items separately for Match and Mismatch items. We show group-averaged Spearman correlations between prediction strength and MEG responses in left hemisphere sensors in Figure 2B (right hemisphere correlations shown in Figure S2B). Consistent with the previous analysis averaged over items (Figure 2A), following Syl2, there was an interaction between prediction strength and congruency from 168 to 300 ms and from 356 to 460 ms: Stronger predictions that matched with speech were associated with weaker MEG responses and this correlation differed from that observed when predictions mismatched (significant timepoints indicated as horizontal cyan bars). Individually, the correlations within the Matching condition were significantly negative (indicated as the grey shaded area). Within the Mismatching condition, correlations were weakly positive and significant in the right hemisphere at a later latency (Figure S2B, 500-600 ms). This pattern of results is more consistent with prediction error than with sharpened signal computations as a

positive effect of prediction strength for Mismatch items is only apparent in prediction error simulations (see Figure 1D). Thus, this analysis demonstrates that the item-averaged effects depicted previously in Figure 2A extend to single items. That is, the strength of lexical level predictions modulates neural responses to heard speech with the magnitude and direction of this effect depending on whether subsequent segments match or mismatch with predictions. The significant interaction shown in both factorial and graded analyses is broadly consistent with both sharpened signal and prediction error accounts although the positive effect of prediction strength for Mismatch items (observed in graded analysis) favours the prediction error account.

To probe time-frequency correlates of prediction strength and congruency, we also assessed changes in spectral power in the 2-48 Hz range. This time-frequency analysis also showed a significant interaction between prediction strength and congruency in the theta band, with a peak frequency at 3 Hz and a peak latency 120 ms after Syl2 onset (left hemisphere results shown in Figure 2D; a similar pattern was observed in right hemisphere sensors). Similar to the time-domain analyses above, stronger predictions resulted in weaker MEG power but only when speech matched with predictions (see summary in Figure 2C). We also analysed high-frequency gamma activity (52-90 Hz) but did not observe reliable condition-wise differences.

# Figure 2

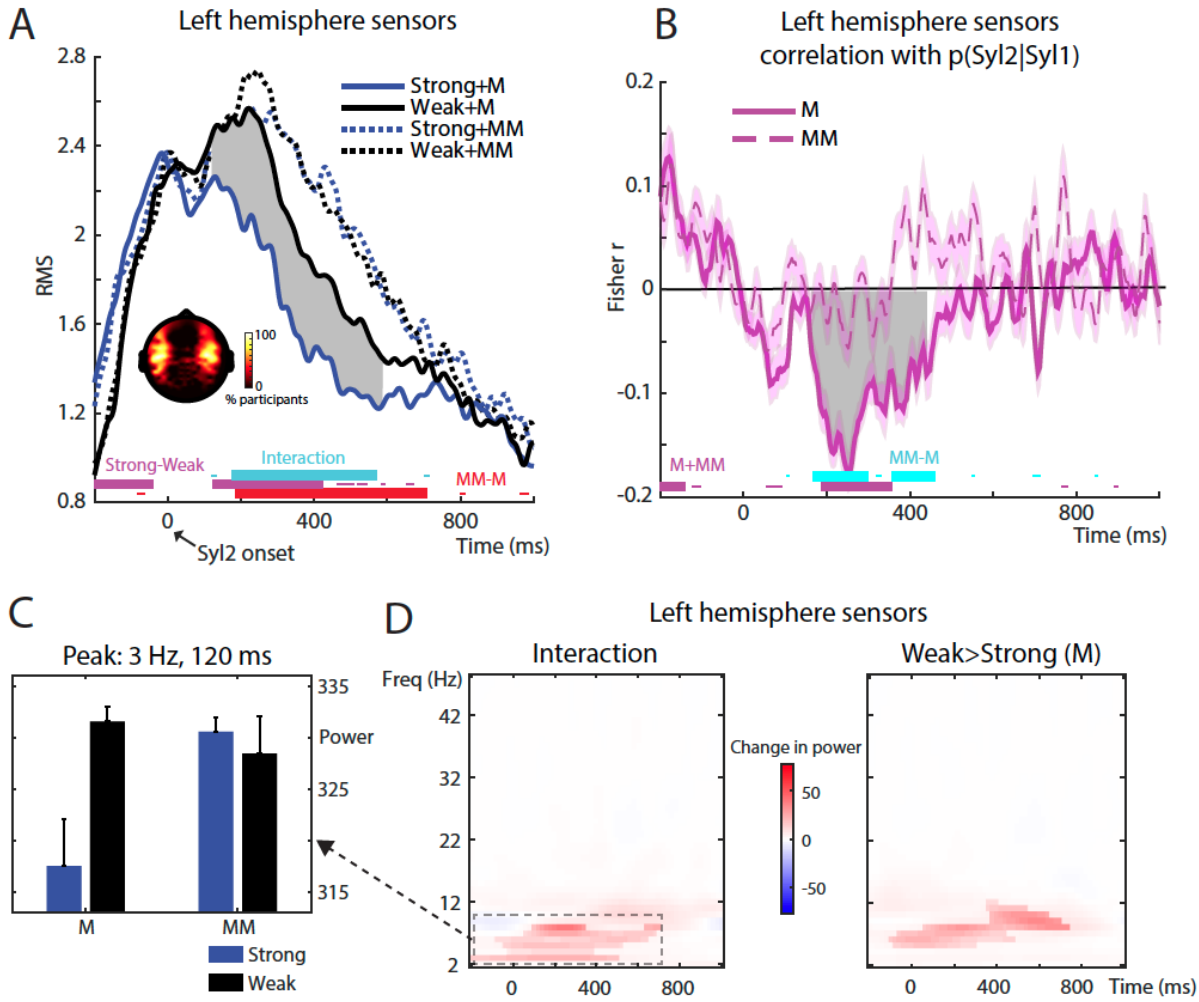



*Figure 2- MEG univariate signal magnitude results. (A) Traces indicate the RMS of the MEG signal across "speech-responsive" sensors in the left hemisphere, timelocked to the onset of the second syllable (Syl2). Sensor selections are indicated in the topography as the percentage of participants for which a sensor was selected. Thick horizontal bars at the bottom of the graph indicate timepoints when statistical contrasts were significant at a clusterwise threshold of p < .05 FWE corrected (thin horizontal bars show significance at an uncorrected p < .05 threshold). Grey shaded area indicates timepoints when the effect of Strong vs Weak in the Match condition was significant at p < .05 (FWE cluster corrected). M = Match. MM = Mismatch. (B) Group-averaged correlations in left hemisphere sensors between p(Syl2|Syl1) and the RMS of the MEG signal. Gray shaded area indicates timepoints when the correlations within Match or Mismatch conditions were significant at p < .05 (FWE cluster corrected). Error bars (transparent area around traces) indicate within-subject errors [28]. (C) Bar graph indicates group means and within-subject standard errors [28] for the four experimental conditions after averaging MEG power over the temporal and frequency extent of the interaction cluster reported in panel D. (D) Image depicts changes in MEG power (between condition differences) across time and frequency in left hemisphere sensors for the interaction (left panel) and simple effect of Weak > Strong in Match trials (right panel) . Nontransparent colors*

*indicate significance at p < .05 (FWE cluster corrected) while transparent colors show p < .05 (uncorrected). Interaction effect was computed as (Strong-Weak [Mismatch)]) – (Strong-Weak[Match]).*

**fMRI signal magnitude**

In a separate group of listeners, we recorded BOLD functional MRI responses to the same stimuli, using a sparse fMRI sequence so that speech was presented in the silent periods between scans (Figure S1A). The procedure (task and number of stimulus repetitions) was otherwise identical to that used in the MEG experiment. This allowed us to localize the neural generators of the MEG effects reported above. Reported effects are all FWE rate corrected using a clusterwise threshold of $p < 0.05$.

As shown in Figure 3A, which presents the univariate fMRI results, the main effect of Mismatch > Match (that was observed following Syl2 onset with MEG) localized to middle/anterior STG bilaterally (see Table 1 for MNI space coordinates). The main effect of Strong>Weak (observed with MEG prior to Syl2 onset) localized to Heschl's gyrus (HG), extending into STG. Despite observing an interaction between prediction and congruency on MEG responses following Syl2 onset, there was no equivalent effect in this fMRI analysis. We suggest this is because of the reduced temporal resolution of fMRI, which makes it more difficult to resolve individual responses to Syl2 and Syl1.

A potentially more sensitive analysis is to correlate prediction strength i.e. p(Syl2|Syl1) with the magnitude of the BOLD signal across individual items (equivalent to the MEG analysis shown in Figure 2B). Employing this approach, fMRI responses in bilateral HG and middle STG again increased with prediction strength (positive correlation) for both Match and Mismatch items (shown in Figure 3B with MNI space coordinates listed in Table 1), demonstrating that the effect depicted previously for items that differ in averaged prediction strength in Figure 3A is also observed for single items which encompass a range of prediction strengths. Because this effect of prediction strength is positive and equally apparent for Match and Mismatch items, we suggest it corresponds to the pre-Syl2 effect observed in MEG i.e. reflecting processing prior to the critical Syl2 event. We therefore will not discuss this effect further because our focus is post-Syl2 processing (as explained above), which can distinguish different computational accounts.

Importantly, we also observed an interaction between p(Syl2|Syl1) and prediction congruency in right anterior STG (shown in Figure 3B), revealing a potential neural generator of the interaction effects previously observed with MEG after Syl2 onset (Figure 2B and Figure S2B). Although this fMRI effect appears driven by a positive correlation in the Mismatch condition (rather than a negative correlation for Match items), the correlation is numerically negative for Match items, which is consistent with earlier MEG analysis (Figure 2B and Figure S2B). We also note that like the present fMRI analysis, right-hemisphere MEG sensors show a positive correlation in the Mismatch condition at later latencies (Figure S2B). As noted earlier for the univariate MEG analysis, this positive effect of prediction strength for Mismatch items is only apparent in prediction error simulations (Figure 1D).

# Figure 3

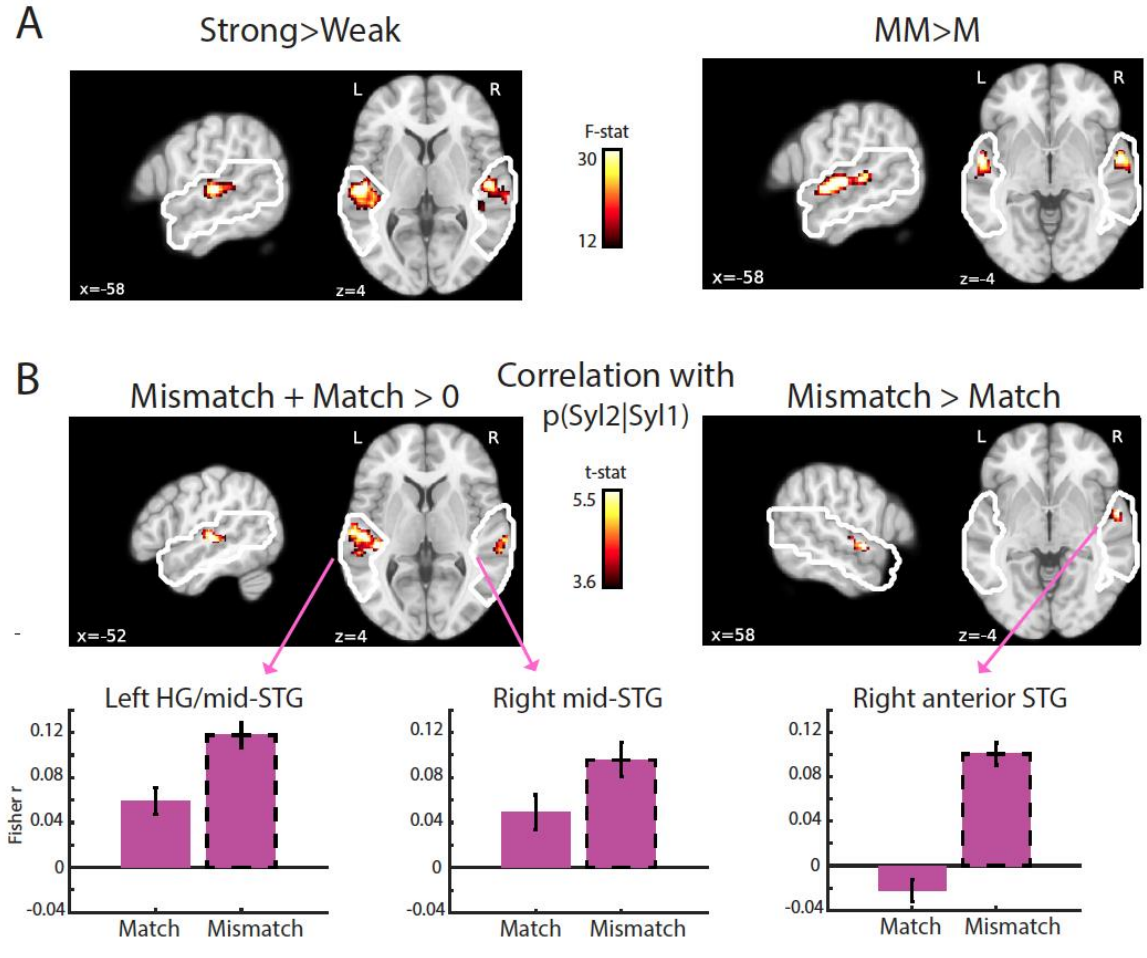



*Figure 3 fMRI univariate signal magnitude results. (A) Repeated measures ANOVA showed main effects of prediction strength (left panel) and congruency (right panel). Statistical maps ($p < .05$ FWE cluster corrected) are shown overlaid onto a template brain in MNI space. White outlines indicate the search volume which included superior and middle temporal regions bilaterally. (B) Statistical maps ($p < .05$ FWE cluster corrected) for correlations between p(Syl2|Syl1) and BOLD magnitude. In HG extending into middle STG (left panel), correlations are significantly positive and do not differ between Match and Mismatch items (equivalent to a main effect of prediction strength). In anterior STG (right panel), correlations are significantly larger for Mismatch than for Match items. Bar graphs show group-averaged correlations (Fisher-transformed r-values) between conditional probability and BOLD magnitude averaged across voxels in the clusters indicated (arrows). Error bars show within-subject standard errors suitable for statistical comparison between conditions.*

**Table 1**

| Location | Extent | Z-stat | x | y | z |
|---|---|---|---|---|---|
| **Univariate Strong > Weak** | | | | | |
| Right Heschl's Gyrus | 551 | 5.74 | 54 | -16 | 2 |
| Right mid-STG | | 4.84 | 64 | -14 | 0 |
| Right mid-STG | | 4.05 | 66 | -26 | 4 |
| Left mid-STG | 551 | 5.41 | -58 | -22 | 4 |
| Left Heschl's Gyrus | | 4.55 | -44 | -24 | 2 |
| Left mid-STG | | 4.55 | -46 | -32 | 4 |
| **Univariate Mismatch > Match** | | | | | |
| Left mid-STG | 569 | 6.00 | -64 | -24 | 6 |
| Left mid-STG | | 5.81 | -60 | -16 | 2 |
| Left anterior-STG | | 5.62 | -58 | -4 | -4 |
| Right anterior-STG | 205 | 5.25 | 58 | -4 | -4 |
| Right mid-STG | | 4.51 | 64 | -16 | 0 |
| Right mid-STG | | 3.54 | 56 | -14 | -2 |
| **Univariate Strong+Noise > Weak+Noise** | | | | | |
| Left anterior-STG | 245 | 4.42 | -54 | -12 | 2 |
| Left Heschl's Gyrus | | 4.21 | -44 | -22 | -2 |
| Left mid-STG | | 3.31 | -60 | -24 | 2 |
| Right posterior-STG | 56 | 4.25 | 64 | -44 | 10 |
| Right posterior-MTG | | 3.58 | 60 | -54 | 6 |
| Right mid-STG | 115 | 4.00 | 54 | -20 | 0 |
| Right Heschl's Gyrus | | 3.30 | 48 | -14 | 0 |
| Right anterior-STG | | 3.10 | 58 | -12 | -4 |
| **Univariate correlations with p(Syl2\|Syl1) Mismatch+Match** | | | | | |
| Left Heschl's Gyrus | 467 | 5.01 | -52 | -22 | 2 |
| Left Heschl's Gyrus | | 4.85 | -42 | -28 | 4 |
| Left mid-STG | | 4.62 | -62 | -18 | 6 |
| Right mid-STG | 105 | 4.04 | 66 | -24 | 6 |
| Right mid-STG | | 3.96 | 64 | -32 | 6 |
| Right mid-STG | | 3.56 | 54 | -26 | 0 |
| **Univariate correlations with p(Syl2\|Syl1) Mismatch>Match** | | | | | |
| Right anterior-STG | 71 | 4.58 | 58 | -2 | -4 |
| Right anterior-STG | | 3.81 | 50 | -4 | -2 |
| Right anterior-STG | | 3.59 | 58 | -10 | 2 |
| **Multivariate Syl2 phonetic distance** | | | | | |
| Right Heschl's Gyrus | 148 | 4.49 | 56 | -6 | 2 |
| Left Heschl's Gyrus | 146 | 4.01 | -44 | -20 | 0 |

*Table 1- MNI coordinates for fMRI analyses. Also shown are the spatial extent (number of voxels) and peak statistic for each cluster.*

## MEG pattern analysis

The univariate signal magnitude analyses above shows that the magnitude of neural responses in superior temporal regions is modulated by syllable prediction strength in different directions depending on whether heard speech matches or mismatches with predictions. While the direction of these neural effects is in line with the magnitude of prediction error in our computational simulations, as mentioned above, our computational simulations indicate that multivariate pattern analysis more clearly distinguishes sharpened signal and prediction error accounts. Simulations of prediction error computations show that representational patterns show a specific cross-over interaction between prediction strength and congruency (Figure 1E) which is clearly absent for sharpened signal computations.

To examine how predictability modulates neural representational patterns in our MEG data, we computed the Euclidean distance between sensor patterns (over the entire 204 array of planar gradiometers) evoked by pairs of items. Taking this approach, we first asked whether the MEG signal contained phonetic information about the syllables by correlating neural pattern distances with the phonetic dissimilarity between the syllables of item pairs (based on a representational dissimilarity matrix derived from the Levenshtein distance between phonetic transcriptions for different syllables, as illustrated in Figure 4A for Syl2 and Figure S5 for Syl1). As shown in Figure 4B, MEG patterns correlated with Syl2 phonetic dissimilarities only after Syl2 onset (reliable effects were observed in two clusters between 168 and 928 ms post-DP, with the peak effect observed at 332 ms; indicated by horizontal red bars). In contrast the peak correlation with Syl1 phonetic dissimilarities was largely confined to periods in which Syl1 was heard, with a cluster from -200 to +188 ms, and peak effects occurring at -112 ms (indicated by horizontal beige bars). This confirms that MEG response patterns represent the phonetic content of the syllables as listeners are hearing them, in line with other, similar demonstrations [29]. Qualitatively similar albeit weaker results are obtained when using cross-validated Mahalanobis distances [30].

We next assessed how neural pattern distances were modulated by syllable predictability by averaging pattern distances over item pairs within each condition and comparing averaged distances between conditions [for a similar approach applied to visual processing, see ref 31]. As shown in Figure 4C, following Syl2 onset, prediction strength and congruency interacted to influence MEG pattern distances: stronger predictions resulted in larger MEG pattern distances but only when speech mismatched with predictions (simple effect of Strong - Weak [Mismatch] is shown as shaded grey area in Figure 4C, and the first interaction cluster peak at 236 ms, is plotted as an inset). This interaction revealed for MEG patterns after Syl2 onset is more consistent with prediction error than with sharpened signal computations (compare with Figure 1E).

Complementing the analysis above, we also examined the correlation between neural pattern distances and Syl2 phonetic dissimilarities, separately for each condition (i.e. the same dissimilarities indicated in Figure 4A but within condition). This provides a second measure of neural representations, one that is potentially more specific to Syl2 phonetic content as opposed to other properties that

could differ between items e.g. acoustic, lexical, semantic etc. However, we did not find any timepoints showing an interaction between prediction strength and congruency nor any main effects (Figure S2D). Given that our simulations provide an apriori hypothesis that the effect of Strong vs Weak predictions should differ for Match and Mismatch conditions (see Figure 1E), we also tested for simple effects of Strong versus Weak prediction within each congruency condition. This revealed greater correlation between MEG pattern and Syl2 phonetic distances for Strong versus Weak items in the Mismatch condition only. This occurred in two clusters, from 44 to 116 ms and from 148 to 220 ms. These results remain broadly consistent with the earlier analysis shown in Figure 4C. While here we did not observe an interaction between prediction strength and congruency, as before we did observe a positive effect of prediction strength within the Mismatch condition. This result favours the prediction error account (see Figure 1E).

Correlating pattern distances with a graded measure of syllable prediction strength (i.e. p(Syl2|Sy1)) revealed further, convergent evidence for prediction error computations (Figure 4D). Stronger predictions for upcoming speech sounds during Syl2 were associated with opposite effects on MEG pattern distances from 360 to 468 ms, depending on prediction congruency (significant timepoints indicated as horizontal cyan bars). During this time period, correlations between prediction strength and pattern distances were positive when predictions mismatched with heard speech (i.e. neural representations were more informative for speech that mismatched with stronger predictions), and negative when predictions matched heard speech (i.e. less informative representations for speech that matched stronger predictions). These findings are as expected for representations that signal prediction error (Figure 1E right). Individually, correlations within the Matching condition were not significantly different from zero. But within the Mismatching condition, correlations were significantly positive from 320 to 544 ms (shown as shaded grey area in Figure 4D).

We also applied our pattern analysis procedure to MEG sensor-space patterns after the data was transformed into a time-frequency representation of signal power (Figure 4E). Here there was again an interaction between prediction strength and congruency with a peak at 560 ms in the theta band (peak at 4 Hz, extending to higher alpha frequencies). This interaction was driven by smaller MEG pattern distances (i.e. less informative neural representations) for Matching syllables

that follow Strong rather than Weak predictions (Figure S2E). Again, this effect is in line with neural responses that signal prediction errors in heard speech. We also applied the same pattern analysis approach to high-frequency gamma activity (52-90 Hz) but did not observe condition-wise differences.

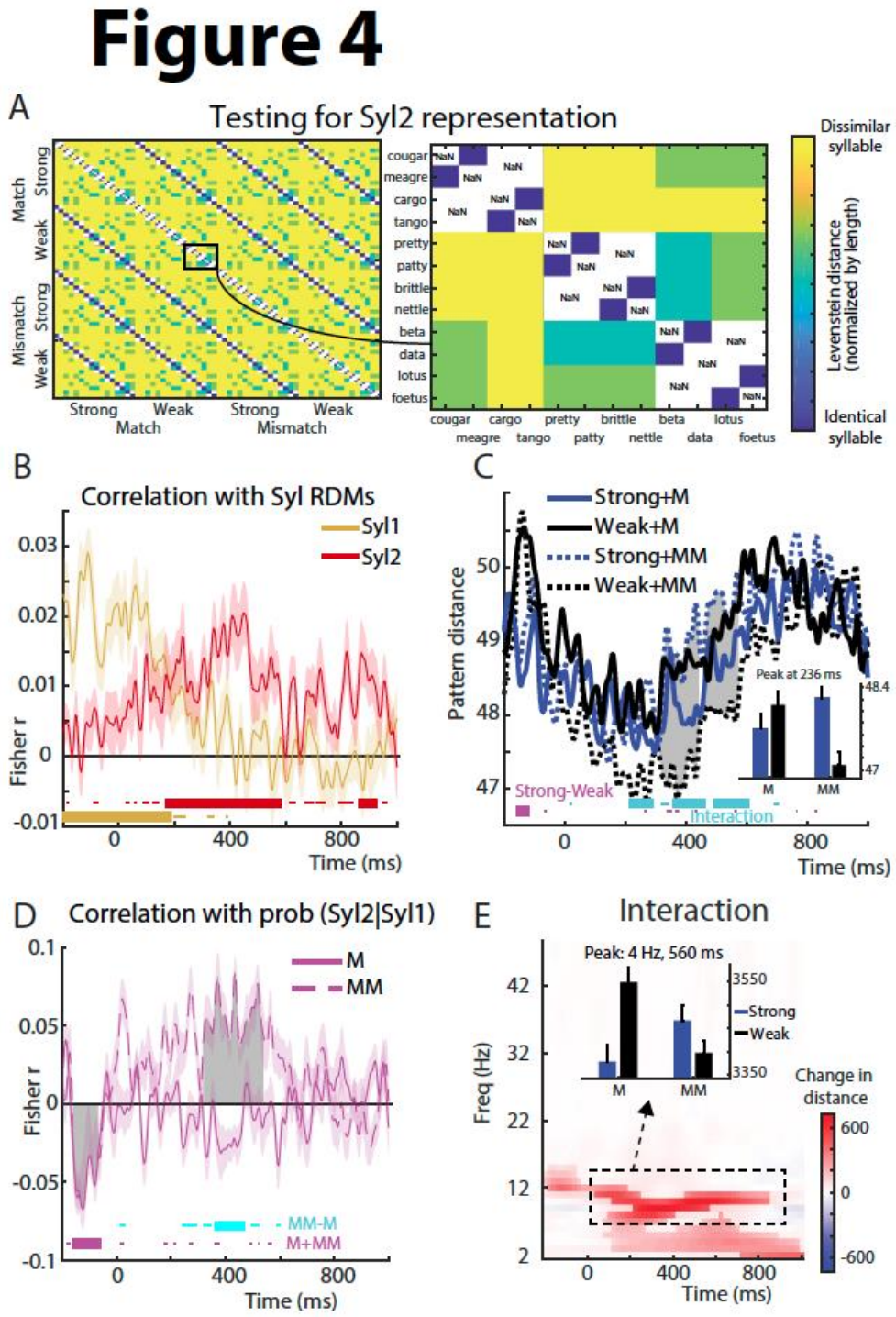



*Figure 4 MEG multivariate pattern analysis results. (A) Phonetic dissimilarity matrix expressing the normalized Levenshtein distance between phonetic transcriptions of Syl2 for different item pairs. Right panel inset highlights*

*predicted dissimilarity used in representational analyses. (B) Group-averaged correlation between MEG pattern distances and phonetic dissimilarities for Syl2 (in red) and Syl1 (in beige). Horizontal bars indicate timepoints when correlations were significantly different from zero (thick lines show corrected significance at p < .05 clusterwise; thin lines show uncorrected significance at p < .05). Error bars (transparent area around traces) indicate standard error of the mean. (C) Group-averaged MEG pattern distances with larger values indicating more dissimilar MEG response patterns between items. Thick horizontal bars at the bottom of the graph indicate timepoints when statistical contrasts were significant at a clusterwise threshold of p < .05 FWE corrected (thin horizontal bars show significance at an uncorrected p < .05 threshold). Grey shaded area indicates timepoints when there was a significant difference between Strong and Weak items in the Mismatch condition (p < .05 FWE clusterwise corrected). The inset bar plot shows group means for each condition (and within-subject standard errors as before) after averaging over the temporal extent of the cluster for the interaction contrast. (D) Group-averaged correlations between p(Syl2|Syl1) and MEG pattern distances. Grey shaded area indicates timepoints when there was a significant positive correlation for Mismatch items (p < .05 FWE clusterwise corrected). Error bars (transparent area around traces) indicate within-subject errors [28]. (E) Group-averaged MEG pattern distances after transforming the MEG signal into time-frequency power. Image depicts the interaction contrast: (Strong-Weak [Mismatch)]) – (Strong-Weak[Match]). Nontransparent colors indicate significance at p < .05 (FWE cluster corrected) while transparent colors indicate p < .05 (uncorrected). Bar graph indicates group means (and within-subject standard errors) after averaging over the temporal and frequency extent of the interaction cluster (black broken outline on the time-frequency image).*

## fMRI pattern analysis

To localize the neural generators of the MEG pattern analysis effects reported above, we applied a similar multivariate pattern analysis approach to spatial searchlights in auditory brain areas in our fMRI data. As with our MEG pattern analysis, we first asked whether fMRI responses contained phonetic information about Syl2 by correlating neural pattern distances with the phonetic dissimilarity between the second syllables of item pairs. As shown in Figure 5A, multivoxel patterns correlated with Syl2 phonetic dissimilarities in HG bilaterally (see Table 1 for MNI space locations).

We next proceeded to examine between-condition differences in neural representations. Rather than directly analysing pattern distances (as with the equivalent MEG analysis shown in Figure 4C), we opted to analyse a more specific measure of the correlation between neural patterns and Syl2 phonetic dissimilarities. This approach is more optimally suited to measure fMRI responses that are specific to Syl2 as opposed to representing both syllables. This is because while the temporally precise MEG signal can easily resolve neural responses to Syl1 and Syl2, this is not the case for fMRI which integrates sensory information over both syllables. However, by analysing the strength of correlation between fMRI pattern distances

and Syl2 phonetic dissimilarities, we can more specifically examine the neural representations of Syl2. Within the left and right HG clusters previously shown in Figure 5A, prediction strength and congruency showed interactive effects on the correlation between multivoxel pattern distances and Syl2 phonetic dissimilarities (shown in Figure 5B; $F(1,20) = 10.00$, $p = .005$). This effect did not reliably differ in magnitude between the two hemispheres (the three-way interaction between prediction strength, congruency and hemisphere was non-significant: $F(1,20) = 0.170$, $p = .684$). Though the interaction effect failed to reach significance in the left hemisphere ($F(1,20) = 3.304$, $p = .084$), the interaction effect was significant in the right hemisphere ($F(1,20) = 7.831$, $p = .011$). While tests of simple effects within each hemisphere were not reliable, the numerical trends are entirely consistent with previous analyses and in line with computational predictions: increasing prediction strength resulted in a numerically weaker correlation for a matching Syl2 ($F(1,20) = 2.85$, $p = .107$) while the opposite numerical difference (i.e. a stronger correlation for stronger predictions) was observed for mismatching Syl2 items ($F(1,20) = 2.70$, $p = .116$).

We also examined whether fMRI responses contained phonetic information about Syl1 by correlating neural pattern distances with the phonetic dissimilarity between the first syllables of item pairs. However, we did not find significant correlations (all clusters $p > .6$ uncorrected).

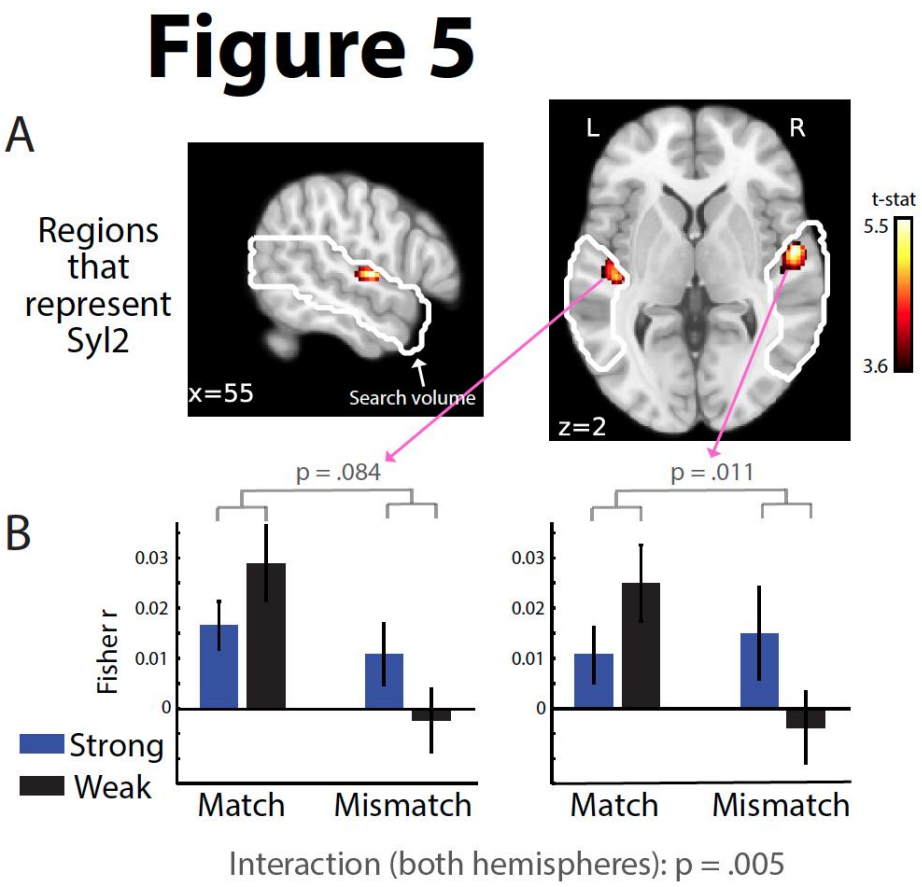



*Figure 5 fMRI multivariate pattern analysis results. (A) Statistical maps showing voxels in which there was a significant correlation between fMRI pattern distances and phonetic dissimilarities for Syl2 ($p < .05$ FWE corrected clusterwise). White outline shows search volumes as in Figure 3. (B) Group-averaged correlations (and within-subject standard errors) after averaging over voxels in the left and right hemisphere clusters shown in (A).*

## Discussion

Across multiple imaging modalities (MEG, fMRI), analysis approaches (univariate signal magnitude and multivariate pattern), and signal domains (phase-locked, evoked activity and induced, time-frequency MEG responses), we show that neural responses to second syllables in spoken words are modulated by the strength and content of prior knowledge (i.e. predictions for expected speech sounds). We further show that these neural influences of prior knowledge begin around 200 ms after the onset of the second syllable, localise to early auditory regions (in fMRI, bilateral Heschl's gyrus and STG) and are also expressed as changes in low-frequency (theta and alpha) power.

**Neural patterns distinguish prediction error from sharpened signal computations**

In the current study, the critical analyses used multivariate pattern analysis procedures that previous work has shown can unambiguously distinguish between sharpened signal and prediction error accounts [18,21]. By examining the pattern and not just the overall strength of neural signals, responses tuned to the stimulus input features as well as those tuned away from the heard stimulus can be jointly modelled. In this analysis framework, if neural response patterns reflect prediction errors they should contain less information about strongly versus weakly predicted speech sounds (see computational simulations in Figure 1E), and this is indeed what we observed for MEG and fMRI responses to familiar words.

While neural representations of second syllables are suppressed for stronger predictions that match heard speech, neural representations of exactly the same syllables show the opposite modulation for second syllables of pseudowords. That is, when predictions mismatch with sensory input, neural representations are enhanced after more strongly- than more weakly-predicting syllables. Our computational simulations and multivariate analysis confirm that the interaction between prediction strength and match/mismatch is entirely consistent with the operation of prediction error computations [8–11] but cannot be explained by sharpened signal accounts in which representations of speech are enhanced by prior knowledge [5–7]. Our study therefore provides convergent neural evidence – from different imaging modalities and multiple analysis approaches – that speech perception is supported by the computation of prediction error representations in auditory brain regions.

It should be noted that in prediction error accounts, computation of prediction errors is not the end goal of the system [as discussed in ref 21]. Rather, they are used to update – and thereby enhance or sharpen – future predictions. In some formulations of predictive coding [8,17], these sharpened predictions are represented in separate neural populations. However, we saw little evidence of sharpened representations in our study; perhaps because these sharpened predictions reside in deeper cortical layers [17] to which MEG [32] and BOLD [33] signals are less sensitive. It has also been proposed that prediction errors and predictions are communicated via distinct oscillatory channels [34,35], although our time-frequency analysis did not provide support for this (discussed below). In future, methodological advances e.g. in layer-specific imaging [36], or more invasive methods [37] may help

to resolve this issue. Insights may also come from studies that simultaneously probe multiple levels of processing (e.g. to observe prediction updating in higher-level semantic stages of processing).

Why are prediction error representations enhanced by prediction strength when predictions mismatch with speech input i.e. for pseudowords? Because the combinations of syllables in these items will never previously have been heard by listeners as English spoken words, the expected probabilities for the second syllables should always approach zero [in information theory terms, surprisal should be maximal; see ref 3]. Therefore, if neural representations were only sensitive to predictions for the heard syllables (and not alternative syllables), we would not expect any neural differences between Strong+Mismatch and Weak+Mismatch items. However, our manipulation of prediction strength reflects not only differences in predictions for the heard second syllables, but also in the uncertainty of predictions over all possible syllables: In information theory, this uncertainty is formulated as entropy [23]. Indeed in natural speech, predictions for specific speech sounds are negatively correlated with entropy [12,14,23].

We suggest our results reflect both aspects of prediction i.e. for the heard syllables as well as the uncertainty of predictions over all possible syllables. In the Strong+Mismatch condition, a focal (low entropy) prediction is misaligned with sensory input. As a result, there is increased information in neural prediction errors because the sensory signal remains “unexplained”. By contrast, weak and more distributed (lower entropy) predictions will tend to align more with speech input representations. This illustrated by the pseudoword *tanger*. After hearing the first syllable /t { N/, predictions are spread weakly over multiple word candidates including the word *tanker* in which the penultimate sound is shared with *tanger*. This overlap results in reduced prediction errors in the Weak+Mismatch condition.

Our findings are consistent with neurophysiological findings showing that in addition to the informativeness of heard speech sounds, brain responses are also sensitive to the uncertainty of predictions made in parallel for alternative sounds [12,14,29,38]. Going beyond this previous work however, our current findings suggest that neural sensitivity to prediction uncertainty reflects the specific neural computation of prediction error. Prediction uncertainty is a parsimonious explanation of the prediction strength effect on pseudoword representations. However, it is also possible that the strong violations of predictions that occur for pseudoword items

may engage additional or qualitatively different mechanisms to those that support processing of familiar words. Bayesian inference for unfamiliar words is ill-defined because a pseudoword by definition has never previously been heard and therefore has a prior probability of zero [2,3]. In this situation, prediction error is maximal [3] and may function as a signal to indicate that the listener's existing internal model is no longer applicable [3,39], triggering the encoding of a new word that may later enter into the lexicon [see below for further discussion and refs 3,40] or perceptual learning if the speaker systematically produces certain speech sounds with an unusual pronunciation [41]. By this view, opposite neural effects of prediction strength during perception of matching and mismatching second syllables arise because distinct neural processes are engaged during perception of words and pseudowords.

**Prediction error computations during natural listening**

Our findings replicate and extend the findings of previous studies that manipulated prior knowledge by presenting matching or nonmatching written words before highly degraded spoken words [18,21,42]. In this previous work, neural representations showed interactive influences of prior knowledge and signal quality, consistent with a prediction error account. We go beyond this previous work by demonstrating prediction error computations in a more natural listening situation. The present experiments are more naturalistic in two respects: Firstly, we presented clear speech rather than the highly degraded (and hence less natural sounding) noise-vocoded words used previously. Secondly, we used an intrinsic manipulation of prior knowledge based on listeners' long-term linguistic knowledge of familiar spoken words rather than an extrinsic manipulation of predictions based on written text. This demonstration of prediction errors during more natural listening situations is important because it directly addresses the concern that previous results reflect a "prediction encouraging" paradigm [22] that would not generalise to other listening situations. Instead, our findings are consistent with the idea that lexical identification operates through prediction error computations and that these are integral to perceptual processing of speech irrespective of the listening situation [3,43–46].

One might argue that our findings fall short of demonstrating prediction error computation in a fully naturalistic listening situation as our stimuli comprised individual spoken words as opposed to connected speech. In other work

investigating speech processing during perception of connected speech (e.g. stories), neurophysiological responses (measured by MEG, EEG or electrocorticography) are shown to correlate with various measures of time-varying predictability, including phoneme surprisal [12–15,27], phoneme entropy [14,27] and lexical entropy [12,15,27]. While these findings clearly demonstrate that speech processing is strongly modulated by prior knowledge, consistent with the notion that speech perception is fundamentally a process of Bayesian inference [2], they do not indicate whether the underlying neural computations involve sharpened signals or prediction errors. Consider for example, the common observation of reduced responses to predictable speech sounds. This response reduction is compatible with either a prediction error account (through the suppression of responses tuned to input features that are predicted) or a sharpened signal mechanism (through the suppression of responses tuned away from the input features; for a more detailed explanation, see Introduction and refs. [5,6,20]). Indeed, this pattern is evident in our simulations as univariate responses are reduced for strongly versus weakly predicted second syllables of words for both sharpened signal and prediction error simulations.

More recently, studies employing connected speech paradigms have shown that neural representations of speech segments and acoustic features are themselves modulated by predictability [16,29,47]. These studies use an approach analogous to the present pattern analysis since they specifically link neural effects of surprisal and entropy to neural representations of specific segments or acoustic features. However, previous findings appear inconsistent with respect to sharpened signal and prediction error accounts since some studies show that neural representations are enhanced when predictions are strong [47] while others show the opposite [16] or even effects in both directions at different latencies [29].

These inconsistent results from story listening paradigms therefore suggest that alternative experimental manipulations of listeners' predictions are needed to distinguish between accounts. Our pseudoword condition creates a second situation in which prediction error and sharpened signal computations can be teased apart. Pseudowords are previously unfamiliar to participants, and hence create a unique listening situation in which predictions entirely mismatch with sensory input. Under a sharpened signal account, the pattern of neural representations of speech sounds forming a pseudoword (as compared with a word) should be suppressed, reflecting

the low prior probability of hearing these speech sounds. In contrast, under a prediction error account, the exact opposite should occur i.e. enhanced representations for pseudowords, particularly following strongly-predicting first syllables. Neural responses to pseudowords are consistent with this latter pattern and hence are likely to reflect prediction error computations. In our signal magnitude analyses, we observed larger MEG and fMRI responses for pseudowords versus words. For pattern analyses, although we did not observe a reliable main effect of prediction congruency, neural representations of speech for pseudoword items were enhanced following strongly- versus weakly-predicting first syllables. Overall, these pseudoword results again favour a prediction error account.

We have previously argued that larger prediction error responses for pseudowords compared to real words provide a potential signal for novelty detection and pseudoword learning [3,interpreting results from 11,48]. Our demonstration of more informative prediction errors for syllables that mismatch with more strongly-predicting first syllables suggests that such pseudowords might be more readily learned. We note that memory studies using visual stimuli have similarly argued for enhanced learning of face-scene associations that generate larger prediction errors at encoding [49]. Consistent with this proposal, are the present behavioural results from the cued memory recall test conducted after the fMRI scans. This showed that listeners' recall for pseudowords is more accurate for items that strongly mismatched with predictions in the scanner.

**Spatiotemporal and spectral profile of prediction error computations**

Our results also establish the timing and location of prediction error computations in response to speech. Previous studies using EEG [13], MEG [12–14,29], fMRI [18,50–53] and electrocorticography [15,54] disagree about when and where prediction errors are computed in the cortical hierarchy. Each of these techniques has their strengths but none combines spatiotemporal resolution with wide coverage of the cortex. Perhaps because of this, it remains unclear whether predictive neural computations emerge only once sensory information has been transformed into non-acoustic (e.g. phonetic) representations [12] or whether acoustic representations themselves can support predictive computations [21]. In the present study, we obtained MEG and fMRI responses to the same stimuli, while participants performed the same incidental (pause detection) task to maintain

attention. This feature of our study enabled us to map out prediction error computations precisely in both time and space.

In our MEG results, effects of prediction strength and congruency occurred around 200 ms following the onset of the critical second syllable. Note however that our word and pseudoword stimuli are identical until approximately 100 ms after the onset of the second syllable (corresponding to the onset of the subsequent segment). Therefore, relative to when the critical acoustic information became available to listeners, the interaction between prediction strength and congruency occurred with a latency of approximately 100 ms. This latency is compatible with a low-level acoustic locus of prediction error computation but perhaps not early enough to definitively rule out a higher-level (non-acoustic) locus, if word identification is supported by rapid (< 50 ms) cortical mechanisms [55].

However, our fMRI results enable more definitive conclusions, since searchlight pattern results show prediction effects in primary auditory cortex (Heschl's Gyrus). This localisation to Heschl's Gyrus is striking as it suggests that predictive computations for speech can operate even at the lowest levels of the cortical processing hierarchy, which leaves open the possibility that top-down predictive influences may even extend subcortically [56]. Future work manipulating prediction at different linguistic levels (e.g. sublexical, lexical and supralexical) would be valuable to establish the generality of this finding, and to determine whether these effects are uniquely or differentially localised to specific cortical regions.

In both signal magnitude and pattern analysis of MEG responses, the interaction between prediction strength and congruency occurred in low-frequency neural responses, with a peak in the theta range (3-4 Hz). An influential proposal claims that bottom-up prediction errors are conveyed by high-frequency (gamma) activity while top-down predictions are conveyed by low-frequency (alpha/beta) signals [17,34,57]. This is consistent with the idea that neural representations integrate sensory information over longer timescales at higher cortical levels (where top-down predictions are proposed to originate; see [50,51,58–60]). If the current findings reflect prediction errors as we have argued, then we might have expected modulation in the gamma range, which we did not observe. Instead, our findings are consistent with other work suggesting that prediction errors for speech are conveyed in low-frequency neural signals [14,61]. This does not rule out an additional role for gamma responses. It may be for example that theta and gamma responses are

coupled [62,63] and that gamma effects are not observed in the present study due to the generally lower signal-to-noise ratio of MEG responses at higher frequencies [64].

**Integrating computational models and brain responses to reveal the neural computations supporting speech processing**

The current study follows previous work that also combined computational models and neuroimaging to test sharpened signal and prediction error accounts [18,21]. As has been commented on elsewhere [65], computational models provide the most reliable way to establish the predictions of a particular theory. This observation also applies to the goal of distinguishing sharpened signal and prediction error theories. For example, opposite outcomes are observed in simulated prediction error representations depending on the level of sensory degradation [see refs 18,21 and also computational simulations in Figure S3 and Methods section]. While this result is clearly apparent in simulations, it was not previously obvious from verbal descriptions of prediction error theories. In the current study, we also see opposite outcomes in simulated prediction errors depending on whether predictions match or mismatch sensory input during word and pseudoword processing, respectively. Through computational simulations, we could establish that this is a unique signature of prediction error representations during perception of words and pseudowords and furthermore show that this signature is observed in neural responses.

Importantly, our simulations also provide an explanation for why prediction strength and match/mismatch combine interactively to influence prediction error representations (see discussion above on why prediction error representations are enhanced by prediction strength when predictions mismatch speech input). However, future work is needed to develop a more complete model of prediction error processing [for examples of other ongoing efforts, see refs 66–68]. For example, in the present model implementation, predictions for upcoming speech sounds are subtracted from the input to compute prediction errors but these segmental predictions are never used to update higher-level lexical predictions. Rather, lexical-level predictions are computed afresh as each new segment is presented (by combining word priors and likelihoods using Bayesian inference). A more complete model would update predictions via prediction errors. It would also operate on real (i.e. acoustic) speech input or connected speech without explicitly

requiring word onset locations. We note with interest the success of recent computational models which operate on acoustic speech inputs and are used to predict neural data [69,70]. We hope that the present results can constrain future iterations of these models.

# Methods

## Participants

40 participants (19 in the MEG experiment, 21 in the fMRI experiment) were tested after being informed of the study's procedure, which was approved by the Cambridge Psychology Research Ethics Committee. All were right-handed, native speakers of English, aged between 18 and 40 years and had no self-reported history of hearing impairment or neurological disease. The mean age in the MEG experiment was 25 years (SD = 4.63; 12 female, 7 male) and 23 years in the fMRI experiment (SD = 3.14; 13 female, 8 male).

## Stimuli

Spoken stimuli were 16-bit, 44.1 kHz recordings of a male speaker of southern British English and their duration ranged from 388 to 967 ms (mean = 621, SD = 107). A set of 64 bisyllabic words (Match items) were first selected from the CELEX database [71] based on the probability of the second syllable (Syl2) conditioned on the first syllable (Syl1):

$$p(Syl2|Syl1) = \frac{\sum freq(words\ matching\ both\ syllables)}{\sum freq(words\ matching\ first\ syllable)} \qquad Eq.\ 1$$

For this we used syllabified phonetic transcriptions in CELEX such that the first syllable /t { N/, for example, in words such as *tango* and *tangle* were considered to match the speech signal since all these words share the same first syllable /t { N/. However, words such as "tank" were not considered as matching since this word consists of the single syllable /t { N k/. Words were selected as sets of eight items (one example item set is shown in Figure 1A). Within a set of eight items, each word had a unique Syl1 that strongly or weakly predicted one of two Syl2 syllables. By

cross-splicing the first and second syllables of these words, we were able to create a set of pseudowords (Mismatch items) that contained the same Syl2 syllables as the real words.

To ensure that Syl2 syllables shared between items were acoustically identical, all four recorded syllables were combined using audio morphing using STRAIGHT software [72] implemented in Matlab (The Mathworks Inc). This software decomposes speech signals into source information and spectral information, and permits high quality speech resynthesis based on modified versions of these representations. This enables flexible averaging and interpolation of parameter values that can generate acoustically intermediate speech tokens [48,for application see 73]. Exploiting this capability of STRAIGHT, all words and pseudowords sharing the same Syl2 were constructed to be acoustically identical following Syl1 (by averaging STRAIGHT parameters across the four instances of the same Syl2 syllable within a set and splicing the resulting syllable onto Syl1). To ensure that all syllables sounded natural in these contexts, we constrained word selection such that all Syl2 syllables within a set started with the same speech segment. In this way, we could cross-splice syllables without conflicting co-articulatory information at the onset of Syl2. As a result, the point at which words and pseudowords diverged (the divergence point) occurred not at the onset of Syl2 but one speech segment later (post divergence point speech segments highlighted as bold in Figure 1A). Additional stimuli were derived by replacing the Syl1, Syl2, or both syllables, with spectrally and intensity (RMS) matched noise (data for these stimuli are not reported here). In total there were eight sets of items and therefore 240 stimuli (32 items in each of the following conditions: Strong+Match, Weak+Match, Strong+Mismatch, Weak+Mismatch; 32 each of Strong+Noise, Weak+Noise; 32 of Noise+Syl2; 16 of Noise+Noise). For analysis of stimulus properties, see supplementary Results.

Before the experiment, participants completed a brief practice session lasting approximately three minutes that contained examples of all the stimulus types included (spoken words, pseudowords with and without noise) but using different items to the main experiment. Stimulus delivery was controlled with Psychtoolbox [74] scripts in Matlab.

## Procedure

Stimulus presentation and task procedure was nearly identical for MEG and fMRI listeners. On each trial, listeners heard one of the spoken items separated by a stimulus onset asynchrony (SOA) of 2.5 sec (see Figure S1A). For the MEG experiment, this SOA was jittered by adding a random time interval of ±100 ms. For the fMRI experiment, a constant SOA was used, and stimuli were presented at the offset of the scans (sparse imaging sequence). Participants were asked to look at a fixation cross in the centre of the screen and respond as quickly and accurately as possible whenever they heard a brief pause in the speech stimuli. Pauses consisted of an additional 200 ms silence inserted between Syl1 and Syl2. Pilot testing revealed that pause detection was more difficult when one or both syllables were replaced by noise.

Trials were randomly ordered during each of five presentation blocks of 288 trials. Each block consisted of a single presentation of all stimuli plus ‘Pause’ stimuli. Across all five blocks, the order of ‘Pause’ and ‘No Pause’ items was completely randomised; in total there was one ‘Pause’ item for every six ‘No Pause’ items, with this ratio constant across stimulus type (Strong+Match, Weak+Match etc.). Participants received feedback on performance (% correct) at the end of each block.

## *Computational simulations*

Both sharpened signal and prediction error computations have the goal of implementing Bayes perceptual inference for spoken words [2,9]:

$$p(Word_i|Evidence) = \frac{p(Word_i) \times p(Evidence|Word_i)}{\sum_{j=1}^{j=n} p(Word_j) \times p(Evidence|Word_j)} \qquad Eq.\ 2$$

This equation computes the conditional (posterior) probability of each word, given their prior probabilities (frequency of occurrence) and the available sensory evidence. We modelled uncertainty in the sensory evidence based on the acoustic similarity between individual speech segments. Acoustic representations were obtained from a large set of 1024 words spoken by the same speaker as for the present stimuli. These stimuli comprised recordings from previous studies [75–77] as well as the 32 Strong Match items and 32 Weak Match items from the current study.

From these stimuli, we computed acoustic energy for individual speech segments (phonemes): 48 unique segments in total, averaged over 3659 repeated tokens (median token count = 52). Acoustic energy was expressed as a function of time, frequency and spectrotemporal modulations [78], as computed by the nsltools toolbox in Matlab [http://nsl.isr.umd.edu/downloads.html; for details see 21]. Segment onset times were obtained using a forced-alignment algorithm included as part of BAS software [79; https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSGeneral]. Acoustic representations were vectorized and averaged over multiple productions of the same segment. We then obtained a between-segment similarity matrix using negative Euclidean distances and normalized each row of this matrix using the softmax function (combined with a temperature parameter *T*) such that each row could be interpreted as segment probabilities:

$$p(Segment_i) = \frac{\exp(-dT)}{\sum_{i=1}^{i=n} \exp(-dT)} \qquad \textit{Eq. 3}$$

where *d* denotes the Euclidean distance. From these segment probabilities, we could compute the sensory evidence for a word by applying the product rule over word segments [see 2]. The sensory evidence for an unfolding word will be strong if its constituent segments match internal representations that are acoustically dissimilar to other speech sounds in the language. It will also be strong if there are few word competitors sharing the same speech segments. Given that our stimulus set consisted of clear speech recordings, a reasonable assumption is that individual segment probabilities are close to one (indicating high certainty). Nonetheless, even clear speech segments will be perceived with momentary ambiguities as sub-phonemic information unfolds over the duration of a segment. In the absence of human sub-phonemic confusion data [80] for the British accent of the present stimuli, we scaled acoustic similarities to simulate a range of overall sensory uncertainty levels by varying the Softmax temperature parameter. The same qualitative patterns were present over a range of sensory uncertainty levels (Figure S3). The one exception to this is apparent for simulations with the highest degree of sensory uncertainty for which prediction error representations in the Match condition are enhanced relative to simulations with reduced sensory uncertainty (compare the height of the first bar in each of the bottom row of plots in Figure S3B). The effect of this is that – under high uncertainty – results show a reverse pattern with enhanced

neural representations for Strong versus Weak items. This is consistent with previous modelling and experimental data [18,21] and is explained by an increase in the informativeness of prediction errors that occurs when strong predictions match noisy or ambiguous sensory input: Even though speech segments are accurately predicted, their *acoustic form* is not as expected (deviating from clear speech).

Note that our stimuli were selected based on syllable conditional probabilities p(Syl2|Syl1), which used the syllabified phonetic transcriptions in CELEX. For consistency, we therefore incorporated the syllable boundaries from CELEX into our model simulations. We achieved this by adding a separate segment to the model's inventory of segments, representing the boundary between syllables. Thus, an item like "bingo" was represented by six segments in total (five speech sounds plus the boundary between the two component syllables). In Figure 1B and 1C, we omit the syllable boundary segment for display purposes.

To simulate speech processing as each speech segment in our stimuli was heard, we used Bayes theorem to compute the posterior probability distribution over all words in the lexicon. Then to compute predictions for each upcoming speech segment, we summed posterior probabilities over words sharing the same segment. These segment predictions were then multiplied by individual segment probabilities (derived from acoustic similarity as explained above), to simulate the sharpened signal computation (normalizing to sum to one). For simulating the prediction error computation, segment predictions were instead subtracted from segment probabilities. For analysis of the simulated representations, pattern distances were computed as the Euclidean distance between model signals for pairs of items at the divergence point. For the signal magnitude analysis, we analysed model representations specifically for the speech segment that was heard at the divergence point.

### *MEG acquisition and analysis*

Magnetic fields were recorded with a VectorView system (Elekta Neuromag, Helsinki, Finland) containing two orthogonal planar gradiometers at each of 102 positions within a hemispheric array. Data were also acquired by magnetometer and EEG sensors. However, only data from the planar gradiometers were analysed as these sensors have maximum sensitivity to cortical sources directly under them and are therefore less sensitive to noise artifacts [81]. To monitor eye and heart activity, EOG and ECG signals were recorded with bipolar electrodes.

MEG data were processed using the temporal extension of Signal Source Separation [82] in Maxfilter software to suppress noise sources, compensate for motion, and reconstruct any bad sensors. Also using MaxFilter, the data were transformed to a common coordinate frame for each participant separately (to compensate for changes in head position between blocks). Subsequent processing was done in SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK), FieldTrip (Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands) and The Decoding Toolbox [83] implemented in Matlab.

MEG data were downsampled to 250 Hz and ICA used to remove eye and heart artefacts (based on an automatic procedure correlating the component time-series with EOG and ECG signals). When epoching, we retained only “no pause” trials. MEG recordings were epoched from -1700 ms to 2000 ms relative to the onset of Syl2 and epochs with large amplitude signals were discarded (time- and channel-averaged power larger than 2 standard deviations from the condition-specific mean for signal magnitude analysis; power larger than 3 standard deviations from the pooled mean for pattern analysis). For signal magnitude (evoked response) analyses only, the data were baseline-corrected from -750 to -500 ms relative to Syl2 onset; this corresponds to the silence period before speech onset (since the duration of Syl1 ranged from 130 to 479 ms; see Figure S4C for distribution of Syl1 duration).

For time-frequency analysis, we computed the power between 2 and 48 Hz (in 1 Hz steps) by convolving the data from each trial with Morlet wavelets of 7 cycles. To analyse high frequency activity (52-90 Hz), we used multi-taper estimation (±10 Hz smoothing, 200 ms time-windows with a step size of 20 ms). We analysed differences in power attributable to our experimental manipulations without baseline normalization. For time-domain analyses, the data were additionally low-pass filtered below 30 Hz prior to trial averaging. Note that in the case of signal magnitude analyses, MEG responses reflect the average of 160 trials before outlier removal (5 repetitions × 32 items per stimulus type) while for pattern analyses, MEG responses reflect the average of 5 repetitions (since here we analyze item-specific responses). Subsequent analysis was based on the period immediately before and after Syl2 onset (-200 to 1000 ms).

For signal magnitude (evoked response) analysis, the MEG data across the sensor array were summarized as the root mean square (RMS) over the 20 sensors

with the largest evoked response (separately in each participant and hemisphere). Essentially the same results are obtained when selecting 10 or 40 sensors. The same sensor selections used for the signal magnitude (evoked response) analysis were also used when analysing time-frequency power. For pattern analyses, neural pattern distances were computed by taking the Euclidean distance between sensor patterns (over the entire 204 sensor-array) evoked by each pair of items in the relevant stimulus set. To assess whether MEG responses reflected the phonetic information about the syllables, neural pattern distances were correlated with the phonetic dissimilarities between the syllables of item pairs using Fisher-transformed Spearman's correlations. To compute the phonetic dissimilarities we used the Levenshtein distance between the phonetic transcriptions for each syllable, normalized by the maximum number of segments within the two syllables being compared; shown in Figure 4A for Syl2 and Figure S5 for Syl1 [see 77 for a similar method applied to scoring phonetic responses]. To maximize statistical power for this particular analysis, correlations were computed both across- and within-condition (see Figure 4A). As shown in Figure 4A, we also excluded dissimilarity of item pairs such as "cargar" and "cougo" since similar but opposite prediction errors are evoked during Syl2 of these items. For example, for "cargar", listeners predict /g @U/ but hear /g @ r*/. For "cougo", listeners predict /g @ r*/ but hear /g @U/. The pattern of prediction error in these two cases is identical, albeit opposite in polarity. Given uncertainty on how the polarity of prediction errors is encoded in neural responses [84], we exclude these pseudoword item pairs from pattern analysis of our neural data and model simulations. We also exclude the corresponding word item pairs (e.g. "cougar" and "cargo"; shown as NaNs in Figure 4A).

When correlating syllable conditional probabilities with MEG responses (either magnitude-based or pattern-based measures), we also used Fisher-transformed Spearman's correlations. For pattern analysis, in which the dependent measures reflect pattern distances for pairs of items, we averaged syllable conditional probabilities for the items in each pair of words (after log transformation).

For both signal magnitude (RMS signals, time-frequency power or Spearman correlations) and pattern analyses (pattern distances or Spearman correlations), group-level one-sample t-tests were performed for each timepoint (and frequency bin when relevant) while controlling the family-wise error (FWE) rate using a non-parametric cluster-based permutation procedure based on 5000 iterations [85].

These tests were performed on contrasts of the data, testing for the main effects and interactions in our experimental design, as well as any simple effects. Reported effects were obtained by using a cluster defining height threshold of $p < .05$ with a cluster t-sum threshold of $p < .05$ (FWE corrected), unless otherwise stated.

### *fMRI acquisition and analysis*

Imaging data were collected on a Siemens 3 Tesla Prisma MRI scanner (http://www.siemens.com). A total of 291 echo planar imaging (EPI) volumes were acquired in each of 5 scanning runs, using a 32-channel head coil and a multiband sparse imaging sequence (TR = 2.50 sec; TA = 1.135 sec; TE = 30 ms; 48 slices covering the whole brain; flip angle = 78 deg; in-plane resolution = 3 × 3 mm; matrix size = 64 × 64; echo spacing = 0.5 ms; inter-slice gap = 25%). After the third run, field maps were acquired (short TE = 10 ms; long TE = 12.46 ms). The experimental session commenced with the acquisition of a high-resolution T1-weighted structural MRI scan (TR = 2250 ms; TE = 2.99 ms; flip angle = 9°; 1 mm isotropic resolution; matrix size: 256 × 240 × 192 mm; GRAPPA acceleration factor PE = 2; Reference lines PE = 24).

fMRI pre-processing was performed in SPM12. After discarding the first four volumes to allow for magnetic saturation effects, the remaining images were realigned and unwarped to the first volume to correct for movement of participants during scanning. Also at the unwarping stage, the acquired field maps were used to correct for geometric distortions in the EPI due to magnetic field variations. Realigned images were co-registered to the mean functional image and then subjected to statistical analysis. For signal magnitude analysis, prior to further processing, images were also normalized to the Montreal Neurological Institute (MNI) template image using the parameters from the segmentation of the structural image (resampled resolution: 2 × 2 × 2 mm) and smoothed with a Gaussian kernel of 6 mm full-width at half-maximum. For pattern analysis, normalization and smoothing were performed only after computing pattern distances.

Statistical analysis was based on the general linear model (GLM) of each participant's fMRI time series, using a 1/128 second highpass filter and AR1 correction for auto-correlation. The design matrix comprised the auditory stimulus

events (onset of the spoken words), each modeled as a stick (delta) function and convolved with the canonical haemodynamic response function specified in SPM software. For signal magnitude analysis, separate columns were specified for each of the stimulus types in addition to a column for button presses (replicated for each of the five experimental blocks). For pattern analysis, the design matrix was identical but separate columns were specified for each individual item.

Pattern analysis was performed using searchlight analysis [86] as implemented in the Decoding toolbox, using spheres with a radius of 8 mm and constrained to voxels within the whole-brain mask generated by SPM during model estimation. Pattern distances were computed using the cross-validated Mahalanobis distance [30]. This ‘Crossnobis’ distance is equivalent to the (cross-validated) Euclidean distance, normalized by the noise covariance between voxels (estimated from the GLM residuals). Cross-validation ensures that the expected pattern distance is zero if two voxel patterns are not statistically different from each other, making it a readily interpretable distance measure. Note that because of this cross-validation, the Crossnobis distance can sometimes be negative if it’s true value is close to zero in the presence of noise. The Crossnobis distance has been demonstrated to be a more reliable and accurate metric for characterizing multivoxel patterns than the correlation or Euclidean distance [30]. To assess whether fMRI responses reflected the phonetic information about Syl2, neural pattern distances were Spearman correlated with the phonetic dissimilarities between Syl2 portions of item pairs using the same Levenshtein distance measure as the MEG analysis (shown in Figure 4A).

For both signal magnitude (contrast images from first-level models or Spearman correlations) and pattern analyses (Spearman correlations), group-level one-sample t-tests were performed for each voxel within a mask covering superior and middle temporal regions (shown in Figure 3A). Inference was conducted parametrically, controlling the family-wise error (FWE) rate using random field theory. Reported effects were obtained by using a cluster defining height threshold of $p < .001$ with a cluster extent threshold of $p < .05$ (FWE corrected), unless otherwise stated.

## *Data availability statement*

Stimuli, data and analysis scripts will be made available on OSF upon publication of this manuscript.

## *Code availability statement*

Matlab code for the computational simulations will be made available on OSF upon publication of this manuscript.

# References


1. Norris D, McQueen JM, Cutler A. Prediction, Bayesian inference and feedback in speech recognition. Language, Cognition and Neuroscience. 2015;3798: 1–15. doi:10.1080/23273798.2015.1081703

2. Norris D, McQueen JM. Shortlist B: a Bayesian model of continuous speech recognition. Psychological review. 2008;115: 357–95. doi:10.1037/0033-295X.115.2.357

3. Davis MH, Sohoglu E. Three functions of prediction error for Bayesian inference in speech perception. Cog Neuro. MIT Press; 2020. pp. 177–192.

4. McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986;18: 1–86. doi:10.1016/0010-0285(86)90015-0

5. Lange FP de, Heilbron M, Kok P. How Do Expectations Shape Perception? Trends in Cognitive Sciences. 2018;22: 764–779. doi:10.1016/j.tics.2018.06.002

6. Murray SO, Schrater P, Kersten D. Perceptual grouping and the interactions between visual cortical areas. Neural networks : the official journal of the International Neural Network Society. 2004;17: 695–705. doi:10.1016/j.neunet.2004.03.010

7. Aitken F, Turner G, Kok P. Prior Expectations of Motion Direction Modulate Early Sensory Processing. The Journal of Neuroscience. 2020;40: 6389–6397. doi:10.1523/JNEUROSCI.0537-20.2020

8. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience. 1999;2: 79–87. doi:10.1038/4580

9. Aitchison L, Lengyel M. With or without you: predictive coding and Bayesian inference in the brain. Current Opinion in Neurobiology. 2017;46: 219–227. doi:10.1016/j.conb.2017.08.010

10. Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2005;360: 815–36. doi:10.1098/rstb.2005.1622

11. Gagnepain P, Henson RN, Davis MH. Temporal Predictive Codes for Spoken Words in Auditory Cortex. Current biology : CB. 2012;22: 1–7. doi:10.1016/j.cub.2012.02.015

12. Brodbeck C, Hong LE, Simon JZ. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Current biology : CB. 2018;28: 3976-3983.e5. doi:10.1016/j.cub.2018.10.042

13. Heilbron M, Armeni K, Schoffelen J-M, Hagoort P, de Lange FP. A hierarchy of linguistic predictions during natural language comprehension. Proc Natl Acad Sci USA. 2022;119: e2201968119. doi:10.1073/pnas.2201968119

14. Donhauser PW, Baillet S. Two Distinct Neural Timescales. Neuron. 2020; 1–9. doi:10.1016/j.neuron.2019.10.019

15. Keshishian M, Akkol S, Herrero J, Bickel S, Mehta AD, Mesgarani N. Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. Nat Hum Behav. 2023; 1–14. doi:10.1038/s41562-023-01520-0

16. Tezcan F, Weissbart H, Martin AE. A tradeoff between acoustic and linguistic feature encoding in spoken language comprehension. Peelle JE, King AJ, editors. eLife. 2023;12: e82386. doi:10.7554/eLife.82386

17. Bastos AMM, Usrey WMM, Adams RAA, Mangun GRR, Fries P, Friston KJJ. Canonical Microcircuits for Predictive Coding. Neuron. 2012;76: 695–711. doi:10.1016/j.neuron.2012.10.038

18. Blank H, Davis MH. Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS biology. 2016;14: e1002577. doi:10.1371/journal.pbio.1002577

19. de Lange FP, Schmitt L-M, Heilbron M. Reconstructing the predictive architecture of the mind and brain. Trends in Cognitive Sciences. 2022;26: 1018–1019. doi:10.1016/j.tics.2022.08.007

20. Luthra S, Li MYC, You H, Brodbeck C, Magnuson JS. Does signal reduction imply predictive coding in models of spoken word recognition? Psychonomic bulletin & review. 2021. doi:10.3758/s13423-021-01924-x

21. Sohoglu E, Davis MH. Rapid computations of spectrotemporal prediction error support perception of degraded speech. eLife. 2020;9: 1–25. doi:10.7554/eLife.58077

22. Huettig F, Mani N. Is prediction necessary to understand language? Probably not. Language, Cognition and Neuroscience. 2016;31: 19–31. doi:10.1080/23273798.2015.1072223

23. Gwilliams L, Davis M. Extracting language content from speech sounds: An information theoretic approach. Springer Handbook of Auditory Research. Springer; 2022. pp. 113–139.

24. Grosjean F. Spoken word recognition processes and the gating paradigm. Perception & Psychophysics. 1980;28: 267–283. doi:10.3758/BF03204386

25. Luce PA, Pisoni DB. Recognizing spoken words: the neighborhood activation model. Ear and hearing. 1998;19: 1–36. doi:10.1097/00003446-199802000-00001

26. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nature Neuroscience. 2013;16: 1170–1178. doi:10.1038/nn.3495

27. Brodbeck C, Bhattasali S, Cruz Heredia AA, Resnik P, Simon JZ, Lau E. Parallel processing in speech perception with local and global representations of linguistic context. eLife. 2022;11: e72056. doi:10.7554/eLife.72056

28. Loftus GR, Masson MEJ. Using confidence intervals in within-subject designs. Psychonomic Bulletin & Review. 1994;1: 476–490. doi:10.3758/BF03210951

29. Gwilliams L, King J-R, Marantz A, Poeppel D. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. Nat Commun. 2022;13: 6606. doi:10.1038/s41467-022-34326-1

30. Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage. 2016;137: 188–200. doi:10.1016/j.neuroimage.2015.12.012

31. Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM. The representational dynamics of task and object processing in humans. eLife. 2018;7: 1–21. doi:10.7554/eLife.32816

32. Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of Modern Physics. 1993;65: 413–497. doi:10.1103/RevModPhys.65.413

33. Lankinen K, Ahlfors SP, Mamashli F, Blazejewska AI, Raij T, Turpin T, et al. Cortical depth profiles of auditory and visual 7 T functional MRI responses in human superior temporal areas. Human Brain Mapping. 2023;44: 362–372. doi:10.1002/hbm.26046

34. Arnal LH, Wyart V, Giraud A-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature neuroscience. 2011;14: 797–801. doi:10.1038/nn.2810

35. Bastos AM, Lundqvist M, Waite AS, Kopell N, Miller EK. Layer and rhythm specificity for predictive routing. Proceedings of the National Academy of Sciences of the United States of America. 2020;117: 31459–31469. doi:10.1073/pnas.2014868117

36. 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: 371–376. doi:10.1016/j.cub.2015.12.038

37. Leonard MK, Gwilliams L, Sellers KK, Chung JE, Xu D, Mischler G, et al. Large-scale single-neuron speech sound encoding across the depth of human cortex. Nature. 2023 [cited 1 Feb 2024]. doi:10.1038/s41586-023-06839-2

38. Zhuang J, Tyler LK, Randall B, Stamatakis EA, Marslen-Wilson WD. Optimally efficient neural systems for processing spoken language. Cerebral Cortex. 2014;24: 908–918. doi:10.1093/cercor/bhs366

39. Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005;46: 681–92. doi:10.1016/j.neuron.2005.04.026

40. Davis MH, Gaskell MG. A complementary systems account of word learning: neural and behavioural evidence. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2009;364: 3773–800. doi:10.1098/rstb.2009.0111

41. Blanco-Elorrieta E, Gwilliams L, Marantz A, Pylkkänen L. Adaptation to mis-pronounced speech: evidence for a prefrontal-cortex repair mechanism. Sci Rep. 2021;11: 97. doi:10.1038/s41598-020-79640-0

42. Blank H, Spangenberg M, Davis MH. Neural Prediction Errors Distinguish Perception and Misperception of Speech. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2018;38: 6076–6089. doi:10.1523/jneurosci.3258-17.2018

43. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. 2013;36: 1–24. doi:10.1017/S0140525X12000477

44. Barascud N, Pearce MT, Griffiths TD, Friston KJ, Chait M. Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. Proceedings of the National Academy of Sciences of the United States of America. 2016;113: E616-25. doi:10.1073/pnas.1508523113

45. Friston K. Does predictive coding have a future? Nature Neuroscience. 2018;21: 1019–1021. doi:10.1038/s41593-018-0200-7

46. Kuperberg GR, Jaeger TF. What do we mean by prediction in language comprehension? Language Cognition & Neuroscience. 2016;31: 32–59. doi:10.1080/23273798.2015.1102299

47. Broderick MP, Anderson AJ, Lalor EC. Semantic Context Enhances the Early Auditory Encoding of Natural Speech. The Journal of Neuroscience. 2019;39: 7564–7575. doi:10.1523/JNEUROSCI.0584-19.2019

48. Wang YC, Sohoglu E, Gilbert RA, Henson RN, Davis MH. Predictive Neural Computations Support Spoken Word Recognition: Evidence from MEG and Competitor Priming. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2021;41: 6919–6932. doi:10.1523/JNEUROSCI.1685-20.2021

49. Greve A, Cooper E, Kaula A, Anderson MC, Henson R. Does prediction error drive one-shot declarative learning? Journal of Memory and Language. 2017;94: 149–165. doi:10.1016/j.jml.2016.11.001

50. Schmitt L-M, Erb J, Tune S, Rysop AU, Hartwigsen G, Obleser J. Predicting speech from a cortical hierarchy of event-based time scales. Science Advances. 2021;7: 2020.12.19.423616. doi:10.1126/sciadv.abi6070

51. Caucheteux C, Gramfort A, King J-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat Hum Behav. 2023; 1–12. doi:10.1038/s41562-022-01516-2

52. Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, et al. The neural architecture of language: Integrative modeling converges on predictive processing. Proc Natl Acad Sci USA. 2021;118: e2105646118. doi:10.1073/pnas.2105646118

53. Frank SL, Willems RM. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Language, Cognition and Neuroscience. 2017;32: 1192–1203. doi:10.1080/23273798.2017.1323109

54. Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, et al. Shared computational principles for language processing in humans and deep language models. Nat Neurosci. 2022;25: 369–380. doi:10.1038/s41593-022-01026-4

55. MacGregor LJ, Pulvermüller F, van Casteren M, Shtyrov Y. Ultra-rapid access to words in the brain. Nature communications. 2012;3: 711. doi:10.1038/ncomms1715

56. Tabas A, Mihai G, Kiebel S, Trampel R, von Kriegstein K. Abstract rules drive adaptation in the subcortical sensory pathway. Shinn-Cunningham BG, Griffiths TD, Malmierca MS, editors. eLife. 2020;9: e64501. doi:10.7554/eLife.64501

57. Sedley W, Gander PE, Kumar S, Kovach CK, Oya H, Kawasaki H, et al. Neural signatures of perceptual inference. eLife. 2016;5: 1–13. doi:10.7554/eLife.11476

58. Lerner Y, Honey CJ, Silbert LJ, Hasson U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2011;31: 2906–15. doi:10.1523/JNEUROSCI.3684-10.2011

59. Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, Devinsky O, et al. Slow cortical dynamics and the accumulation of information over long timescales. Neuron. 2012;76: 423–34. doi:10.1016/j.neuron.2012.08.011

60. Norman-Haignere SV, Long LK, Devinsky O, Doyle W, Irobunda I, Merricks EM, et al. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. Nature Human Behaviour. 2022. doi:10.1038/s41562-021-01261-y

61. Di Liberto GM, Crosse MJ, Lalor EC. Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. eNeuro. 2018;5: ENEURO.0084-18.2018. doi:10.1523/ENEURO.0084-18.2018

62. Giraud A-L, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience. 2012 [cited 19 Mar 2012]. doi:10.1038/nn.3063

63. Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Berger MS, et al. High gamma power is phase-locked to theta oscillations in human neocortex. Science (New York, NY). 2006;313: 1626–8. doi:10.1126/science.1128115

64. Sedley W, Teki S, Kumar S, Overath T, Barnes GR, Griffiths TD. Gamma band pitch responses in human auditory cortex measured with magnetoencephalography. Neuroimage. 2012;59: 1904–1911. doi:10.1016/j.neuroimage.2011.08.098

65. Norris D. How do Computational Models Help us Develop Better Theories? Twenty-first century psycholinguistics: Four cornerstones. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2005. pp. 331–346.

66. Nour Eddine S, Brothers T, Wang L, Spratling M, Kuperberg GR. A predictive coding model of the N400. Cognition. 2024;246: 105755. doi:10.1016/j.cognition.2024.105755

67. Hovsepyan S, Olasagasti I, Giraud A. Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. Nature Communications. 2020;11: 3117. doi:10.1038/s41467-020-16956-5

68. Yildiz IB, von Kriegstein K, Kiebel SJ. From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems. PLoS Computational Biology. 2013;9. doi:10.1371/journal.pcbi.1003219

69. Magnuson JS, You H, Luthra S, Li M, Nam H, Escabí M, et al. EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. Cognitive science. 2020;44: e12823. doi:10.1111/cogs.12823

70. Kell AJE, Yamins DLK, Shook EN, Norman-haignere SV, Mcdermott JH, Kell AJE, et al. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron. 2018;98: 1–15. doi:10.1016/j.neuron.2018.03.044

71. Baayen RH, Piepenbrock R, Van Rijn H. The CELEX lexical database (CD-ROM). Philadelphia Linguistics Data Consortium University of Pennsylvania. 1993.

72. Kawahara H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoust Sci & Tech. 2006;27: 349–353. doi:10.1250/ast.27.349

73. Rogers JC, Davis MH. Inferior Frontal Cortex Contributions to the Recognition of Spoken Words and Their Constituent Speech Sounds. Journal of cognitive neuroscience. 2017;25: 1–18. doi:10.1162/jocn_a_01096

74. Brainard DH. The Psychophysics Toolbox. Spatial Vision. 1997;10: 433–436. doi:10.1163/156856897X00357

75. Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive Top-Down Integration of Prior Knowledge during Speech Perception. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2012;32: 8443–8453. doi:10.1523/JNEUROSCI.5069-11.2012

76. Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Top-down influences of written text on perceived clarity of degraded speech. Journal of Experimental Psychology: Human Perception and Performance. 2014;40: 186–199. doi:10.1037/a0033206

77. Sohoglu E, Davis MH. Perceptual learning of degraded speech by minimizing prediction error. Proceedings of the National Academy of Sciences. 2016;113: E1747–E1756. doi:10.1073/pnas.1523266113

78. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. The Journal of the Acoustical Society of America. 2005;118: 887–906. doi:10.1121/1.1945807

79. Kisler T, Reichel U, Schiel F. Multilingual processing of speech via web services. Computer Speech & Language. 2017;45: 326–347. doi:10.1016/j.csl.2017.01.005

80. Warner N, McQueen JM, Cutler A. Tracking perception of the sounds of English. The Journal of the Acoustical Society of America. 2014;135: 2995–3006. doi:10.1121/1.4870486

81. Hämäläinen MS. Functional localization based on measurements with a whole-head magnetometer system. Brain topography. 1995;7: 283–9. doi:10.1007/BF01195254

82. Taulu S, Simola J, Kajola M. Applications of the signal space separation method. IEEE Transactions on Signal Processing. 2005;53: 3359–3372. doi:10.1109/TSP.2005.853302

83. Hebart MN, Görgen K, Haynes J-D. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Frontiers in Neuroinformatics. 2015;8. Available: https://www.frontiersin.org/articles/10.3389/fninf.2014.00088

84. Keller GB, Mrsic-Flogel TD. Predictive Processing: A Canonical Cortical Computation. Neuron. 2018;100: 424–435. doi:10.1016/j.neuron.2018.10.003

85. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. Journal of neuroscience methods. 2007;164: 177–90. doi:10.1016/j.jneumeth.2007.03.024

86. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America. 2006;103: 3863–3868. doi:10.1073/pnas.0600244103

# SI Methods

## Cued recall task

After the fMRI scans, listeners completed a cued recall test assessing memory for the pseudowords. Data were not available for this task for four participants either because of technical issues or because the participant did not understand the task instructions. The final sample size consisted of 17 participants for this test. We cued participants with the first syllable of each pseudoword item by presenting the same stimuli as the MEG and fMRI experiments but replacing the second syllables with spectrally and intensity (RMS) matched noise. We refer to these as Strong+Noise and Weak+Noise items. Participants were told that the nonsense (pseudo) words were “partly hidden by noise” and were asked to recall the nonsense word by saying it aloud. Responses were scored for accuracy based on counting the percentage of trials in which participants generated the correct Syl2 used in the Mismatch items. The order of stimulus presentation was randomised for each participant.

## Web-based gating study

Our measure of p(Syl2|Syl1) is derived from frequency counts and phonological transcriptions in the CELEX database. To validate this measure, we conducted a web-based gating task using the Strong+Noise and Weak+Noise stimuli (each item presented once in a random order). Participants were British English speakers (N=110, recruited via https://www.prolific.co and non-overlapping with participants in the MEG/fMRI studies), aged 18 to 40 years and self-reported as having no hearing difficulties or language-related disorders. They were told that they would hear real English words “partly hidden by noise” and asked to report the identity of the word by typing it out. For each item, we calculated the proportion of participants reporting the source word i.e. the matching word prior to replacing Syl2 with noise.

# SI Results

## Stimulus properties

A summary of stimulus properties is shown in Figure S4A and S4B. Based on frequency counts, and phonological transcriptions in CELEX, the average conditional probability of syllable 2 given syllable 1 (p(Syl2|Syl1)) for Strong prediction items was 0.574 (SD = 0.332) while for Weak items it was 0.019 (SD = 0.029). Strong items were also associated with smaller syllable entropy (calculated as the negative sum of log-transformed p(Syl2|Syl1) over all possible syllables; mean = 1.34, SD = 1.11) than Weak items (mean = 2.99, SD = 0.99). Strong and Weak items had similar log frequencies (Strong: mean = 4.90, SD = 1.56; Weak: mean = 3.89, SD = 1.62) although the difference was significant (independent samples $t(62) = 2.53$, $p = 0.014$). However, all reported MEG and fMRI effects involving correlations with conditional probabilities (Figures 2B, 3B and 4D) survive when controlling for differences in word frequency (using partial Spearman correlations). Moreover, p(Syl2|Syl1) is the strongest predictor of listeners' behaviour in a free report gating task (conducted in a separate group of listeners; see Figure S1E and SI Methods for more details), as compared with either word frequency or syllable entropy (see web-based gating study results reported below). Thus, we are confident that our effects of prediction strength after Syl2 onset stem from differences in conditional probability rather than differences in word frequency.

For the fMRI experiment (Figure S4A), the Syl1 portion of Strong items was lower in acoustic intensity ($t(62) = -2.14$, $p = 0.036$) and longer in duration ($t(62) = 3.47$, $p < .001$) than Weak items. The RMS intensity for Syl1 and Syl2 was separately normalised over items for the MEG experiment (which was conducted after the fMRI experiment) and did not differ between conditions (Figure S4B). All other stimulus properties did not differ between the fMRI and MEG experiments. Importantly, there were no acoustic differences between conditions for the Syl2 portion of our stimuli which were acoustically identical in all conditions. Neural responses to Syl2 are the key focus of this study since it is representations of Syl2 that distinguish sharpened signal and prediction error accounts.

## Cued recall task

After the fMRI scans, we assessed listeners' memory for the pseudowords in a cued recall task (see SI Methods). Recall accuracy (percentage correct) was around 10% on average over participants, with most participants achieving 5% or higher (see Figure S1D). These data indicate the recall task was difficult. Importantly however, they also suggest that participants could perform the task to some degree and hence were attentively processing the pseudoword stimuli. Recall accuracy was better for pseudowords formed with Strongly- versus Weakly-predicting first syllables (one-tailed $t(16) = 1.768$, $p = .048$), consistent with the proposal that items evoking strong prediction errors are learned more readily than items evoking weak prediction errors [see Discussion].

## Web-based gating study

Listeners' reports of the identity of spoken words with second syllables replaced by noise (Strong+Noise and Weak+Noise stimuli, see SI Methods) were combined over participants (N=110) to generate the proportion of responses that correspond to the target words. For these items, p(Syl2|Syl1) was strongly correlated with the proportion of listeners' responses matching the target word ($r = .636$, $p < .001$, N = 64 items; see Figure S1E) and this effect remains significant when other lexical predictors (entropy and word frequency) are factored out ($r = 0.341$, $p < 0.01$, N = 64 items). This finding shows that more strongly predicted second syllables were more likely to be reported by listeners than weakly predicted syllables when replaced with noise. Moreover, while both entropy and word frequency also correlate with gating responses ($r = -0.513$, $p < .001$ N = 64 items; and $r = 0.259$, $p < .05$, N = 64 items), neither of these factors predict gating responses when p(Syl2|Syl1) is factored out (both $p > .05$, see Figure S1E). These findings confirm that our experimental manipulation of prediction strength was successful, and that measures of prediction strength based on conditional probability computed from lexical statistics are an accurate predictor of listening behaviour.

# Supplementary figures

## Figure S1

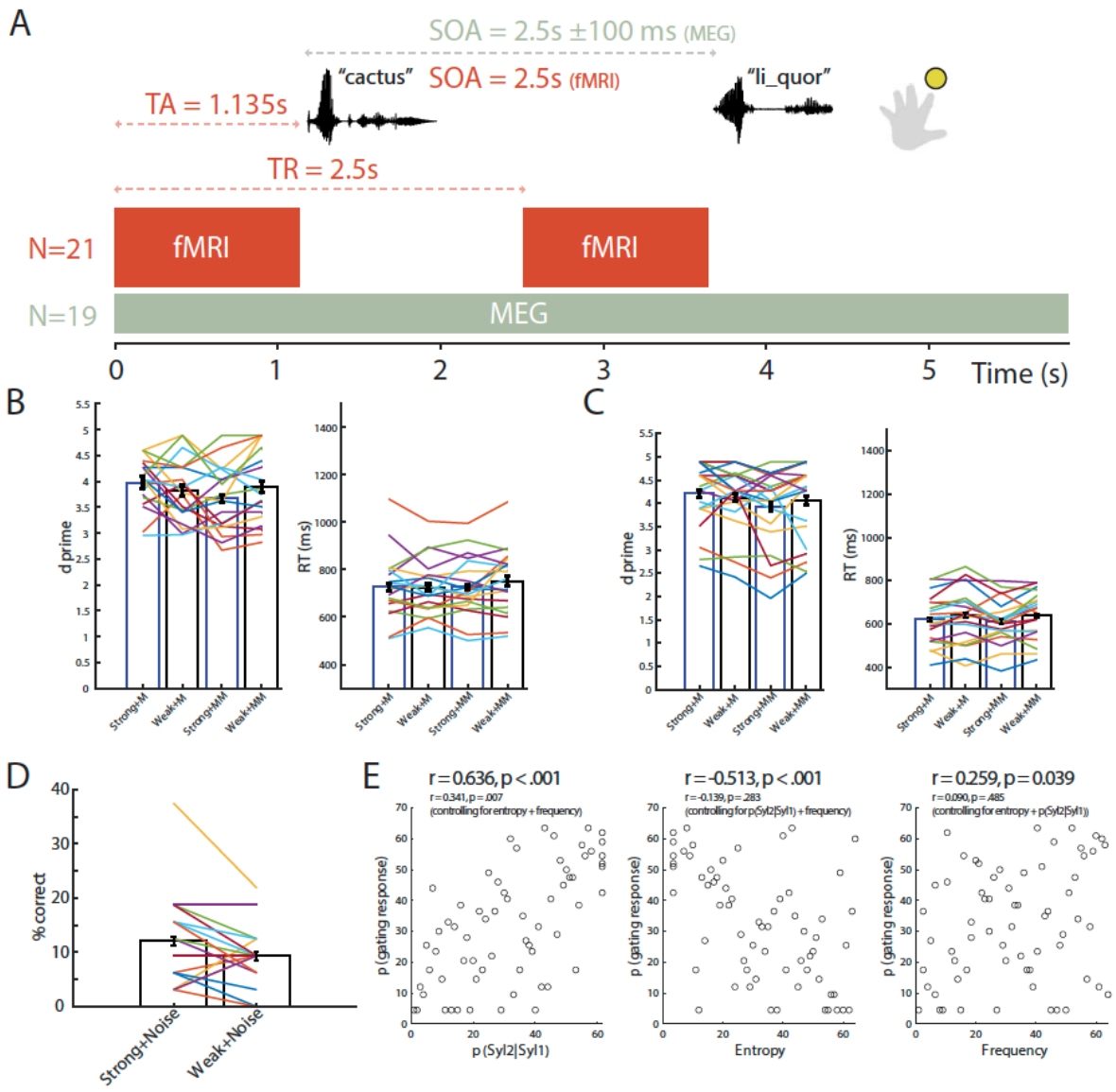



*Figure S1- Method and behavioural results. (A) Participants listened to the spoken stimuli and detected a brief pause that appeared occasionally between Syl1 and Syl2 by responding with a button press. MEG and fMRI were recorded in two separate groups of participants (N = sample size for each group). Two example trials are indicated with stimulus and scan timings. TA = time of fMRI scan acquisition. TR = repetition time for fMRI scans. SOA = Stimulus Onset Asynchrony. s = seconds. ms = milliseconds. (B) Bar graph showing group-averaged d-primes and response times (target-present trials) for the pause detection task in the MEG experiment (with within-subject standard errors). Individual lines represent individual participants. (C) Same as panel B but for pause detection responses in the fMRI experiment. (D) Percentage correct accuracy of pseudoword memory in a cued recall test conducted after the fMRI scans. For this recall test, participants were presented with the stimuli after the second syllables were replaced with noise ('Strong+Noise' and 'Weak+Noise' items) and asked to recall previously heard pseudowords that started with the same first syllable. (E) Scatter plots of the proportion of participants identifying target items in a web-based gating pre-test study plotted against lexical statistics for p(Syl2|Syl1), entropy and word frequency (each circle representing one of the 64 word items used in the*

*experiment). Correlation coefficients (and p values) for each relationship are shown above each graph with further partial correlations reported within brackets.*

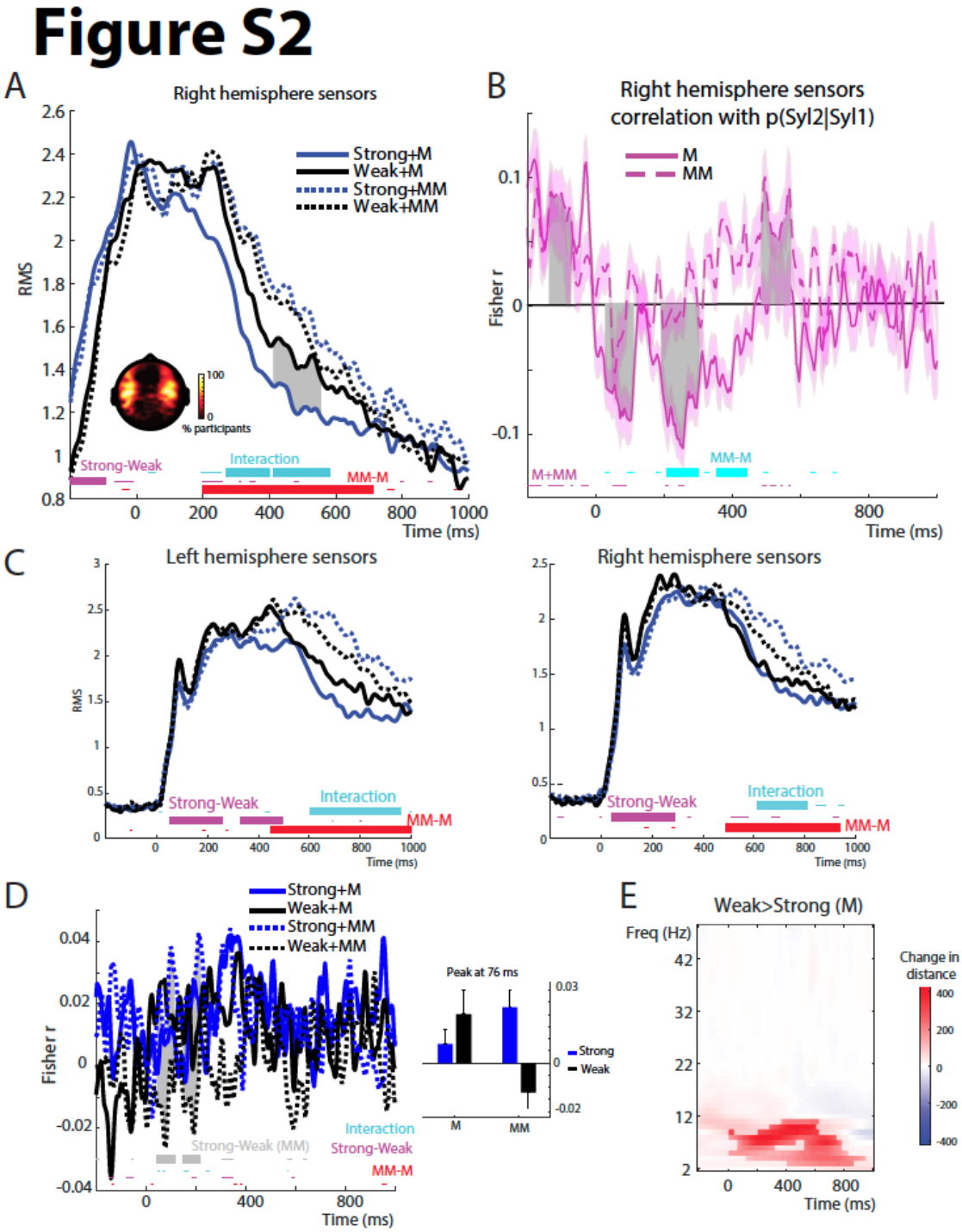



*Figure S2- Supplementary MEG analyses. (A) Same as Figure 2A but for right hemisphere sensors. (B) Same as Figure 2B but for right hemisphere sensors. (C) Same as Figure 2A for MEG signal timelocked to Syl1 onset. (D) Group-averaged correlation between MEG pattern distances and Syl2 phonetic dissimilarities in each condition. Thin horizontal bars at the bottom of the graph indicate timepoints when main effects and the interaction were significant at an uncorrected p < .05 threshold (no timepoints survived correction for multiple comparisons). Grey shaded areas indicate the two clusters during which there was a significant difference between Strong and Weak*

*items in the Mismatch condition (p < .05 FWE clusterwise corrected). The inset bar plot shows group means for each condition (and within-subject standard errors as before) after averaging over the temporal extent of the first cluster (peaking at 76 ms). (E) Pattern analysis showing simple effect of Weak > Strong (Match) on MEG pattern distances.*

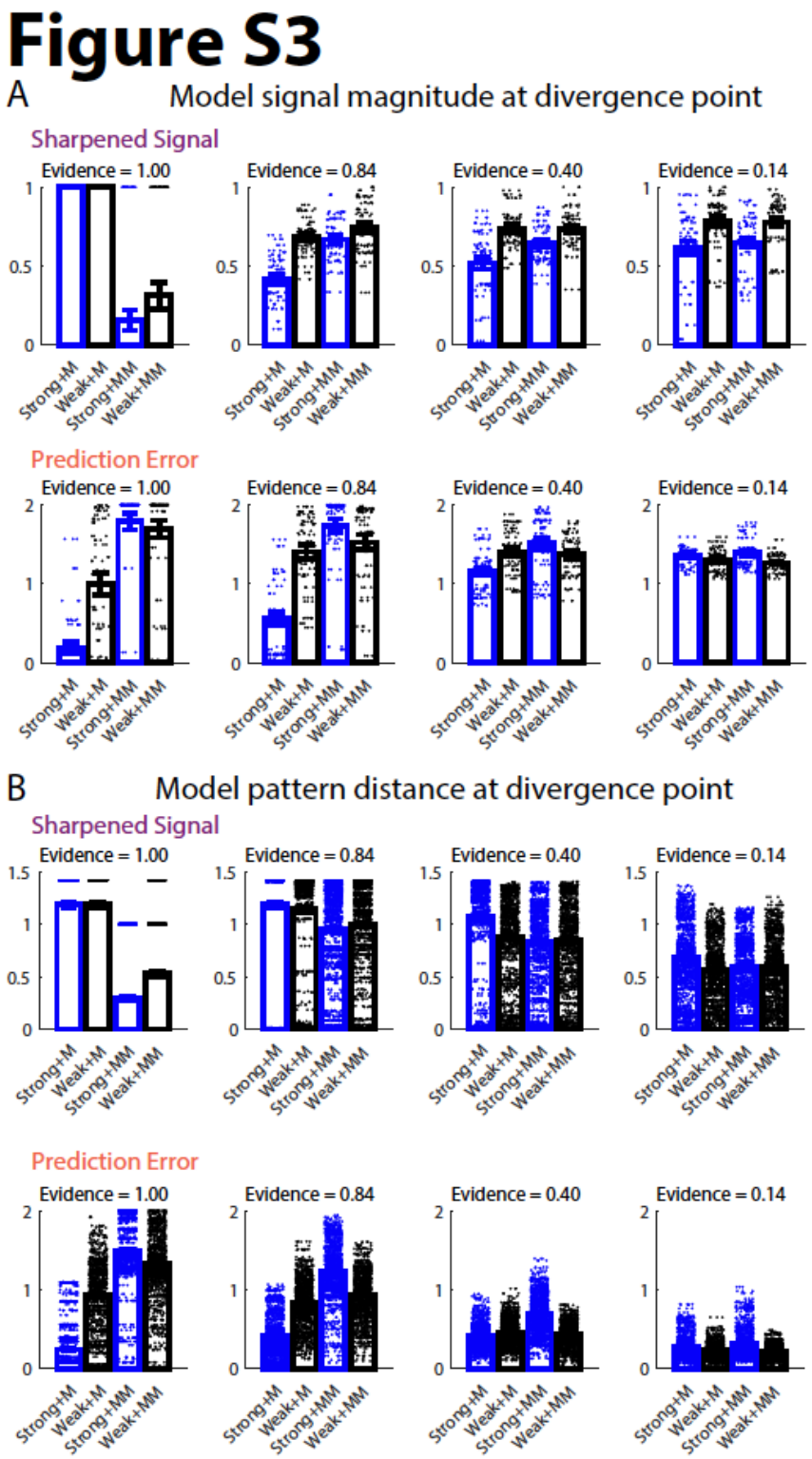



*Figure S3- Same as Figure 1D and 1E but showing the impact of varying sensory uncertainty levels (using the Softmax Temperature parameter) for Sharpened Signal (upper panels) simulations and Prediction Error (lower*

*panels) for simulations of (A) Univariate signal magnitude and (B) Multivariate pattern distance results. For univariate simulations, sharpened signals were summarised as the sum of log activity over segment units while prediction errors were summarised as the sum of absolute activity. To give an indication of the degree of sensory uncertainty for each simulation, above each graph we show the average along the diagonal of the acoustic similarity matrix used for modelling the sensory evidence. This quantity can be interpreted as the average probability of correctly identifying speech segments. The simulation results depicted in Figure 1D and 1E correspond to the graphs in the second column (i.e. Evidence = 0.84).*

# Figure S4

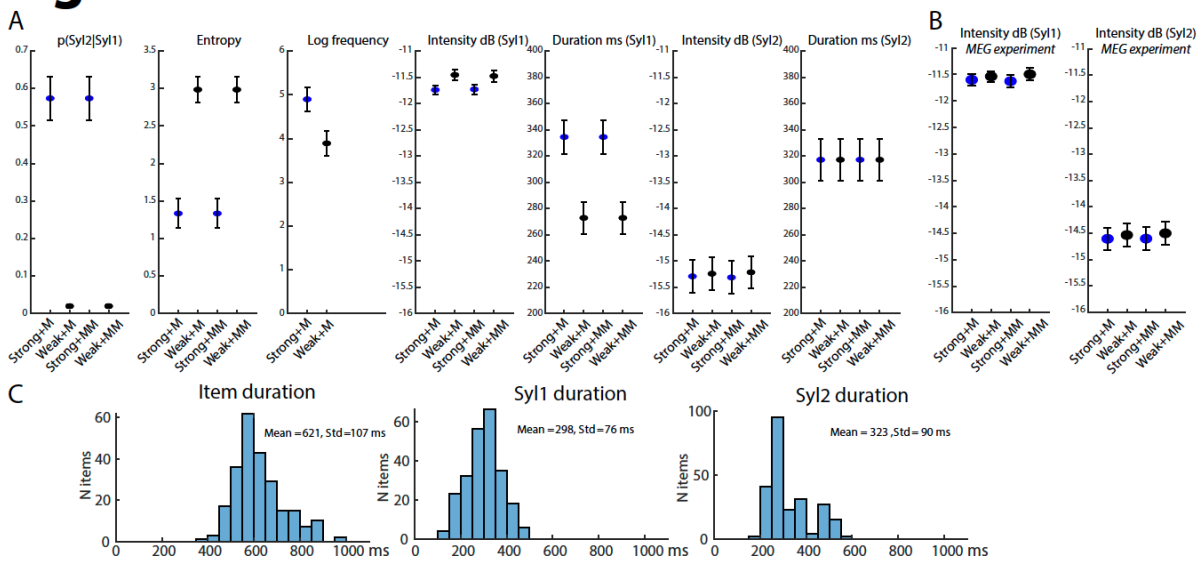



*Figure S4- (A) Stimulus properties for the fMRI experiment (means and standard errors of the mean). (B) The RMS intensity of the stimuli for the MEG experiment, which were slightly different to those for the fMRI experiment (all other stimulus properties did not differ) (C) Histograms showing the distribution of item durations in our stimuli as well as Syl1 and Syl2 durations.*

# Figure S5

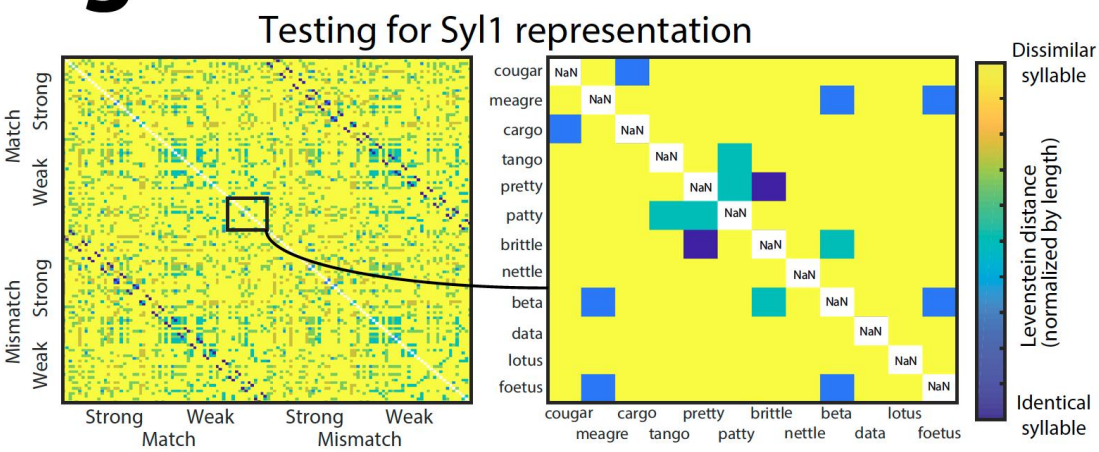



*Figure S5- Same as Figure 4A. Phonetic dissimilarity matrix expressing the normalized Levenshtein distance between phonetic transcriptions of Syl1 in different items.*

# Figure S6

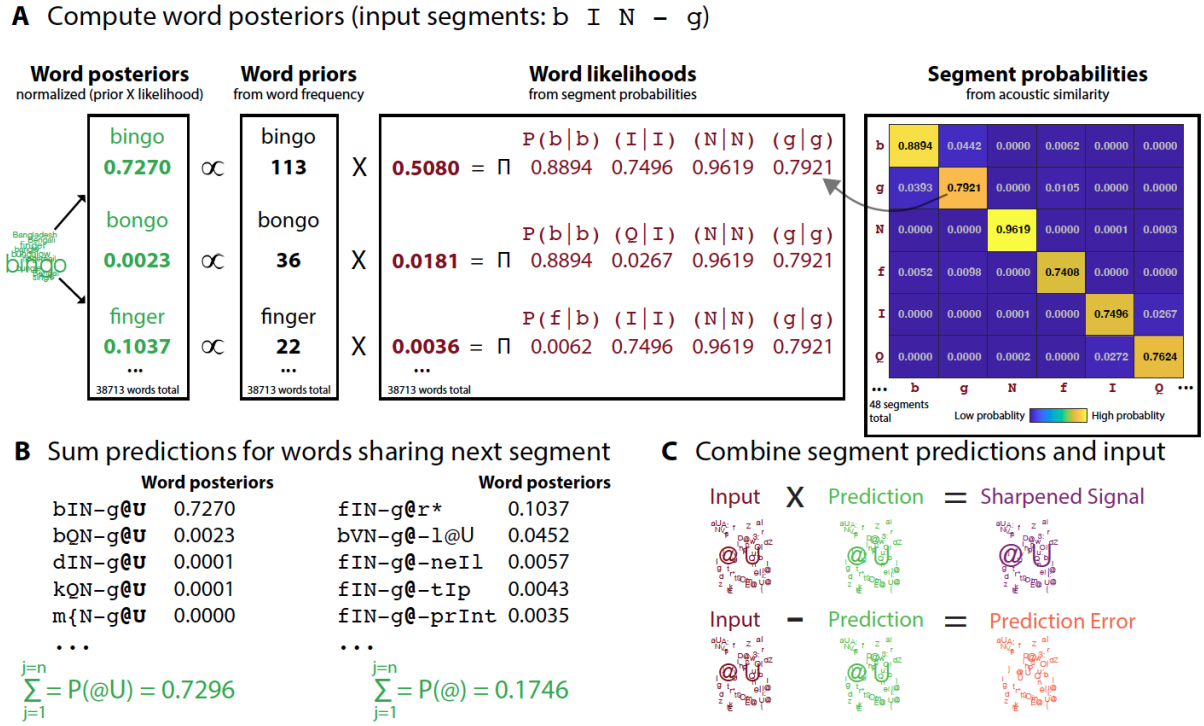
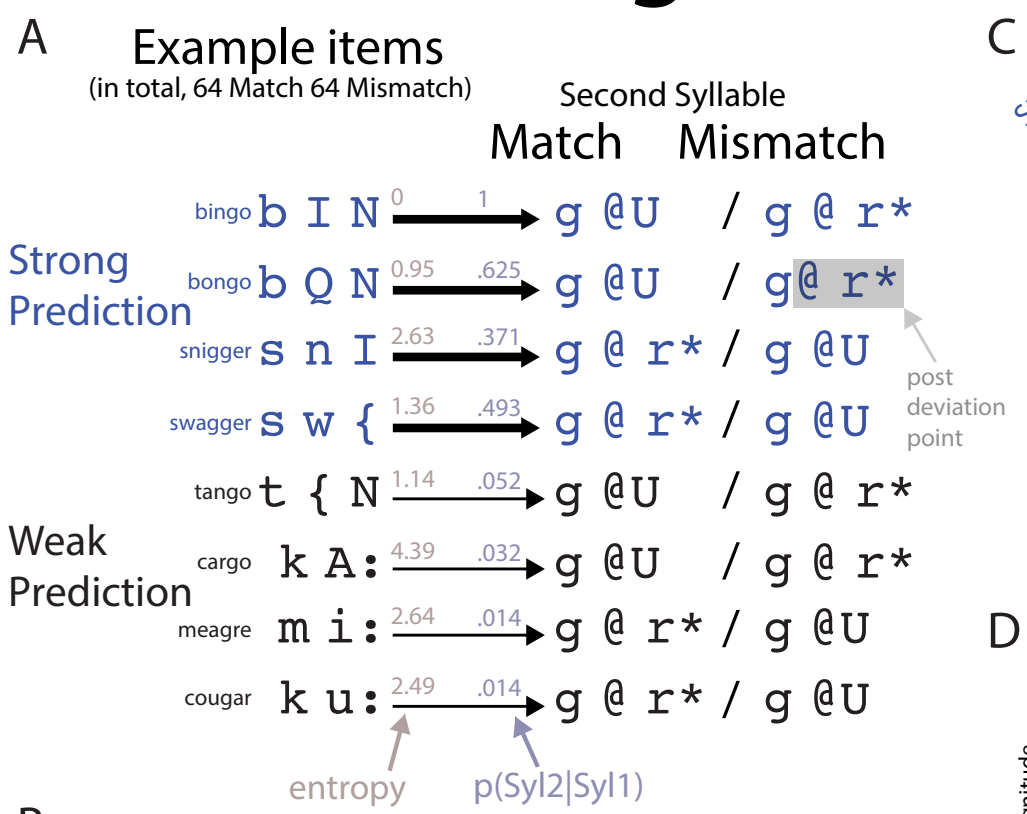
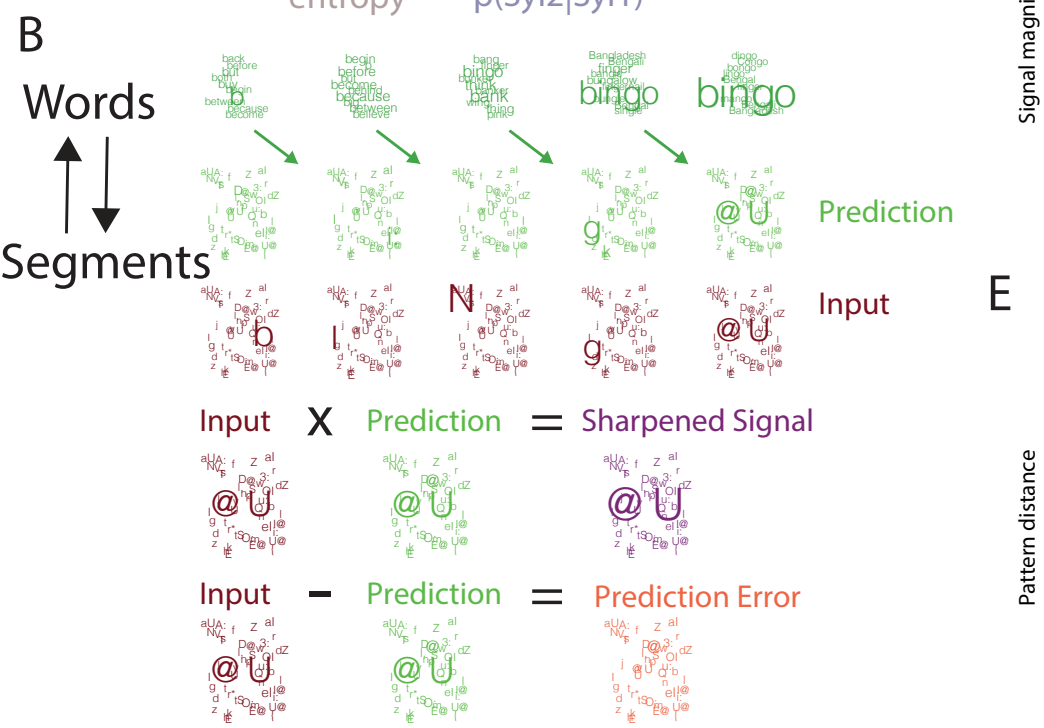
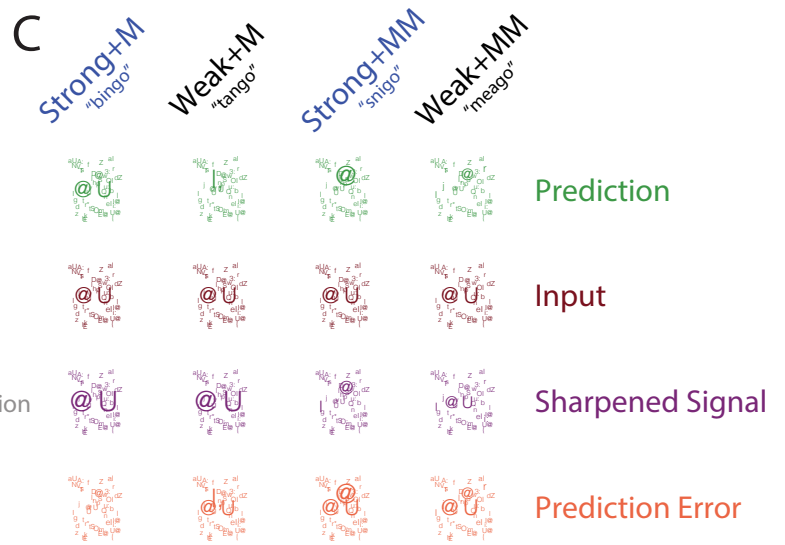


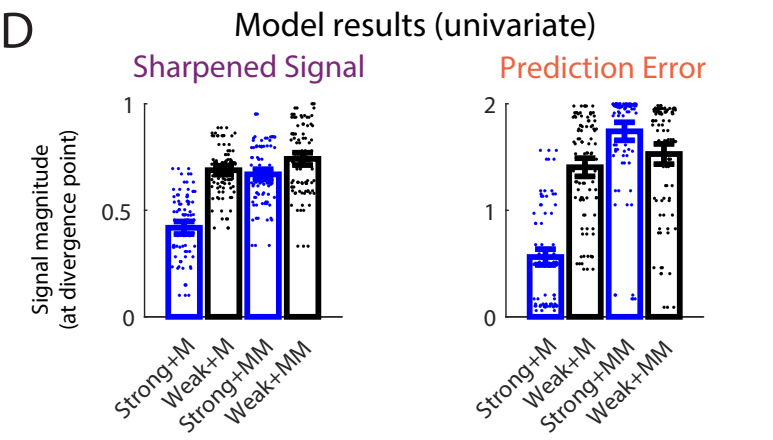

*Figure S6- Illustration of each step in simulations for the example sequence of segments /b I N – g/ (i.e. after the first segment in the second syllable of "bingo"). A) The first step in the simulations is to compute the posterior probability of each word (shown in green for three selected words), by multiplying word priors (frequency of usage as shown, each of which is divided by the total word count for the corpus, 18585424) with word likelihoods (shown in brown) and normalising the result to sum to one. We compute word likelihoods as the product of segment probabilities, which are based on the acoustic similarity between segments. Selected segment probabilities are shown in the probability matrix in top right of figure. Each row in this matrix can be interpreted as the sensory evidence for different segments given an input segment (with yellow indicating strong evidence and blue indicating weak evidence). For example, given the input segment /b/, the sensory evidence for /b/ is high (P=0.8894), but also to some degree, there is evidence for acoustically similar segments /g/ (P=0.0442) and /f/ (P=0.0062). The non-zero probability of segment confusions is sufficient to make the likelihood of other, similar-sounding words non-zero (e.g. "finger"). B) In the second step, we compute predictions for each upcoming speech segment by summing posterior probabilities over words sharing the same next segment. Here we show posterior probabilities for two example segments /@U/ in "bingo", "bongo", etc, and /@/ in "finger", "bungalow", etc. Probabilities are similarly computed for each possible subsequent segment. C) In the final step, segment predictions (green distribution) are multiplied by individual segment probabilities (brown distribution) and normalized to sum to one to simulate the sharpened signal computation. To simulate the prediction error computation, segment predictions are instead subtracted from segment probabilities. The resulting distributions show how representations of expected segments (/@U/) are enhanced in sharpened signal computations (purple),and suppressed in prediction error computations (orange).*

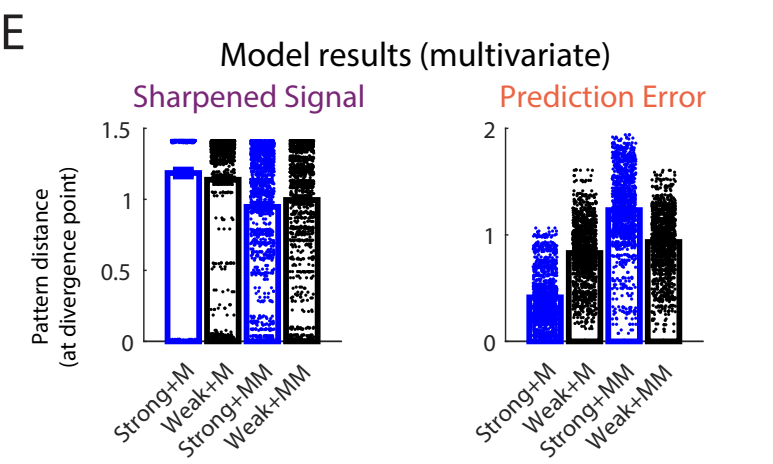
# Figure 1


A
Example items
(in total, 64 Match 64 Mismatch)
Second Syllable
Match
Mismatch
Strong Prediction
Weak Prediction
bingo b I N 0 1 g @U / g @ r*
bongo b Q N 0.95 .625 g @U / g @ r*
snigger s n I 2.63 .371 g @ r* / g @U
swagger s w { 1.36 .493 g @ r* / g @U
tango t { N 1.14 .052 g @U / g @ r*
cargo k A: 4.39 .032 g @U / g @ r*
meagre m i: 2.64 .014 g @ r* / g @U
cougar k u: 2.49 .014 g @ r* / g @U
post deviation point
entropy
p(Syl2|Syl1)

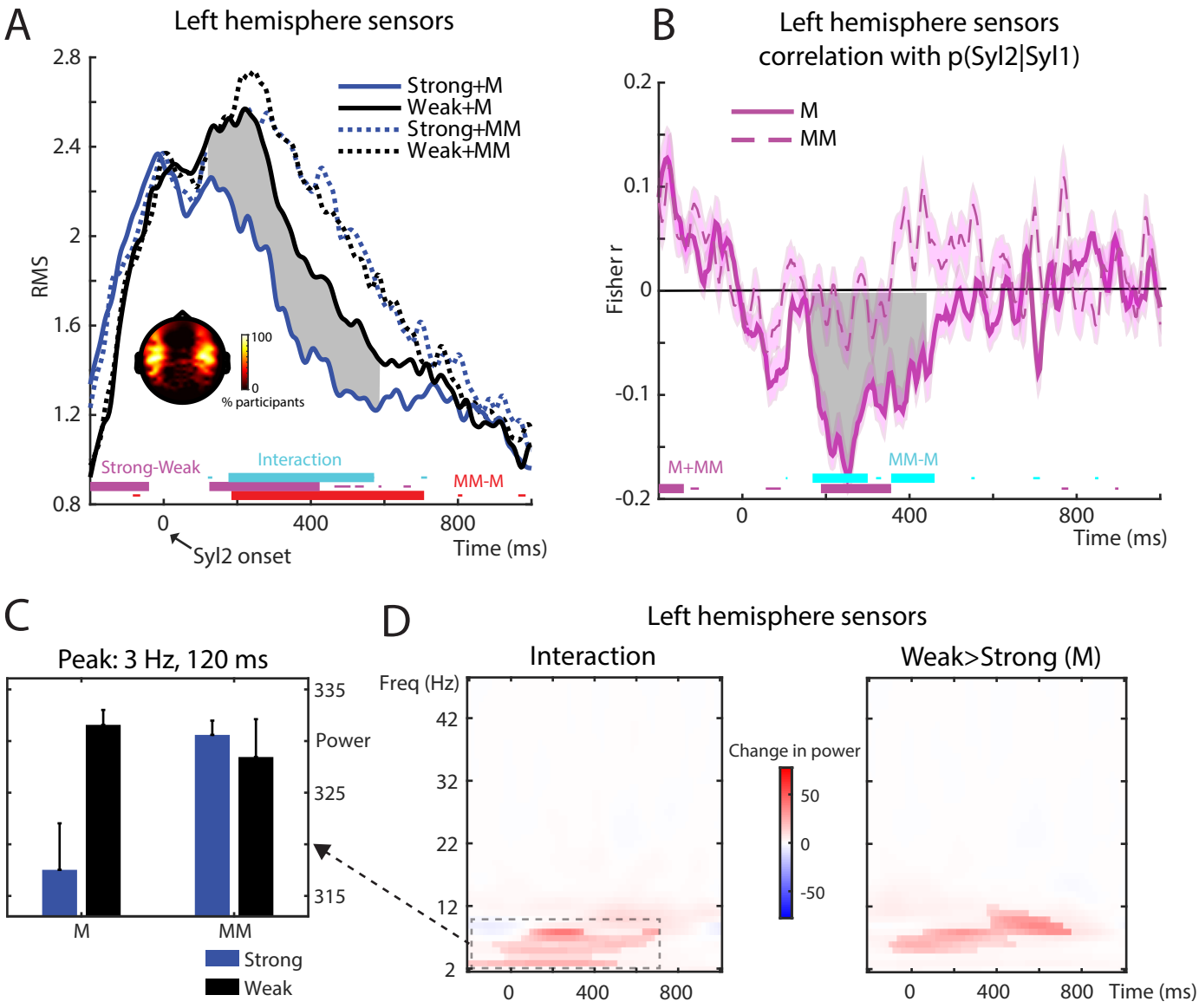

B
Words
Segments
Prediction
Input
Input X Prediction = Sharpened Signal
Input - Prediction = Prediction Error


C
Strong+M "bingo"
Weak+M "tango"
Strong+MM "snigo"
Weak+MM "meago"
Prediction
Input
Sharpened Signal
Prediction Error


D
Model results (univariate)
Sharpened Signal
Prediction Error
Signal magnitude (at divergence point)
Strong+M
Weak+M
Strong+MM
Weak+MM


E
Model results (multivariate)
Sharpened Signal
Prediction Error
Pattern distance (at divergence point)
Strong+M
Weak+M
Strong+MM
Weak+MM

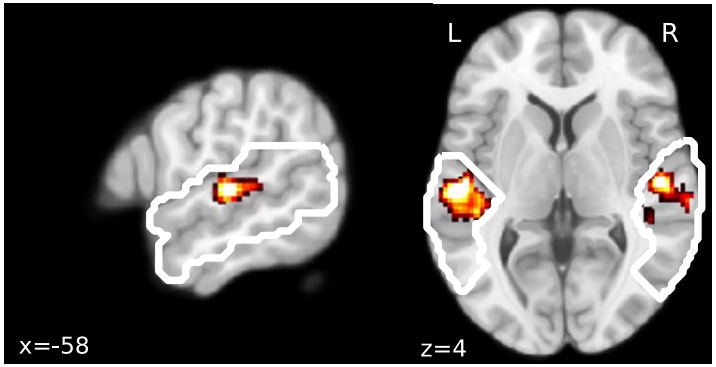
# Figure 2


A
Left hemisphere sensors
RMS
2.8
2.4
2
1.6
1.2
0.8
Strong+M
Weak+M
Strong+MM
Weak+MM
100
0
% participants
Strong-Weak
Interaction
MM-M
0
400
800
Time (ms)
Syl2 onset
B
Left hemisphere sensors
correlation with p(Syl2|Syl1)
Fisher r
0.2
0.1
0
-0.1
-0.2
M
MM
M+MM
MM-M
0
400
800
Time (ms)
C
Peak: 3 Hz, 120 ms
335
Power
325
315
M
MM
Strong
Weak
D
Left hemisphere sensors
Interaction
Weak>Strong (M)
Freq (Hz)
42
32
22
12
2
Change in power
50
0
-50
0
400
800
0
400
800
Time (ms)

# Figure 3

A

Strong>Weak

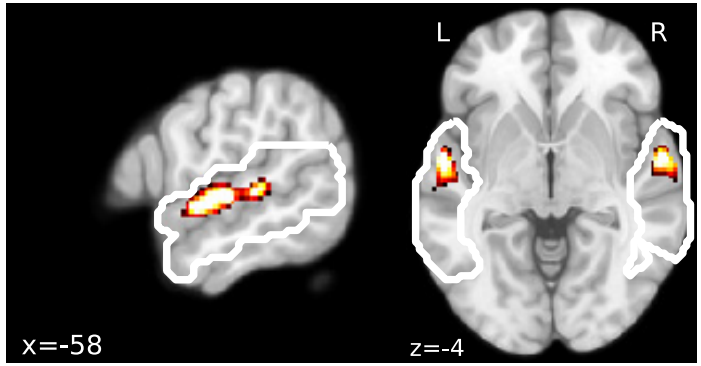



F-stat

30

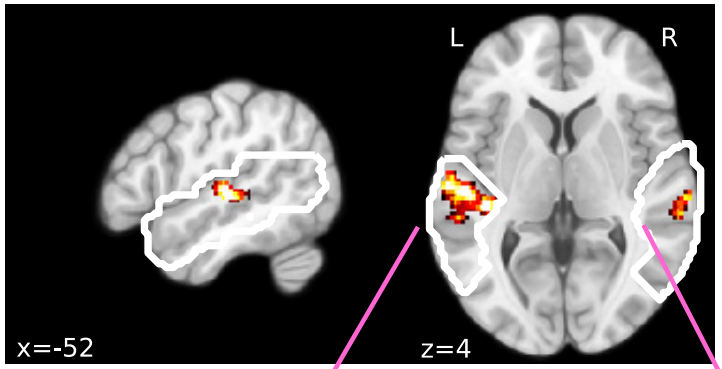
12

MM>M




B

Mismatch + Match > 0



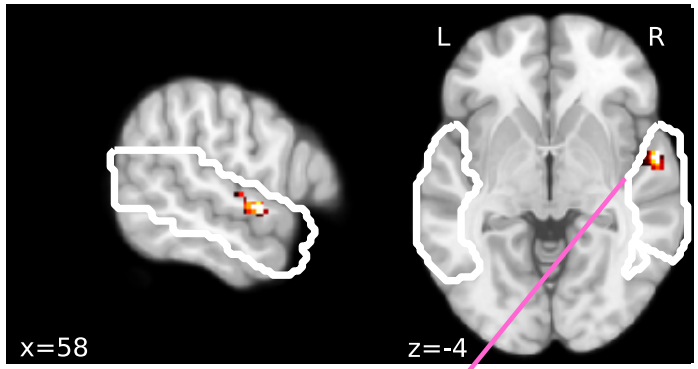

Correlation with p(Syl2|Syl1)

t-stat

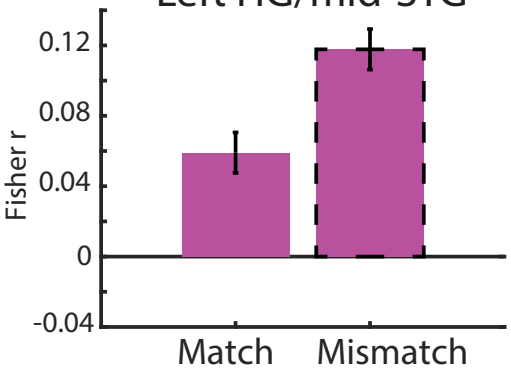
5.5

3.6

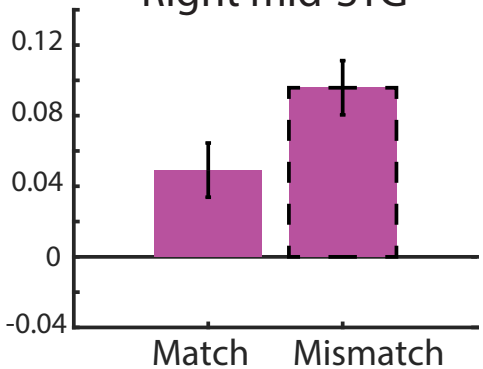
Mismatch > Match



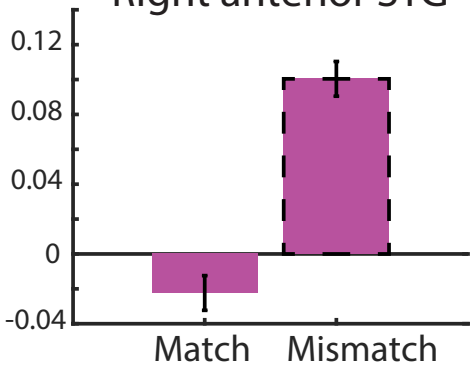

Left HG/mid-STG



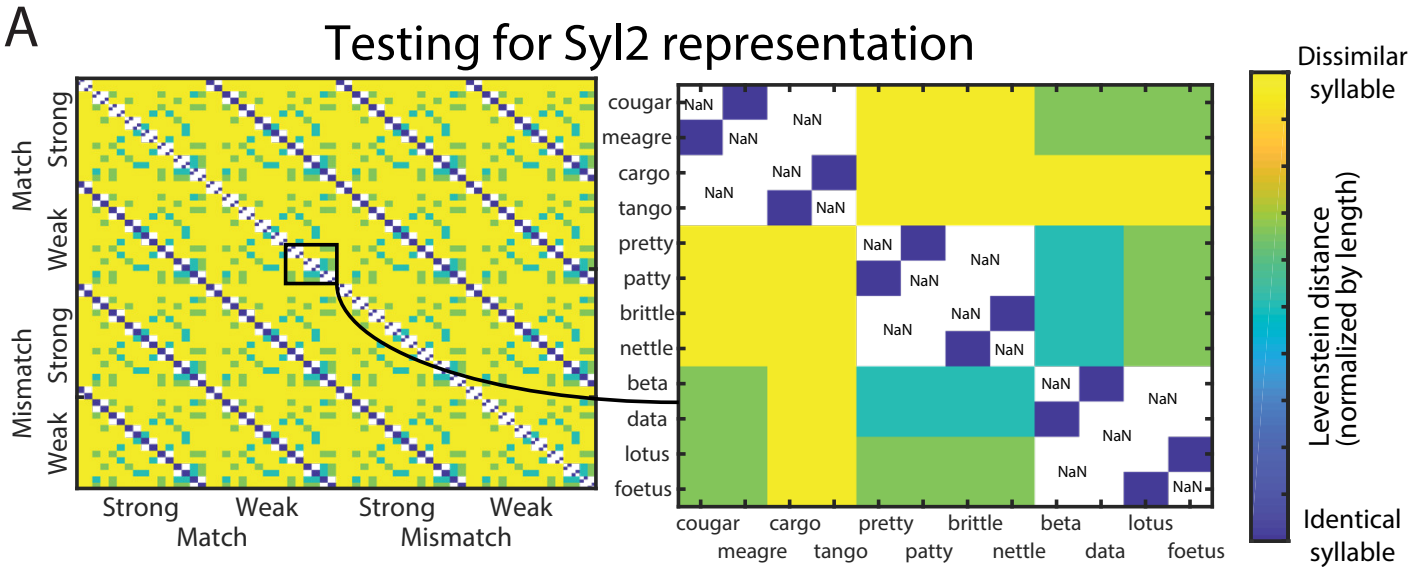

Right mid-STG

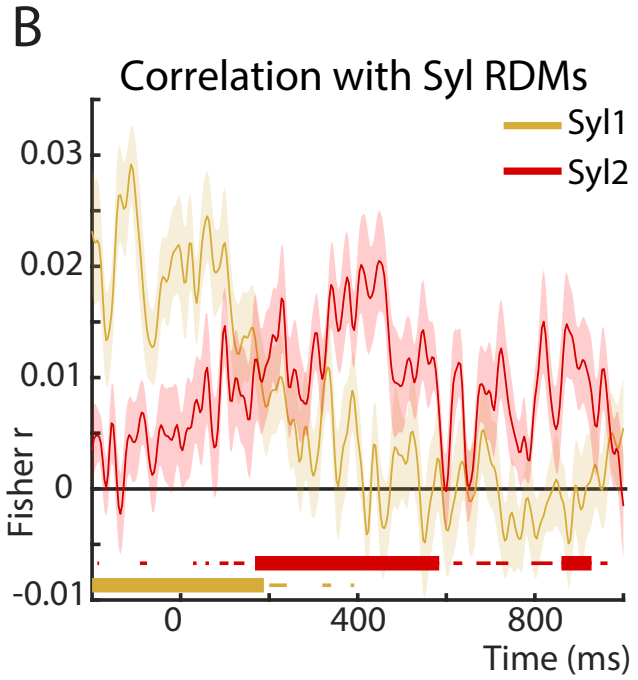


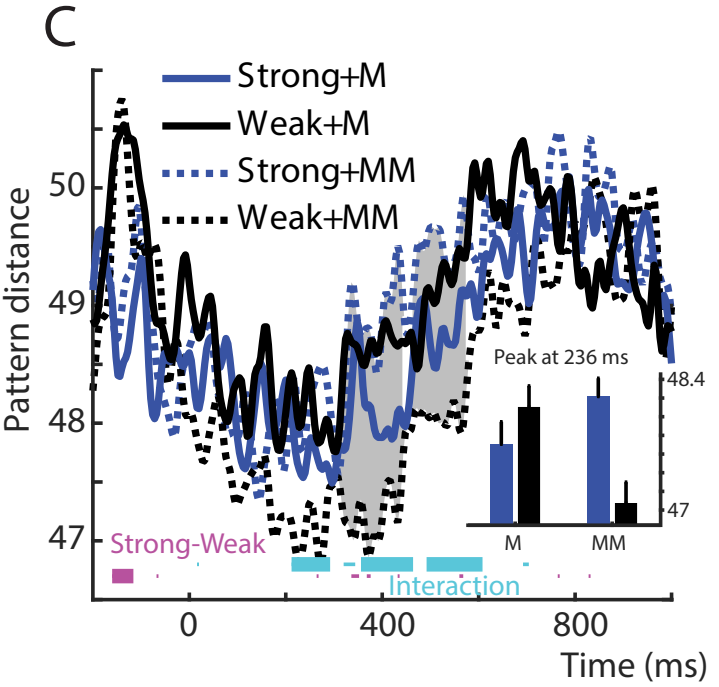

Right anterior STG

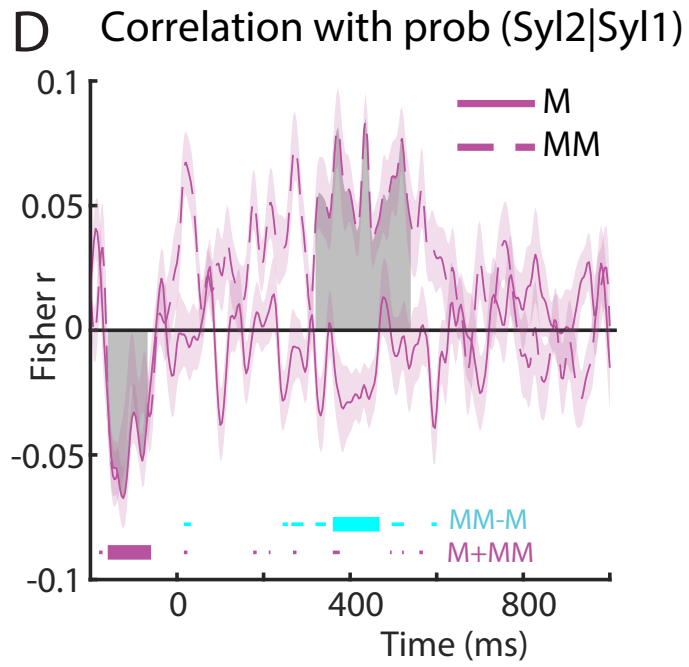


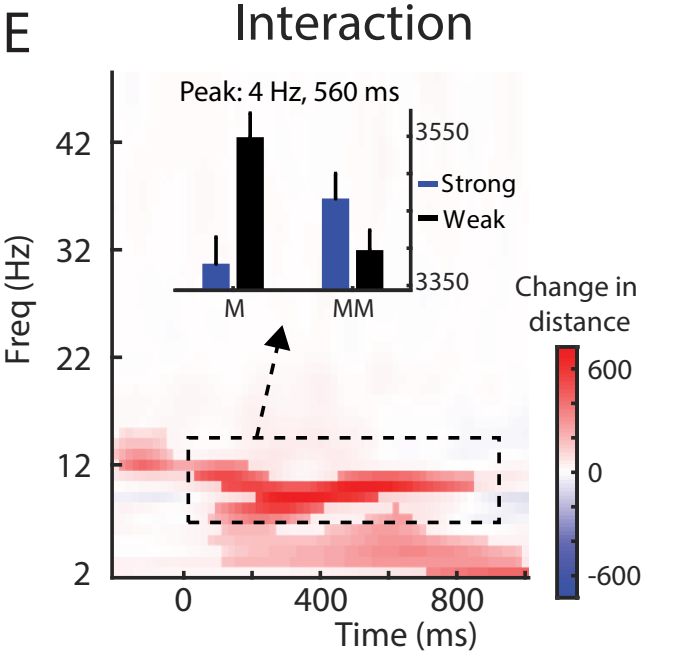
# Figure 4


A
Testing for Syl2 representation
Match
Strong
Weak
Mismatch
Strong
Weak
Strong
Weak
Match
Mismatch
cougar
meagre
cargo
tango
pretty
patty
brittle
nettle
beta
data
lotus
foetus
NaN
Dissimilar syllable
Levenstein distance (normalized by length)
Identical syllable

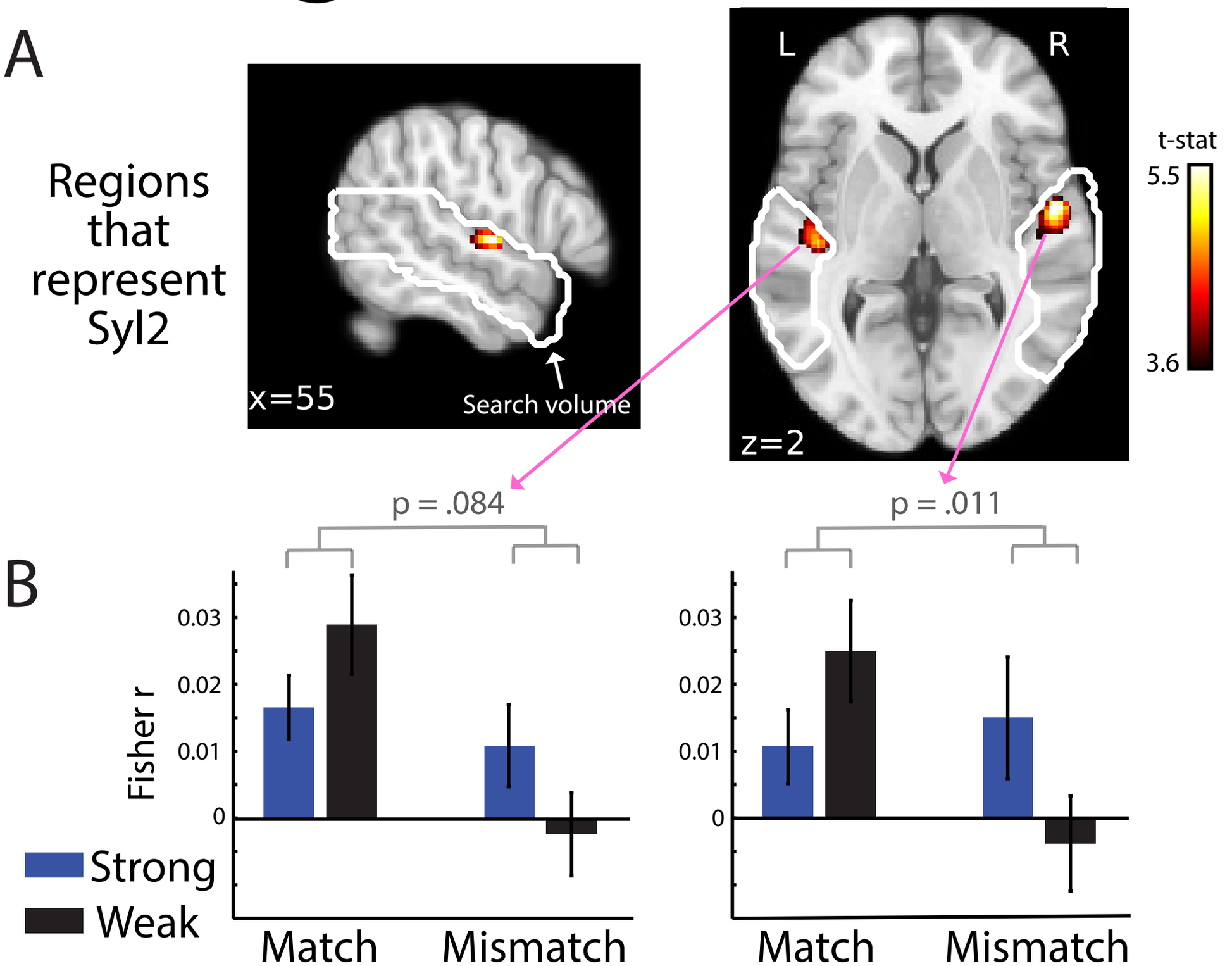

B
Correlation with Syl RDMs
Syl1
Syl2
Fisher r
0.03
0.02
0.01
0
-0.01
0
400
800
Time (ms)


C
Strong+M
Weak+M
Strong+MM
Weak+MM
Pattern distance
50
49
48
47
Peak at 236 ms
48.4
47
M
MM
Strong-Weak
Interaction
0
400
800
Time (ms)

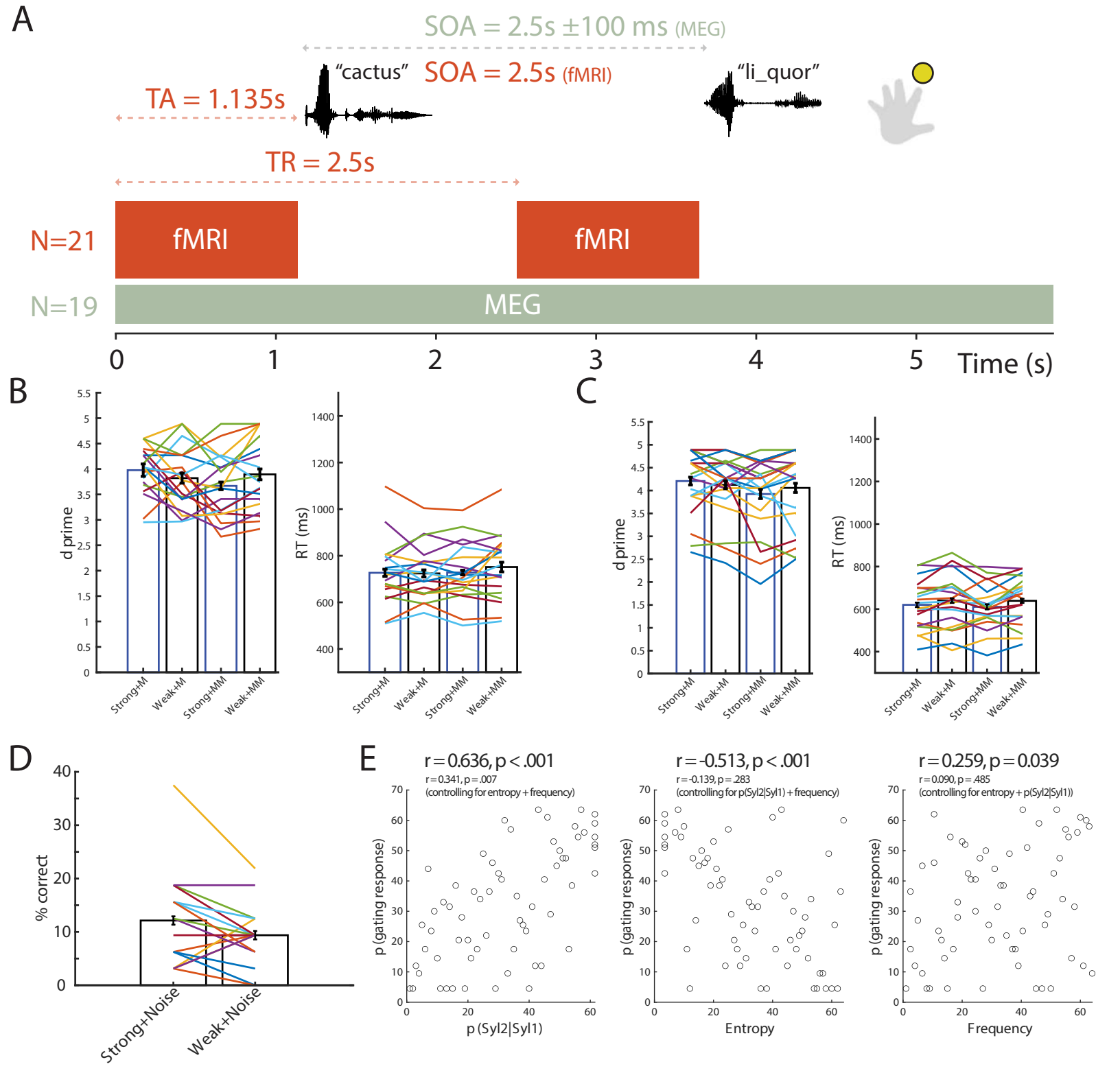

D
Correlation with prob (Syl2|Syl1)
M
MM
Fisher r
0.1
0.05
0
-0.05
-0.1
MM-M
M+MM
0
400
800
Time (ms)


E
Interaction
Peak: 4 Hz, 560 ms
3550
3350
Strong
Weak
M
MM
Freq (Hz)
42
32
22
12
2
Change in distance
600
0
-600
0
400
800
Time (ms)

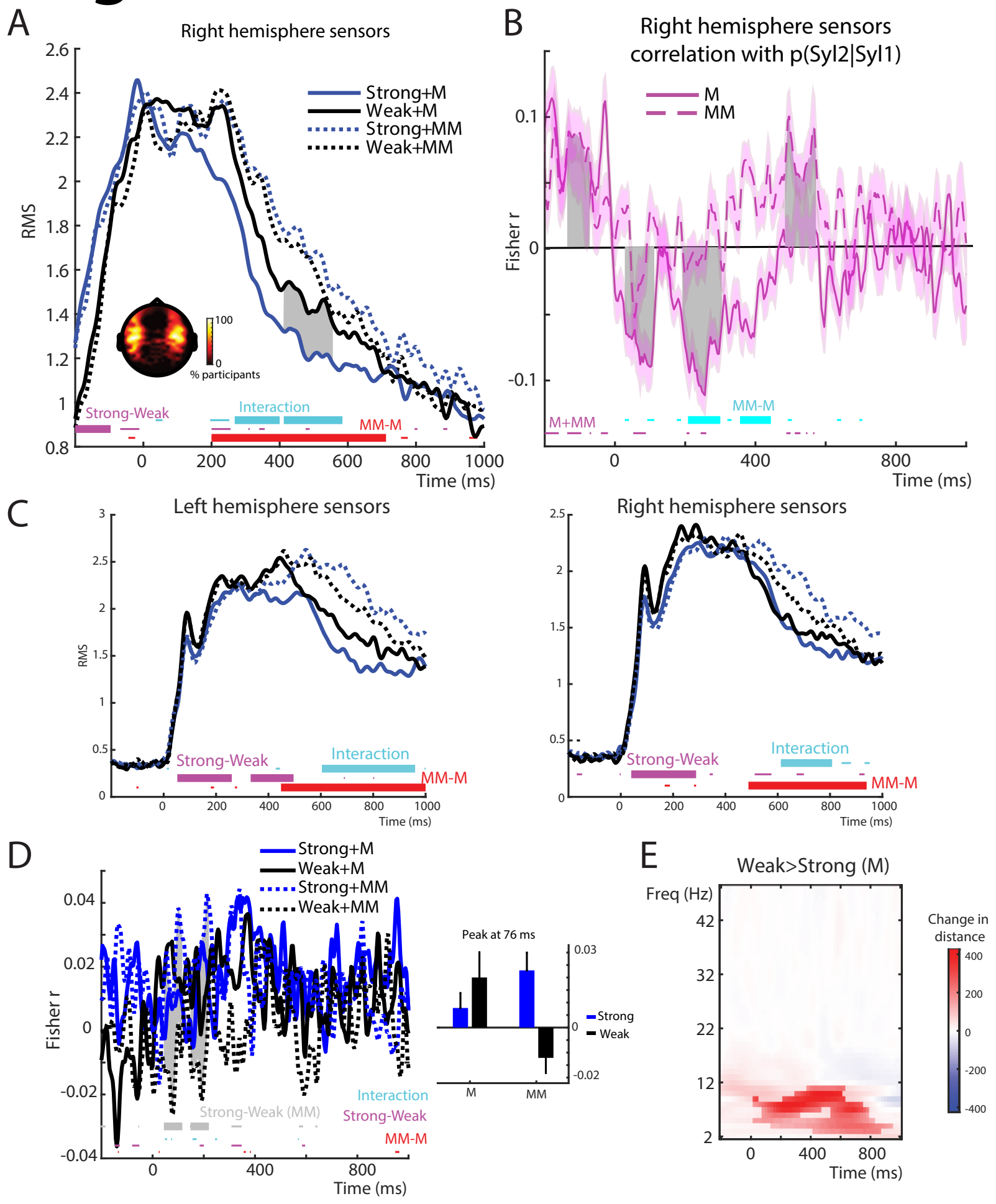
# Figure 5


A
Regions that represent Syl2
x=55
Search volume
L
R
z=2
t-stat
5.5
3.6
B
p = .084
p = .011
Fisher r
0.03
0.02
0.01
0
Strong
Weak
Match
Mismatch
Match
Mismatch
Interaction (both hemispheres): p = .005

# Figure S1


A
SOA = 2.5s ±100 ms (MEG)
"cactus"
SOA = 2.5s (fMRI)
"li_quor"
TA = 1.135s
TR = 2.5s
N=21
fMRI
fMRI
N=19
MEG
0
1
2
3
4
5
Time (s)
B
d prime
RT (ms)
Strong+M
Weak+M
Strong+MM
Weak+MM
C
d prime
RT (ms)
Strong+M
Weak+M
Strong+MM
Weak+MM
D
% correct
Strong+Noise
Weak+Noise
E
r = 0.636, p < .001
r = 0.341, p = .007
(controlling for entropy + frequency)
p (gating response)
p (Syl2|Syl1)
r = -0.513, p < .001
r = -0.139, p = .283
(controlling for p(Syl2|Syl1) + frequency)
p (gating response)
Entropy
r = 0.259, p = 0.039
r = 0.090, p = .485
(controlling for entropy + p(Syl2|Syl1))
p (gating response)
Frequency

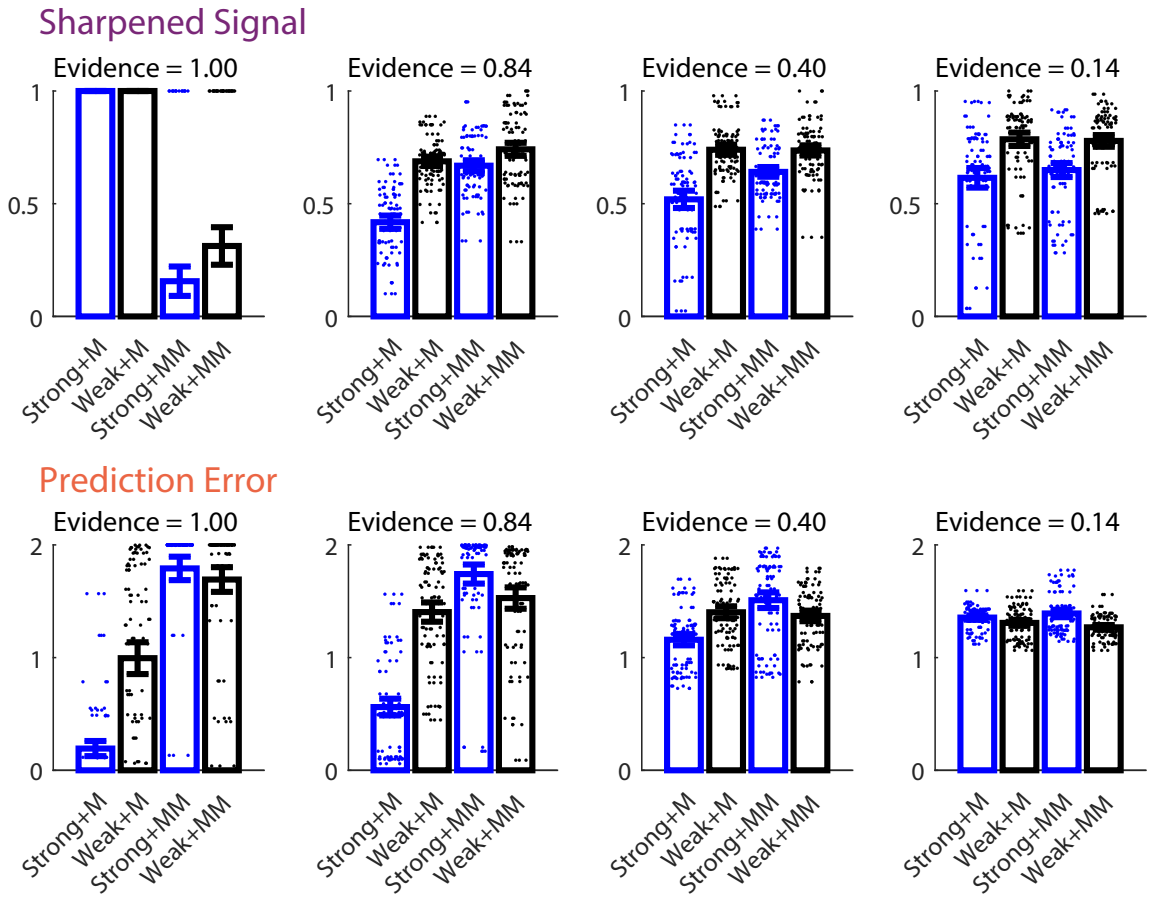
# Figure S2


A
Right hemisphere sensors
RMS
Time (ms)
Strong+M
Weak+M
Strong+MM
Weak+MM
100
0
% participants
Strong-Weak
Interaction
MM-M
B
Right hemisphere sensors
correlation with p(Syl2|Syl1)
Fisher r
M
MM
M+MM
MM-M
Time (ms)
C
Left hemisphere sensors
Right hemisphere sensors
RMS
Time (ms)
Strong-Weak
Interaction
MM-M
D
Strong+M
Weak+M
Strong+MM
Weak+MM
Fisher r
Time (ms)
Strong-Weak (MM)
Interaction
Strong-Weak
MM-M
Peak at 76 ms
Strong
Weak
M
MM
E
Weak>Strong (M)
Freq (Hz)
Time (ms)
Change in distance

# Figure S3

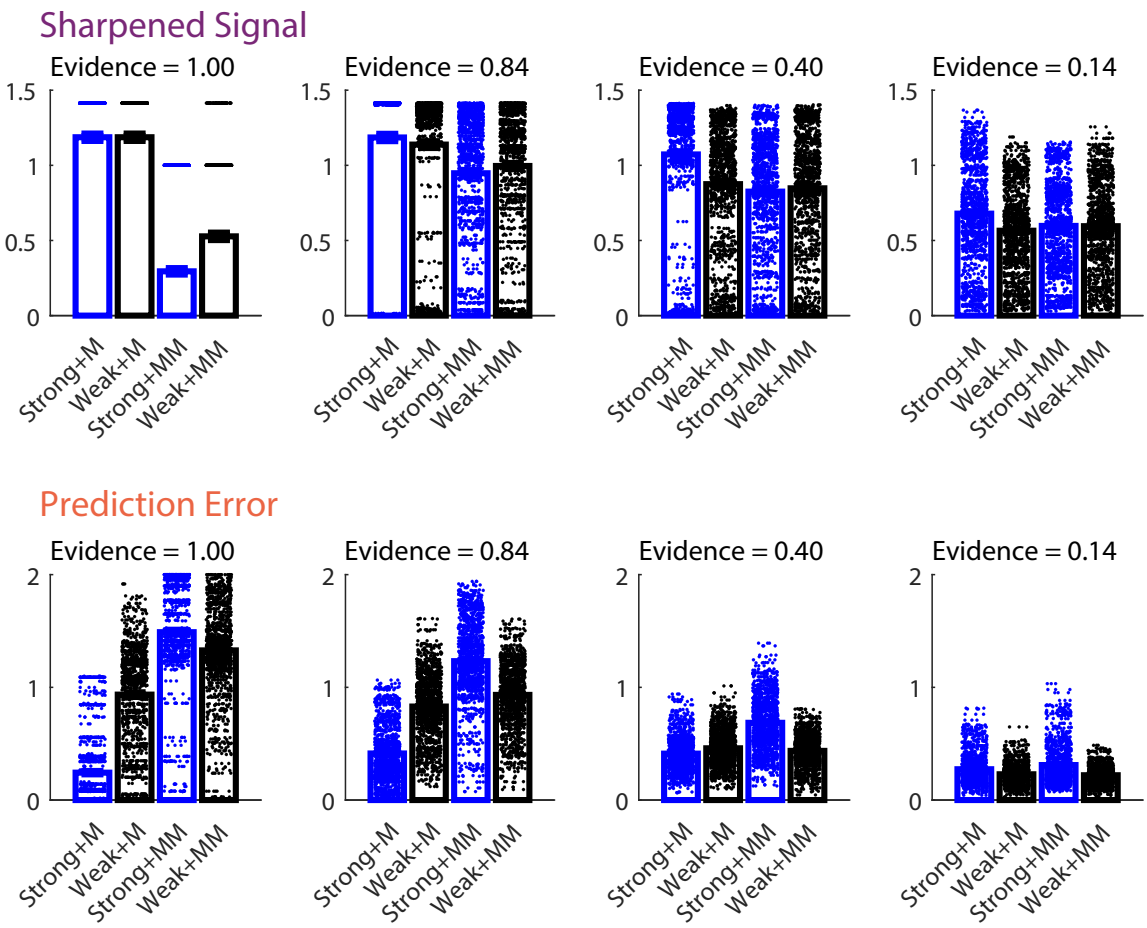
A

## Model signal magnitude at divergence point

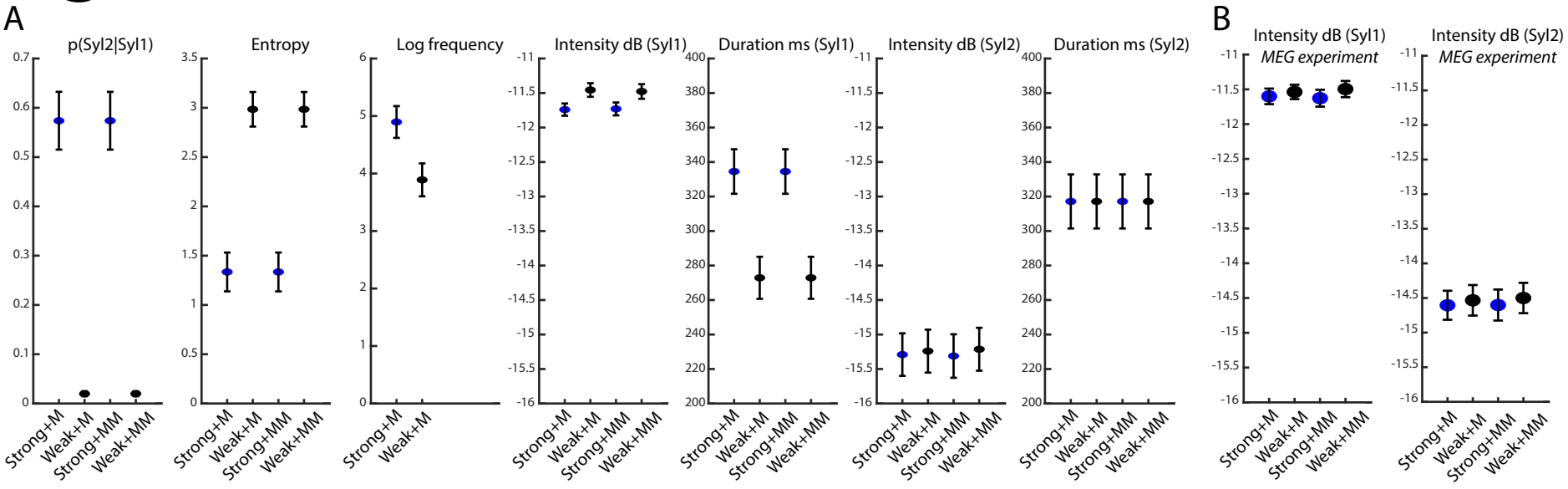


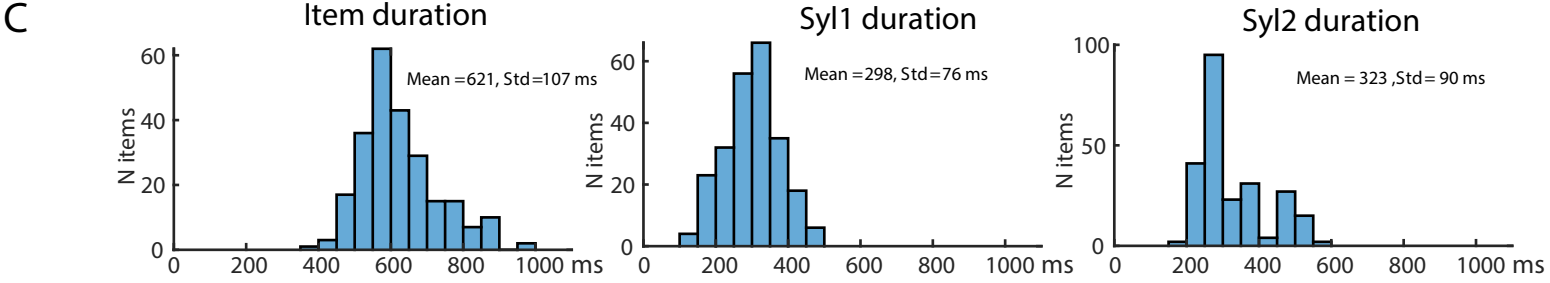

B

## Model pattern distance at divergence point



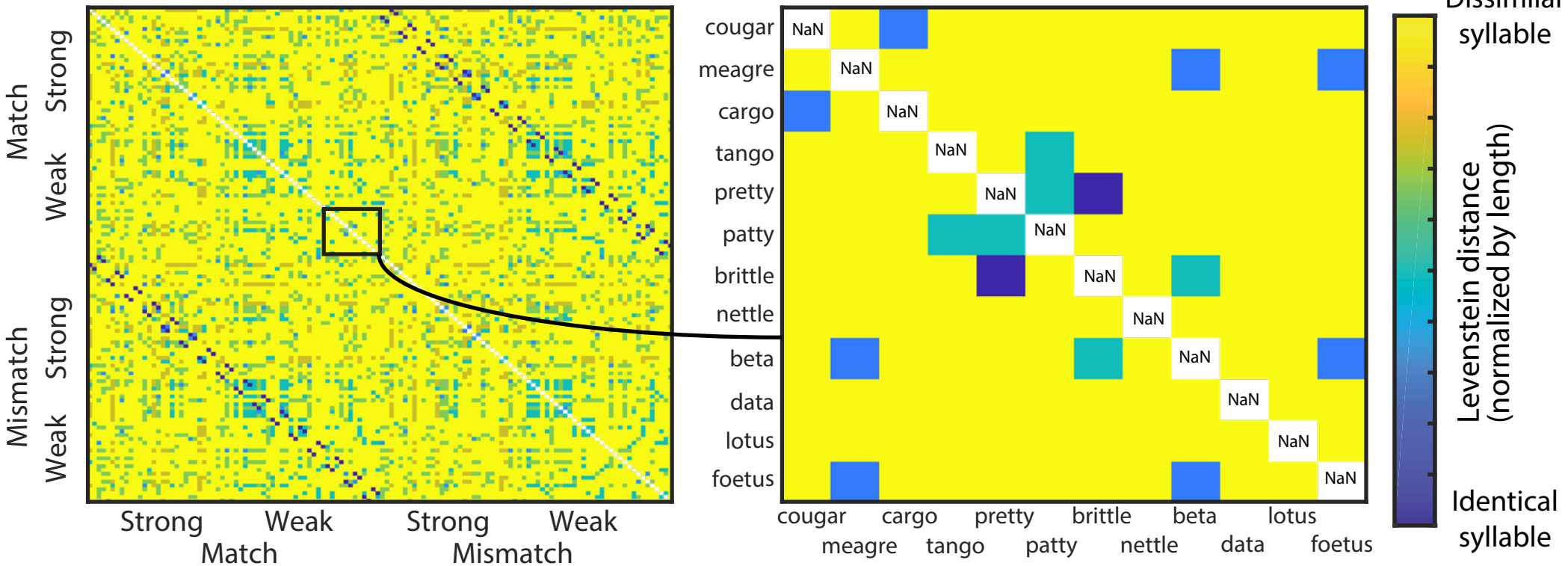
# Figure S4


A
p(Syl2|Syl1)
Entropy
Log frequency
Intensity dB (Syl1)
Duration ms (Syl1)
Intensity dB (Syl2)
Duration ms (Syl2)
Strong+M
Weak+M
Strong+MM
Weak+MM
B
Intensity dB (Syl1)
MEG experiment
Intensity dB (Syl2)
MEG experiment


C
Item duration
Mean =621, Std=107 ms
Syl1 duration
Mean =298, Std=76 ms
Syl2 duration
Mean = 323 ,Std= 90 ms
N items
ms

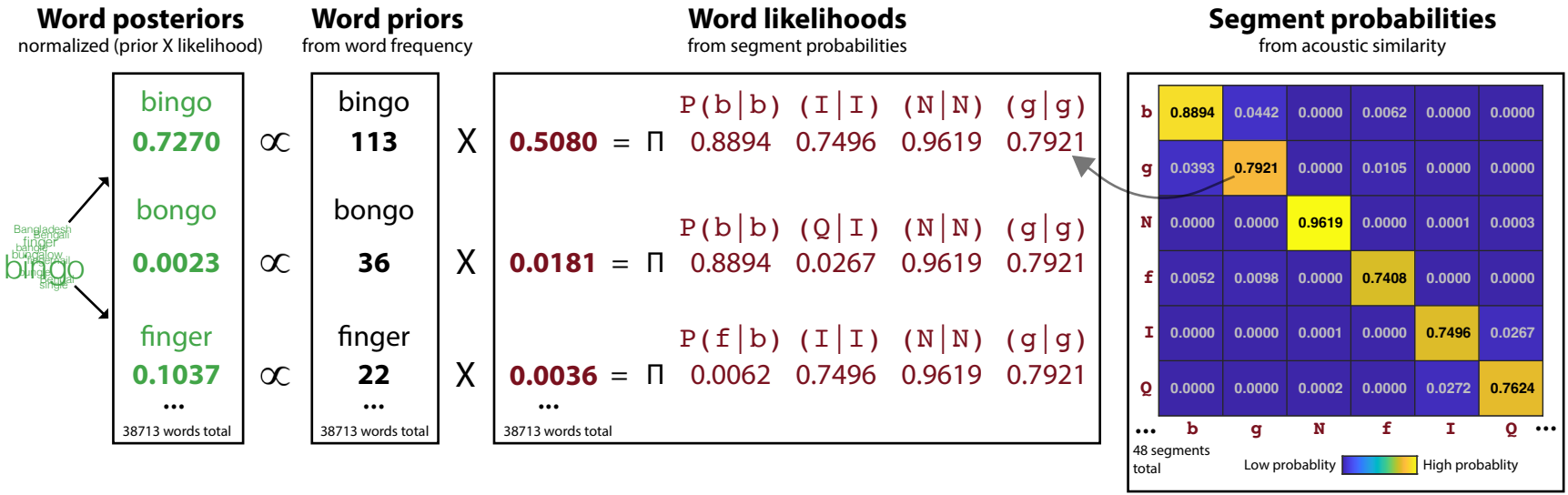
# Figure S5

## Testing for Syl1 representation



# Figure S6

## A Compute word posteriors (input segments: b I N – g)



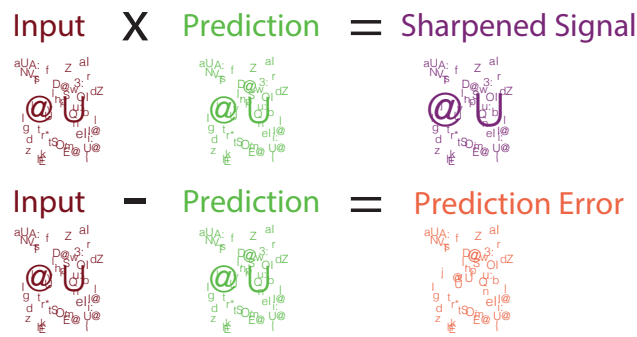

## B Sum predictions for words sharing next segment

| Word posteriors | | Word posteriors | |
|---|---|---|---|
| bIN-g@U | 0.7270 | fIN-g@r* | 0.1037 |
| bQN-g@U | 0.0023 | bVN-g@-l@U | 0.0452 |
| dIN-g@U | 0.0001 | fIN-g@-neIl | 0.0057 |
| kQN-g@U | 0.0001 | fIN-g@-tIp | 0.0043 |
| m{N-g@U | 0.0000 | fIN-g@-prInt | 0.0035 |
| ... | | ... | |

$$\sum_{j=1}^{j=n} = P(@U) = 0.7296$$

$$\sum_{j=1}^{j=n} = P(@) = 0.1746$$

## C Combine segment predictions and input



# Table 1

| Location | Extent | Z-stat | x | y | z |
|---|---|---|---|---|---|
| **Univariate Strong > Weak** | | | | | |
| Right Heschl's Gyrus | 551 | 5.74 | 54 | -16 | 2 |
| Right mid-STG | | 4.84 | 64 | -14 | 0 |
| Right mid-STG | | 4.05 | 66 | -26 | 4 |
| Left mid-STG | 551 | 5.41 | -58 | -22 | 4 |
| Left Heschl's Gyrus | | 4.55 | -44 | -24 | 2 |
| Left mid-STG | | 4.55 | -46 | -32 | 4 |
| **Univariate Mismatch > Match** | | | | | |
| Left mid-STG | 569 | 6.00 | -64 | -24 | 6 |
| Left mid-STG | | 5.81 | -60 | -16 | 2 |
| Left anterior-STG | | 5.62 | -58 | -4 | -4 |
| Right anterior-STG | 205 | 5.25 | 58 | -4 | -4 |
| Right mid-STG | | 4.51 | 64 | -16 | 0 |
| Right mid-STG | | 3.54 | 56 | -14 | -2 |
| **Univariate Strong+Noise > Weak+Noise** | | | | | |
| Left anterior-STG | 245 | 4.42 | -54 | -12 | 2 |
| Left Heschl's Gyrus | | 4.21 | -44 | -22 | -2 |
| Left mid-STG | | 3.31 | -60 | -24 | 2 |
| Right posterior-STG | 56 | 4.25 | 64 | -44 | 10 |
| Right posterior-MTG | | 3.58 | 60 | -54 | 6 |
| Right mid-STG | 115 | 4.00 | 54 | -20 | 0 |
| Right Heschl's Gyrus | | 3.30 | 48 | -14 | 0 |
| Right anterior-STG | | 3.10 | 58 | -12 | -4 |
| **Univariate correlations with p(Syl2\|Syl1) Mismatch+Match** | | | | | |
| Left Heschl's Gyrus | 467 | 5.01 | -52 | -22 | 2 |
| Left Heschl's Gyrus | | 4.85 | -42 | -28 | 4 |
| Left mid-STG | | 4.62 | -62 | -18 | 6 |
| Right mid-STG | 105 | 4.04 | 66 | -24 | 6 |
| Right mid-STG | | 3.96 | 64 | -32 | 6 |
| Right mid-STG | | 3.56 | 54 | -26 | 0 |
| **Univariate correlations with p(Syl2\|Syl1) Mismatch>Match** | | | | | |
| Right anterior-STG | 71 | 4.58 | 58 | -2 | -4 |
| Right anterior-STG | | 3.81 | 50 | -4 | -2 |
| Right anterior-STG | | 3.59 | 58 | -10 | 2 |
| **Multivariate Syl2 phonetic distance** | | | | | |
| Right Heschl's Gyrus | 148 | 4.49 | 56 | -6 | 2 |
| Left Heschl's Gyrus | 146 | 4.01 | -44 | -20 | 0 |