# How are visual words represented? Insights from EEG-based visual word decoding, feature derivation and image reconstruction

Running title: EEG-based visual word representations

Shouyu Ling[1], Andy C.H. Lee[1,2], Blair C. Armstrong[1,3], Adrian Nestor[1*]

1. Department of Psychology at Scarborough, University of Toronto, Toronto, Ontario, Canada, M1C1A4
2. Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada, M6A 2E1
3. BCBL. Basque Center on Cognition, Brain, and Language, Donostia - San Sebastián, Spain, 20009

*To whom correspondence should be addressed:

Email: anestor@utsc.utoronto.ca; Tel: 1 (416) 208-2998

Department of Psychology at Scarborough, University of Toronto, 1265 Military Trail, Scarborough, Ontario, M1C1A4, Canada

**Acknowledgements:** This research was supported by the Natural Sciences and Engineering Research Council of Canada (to A.N., A.C.H.L. and B. C. A.).

**Abstract.** Investigations into the neural basis of reading have shed light on the cortical locus and the functional role of visual-orthographic processing. Yet, the fine-grained structure of neural representations subserving reading remains to be clarified. Here, we capitalize on the spatiotemporal structure of electroencephalography (EEG) data to examine if and how EEG patterns can serve to decode and reconstruct the internal representation of visually presented words in healthy adults. Our results show that word classification and image reconstruction were accurate well above chance, that their temporal profile exhibited an early onset, soon after 100ms, and peaked around 170ms. Further, reconstruction results were well explained by a combination of visual-orthographic word properties. Last, systematic individual differences were detected in orthographic representations across participants. Collectively, our results establish the feasibility of EEG-based word decoding and image reconstruction. More generally, they help to elucidate the specific features, dynamics, and neurocomputational principles underlying word recognition.

**Keywords:** EEG, multivariate analysis, reading, word processing

Extensive work has been dedicated to elucidating the neural basis of reading and its reliance on visual-orthographic representations. For instance, much is known about the role played by the ventral occipital-temporal cortex (vOT) in deriving such representations (Dehaene & Cohen, 2011; Glezer, Jiang, & Riesenhuber, 2009; Price & Devlin, 2011; Rauschecker, Bowen, Parvizi, & Wandell, 2012; Striem-Amit, Cohen, Dehaene, & Amedi, 2012; Taylor, Rastle, & Davis, 2013). Also, the speed and efficiency of processing visual-orthographic representations, as revealed by their time course, has provided important theoretical insights (Araújo, Faísca, Bramão, Reis, & Petersson, 2015; Chen, Davis, Pulvermüller, & Hauk, 2015; Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006). Yet, the nature and the visual structure of such representations remain to be clarified.

One longstanding challenge, with considerable theoretical and practical implications, is whether visual words could be discriminated from one another based on the neural activity that they elicit (Suppes, Lu, & Han, 1997). Recently, this challenge has been addressed with the aid of pattern analyses (e.g., classification) as applied to functional Magnetic Resonance Imaging (fMRI) (Baeck, Kravitz, Baker, & de Beeck, 2015; Nestor, Behrmann, & Plaut, 2013), electrocorticography (ECoG) (Hirshorn et al., 2016) or combinations of magnetoencephalography (MEG) and EEG data (Chan, Halgren, Marinkovic, & Cash, 2011). These attempts have shed light on the visual-orthographic representational space underlying reading, on its cortical locus, and on the extended time course of visual word discrimination. However, the precise nature of the information that facilitates discrimination, as well as its robustness and its variability across individuals remain to be elucidated.

Relevantly here, neural-based image reconstruction (Chang & Tsao, 2017; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009; Nestor, Plaut, & Behrmann, 2016; Nishimoto et al., 2011; Shen, Horikawa, Majima, & Kamitani, 2017) aims to reveal the content of fine-grained visual representations by retrieving the appearance of visual objects from neural activity prompted by their processing. For instance, several fMRI studies have addressed the challenge of reconstructing the appearance of single letters from fMRI patterns associated with their reading (Miyawaki et al., 2008; Schoenmakers, Barth, Heskes, & van Gerven, 2013; Thirion et al., 2006). Broadly, image reconstruction informs the nature of the mapping between the visual world and neural representations: how exactly a visual pattern (e.g., corresponding to a stimulus) is converted into a neural pattern and vice-versa (Naselaris, Kay, Nishimoto, & Gallant, 2011). Critical to our purposes, reconstruction can help to characterize the fidelity and the robustness of visual representations underlying reading. Yet, to date, the application of this methodology to single characters, rather than entire words, has limited its psycholinguistic implications.

Here, we used pattern analysis of electroencephalography (EEG) data and image reconstruction to uncover the structure of visual word representations, their temporal dynamics, as well as individual differences associated with their processing. To be clear, while pattern analysis may be able to shed light on multiple types of psycholinguistic processing (e.g., semantic), the present work focuses mainly on visual and orthographic processing. To this aim, here we collected EEG recordings associated with reading 80 high-frequency nouns in healthy adults and, then, we exploited spatiotemporal patterns associated with these words to decode and to reconstruct their visual appearance from neural data. A key aspect of the method concerns the use of representational similarity (Kriegeskorte, Mur, & Bandettini, 2008), applied here to EEG patterns, as a way to probe the structure of a visual word representational space and, also, as a step in our reconstruction procedure. Of note, both neural-based similarity and objective image similarity are considered in the process of deriving human and theoretical observer (TO) reconstructions. This approach facilitates an evaluation of the veracity of visual representations and/or their divergence from an image-based groundtruth.

Several hypotheses motivate the current work. First and foremost, our study tested the hypothesis that EEG-based decoding and reconstruction of visual words are feasible by virtue of their ability to capture both visual and orthographic aspects of neural word representations. Second, we hypothesized that word decoding and reconstruction exploit an extensive temporal window, though dominated by specific temporal intervals (e.g., around the N170 component) in agreement with previous ERP research. Third, we surmised that reconstruction may be able to identify individual differences in visual-orthographic representations (e.g., with regard to the shape of specific letters).

Overall, our results show that: (i) pairwise word classification is well above chance across participants (61-80% accuracy against 50% chance level) and that image reconstruction can be achieved with a level of accuracy closely matching that of word classification; (ii) the time course of classification/reconstruction peaks in the proximity of the N170 component, though complementary information can be found across an extensive temporal interval, and (iii) the structure of visual representations varies systematically across participants. More generally, these results speak to the underexploited wealth of information available in the EEG signal, accessible through pattern analyses, and to its ability to shed light on the fine-grained structure of visual-orthographic representations.

## Materials and Methods

### Participants

Eighteen healthy Caucasian adults were recruited from the University of Toronto community in exchange for monetary compensation. One participant was excluded due to technical difficulties with the EEG recordings while three other participants were excluded due to left-handedness. The remaining fourteen right-handed participants (nine females; age range: 20–26 years) were included in the analyses. Participants listed English as their first language and the only language in which they were fluent in speaking and writing. All participants had normal or corrected-to-normal vision and reported no history of cognitive or neurological impairment. All participants provided informed consent and all experimental procedures were approved by the Research Ethics Board at University of Toronto.

### Stimuli

Eighty word images of concrete nouns with consonant-vowel-consonant (CVC) structure were used as experimental stimuli. The words were selected from the UNION database (www.blairarmstrong.net/tools/index.html) which includes words with frequencies higher than or equal to 1 in the SUBTL word frequency norms (Brysbaert & New, 2009) and words with syllabified pronunciations from the CMU pronunciation dictionary (Bartlett, Kondrak, & Cherry, 2009). Stimuli were selected so as to balance the number of occurrences of each letter at each position as much as possible in the context of the experimental dataset (e.g., each letter appeared at least twice in each position). Psycholinguistic covariates explored including positional letter frequency (M = 194.63, SD = 33.32, range: 133 – 256), positional letter-bigram frequency (M = 16.64, SD = 4.38, range: 7 - 26), SUBTL word frequency (M = 70.06, SD = 130.22, range: 1.22 – 569.92), orthographic Levenshtein distance (M = 1.16, SD = 0.19, range: 1.00 – 1.75), number of orthographic neighbors(M = 16.61, SD = 4.51, range: 9 – 26), phonological Levenshtein distance (M = 1.08, SD = 0.17, range: 1.00 – 1.70) and number of phonological neighbors (M = 22.55, SD = 6.81, range: 8 - 38).

Word stimuli were presented on a black background using monospaced font Consolas lower-case font with white strokes. Word images were created with a font size of 150, resulting in 247x151 pixel images. Stimuli were presented at the center of the screen against a black background and subtended a visual angle of 4.87º x 2.86º from a distance of 80 cm.

## Data Collection

During the experiment participants were seated in a dimly lit room in front of an LCD monitor (resolution: 1920 X 1080, refresh rate: 60Hz). Participants were presented with sequences of experimental stimuli and were asked to complete a go/no-go one-back image task by pressing a designated key every time they noticed that a stimulus was presented twice in a row. The experiment consisted of two sessions conducted on two separate days. Each session contained 16 experimental blocks preceded by one training block that aimed to familiarize participants with the task and the stimuli as well as to direct their focus to the perceptual properties as opposed to the semantic properties of the stimuli. Due to fatigue, one participant completed only 14 blocks on the second session, resulting in a total of 30 completed blocks.

Specifically, each experimental block consisted of a sequence of 270 trials: 30 go trials and 240 no-go trials consisting of three repetitions of each stimulus. Trial order was pseudorandomized so that repetitions of one word, other than those on go trials, were separated by at least 40 intervening trials. On each trial a stimulus was displayed for 300ms, then it was replaced by a white noise mask for 100ms and it was followed by a fixation cross for a duration ranging randomly between 500-600ms. The blocks were separated by self-paced breaks. Each experimental session, including participant and equipment setup, lasted around 2.5 hours. Stimulus presentation and response recording relied on Matlab (Mathworks, Natick, MA) and Psychtoolbox 3.0.8 (Brainard, 1997; Pelli, 1997).

## EEG acquisition and pre-processing

High-density EEG was recorded using a Biosemi ActiveTwo system with 64 gelled electrodes mounted on an elastic cap using the 10/20 System. This system replaces conventional ground electrodes with the Common Mode Sense (CMS) active electrode and the Driven Right Leg (DRL) passive electrode. These two electrodes form a feedback loop which drives the average potential of the subject to be roughly equivalent to the Analogue Digital Converter (ADC) reference voltage, which serves as the amplifier's “zero”. Electrodes CMS and DRL served as the online reference while AFz served as the ground. The reference was computed offline based on the average of all electrodes. The EEG signal was amplified at a sampling rate of 512 Hz. The electrode offset was kept below 40 mV. The EEG were low-pass filtered using a fifth order sinc filter with a half-power cutoff at 204.8 Hz and then digitized at 512 Hz with 24 bits of resolution. All data were digitally filtered offline (zero-phase 24 dB/octave Butterworth filter) with a bandpass of 0.1–40 Hz. Then, data were separated into epochs, from 100ms prior to stimulus presentation until 900ms later, and baseline-corrected. Specifically, the pre-stimulus period (-100 to 0ms) signal served as baseline and was subtracted from each trial.

‘Go’ trials as well as false alarm trials were excluded from analyses. Further, epochs with voltage exceeding +/- 150 μV at any electrode were excluded. After removing trials containing artifacts and/or false alarms, an average of 99.4% of trials (range: 97.8% - 99.9% across participants) were selected for further analysis. In particular, we note that relatively few trials contained false alarms as participants performed the go/no-go recognition task at ceiling (accuracy range: 95.8% - 99.7%; reaction time: 593ms – 774ms across participants). Of note, neither accuracy, nor reaction time correlated significantly with decoding or reconstruction accuracy across participants (p’s > 0.32).

Further, noisy electrodes were interpolated if necessary (no more than 2 electrodes per subject) and ocular artifacts (i.e., blinks) were removed using independent component analysis (exactly one component was removed from each participant).

All EEG analyses were carried out using Letswave 6 (Mouraux & Iannetti, 2008, RRID:SCR_016414), and MATLAB 9.0.

## Stimulus classification

Decoding relied on spatiotemporal patterns across 12 bilateral OT electrodes (left: P5, P7, P9, PO3, PO7, O1 and right: P6, P8, P10, PO4, PO8, O2). Their selection was motivated by their relevance for word processing (e.g., robust N170 amplitudes) (Bentin, Mouchetant-Rostaing, Giard, Echallier, & Pernier, 1999; Maurer, Zevin, & McCandliss, 2008).

To derive spatiotemporal patterns for classification purposes, EEG signals were first normalized across all trials by z-scoring data separately for each electrode and each time bin. To be clear, normalization, along with subsequent pattern classification steps, were conducted separately for each participant allowing the evaluation of decoding performance separately for each participant. Then, the data were averaged for each stimulus across all epochs from two consecutive blocks (i.e., for a maximum of 6 trials) in order to boost the signal-to-noise ratio (SNR) of spatiotemporal patterns for classification purposes (Grootswagers, Wardle, & Carlson, 2017; Nemrodov, Niemeier, Patel, & Nestor, 2018) and to speed up processing times. This procedure aimed to find the right balance between the number of observations per class, on the one hand, versus the number of trials that are averaged into a single observation, on the other. The averaging parameters (i.e., yielding 16 observations per class and 6 trials averaged per observations) were guided by previous explorations of experimental data not included in the current study.

Next, data were concatenated across 12 electrodes and multiple time points to capture spatiotemporal information present in the EEG signal. Specifically, data were concatenated across a large 50-650ms window, for temporally cumulative analyses aimed at boosting classification accuracy. In addition, for the purpose of complementary analyses aimed at elucidating the temporal profile of word decoding rather than boosting overall accuracy, data were concatenated across consecutive 10ms windows (5 bins*1.95ms≈10ms) between -100 and 900ms. These procedures both delivered 16 observations per word for each participant either across the overall time course, in the former case, or for each position of the sliding window, in the latter.

To assess word discrimination thoroughly we considered the ability to classify each word (out of 80) from every other word (yielding a total number of 3160 word pairs). Pattern classification was conducted for each pair of words for each participant, with the aid of linear SVM (c = 1) and leave-one-out cross-validation (i.e., one out of 16 pairs of observations was systematically left out for testing while the remaining 15 were used for training). Classification accuracy was then assessed both parametrically at the group level (one-sample two-tailed t-tests against 50% chance level) and non-parametrically via permutation tests separately for each participant (i.e., based on 1000 random shuffles of classification labels). Multiple comparison correction was carried out via FDR in the case of 10ms-based estimates across the entire time course.

Cross-time classification followed a similar approach except that the classifier was trained on any given 10ms window and then tested on every 10ms window. Significance testing was carried out in this case via two-tailed t-tests against chance followed by FDR correction.

## Image reconstruction

The current procedure builds upon a recent approach to facial image reconstruction designed to exploit spatiotemporal information in neuroimaging patterns (Nemrodov et al., 2018; Nestor et al., 2016). Here, we deployed this procedure to capture the structure of an EEG-derived word space and its ability to support word image reconstruction. This procedure consisted of a sequence of steps as follows - see **Figure 1**. First, a word similarity space was derived from the pairwise classification of 79 words, after leaving out the reconstruction target. Specifically, a 20-dimensional similarity space was estimated through metric MDS, given that this number of dimensions accounted for a significant proportion of the data variance for any participant (e.g., over 70% for temporally cumulative analyses).

Second, a corresponding number of visual features (i.e., one for each dimension of MDS-derived space) were computed for each dimension through an approach akin to reverse correlation/image classification (see (Murray, 2011) for a review). Notably, this approach aims to synthesize stimulus features responsible for stimulus space topography through a linear combination of stimulus images. Specifically, images were processed with a Gaussian filter with a 5-pixel kernel size (previously optimized to boost reconstruction accuracy for the theoretical observer). Then, a weighted sum of these images was computed proportionally to the coordinates of the corresponding words on any given dimension. Thus, the outcome of these computations delivers, for each dimension, a single feature, or 'classification image' (CIM).

Third, we considered the possibility that not all stimulus space dimensions encode visual information (e.g., as opposed to higher-level semantic information or just noise). Hence, to identify relevant features, a permutation test was conducted to assess the presence of significant information. Specifically, word identities were randomly shuffled with respect to their coordinates on each dimension and a corresponding feature was recomputed for a total of 1000 permutations. Then, each true feature was compared to all permutation-based features, pixel by pixel (two-tailed permutation test; FDR correction across pixels; $q < 0.05$). Following this procedure, only features that contained significant pixels were selected for reconstruction purposes.

Fourth, the target word was projected into the existing similarity space. To this end, a new MDS solution was constructed for all 80 identities and aligned with the original one via Procrustes analysis using the 79 common words between the two spaces. The resulting alignment provides us with a mapping between the two spaces that allows us to project the target word and to retrieve its coordinates in the original space, for which visual features were derived. Of note, this procedure enforces non-circularity by excluding the reconstruction target from the estimation visual word features.

Last, informative features were linearly combined proportionally to the coordinates of the target word on each corresponding dimension. Then, their sum was added to the average of the 79 stimuli used for feature derivation into an image reconstruction of the target.

The reconstruction procedure above was carried out in two complementary manners: by considering word classification estimates separately for consecutive 10ms windows between -100 – 800ms or by considering a single larger window between 50 – 650ms. Further, the results of each participant were either considered separately or averaged across similarity matrices and then treated in the same manner as the data of any single participant.

**Evaluation of reconstruction results**

Reconstruction accuracy was assessed by comparing each reconstructed stimulus with every filtered stimulus, with the aid of an L2 pixelwise metric, and determining in each case whether the reconstruction is closer to its intended target than to any other stimulus. This procedure was carried out for entire words or separately for each letter position (i.e., the first consonant, the middle vowel and the third consonant) – in the latter case each reconstructed letter was compared against the corresponding image fragment.

Further, a single set of reconstructions, based on temporally cumulative group-based data, was subjected to experimental evaluation in a separate behavioral test. To this end, 20 new participants (6 males and 14 females, age range: 16 – 27 years), who were all proficient English speakers and whose first language relied on the Roman alphabet, were requested to match image reconstructions to their targets in a two-alternative forced choice (2AFC) task. Specifically, each of 80 word reconstructions was presented in the company of two stimuli, one of which was the actual target and the other another word stimulus. Thus, on each trial, a display was shown containing a reconstructed image, at the top, and two stimuli side by side, at the bottom. Each display was presented until participants made a response to decide which stimulus was more similar to the top image by pressing a designated left/right key. For each participant, any reconstructed image was presented twice in the company of different foils; thus, across participants, all 79 possible foils for a given reconstruction were exhausted. Stimulus order was pseudorandomized so that different reconstructed images appeared on consecutive trials while target stimuli appeared equally often on the left/right side. Each experimental session was completed over the course of 30 minutes.

Experimental-based estimates of reconstruction accuracy results were measured as the proportion of correct matches across participants and tested for significance tested against chance (50%) using a one-sample two-tailed t-test.

## Word similarity and visual theoretical observer

Multiple sources of pairwise word similarity were considered as follows: (i) visual similarity based on L2 image distances across pairs of stimuli; (ii) orthographic similarity measured as the number of shared letters at each letter position; (iii) phonological similarity based on estimates of pairwise phoneme confusability (Cutler, Weber, Smits, & Cooper, 2004) averaged across letter positions, and (iv) semantic similarity computed as the Euclidean distance between pairs of words based on GloVe vectors (Pennington, Socher, & Manning, 2014).

The pairwise discriminability for every 10ms interval was correlated with the corresponding estimates of pairwise word similarity above. Temporally cumulative word discriminability was also examined with the aid of multiple linear regression using the similarity estimates above.

In addition, a visual theoretical observer was constructed by using the objective measures of visual similarity above as inputs for the reconstruction procedure. Its accuracy was then computed for entire words and, also, separately for each letter position.

# Results

## Visual word classification

Participants viewed 80 word stimuli, consisting of high-frequency nouns with a three-letter consonant-vowel-consonant (CVC) structure - see prior work (Laszlo & Federmeier, 2011) for a characterization of the EEG signal elicited by such stimuli. Pattern classification was conducted across ERP traces corresponding to these stimuli across multiple electrodes – we detail here results obtained for 12 bilateral occipitotemporal (OT) electrodes since they yielded equivalent or better results to those obtained for all electrodes, as described below. Specifically, we aimed to estimate the discriminability of each pair of word images for each participant from spatiotemporal (i.e., channels x temporal points) patterns – see **Figure 1** for a flowchart of the decoding and reconstruction procedure.

First, classification was conducted on temporally cumulative data from a large interval ranging between 50 – 650ms post-stimulus onset. The average classification accuracy across participants (M = 71.5%, SD = 5.9%) was higher than chance (two-tailed one-sample t-test against 50% accuracy: $t(13) = 13.59$, $p < 0.001$) – see **Figure 2**. Additional permutation tests confirmed that decoding accuracy was above chance for every single participant ($p = 0.001$).

To examine the temporal profile of word discrimination, pattern classification was conducted next separately for ~10ms windows (i.e., 5 time bins x 1.95 ms) between -100ms and 800ms relative to stimulus onset. The resulting classification time course evinced a long interval of above-chance classification (two-tailed t-tests against chance; FDR-corrected; $q < 0.05$) – see **Figure 3a**. Classification reached significance around 100ms and it peaked at

200ms (M = 61.4%, SD = 4.4%), in the proximity of the N170 ERP component - see **Figure S1** for ERP traces.

Given that the temporally cumulative analysis above, which considered a single large temporal interval, resulted in higher classification accuracy than the peak performance across multiple smaller temporal windows, it is likely that complementary information about word decoding exists at different points in time. To assess this hypothesis, we evaluated cross-time generalization by training a classifier on data from any given 10ms window and, then, testing it on every 10ms window. This analysis revealed above-chance classification across time, especially between 100-600ms (two-tailed one-sample t-tests against chance; FDR correction, q < 0.01), indicating that relevant information is maintained over time and – see **Figure S2**, thus, some degree of redundancy. However, off-diagonal cells, corresponding to different temporal windows for training and testing, yielded relatively low levels of accuracy, consistent with poor generalization across time and, thus, with the presence of complementary information over time.

To assess our choice of electrodes, we conducted the temporally cumulative analysis on all 64 electrodes and compared the results with those obtained from 12 OT electrodes described above. On average, decoding accuracies based on all electrodes were slightly lower (M = 70.1%, SD = 5.1%) and the difference was marginally significant (*t*(13) = 1.84, *p* = 0.09). In light of these findings, all subsequent results are based on data recorded from OT electrodes.

**Representational similarity analyses and visual similarity space**

To evaluate the similarity structure of word decoding results, pairwise word classification estimates were averaged across participants and compared against other measures of word similarity. Specifically, EEG-based estimates were compared against visual, orthographic, phonological and semantic measures of word similarity (see Methods).

First, we conducted a multiple linear regression with pairwise EEG-based word discriminability obtained from the temporally cumulative analysis as outcome, and visual, orthographic, phonological, and semantic similarities as predictors. Visual similarity (b = 0.003, t(3155) = 29.68, $p < 0.001$) and orthographic similarity (b = 0.06, t(3155) = 9.27, $p < 0.001$), but not phonological or semantic similarity, made significant independent contributions to predicting EEG-based word discriminability.

Next, in order to examine the temporal profile of word recognition, we correlated each psycholinguistic similarity measures with EEG-based word discriminability for every 10ms windows between -100 – 800ms relative to stimulus onset. An evaluation of these estimates across time showed significant correlations between the EEG data on one hand, and visual similarity, orthographic similarity, and phonological similarity on the other (see **Figure 4**; Pearson correlation; FDR-corrected; $q < 0.01$). These correlations appear to peak for visual, orthographic, and phonological similarity, in the proximity of the N170 component – see Discussion for the relationship between orthographic and phonological similarity. For semantic similarity, the correlation reached significance only for two brief intervals (372 – 392ms and 748 – 758ms).

As expected, given the nature of the experimental task and the location of the signals considered, the largest correlations were found between EEG-based estimates and measures of visual similarity. To clarify and to visualize the nature of the specific information structuring the EEG-based similarity space we proceeded in two steps. First, we constructed a visual word space by applying metric multidimensional scaling (MDS) to pairwise word classification – see **Figure 5a** for an example based on the data of a single representative participant. Then, we synthesized classification images (CIMs), through a linear combination of stimulus images, separately for each of 20 dimensions of this space, with the aim of capturing the visual information underlying the topography of the space.

An examination of the corresponding CIMs showed their potential value in encoding orthographic information - for instance, the first dimension in **Figure 5b** appears to encode the difference between the vowel “i” on the one hand, and the vowels “o” and “u” on the other. Overall though, CIMs appear to summarize visual features that go beyond the shapes of letters present at a single position.

## Visual word image reconstruction

Word image reconstruction was carried out next by linear combinations of CIMs in an effort to approximate the visual appearance of novel stimuli (i.e., CIMs were systematically derived from 79 stimuli and then used to reconstruct one left-out stimulus). Then, reconstruction accuracy was assessed objectively based on pixelwise image similarity between reconstructions and stimuli (see **Figure 6** for examples of reconstructions).

This analysis was carried out, first, for temporally cumulative data between 50 and 650ms separately for each participant. Mean reconstruction accuracy across participants was 71.2% (SD = 6.3%; *t*(13) = 12.55, *p* < 0.001). In addition, permutation tests confirmed that each of the 14 participants yielded above-chance reconstruction accuracies (*p's* < 0.01) An examination of classification accuracy and reconstruction accuracy also revealed that the two estimates were highly correlated across participants (*r* = .86, *p* = 0.0001) – see **Figure 2**. (For an additional evaluation of reconstruction accuracy, its robustness and its relationship with pairwise visual word similarity see Supporting information, Visual similarity and image reconstruction.)

Further, the time course of reconstruction accuracy was examined for consecutive 10ms windows between -100 – 800ms. In agreement with the time course of word classification, reconstruction performance reached significance shortly after 100ms and peaked at 190ms in the proximity of the N170 ERP component, M = 63.5%, SD = 5.7% - see **Figure 3b**. For an illustration of word reconstruction across time see also **Movie 1**.

To further boost accuracy, we considered the possibility that averaging the similarity matrices of the participants may increase the signal-to-noise ratio (SNR) of the data used for reconstruction purposes (Cowen, Chun, & Kuhl, 2014; Nemrodov et al., 2018). Specifically, a single average similarity matrix across the 14 participants was used for word space derivation, feature synthesis and word reconstruction. This manipulation led to robust performance over time; for instance, peak performance reached 70.4% (**Figure S3**) compared to the 63.5% average obtained for single-participant reconstructions. In addition, temporally-cumulative reconstruction reached 84.5% accuracy (*p* = 0.001, permutation test) which is significantly higher than the corresponding results of any single participant (all *p*'s < 0.001, permutation test).

To further explore the generalizability of our reconstruction results, we compared the reconstructed words not only to the 80 words in our stimuli set, but to all possible CVC pseudo/words constructed by considering all possible combinations of letters occurring in each position in our stimuli set, for a total of 750 pseudo/words. The average reconstruction accuracy was slightly lower (M = 68.8%, SD = 5.1%; *t*(13) = 7.37, *p* < 0.001), but still well above chance.

A complementary assessment of group-based reconstruction results also considered experimental data, instead of objective pixelwise similarity, from a novel group of 20 naïve participants. Specifically, data from a two-alternative forced choice (2AFC) task involving the match of word reconstructions to their stimulus targets (vs any possible stimulus foil) confirmed that reconstructions were successful (M = 81.8%, SD = 6.6%; two t-test against 50% chance across participants, *t*(19) = 21.56, *p* < 0.001).

### Visual letter reconstruction

To bridge our results with previous investigations into single-letter reconstructions, we proceeded to compute the reconstruction accuracy for each letter position. Of note, this analysis can reveal potential differences in accuracy across different letter positions and, also, facilitate an examination of the contribution of each letter position to whole-word image reconstruction.

To this end, we assessed group-based reconstruction accuracies separately for each position. This analysis revealed that the middle vowel has the highest reconstruction accuracy relative to the first consonant (paired permutation test, *p* = 0.001) and the last consonant (paired permutation test, *p* = 0.001) – see **Figure 7**.

The result above is particularly intriguing given the importance of consonants for word recognition (Vergara-Martínez, Perea, Marín, & Carreiras, 2011). Two possible mechanisms might be responsible for this difference across letter positions. The first possibility concerns the central position of the vowel at fixation and thus, its privileged encoding in the EEG signal. In other words, the vowel may be better reconstructed because there is more information related to its visual processing in the EEG signal. Another explanation stems from the fact that vowels might be objectively more discriminable than consonants, for instance, because there are fewer vowels than consonants in Roman scripts.

To examine this latter possibility, a visual theoretical observer was constructed based on a similarity matrix derived from the objective pixelwise image similarity of the original stimuli (see Methods). The theoretical observer assumes access to all visual information, thus providing a theoretical upper limit for EEG-based reconstruction. The overall accuracy of this theoretical observer for entire words was 98.9% - see **Figure 7**.

More relevant to the current question, we computed the reconstruction accuracy for each letter position, and found, in this case, no apparent advantage of the middle vowel relative to the first and last consonant of the word (**Figure 7**). Hence, differences in accuracy across position emerge from the structure of empirical data rather than from the nature of the method or from the visual properties of the stimuli. In particular, performance for vowels was not superior because vowels are more visually discriminable than consonants.

Another possibility we considered is that letters more frequent in the stimulus set at a given position are more accurately reconstructed due to the overrepresentation of their visual features in the derived CIMs. However, a correlation between relative letter frequency and letter reconstruction accuracy did not reveal any significant correlation at any position (all *p*'s > 0.05). Therefore, the higher EEG-based reconstruction accuracy of the middle vowel appears to be due to its central placement in the visual field rather than a direct outcome of the objective properties of the stimulus set. The central position of the vowel along with the smaller number of vowels relative to consonants may lead participants to assign more weight to vowel information. At the same time though we do point out that the results above indicate above-chance sensitivity to all letter positions.

Relevantly here, while vowels yield higher reconstruction accuracies relative to consonants, they may have lower discriminative value for word identification and reconstruction. Specifically, reconstructed vowels only distinguish between 5 sets of words containing 5 different potential vowels while consonants can distinguish between substantially more sets of words containing 15 and 10 different potential consonants in the first and the third position, respectively. To address this possibility, we have conducted an additional analysis aimed at clarifying the contribution of each letter position to word reconstruction. Specifically, we have correlated reconstruction accuracy for each letter position with word reconstruction while partialling out the contribution of the other two positions. This analysis was conducted across the 80 word stimuli for participant-averaged reconstruction estimates. Interestingly, the results showed that both the first consonant and the last made significant contributions to word reconstruction ($r$ = 0.72, $p$ < 0.001 and $r$ = 0.61, $p$ < 0.001, respectively) while the vowel only made a marginally significant contribution ($r$ = 0.19, $p$ = 0.097).

Thus, while vowel reconstruction shows the highest levels of reconstruction accuracy per position, it contributes the least to word reconstruction. This result provides further evidence for the ability of reconstruction to capture information across multiple letter positions and, also, it provides convergence with the importance of consonants for word recognition, as noted above (Vergara-Martínez et al., 2011).

## Individual differences

While the analyses above capitalize on the similarity of data structure across participants to boost overall reconstruction accuracy, conversely, it is important to consider individual variability and the source of such variability in our data. From a methodological standpoint, this analysis could also inform the ability of reconstruction techniques to shed light on individual differences in perception more generally.

To this end, first, we computed typicality estimates based on the reconstruction accuracies of each participant. Specifically, the typicality of one participant was measured as the correlation between the reconstruction accuracies of all 80 stimuli from that participant and the average reconstruction accuracies from all other participants. All typicality estimates were above chance (all $p$'s < 0.001) in agreement with the presence of similar data structure across participants, as noted above. At the same time, an examination of typicality and accuracy across participants (**Figure S4**) showed no systematic relationship (Spearman correlation, $p$ = 0.45). Thus, the reconstruction procedure is effective even for less typical participants and its success is not impacted by participant typicality.

For completeness, we also estimated the typicality of group-based reconstructions, relying on an average confusability matrix, and of the theoretical observer. Specifically, these estimates were computed as the correlation between the corresponding reconstruction accuracies across 80 words and the average reconstruction results across all 14 participants. As expected, group-based data scored high on typicality given that they rely primarily on a data structure common across participants (Pearson correlation, $r$ = 0.89, $p$ < 0.001). In contrast, the theoretical observer, while still significant, scored low on typicality ($r$ = 0.38, $p$ < 0.001). This is consistent with our results above indicating that the theoretical observer stands out from human data, for instance, through better access to visual information relating to the first and last consonant of a word.

Further, to identify and to visualize individual differences in the representation of words across participants, we computed the average reconstruction accuracy of all words separately for each pixel and each participant. Then, PCA was conducted across the heatmaps of all participants – see **Figure 8a**. Lastly, we computed averages of these maps, separately for each PCA dimension, weighted proportionally to the z-scored coefficient corresponding to each participant on a given dimension. Thus, such weighted sums provide new CIMs illustrating different sources of participant variability. An examination of these CIMs (**Figure 8b**) indicate that individuals vary primarily in their ability to capture information in the central position, as illustrated for the first principal component. However, additional visual cues, such as the lower part of the last consonant illustrated for the second component, are also a source of individual variability.

### Additional psycholinguistic analyses

While our investigation is primarily focused on visual-orthographic processing of single words, an exploration of multiple psycholinguistic variables and their impact on word decoding could be informative. Specifically, such an exploration may provide a more complete picture of the perceptual and linguistic processes underlying reading and pave the way for dedicated studies of such processes relying on pattern analyses of EEG signals.

Accordingly, to explore the dependence of word classification on a variety of psycholinguistic variables, multiple regression analysis was conducted to account for the average EEG-based discriminability of each word across participants. To this aim, we considered seven psycholinguistic measures estimated across English words irrespective of length and format (i.e., not just CVC). Specifically, we considered: positional letter frequency, positional letter-bigram frequency (i.e., sublexical covariates), word frequency, orthographic Levenshtein distance, number of orthographic neighbors, phonological Levenshtein distance and number of phonological neighbors. Each of these measures was correlated with the discriminability of each word, computed as the average accuracy of its EEG-based classification across a 50-650ms interval from all other 79 words. Of note, we considered here EEG classification rather than reconstruction results, since the latter depend on the former. Also, we reasoned that reconstruction captures primarily visual aspects of neural processing while decoding may be facilitated by multiple linguistic properties of the stimuli and, thus, contain a richer and more diverse structure.

The results of this analysis pointed to word frequency (b = 5.04e-05, *t*(72) = 2.48, *p* = 0.015) and the number of orthographic neighbors (b = 0.01, *t*(72) = 2.05, *p* = 0.044) as significant predictors making an independent contribution to accounting for EEG data. To assess the robustness of these results we performed this analysis again using psycholinguistic measures estimated exclusively across CVC words. This analysis rendered qualitatively similar results, though the number of orthographic neighbors only provided a marginally significant contribution this time (b = 0.02, *t*(72) = 1.87, *p* = 0.066). (For an additional examination of these measures with respect to their impact on individual differences, see Supporting information, Psycholinguistic variables and individual differences.)

To align our current results with the literature, we repeated the multiple regression analysis for the average EEG-based discriminability of each word across participants with psycholinguistic measures obtained from the UNION database while taking the natural logarithm of SUBTL word frequency. This analysis showed similar numerical trends to the raw frequency data but did not reach statistical significance.

The pairwise EEG-based word discriminability for every 10ms interval was also correlated with the corresponding estimates of semantic similarity obtained from the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013). No significant correlations were found.

## Discussion

Reading relies on the ability to identify words quickly and reliably by access to their visual-orthographic characteristics (Carreiras, Armstrong, Perea, & Frost, 2014; Perfetti, 2007; Verhoeven, Reitsma, & Siegel, 2011). The present work aims to uncover the structure of underlying word representations with the aid of pattern analysis and reconstruction techniques as applied to EEG data. Our results demonstrate the feasibility of decoding and reconstructing visual words from neural data. These results evince several noteworthy aspects, as follows.

First, word decoding reveals a representational space shaped by visual and orthographic features consistent with that found by fMRI investigations of the visual word form area (vWFA) (Baeck et al., 2015; Nestor et al., 2013). Specifically, sensitivity to letter identity is found for every letter position across a relatively large and well-controlled pool of words (e.g., having the same length and CVC structure). Of note, orthographic similarity accounts for the structure of the data beyond pure visual similarity, suggesting sensitivity to visual forms more abstract than the pictorial content of a given stimulus (Carreiras, Armstrong, & Dunabeitia, 2018). Additional correlations between decoding accuracy, on one hand, and word frequency and the number of orthographic neighbors, on the other, also confirm the impact of linguistic processing on our results.

Second, visual word features were derived directly from the structure of the EEG data and used for the purpose of word image reconstruction. Previous work has reconstructed single characters such as letters from fMRI patterns in visual cortex (Schoenmakers et al., 2013; Shen et al., 2017; Thirion et al., 2006) or visualized their representation through psychophysical methods (Gosselin & Schyns, 2003) – see also complementary work (Pasley et al., 2012) targeting ECoG-based speech reconstruction. In contrast, the current results demonstrate, for the first time to our knowledge, the ability to reconstruct the visual appearance of whole words from neural recordings. Specifically, accuracy was above chance for every letter position confirming that reconstruction retrieves the appearance of the entire word rather than of a single later. Of note, reconstruction accuracy was well above chance for every participant (range: 58% - 77%) and even higher when combining the data of multiple participants (84.50%). Thus, reconstruction results are quite robust and, moreover, they serve to clarify and to visualize the information underlying neural decoding.

Third, we find that the time course of decoding and reconstruction peaks around 200ms after stimulus onset, in the proximity of the N170 component, but reaches significance earlier, soon after 100ms. These findings are consistent with access to lexical orthographic information for familiar words between 100 and 200ms (Araújo et al., 2015; Dufau, Grainger, Midgley, & Holcomb, 2015; Hauk et al., 2006; Sereno, Rayner, & Posner, 1998) as well as with the significance of the N170 component for orthographic processing, presumably driven by a vWFA neural generator (Brem et al., 2006). Interestingly though, cross-temporal generalization as well as temporal cumulative analyses suggest the presence of complementary information across an extended temporal interval, roughly between 100-600 ms. One likely explanation for this result is a quick and efficient reading mechanism that allows subsequent refinement, as illustrated by the need to distinguish between highly confusable words (Hirshorn et al., 2016).

Fourth, we find that participants vary considerably in how typically they represent words relative to one another, yet that does not determine reconstruction success. More importantly, we extract visual templates that account for individual differences in word representations. These templates reveal differences in sensitivity to the visual encoding of the middle vowel as well as of the lower part of the last consonant, possibly related to different reading strategies and/or different types and degrees of language experience (Seidenberg & MacDonald, 2018). Thus, neural-based image reconstruction can shed light on visual-orthographic differences in reading and, in doing so, complement the extensive work on phonological and semantic individual differences (Brady, Braze, & Fowler, 2011).

More generally, from a methodological standpoint, the current findings demonstrate and illustrate the ability of EEG signals to support the recovery and the visualization of fine-grained neural representations, such as those supporting reading. Recent work(Nemrodov et al., 2018) has demonstrated the feasibility of EEG-based image reconstruction for human face stimuli. Here, we confirm this demonstration by appeal to a new class of visual stimuli and, thus, open the door to more extensive and varied applications of image reconstruction to EEG data.

Of particular interest in this sense is clarifying the nature of the representations accessible through reconstruction. The differential retrieval of information across letter positions, as reported above, may speak to this issue. Specifically, the privileged encoding of the middle vowel, likely driven by its central fixation, suggests access to more general, early visual representations. Given the importance of consonants for word recognition (Vergara-Martínez et al., 2011), it is possible that such representations are subsequently refined into more abstract ones, subject to language-specific constraints, such as the need to identify consonants correctly. At the same time, we note that orthographic word processing relies on flexible representations sensitive to task demands (Chen et al., 2015; Yang & Zevin, 2014). Hence, a different experimental task involving deeper lexical-semantic processing than the one-back memory task used here, may provide access to higher-level word representations.

Relevantly here, an important challenge for future work concerns the ability to reconstruct the appearance of entire sentences rather than single words through the use of image reconstruction methods relying on more complex combinations of visual and psycholinguistic features. This would allow investigating the interplay of multiple factors impacting discourse (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005), including semantics and phonology, which largely fell outside the scope of the present work. In particular, the nature of the experimental task as well as the large number of word repetitions likely diminished our ability to capture semantic effects (Rossell, Price, & Nobre, 2003; Rugg, 1985). Also, in the absence of words with irregular pronunciation, the correlation of phonological and orthographic properties made difficult disentangling their distinct contributions to neural processing. Thus, it is possible that the structure of the EEG data also reflects phonological effects, especially given the role of rapid phonological feedback to posterior visual areas in stabilizing grapheme string representations. The extension of our present findings to different stimulus sets and languages with more complex grapheme-phoneme mapping will be particularly relevant in this respect and, also, help assess their cross-linguistic validity (Rueckl et al., 2015; Share, 2008).

Importantly, the evaluation of individual differences, as illustrated above, carries relevance for the study of dyslexia. Impaired visual expertise for print appears to play a role in the development of at least some subtypes of dyslexia (Helenius, Tarkiainen, Cornelissen, Hansen, & Salmelin, 1999; Maurer et al., 2007; Paulesu et al., 2001). Hence, image reconstruction could provide a valuable means of revealing impaired visual processing and representations in individuals with dyslexia and of refining our understanding of the subtypes of this disorder (Zoubrinetzky, Bielle, & Valdois, 2014).

To conclude, our work illustrates the benefit of a new approach to the study of visual word representations. Theoretically, our results help to uncover the visual-orthographic structure of such representations as well as the temporal dynamics of their processing. Methodologically, they showcase the ability of pattern analyses as applied to EEG data to reveal the fine-grained structure of neural representations. More generally, the current work paves the way to in-depth studies of reading, via EEG-based image reconstruction, in healthy individuals as well as in those with visual deficits.

**Data availability statement:**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Conflict of interest**: The authors declare no conflict of interest.

# References

Araújo, S., Faísca, L., Bramão, I., Reis, A., & Petersson, K. M. (2015). Lexical and sublexical orthographic processing: An ERP study with skilled and dyslexic adult readers. *Brain and Language*, *141*, 16–27.

Baeck, A., Kravitz, D., Baker, C., & de Beeck, H. P. O. (2015). Influence of lexical status and orthographic similarity on the multi-voxel response of the visual word form area. *Neuroimage*, *111*, 321–328.

Bartlett, S., Kondrak, G., & Cherry, C. (2009). On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 308–316). Association for Computational Linguistics.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M.-H., Echallier, J.-F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*(3), 235–260.

Brady, S. A., Braze, D., & Fowler, C. A. (2011). *Explaining individual differences in reading: Theory and evidence*. Psychology Press.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Brem, S., Bucher, K., Halder, P., Summers, P., Dietrich, T., Martin, E., & Brandeis, D. (2006). Evidence for developmental changes in the visual word processing network beyond adolescence. *NeuroImage*, *29*(3), 822–837. https://doi.org/10.1016/j.neuroimage.2005.09.023

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Carreiras, M., Armstrong, B. C., & Dunabeitia, J. A. (2018). Reading. In *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition*. John Wiley & Sons.

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, *18*(2), 90–98.

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage*, *54*(4), 3028–3039.

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Chen, Y., Davis, M. H., Pulvermüller, F., & Hauk, O. (2015). Early visual word processing is flexible: Evidence from spatiotemporal brain dynamics. *Journal of Cognitive Neuroscience*, *27*(9), 1738–1751.

Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: Reconstructing face images from evoked brain activity. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2014.03.018

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262.

Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological Science*, *26*(12), 1887–1897.

Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for highly selective neuronal tuning to whole words in the "visual word form area." *Neuron*, *62*(2), 199–204.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *14*(5), 505–509.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Helenius, P., Tarkiainen, A., Cornelissen, P., Hansen, P. C., & Salmelin, R. (1999). Dissociation of normal feature analysis and deficient processing of letter-strings in dyslexic adults. *Cerebral Cortex*, *9*(5), 476–483.

Hirshorn, E. a., Li, Y., Ward, M. J., Richardson, R. M., Fiez, J. a., & Ghuman, A. S. (2016). Decoding and disrupting left midfusiform gyrus activity during word reading.

*Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 201604126. https://doi.org/10.1073/pnas.1604126113

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176–186.

Maurer, U., Brem, S., Bucher, K., Kranz, F., Benz, R., Steinhausen, H.-C., & Brandeis, D. (2007). Impaired tuning of a fast occipito-temporal response for print in dyslexic children learning to read. *Brain*, *130*(12), 3200–3210.

Maurer, U., Zevin, J. D., & McCandliss, B. D. (2008). Left-lateralized N170 effects of visual expertise in reading: evidence from Japanese syllabic and logographic scripts. *Journal of Cognitive Neuroscience*, *20*(10), 1878–1891.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proc. of HLT-NAACL*. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915–929. https://doi.org/10.1016/j.neuron.2008.11.004

Mouraux, A., & Iannetti, G. D. (2008). Across-trial averaging of event-related EEG responses and beyond. *Magnetic Resonance Imaging*, *26*(7), 1041–1054.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5), 2.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*. https://doi.org/10.1016/j.neuron.2009.09.006

Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The Neural Dynamics of Facial Identity Processing: insights from EEG-Based Pattern Analysis and Image Reconstruction. *Eneuro*, ENEURO.0358-17.2018. https://doi.org/10.1523/ENEURO.0358-17.2018

Nestor, A., Behrmann, M., & Plaut, D. C. (2013). The neural basis of visual word form processing: a multivariate investigation. *Cerebral Cortex*, *23*(7), 1673–1684.

Nestor, A., Plaut, D. C., & Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences*, *113*(2), 201514551. https://doi.org/10.1073/pnas.1514551112

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*. https://doi.org/10.1016/j.cub.2011.08.031

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., … Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001251

Paulesu, E., Démonet, J.-F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., … Frith, C. D. (2001). Dyslexia: cultural diversity and biological unity. *Science*, *291*(5511), 2165–2167.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383.

Price, C. J., & Devlin, J. T. (2011). The Interactive Account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*(6), 246–253. https://doi.org/10.1016/j.tics.2011.04.001

Rauschecker, A. M., Bowen, R. F., Parvizi, J., & Wandell, B. A. (2012). Position sensitivity in the visual word form area. *Proceedings of the National Academy of Sciences*, *109*(24), E1568–E1577.

Rossell, S. L., Price, C. J., & Nobre, A. C. (2003). The anatomy and time course of semantic priming investigated by fMRI and ERPs. *Neuropsychologia*, *41*(5), 550–564.

Rueckl, J. G., Paz-Alonso, P. M., Molfese, P. J., Kuo, W.-J., Bick, A., Frost, S. J., … Duñabeitia, J. A. (2015). Universal brain signature of proficient reading: Evidence from four contrasting languages. *Proceedings of the National Academy of Sciences*, *112*(50), 15510–15515.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, *22*(6), 642–647.

Schoenmakers, S., Barth, M., Heskes, T., & van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. *NeuroImage*, *83*, 951–961. https://doi.org/10.1016/j.neuroimage.2013.07.043

Seidenberg, M. S., & MacDonald, M. C. (2018). The Impact of Language Experience on Language and Reading. *Topics in Language Disorders*, *38*(1), 66–83.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport*, *9*(10), 2195–2200.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: the perils of overreliance on an" outlier" orthography. *Psychological Bulletin*, *134*(4), 584.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017). Deep image reconstruction from human brain activity. *BioRxiv*, 240317.

Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. *Neuron*, *76*(3), 640–652.

Suppes, P., Lu, Z.-L., & Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, *94*(26), 14965–14969.

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin*. https://doi.org/10.1037/a0030266

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, *33,* 1104–1116. https://doi.org/10.1016/j.neuroimage.2006.06.062

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/0278-7393.31.3.443

Vergara-Martínez, M., Perea, M., Marín, A., & Carreiras, M. (2011). The processing of consonants and vowels during letter identity and letter position assignment in visual-word recognition: An ERP study. *Brain and Language*. https://doi.org/10.1016/j.bandl.2010.09.006

Verhoeven, L., Reitsma, P., & Siegel, L. S. (2011). Cognitive and linguistic factors in reading acquisition. *Reading and Writing*. https://doi.org/10.1007/s11145-010-9232-4

Yang, J., & Zevin, J. (2014). The impact of task demand on visual word recognition. *Neuroscience*, *272*, 102–115.

Zoubrinetzky, R., Bielle, F., & Valdois, S. (2014). New insights on developmental dyslexia subtypes: heterogeneity of mixed reading profiles. *PloS One, 9*(6), e99337.

## Figure legends

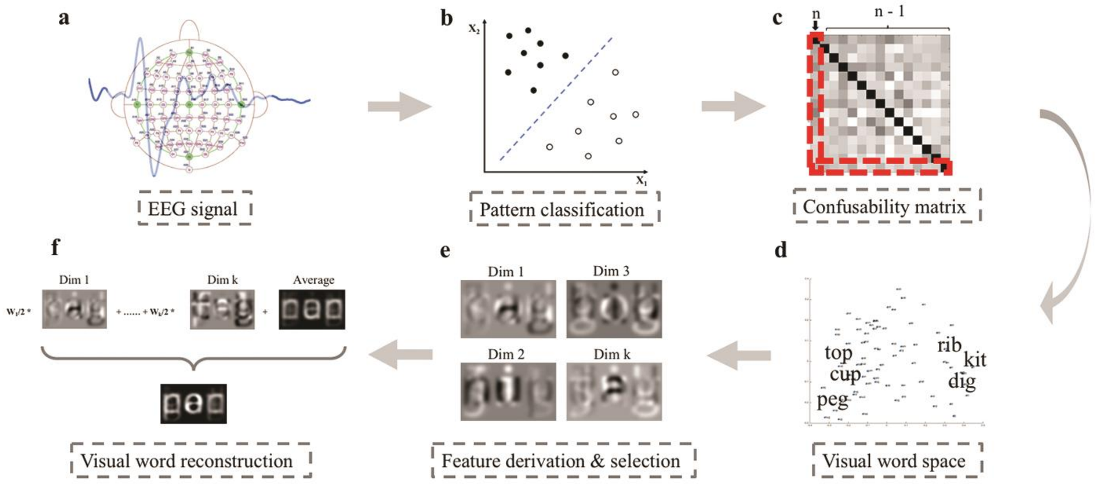

**Figure 1.** Procedure for visual word decoding and reconstruction: (a) ERP traces across 12 bilateral occipitotemporal (OT) electrodes were recorded for each; (b) linear classification was conducted across the corresponding spatiotemporal patterns; (c) discriminability estimates were summarized by a similarity matrix; (d) a 20-dimension word similarity space was estimated from the similarity structure of the data using a leave-one-out procedure (only two dimensions are displayed here for visualization purposes); (e) visual features were derived for each dimension and evaluated for the presence of significant visual information, and (f) a word image corresponding to the left-out stimulus was reconstructed though a linear combination of significant features.

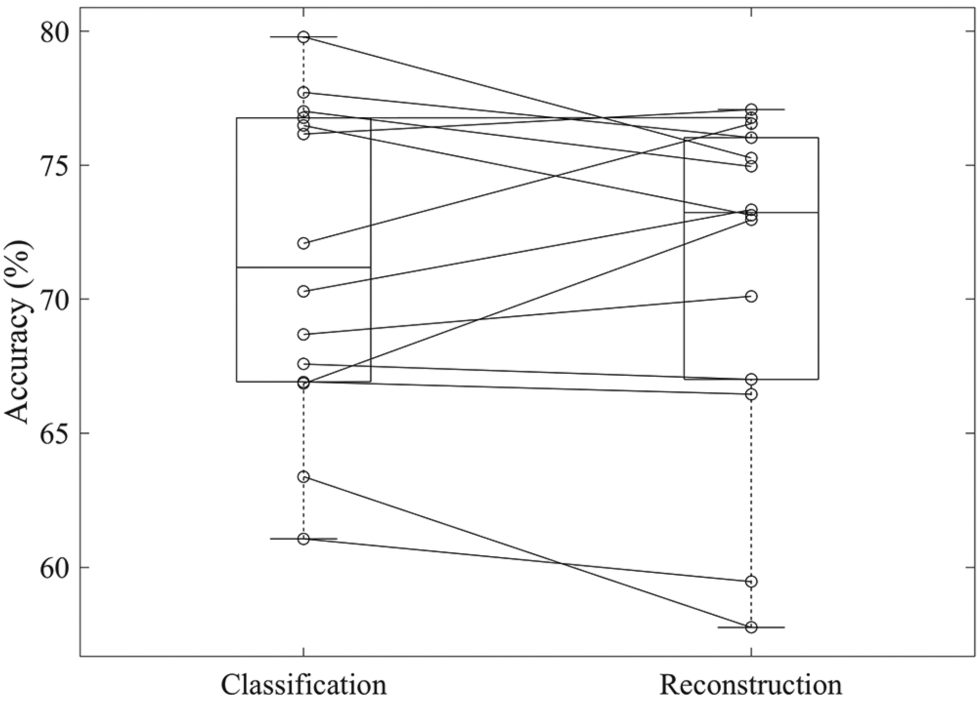

**Figure 2.** Accuracy of word classification and image reconstruction, based on a 50-650ms temporal window, for each of 14 participants. Estimates were above chance for all participants (*p's* < 0.01, permutation test). Classification and reconstruction accuracies were comparable in magnitude and correlated across participants (*r* = 0.86, *p* = 0.0001).

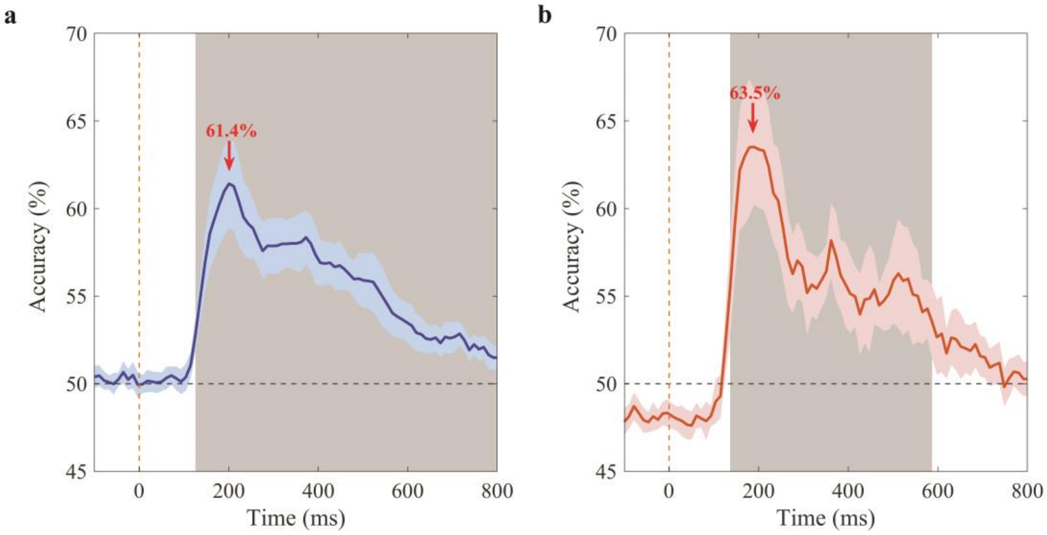

**Figure 3.** (a) The time course of word discrimination revealed by pattern classification for

9.75 ms windows between -100 – 800ms. Classification reached significance at 114ms post-stimulus onset and peaked at 200ms. (b) The time course of reconstruction obtained by performing image reconstruction for 9.75 ms windows between -100 – 800ms. Performance reached significance at 125ms post-stimulus onset and peaked at 190ms (gray shading marks intervals of above-chance accuracy, two-tailed one-sample t-test, q < 0.01; blue/red shading marks 95% confidence intervals across participants).

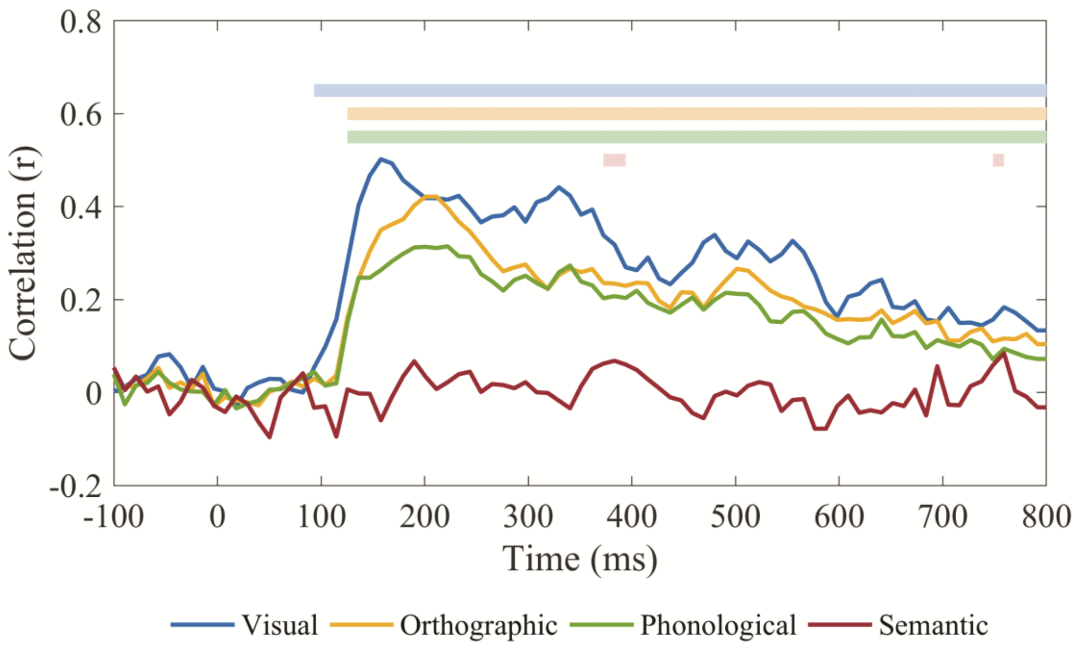

**Figure 4.** Correlations between EEG-based word discriminability (i.e., average accuracy of pairwise word classification) and estimates of visual, orthographic, phonological and semantic similarity. Word discriminability, estimated for 9.75 ms windows between -100 – 800ms, was significantly correlated with the first three measures across extensive intervals, but only briefly with semantic similarity (color bars at the top mark intervals of significant correlation, q < 0.01).

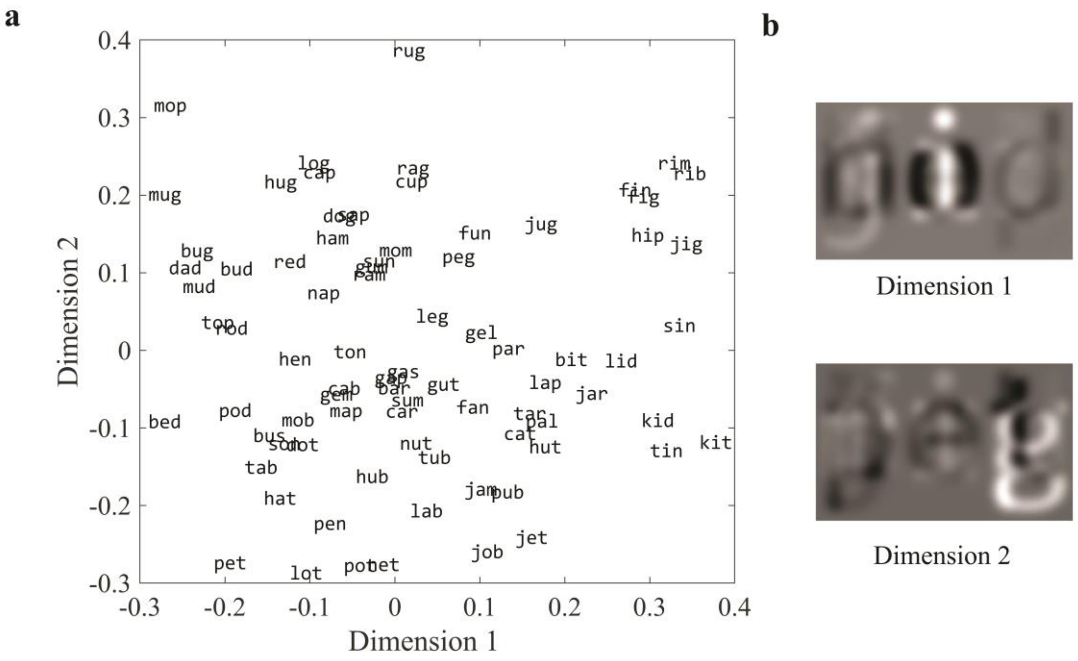

**Figure 5.** Example of (a) multidimensional word space derived from a 50-650ms temporal window, and (b) CIMs corresponding to the first two dimensions synthesized from this word space through image classification. For convenience, the figure shows only the first two dimensions for one representative participant (the two dimensions account for 7.6% and 6.5% of the variance, respectively).

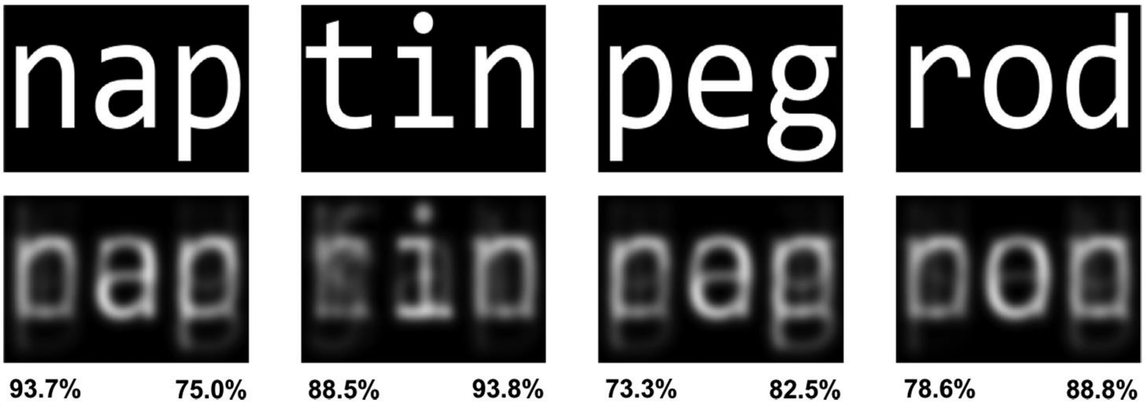

**Figure 6.** Examples of stimuli and reconstructed words based on a 50-650ms temporal window from a single representative participant. The first row shows word stimuli and the

second row displays corresponding reconstructed word images. The values at the bottom left of each image indicate objective accuracy based on pixel-wise image similarity. The values at the bottom right indicate experimental estimates from a separate group of participants. The superior performance of reconstruction in the vowel position, relative to the two consonant positions, can be observed in the figure.

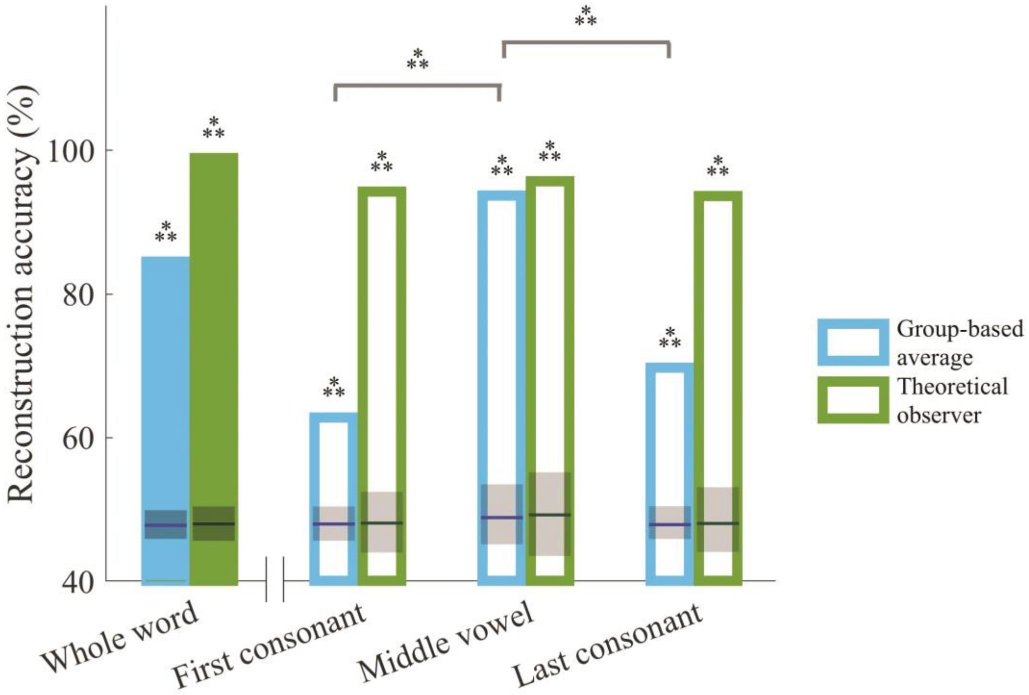

**Figure 7.** Reconstruction accuracy for whole words was first calculated from the group-based average data, based on a 50-650ms temporal window, and, also, based on the theoretical observer. Reconstruction accuracy was then calculated separately for each letter position. The advantage of the middle vowel was apparent for EEG data but not for the theoretical observer (permutation test, ***, $p < 0.001$).

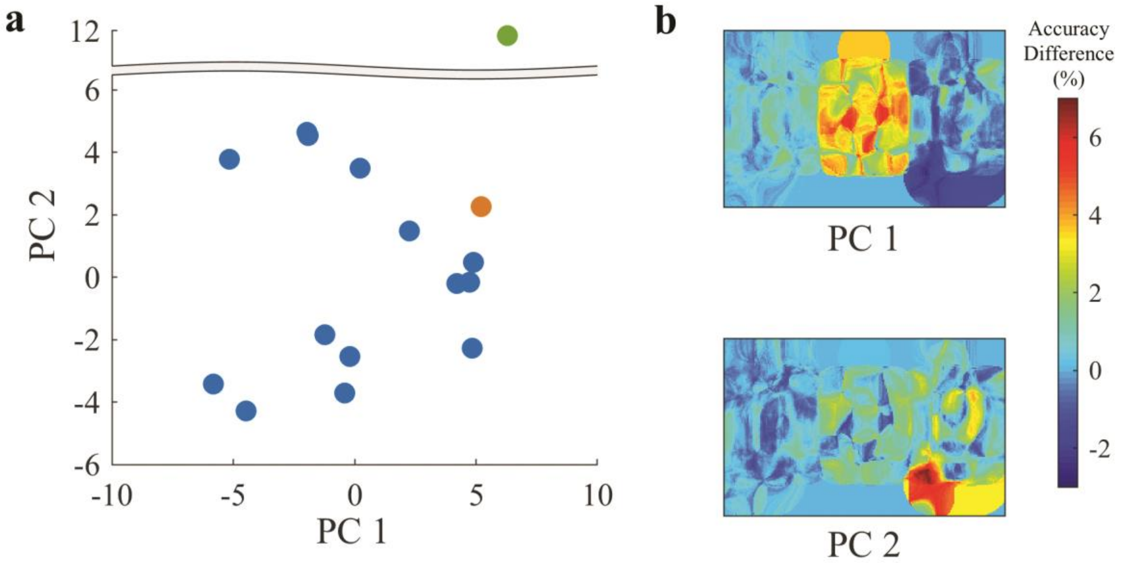

**Figure 8.** Individual differences in word reconstruction based on a 50-650ms temporal window. (a) PCA was applied to pixelwise accuracy heatmaps (for convenience, only the first two PCs are plotted). Each blue dot represents one of 14 participants while orange marks group-averaged data and green marks the theoretical observer. (b) Classification images were computed for each component to illustrate sources of individual variability: heatmaps illustrate components of variability in pixelwise reconstruction accuracy across participants. Specifically, PC1 indicates that participants vary primarily in how accurately they represent the vowel in the central position while PC2 indicates that participants also vary in how accurately they represent the bottom part of the second consonant.