COGNITIVE NEUROPSYCHOLOGY, 2016
VOL. 33, NOS. 3–4, 265–275
http://dx.doi.org/10.1080/02643294.2016.1195343

SHORT REPORT

Using precise word timing information improves decoding accuracy in a multiband-accelerated multimodal reading experiment

An T. Vu[a*], Jeffrey S. Phillips[b], Kendrick Kay[c**], Matthew E. Phillips[d], Matthew R. Johnson[e], Svetlana V. Shinkareva[f], Shannon Tubridy[g], Rachel Millin[d], Murray Grossman[b], Todd Gureckis[g], Rajan Bhattacharyya[d] and Essa Yacoub[a]

[a]Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN, USA; [b]Department of Neurology, University of Pennsylvania, Philadelphia, PA, USA; [c]Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA; [d]HRL Laboratories LLC, Malibu, CA, USA; [e]Department of Psychology, Yale University, New Haven, CT, USA; [f]Department of Psychology, University of South Carolina, Columbia, SC, USA; [g]Department of Psychology, New York University, New York, NY, USA

ABSTRACT
The blood-oxygen-level-dependent (BOLD) signal measured in functional magnetic resonance imaging (fMRI) experiments is generally regarded as sluggish and poorly suited for probing neural function at the rapid timescales involved in sentence comprehension. However, recent studies have shown the value of acquiring data with very short repetition times (TRs), not merely in terms of improvements in contrast to noise ratio (CNR) through averaging, but also in terms of additional fine-grained temporal information. Using multiband-accelerated fMRI, we achieved whole-brain scans at 3-mm resolution with a TR of just 500 ms at both 3T and 7T field strengths. By taking advantage of word timing information, we found that word decoding accuracy across two separate sets of scan sessions improved significantly, with better overall performance at 7T than at 3T. The effect of TR was also investigated; we found that substantial word timing information can be extracted using fast TRs, with diminishing benefits beyond TRs of 1000 ms.

ARTICLE HISTORY
Received 15 May 2015
Revised 5 May 2016
Accepted 19 May 2016

KEYWORDS
7T; decoding; functional magnetic resonance imaging; multiband; timing; words

Introduction

The blood-oxygen-level-dependent (BOLD) signal measured in functional magnetic resonance imaging (fMRI) experiments is generally regarded as sluggish and poorly suited for probing neural function at the faster timescales where sentence processing and lexical semantics come into play. As such, many fMRI studies on sentence comprehension utilize a relatively long repetition time (TR, ~2000 ms) and limit experiments to single word presentation per trial (Correia et al., 2014; Flegal, Marin-Gutierrez, Ragland, & Ranganath, 2014; Meltzer-Asscher, Mack, Barbieri, & Thompson, 2015). The majority of studies that do present full sentences typically only model sentence onsets, assuming all words within a sentence are read simultaneously and/or processed throughout the trial (Berken et al., 2015; He et al., 2015; Moisala et al., 2015; Vitello, Warren, Devlin, & Rodd, 2014). While such a model is expedient, it does not fully capture, for example, syntactic differences between sentences, which involve subtle word order variations (Dapretto & Bookheimer, 1999; Wu, Vissiennon, Friederici, & Brauer, 2016). Modelling of individual words and their timing within sentences should enable more in-depth investigations of such dynamic sentence processing and uncover additional brain regions performing sentence computations of interest. With the recent advancements in multiband (MB) acceleration allowing for sub-second whole-brain imaging (Feinberg et al., 2010; Moeller et al., 2010), determining how much word timing information is available in the BOLD response becomes exceedingly relevant.

Early work on the effect of repetition time (TR) on fMRI experimental paradigms found that shorter TRs (on the order of 1000 ms) provide optimal statistical power (Constable & Spencer, 2001). Unfortunately, few studies to date have been able to take advantage of such short TRs since they typically require scanning only a fraction of the brain (for example, see Kay,

CONTACT An T. Vu anvu@berkeley.edu

Supplemental data for this article can be accessed at 10.1080/02643294.2016.1195343

[*]Current address: An Vu, Center for Imaging of Neurodegenerative Diseases, VAMC San Francisco, CA, since 06/2016
[**]Current address: Kendrick Kay, Department of Radiology, University of Minnesota, Minneapolis, USA, since 08/2015

Naselaris, Prenger, & Gallant, 2008). However, more recent studies have begun to show the value of acquiring slice-accelerated, whole-brain data with very short TRs. In addition to expected gains in BOLD contrast to noise ratio (CNR) through effective averaging of additional time points (Feinberg et al., 2010; Smith et al., 2013; Xu et al., 2013) and increasing benefits of physiological denoising (Tong & Frederick, 2014), use of accelerated imaging has also been shown to provide additional fine-grain temporal information (Chang et al., 2013; Chen et al., 2015).

While the delay between stimulus onset and peak BOLD response can be quite long (Hulvershorn, Bloy, Gualtieri, Leigh, & Elliott, 2005; Lee, Glover, & Meyer, 1995), the BOLD response for a particular brain region is impressively time invariant and temporally precise. Shifts in the temporal profile of the BOLD response corresponding to stimulus shifts on the order of 100 ms have been reliably detected using sub-second TRs (Chang et al., 2013; Menon, Luknowsky, & Gati, 1998). When applied to visual stimulation paradigms, Chen et al. (2015) showed that slice acceleration factors between 8 and 12 (corresponding to TRs between 300 and 600 ms) were optimal in terms of CNR and decodable information (as measured by movie frame classification accuracy). These results suggest that the timing of individual words within a sentence is recoverable from the BOLD response, especially with the shorter TRs achievable with multiband imaging.

Perhaps the primary concern regarding recovery of word timing information from sentences is the nonlinearity of the BOLD response elicited by multiple stimuli (e.g., words) presented in quick succession (Mukamel, Harel, Hendler, & Malach, 2004; Ogawa et al., 2000; Zhang, Zhu, & Chen, 2008). Fortunately, it has been shown that words elicit very different patterns of BOLD activity depending on the semantics of the individual words (Mitchell et al., 2008). Therefore, for simple sentences where each semantically rich word occurs only once per sentence (as in the study presented here), higher order brain regions involved in semantic processing should not be strongly impacted by this nonlinearity. Furthermore, use of higher field strength (e.g., 7T) also greatly reduces nonlinearities in the BOLD response (Pfeuffer, McCullough, Van de Moortele, Ugurbil, & Hu, 2003), potentially due to the relatively larger contribution from the microvasculature at higher field. Thus, in addition to the gains in CNR expected from higher field strength (Ugurbil et al., 2013; Ugurbil, 2014; Vaughan et al., 2001), improvements in BOLD linearity should further enhance extraction of word timing information.

Here, we investigate how much word-specific timing information is recoverable from the BOLD signal in a multi-modal sentence reading experiment. By pushing the limits of slice-accelerated (also known as multiband) fMRI, we are able to achieve whole-brain scans at 3-mm resolution with a TR of just 500 ms. Importantly, this enables us to sample close to the reading rate of ~300 ms per word in an experiment where 240 unique sentences (constructed from a set of 261 words) were randomly presented to subjects at a rate of roughly one sentence every 5 s. To evaluate our data for individual word timing information, we tested two general linear model (GLM) models of the sentence data (schematics shown in Figure 1). The more traditional static word model includes a regressor for each of the 261 words and assumes that all words of a given sentence are read simultaneously at trial onset and processed throughout the trial. The dynamic word model is the same as the static model but with the timing of individual words preserved. Importantly, both models use an identical number of regressors. We used a metric of BOLD variance explained (e.g., *F*-value) and word pair classification accuracy (where the probability of each true word is contrasted with the probability of each possible word) as measures of model performance at various TRs (500 ms, 1000 ms, and 2000 ms) and field strengths (3T and 7T). The degree to which the dynamic model outperforms the static model reflects the amount of word timing information available in the BOLD response.

Method

Four subjects (two male) provided informed consent and participated in a multimodal reading study where sentences were delivered both visually and aurally. All experimental protocols were approved by the Committee for the Protection of Human Subjects at the University of Minnesota as well as the United States Air Force Research Laboratory (AFRL).

Stimulus paradigm

Stimuli consisted of the 240 sentences provided by the Intelligence Advanced Research Projects Activity

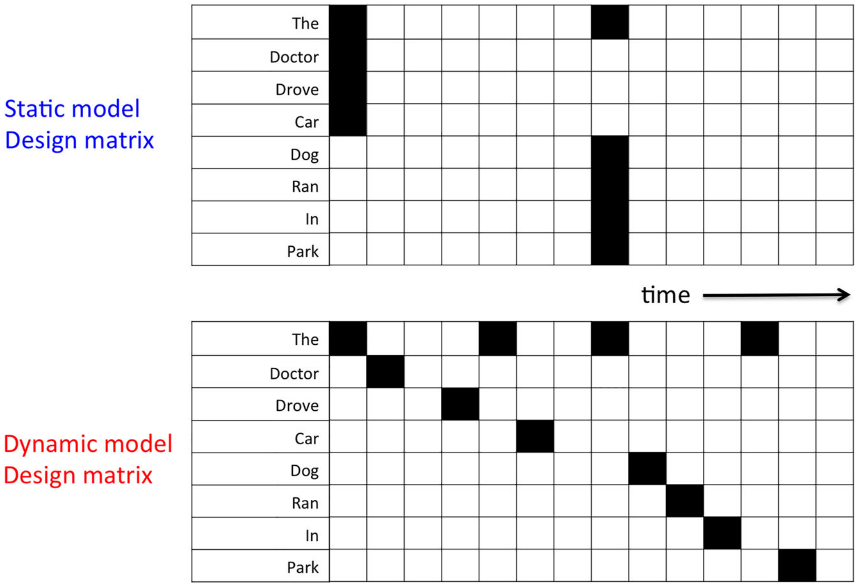

Figure 1. Schematic of general linear model (GLM) models for example sentence data ("The doctor drove the car." and "The dog ran in the park."). The static word model assumes that all words of a given sentence are read simultaneously at trial onset. The dynamic word model is the same as the static model but with the timing of individual words preserved. [To view this figure in colour, please see the online version of this Journal.]

(IARPA) for the Knowledge Representation in Neural Systems (KRNS) programme (Glasgow, Roos, Haufler, Chevillet, & Wolmetz, 2016). Some example sentences were as follows: "The driver wanted cold tea.", "The tired patient slept in the dark hospital.", and "The bird was red.". Subjects were instructed to think about the meaning of each sentence. Behavioural compliance was verified by means of probe questions where subjects must make a yes/no button-press response to the question regarding the previously presented sentence. Probe questions occurred randomly four times per ~5-min run, and subjects were provided score feedback at the end of each run. All 240 sentences were presented in pseudo-random order within five runs of 54 trials (including 4 probe trials and 2 fixation trials) with each trial lasting 5 seconds. Five sessions, with 10 runs (2 repeats per sentence) each, were acquired per subject. Each run began and ended with 10 s of fixation.

All stimuli were visually presented in the centre of a screen spanning 1–3 lines of text. Sentences were converted to audio via the txt2mp3mac software and were simultaneously presented to subjects at a rate of approximately 300 ms per word via Sensimetric earbuds, which also provided acoustic sound protection. The duration of visual presentation was matched to the audio (maximum duration was no greater than 3 s, which was then followed by a fixation cross) to ensure consistency of when individual words were read. Initial pilot studies confirmed that this multimodal duration-matched stimulus presentation resulted in the strongest/most reliable BOLD responses (when compared to text or sound alone or text plus sound where the text was left on for three seconds regardless of auditory stimulus duration). The font face and voice varied randomly from one run to the next to reduce the risk that decoding analyses would pick up on low-level perceptual details of the stimuli. Stimuli were presented via an iMac computer using Matlab Psychtoolbox, and the display screen was projected from behind the MRI scanner.

Data acquisition

Data were acquired from Siemens 3T Prisma and 7T Magnetom scanners using standard 32-channel receive array coils. The fMRI protocols were separately optimized at each scanner for CNR per unit time while maintaining reasonable spatial resolution (~3 mm) and a TR of 500 ms. Given that the optimal echo times (TEs) at 7T are shorter (19 ms versus 38 ms), we are able to acquire more slices per unit time and thus use a slightly lower multiband factor at 7T (6 versus 7 at 3T). The number of slices was set to the smallest multiple of MB that would cover >130 mm in the slice direction to ensure whole-brain coverage for our subjects (54 versus 49 slices at 3T). The slice thickness at 7T was made slightly thinner (2.5 mm versus 3.0 mm at 3T) to minimize B0 dropout, while also offsetting the relatively stronger T2* blurring along the phase encode direction, and to utilize the greater time efficiency at 7T. The in-plane field of view (FOV) was 210 mm, and Ernst angle (40 versus 47 at 3T) was used for both protocols. FOV shift factor of 3 was used to minimize slice-accelerated leakage and g-factors (Setsompop et al., 2012). Two of the subjects completed the entire experiment at both 3T and 7T, with sessions at each field strength interleaved across days. One of the 3T-only subjects was removed from analysis due to poor behavioural performance (session average of 50% accuracy or less on behavioural probe questions).

Data processing

fMRI data were preprocessed using FSL (www.fmrib.ox.ac.uk/fsl/) for motion correction and co-registration across sessions within subjects. No spatial or temporal smoothing was applied. Custom Matlab algorithms were used to perform regularized linear regression (Nishimoto et al., 2011) on individual voxels for both the static and dynamic word models after convolution with a canonical haemodynamic response function (HRF; Kay, 2014). For simplicity, word onset times in the dynamic word model were rounded to the nearest TR. Run-specific regressors for drift were modelled by zero-, first-, second-, and third-degree nuisance polynomials (Kay, David, Prenger, Hansen, & Gallant, 2008).

Model performance was evaluated by the *F*-value (Freedman, 2009), which is defined as variance explained by the model divided by variance of the residuals. The models were estimated on a per session basis to reduce computational memory requirements and produced one *t*-value estimate (beta divided by beta error; Freedman, 2009) per word per voxel. For each of the 261 words, 260 pairwise word classifications were performed using the correlation similarity metric between the training set *t*-values for the word A and testing set *t*-values for the words A and B. The 10,000 voxels with highest *F*-values in the training set were selected for calculating these correlations. The mean performance across all 261 × 260 classifications is reported. To avoid overfitting, voxel selection *F*-values and training *t*-values were based on the average values across two sessions. The average *t*-values across the remaining three sessions were used as the classification test set. This set-up allowed for 10 unique permutations of train and test sets, which were then used to estimate standard error of the mean (*SEM*) error bars. To evaluate the effect of TR on the above metrics, the preprocessed (motion corrected and co-registered) data as well as the post HRF convolved design matrices were temporally down-sampled after 2 and 4 TR window averaging to achieve effective TRs of 1000 ms and 2000 ms. Window averaging was used to emulate (a) the CNR advantage analogous to the CNR advantage that one could achieve at slower TRs (due to larger Ernst flip angle and lower multiband acceleration related g-factor noise) and (b) the increased chance for similar temporal blurring due to the traditional interleaved slice acquisition order where adjacent slices are acquired roughly half a TR apart in time. While window averaging also helps to low-pass filter the time series to some degree, it is noted that more elegant low-pass filters could be used to further reduce aliasing of signals above the Nyquist rate (e.g., via Matlab's decimation function). The impact of these various filters on BOLD temporal frequency information should be investigated in future studies.

Results

Figure 2 shows the mean difference between dynamic model *F*-values and static model *F*-values for a representative subject at 3T (thresholded for differences greater than 0.25, $p < .05$). Improvements in *F*-values can be seen in language processing areas such as the superior temporal gyrus (including primary auditory cortex), ventrolateral extreme of the central sulcus (Brodmann area 43), Broca's area, frontal eye fields, and portions of visual cortex. Negative differences were observed; however, they tended to be smaller and towards the edge of the brain, suggesting that simpler models that only account for the onset of each trial are sensitive to potentially non-interesting variance due to trial onset arousal/motion (not shown). These improvements are summarized in Figure 3 as the ratio of dynamic and static model testing set *F*-values (averaged across voxels having a ratio of at least 1.1, i.e., 10% increase, and a voxel

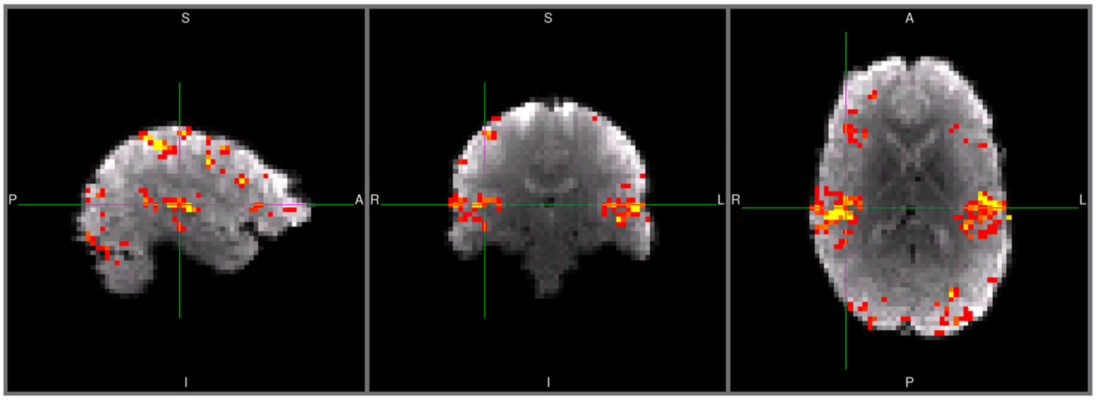

Figure 2. Mean difference between dynamic and static model *F*-values for a representative subject at 3T. Colour range from red to yellow represents *F*-value differences from 0.25 to 2, respectively. [To view this figure in colour, please see the online version of this Journal.]

cluster size of at least 10 contiguous voxels as determined from the separate training data set). On average across the three subjects and two field strengths, the use of the dynamic model increased *F*-values (i.e., explained variance) by a factor of ~1.2 or 20% relative to when the static model was used, which is significantly greater than a chance factor of 1 (*t*-test, *df* = 4, $p < .0001$). Down-sampling to an effective TR of 1000 ms highlighted even more the advantages of the dynamic model (paired *t*-test, *df* = 4, $p < .01$). However, further down-sampling to an effective TR of 2000 ms significantly reduced the benefit of the dynamic model relative to the 1000 ms TR (paired *t*-test, *df* = 4, $p < .05$). Ratio of dynamic to static model *F*-values, as opposed to raw *F*-values, were used here since the *F*-statistic is weighted by the number of degrees of freedom. Without additional special consideration, the *F*-statistic yields smaller standard-error-of-the-mean estimates and inflated *F*-values in datasets with shorter TRs/more time-points. To avoid this confound, we decided to use classification performance to evaluate differences in TR.

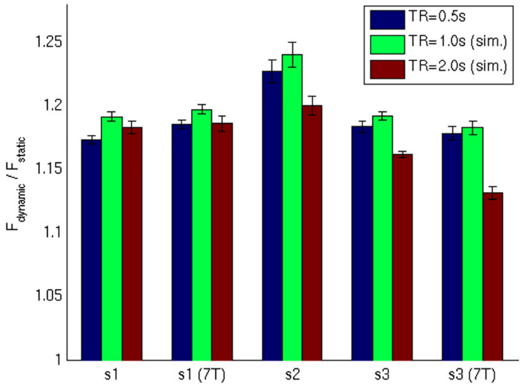

Figure 3. Ratio of dynamic and static model *F*-values. Data sets using 1.0-s and 2.0-s repetition times (TRs) were simulated from the 0.5-s TR data by down-sampling after window averaging. [To view this figure in colour, please see the online version of this Journal.]

Figure 4A shows the classification performance of both the static and the dynamic word models. While both models performed above chance (*t*-test, *df* = 4, $p < .01$), the dynamic model enabled significantly more accurate classifications (on average 70% versus 55%; paired *t*-test, *df* = 4, $p < .0001$). Furthermore, there was a significant (~5%) improvement in dynamic model word classification performance at 7T versus 3T (individually for Subjects 1 and 3 across 10 permutations; paired *t*-test, *df* = 9, $p < 10^{\wedge -9}$). Figure 4B shows the effect of TR on classification performance. While classification accuracy was quite robust to changes in TR, using the slowest TR of 2000 ms did result in significantly poorer classification performance (paired *t*-test, *df* = 4, $p < .01$). No significant difference was found between the shorter TRs of 500 ms and 1000 ms.

Table 1 shows the top 50 most accurately classified words rank-ordered by the accuracy scores achieved with the static model (averaged across all subjects and field strengths). Dynamic model accuracy scores set in bold indicate a tendency for higher classification scores. Scores set in italics indicate similar classification accuracy (within 1%). Notably, both models are able to classify sentence onset (signified by the word "The") with 100% accuracy. However, even for these 50 words best classified with the static model, the majority (66%) of them were classified with similar or better accuracy

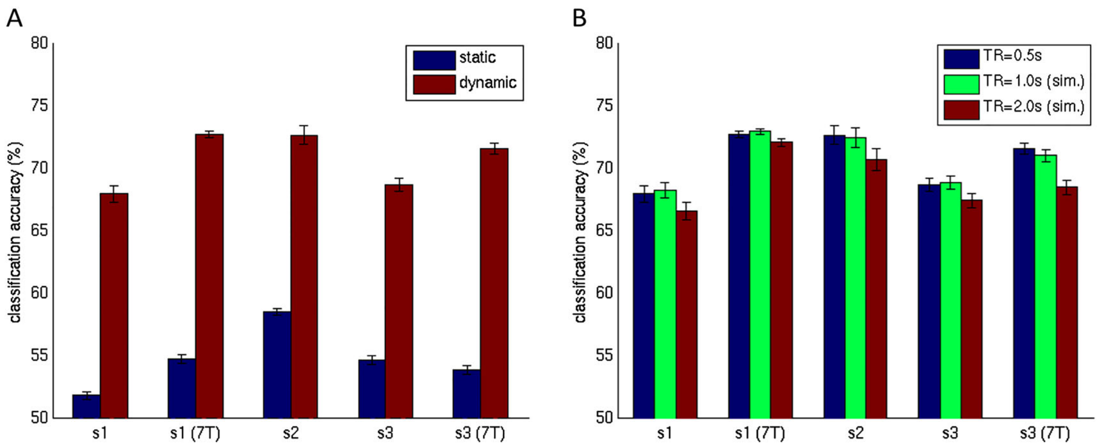

Figure 4. Pairwise word classification performance. (A) Static versus dynamic model. (B) Dynamic model with repetition times (TRs) = 0.5, 1.0, and 2.0 s. Data sets using 1.0-s and 2.0-s TRs were simulated from the 0.5-s data by down-sampling after window averaging. [To view this figure in colour, please see the online version of this Journal.]

using the dynamic model. Table 2 shows the top 50 most accurately classified words rank-ordered by the accuracy scores achieved with the dynamic model. In this case, all words (except "The") were, on average, more accurately classified by the dynamic word model.

Given that some words occurred more often than others across the 240-sentence stimulus set, we wanted to see whether this had any impact on individual word classification. To avoid bias in correlations, the top 7 most frequently occurring words, which occurred roughly an order of magnitude more often than the other 254 words, were removed from this analysis. The removed words were: "on", "to", "at", "in", "was", "the", and "The". Figure 5A plots the static model classification accuracy against number of word occurrences (within a given scan session) for these remaining 254 words (with each word represented by a circle). No significant correlation was found in this case ($r = .05$, $p = .46$). However, with the dynamic model (Figure 5B), there was a significant positive

Table 1. Top 50 most accurately classified words rank-ordered by static model accuracy.

	Static	Dynamic		Static	Dynamic
""The"	100.00	*100.00*	"minister"	67.68	*67.30*
"the"	99.98	80.51	"empty"	67.32	**73.57**
"was"	93.39	**96.84**	"cash"	67.08	62.51
"tea"	79.49	54.65	"black"	67.08	53.33
"glass"	78.18	**79.78**	"yellow"	66.74	62.35
"over"	77.02	**84.50**	"fish"	66.39	**79.55**
"hall"	76.30	56.57	"hiked"	65.72	**78.51**
"long"	76.25	58.34	"vacation"	65.66	**88.44**
"bicycle"	76.04	*75.83*	"girl"	65.56	**76.91**
"interviewed"	73.64	**92.28**	"wrote"	65.34	57.69
"left"	73.13	**96.17**	"old"	65.11	59.68
"spring"	73.12	72.13	"book"	65.01	**79.54**
"spoke"	73.04	**78.52**	"tourist"	64.87	**77.24**
"dangerous"	72.88	59.30	"dusty"	64.63	37.56
"during"	72.64	**82.22**	"used"	64.52	**76.50**
"grew"	72.08	59.13	"green"	64.35	60.49
"threw"	71.68	**87.43**	"scientist"	64.25	**69.79**
"about"	71.31	66.28	"liked"	64.21	**86.80**
"flew"	71.30	62.69	"from"	64.03	54.02
"night"	70.74	**71.91**	"blue"	63.99	*62.99*
"diplomat"	70.67	**83.36**	"flower"	63.74	**73.96**
"in"	69.62	**95.97**	"planned"	63.68	**81.85**
"stole"	69.27	**87.83**	"lived"	63.53	**74.25**
"negotiated"	68.83	**81.12**	"judge"	63.27	**79.76**
"accident"	68.10	**74.08**	"parent"	63.26	**75.78**

Note: Scores in bold indicate that dynamic model improves accuracy; scores in italics indicate that dynamic model has similar accuracy.

Table 2. Top 50 most accurately classified words rank-ordered by dynamic model accuracy.

	Static	Dynamic		Static	Dynamic
"The"	100.00	*100.00*	"gave"	61.51	**88.67**
"was"	93.39	**96.84**	"school"	41.85	**88.55**
"left"	73.13	**96.17**	"vacation"	65.66	**88.44**
"damaged"	62.49	**95.98**	"embassy"	60.31	**87.97**
"in"	69.62	**95.97**	"stole"	69.27	**87.83**
"kicked"	62.81	**95.97**	"slept"	50.90	**87.44**
"crossed"	50.92	**94.40**	"threw"	71.68	**87.43**
"lost"	54.41	**93.13**	"protest"	46.11	**87.19**
"bought"	58.67	**93.13**	"desk"	56.82	**87.10**
"approached"	43.29	**92.57**	"car"	49.53	**86.82**
"interviewed"	73.64	**92.28**	"liked"	64.21	**86.80**
"visited"	62.71	**92.12**	"family"	48.01	**86.74**
"hospital"	58.19	**92.12**	"saw"	55.59	**86.42**
"restaurant"	61.25	**91.85**	"found"	60.90	**86.22**
"took"	57.61	**91.65**	"victim"	62.45	**86.16**
"theatre"	46.05	**91.43**	"celebrated"	55.97	**85.85**
"trial"	55.48	**91.15**	"broke"	46.02	**85.57**
"beach"	53.36	**90.90**	"on"	60.44	**85.57**
"forest"	44.84	**90.37**	"flood"	48.59	**85.13**
"put"	59.65	**90.13**	"boat"	59.97	**85.07**
"park"	47.30	**90.09**	"lab"	43.38	**84.95**
"destroyed"	42.19	**89.86**	"office"	37.32	**84.90**
"survived"	56.87	**88.90**	"soldier"	45.87	**84.87**
"watched"	36.58	**88.87**	"read"	54.84	**84.77**
"arrested"	58.74	**88.79**	"artist"	56.42	**84.73**

Note: Scores in bold indicate that dynamic model improves accuracy; scores in italics indicate that dynamic model has similar accuracy.

correlation found ($r = .36$, $p < 10^{\wedge-8}$) such that words occurring more often tended to be classified more accurately. This suggests that words that are presented infrequently may classify better with the static model. This is confirmed in Figure 5C, where the difference in classification accuracy between dynamic and static models is shown to have a significant correlation with number of word occurrences ($r = .28$, $p < 10^{\wedge-5}$).

Similar plots comparing 3T and 7T dynamic model performance versus number of word occurrences are shown in Figures 5D–5F. Both 3T and 7T performances were significantly correlated with word occurrence ($r = .31$, $p < 10^{\wedge-6}$; $r = .34$, $p < 10^{\wedge-7}$, respectively). Permutation testing did not find the correlation difference between field strengths to be significant (paired *t*-test, $df = 9$, $p = .14$). The difference between 7T and 3T performance (Figure 5F) was not significantly correlated with word occurrence ($r = .007$, $p = .91$). However, the two words with greatest difference in accuracy in favour of the static model occurred relatively few times, suggesting they may be outliers due to random lapses in behavioural attention. In contrast, the two words with the greatest difference in accuracy in favour of the dynamic model occurred relatively often, suggesting a reliable benefit of 7T in terms of classification accuracy. For reference, the 50 words with greatest difference in accuracy (7T versus 3T) are listed in Table 3, with the 25 most in favour of the 7T dynamic model shown in the left-hand columns and the 25 most in favour of the 3T dynamic model shown in the right-hand columns.

Finally, to address the degree to which our classification results reflect lexical semantic information (as opposed to lower level stimulus features), classification accuracy for a given word pair was correlated with their semantic similarity and their word length difference. It was found that, for the dynamic model, words more semantically similar (as represented by human behavioural ratings in a 21-feature space; see Supplemental Data for details) resulted in significantly more misclassifications

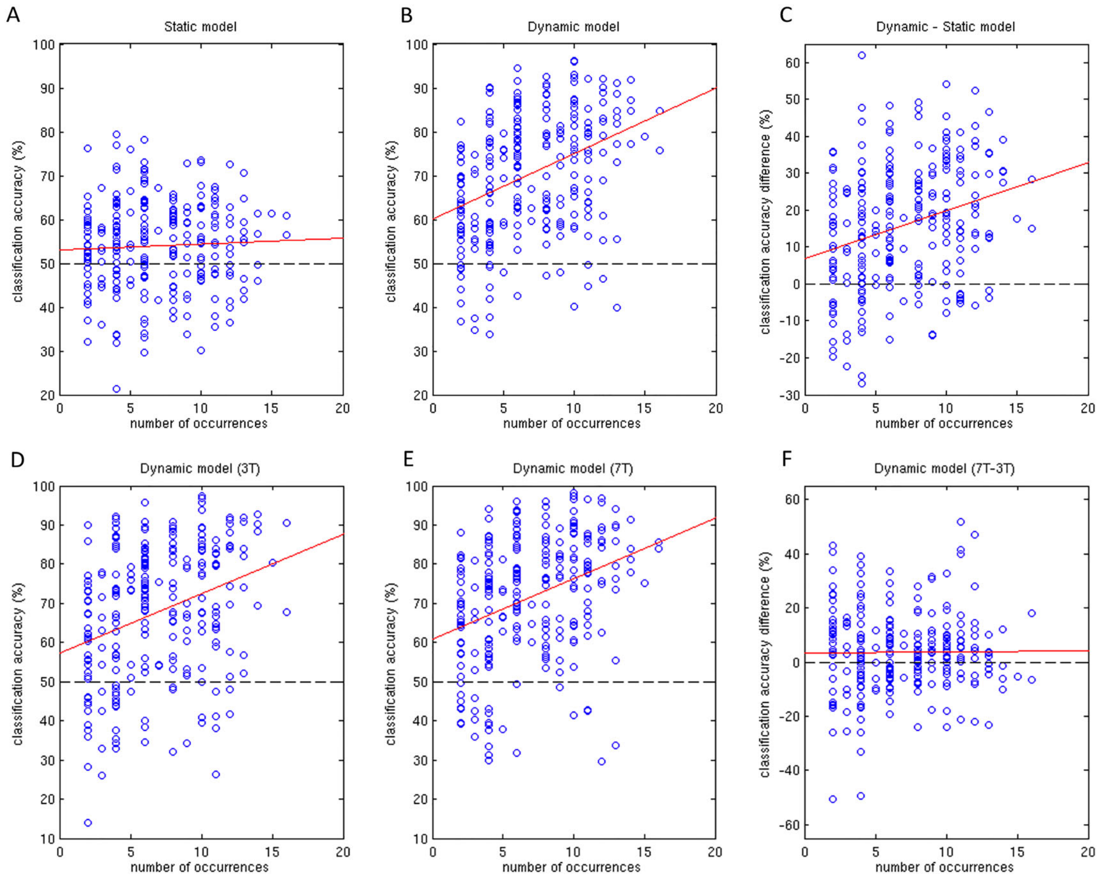

Figure 5. Classification performance versus number of word occurrences. (A) Static model. (B) Dynamic model. (C) Dynamic minus static model. (D) Dynamic model 3T only. (E) Dynamic model 7T only. (F) Dynamic model 7T minus 3T.

Table 3. Words with most extreme differences in classification accuracy: 7T versus 3T.

	7T − 3T		7T − 3T
"angry"	51.65	"shiny"	−50.40
"wealthy"	46.93	"white"	−49.27
"wrote"	43.20	"ran"	−32.97
"yellow"	41.57	"ended"	−26.03
"worker"	40.52	"store"	−25.86
"editor"	39.79	"mouse"	−25.80
"about"	38.89	"ball"	−25.63
"lonely"	35.73	"shouted"	−23.91
"fixed"	33.83	"minister"	−23.91
"flew"	33.39	"listened"	−23.81
"new"	32.68	"small"	−23.10
"water"	32.51	"red"	−21.92
"tired"	31.67	"green"	−21.40
"loud"	30.57	"ticket"	−21.09
"church"	29.10	"jury"	−19.06
"with"	28.74	"black"	−18.28
"peaceful"	28.07	"bird"	−17.99
"stayed"	27.87	"woman"	−17.78
"lawyer"	27.82	"drew"	−17.01
"to"	27.11	"met"	−16.82
"heavy"	25.34	"fence"	−16.80
"near"	25.31	"marched"	−16.63
"street"	24.87	"morning"	−15.94
"priest"	24.83	"island"	−15.84
"long"	24.81	"army"	−15.65

($r = .15$, $p < 10^{\wedge -15}$). This was not the case for the static model ($r = .001$, $p = .7$). Encouragingly, the effect of word length difference was much weaker than the semantic similarity effect with the static model ($r = .01$, $p = .003$) and not significant with the dynamic word model ($r = .003$, $p = .5$). Full description of the analysis and scatter plots may be found in the Supplemental data.

Discussion

We found that including word timing information in models of the BOLD response in a multi-modal reading experiment significantly increases the amount of explained BOLD variance (~20% on average; Figure 3) in key language- and reading-related brain regions. Furthermore, pairwise word classification accuracy, across the 261 words presented in this study, improved by 27% (from 55% to 70% accuracy; Figure 4A). These results suggest that a substantial amount of word timing information is recoverable from the BOLD response. The advantage of the dynamic model over the static model appeared to be relatively robust to changes in effective TR (Figures 3 and 4B), but was significantly negatively impacted by TRs greater than 1000 ms. The greatest *F*-value improvement was found at a TR of 1000 ms; however, this did not result in significantly better classification performance when compared to the original TR of 500 ms. This suggests that the factor of two times more time samples aided in classification enough to compensate the gain in CNR (and hence *F*-value) achieved by down-sampling. However, it is important to note that actual acquisition of data at the slower TRs would probably achieve worse results than what was achieved through the window averaged down-sampling shown here. While acquisition of slower TRs would gain CNR from additional T1 relaxation (larger Ernst angle) and lower multiband acceleration (Moeller et al., 2010; Setsompop et al., 2012; Xu et al., 2013), these effects are typically much smaller in comparison to down-sampling through averaging of time-samples (Chen et al., 2015; Feinberg et al., 2010).

Interestingly, we also found that, in terms of classification accuracy, the dynamic model was relatively more sensitive to the number of occurrences of a given word. While this could simply reflect the poorer performance of the static model in general, it also suggests that the dynamic model benefits from multiple samplings of specific words at different locations within sentences (which reduces co-linearity of the design matrix and beta estimate sensitivity to noise). This notion is consistent with the fact that words that the static model more accurately classified (relative to the dynamic model) tended to be words that occurred less frequently (Figure 5C).

In our study, two subjects underwent the exact same experiment at both 3T and 7T. We found that classification accuracy improved significantly at the higher field strength, with 65% of all words being classified with similar or better accuracy and an average accuracy improvement of 6% across all words, from 68% to 72% (Figures 4 and 5). Given that we found no significant effect of field strength on relative (Figure 2) or absolute (not shown) improvements in *F*-value (dynamic versus static model), it is likely that our protocol was physiological noise dominated due to the relatively large voxel size, 32-channel coil array, and high field strength employed (Triantafyllou, Polimeni, & Wald, 2011). Traditionally, it is thought that there is little CNR advantage going to lower spatial resolutions at 7T, due to amplification of both signal and physiological noise, and thus no net CNR gain. However, by trading spatial resolution for temporal resolution, the relative contribution of thermal noise can be increased (i.e., given the smaller Ernst angle), making relatively lower spatial resolutions potentially relevant at high field strength – especially in the context of studying dynamic processes such as sentence comprehension. Future work to explore different combinations of temporal and spatial resolutions at high field, in conjunction with physiological denoising (Griffanti et al., 2014; Tong & Frederick, 2014), will be necessary for determining the protocol with optimal word decoding performance.

Our results suggest that there is more temporal information available at higher field strength, even for the same sampling rate (500 ms) and similar CNR conditions. Much of the improvements in classification accuracy at 7T, in our study, may have come from the improved BOLD linearity (Pfeuffer et al., 2003) as well as the increased sensitivity to capillaries (Duong et al., 2003; Yacoub et al., 2001) found at higher field strengths. Given that capillaries respond faster (Hulvershorn et al., 2005) and more linearly (Zhang, Yacoub, Zhu, Ugurbil, & Chen, 2009) than large veins, it would make sense that increasing sensitivity to them by going to higher field strength would improve decodability of fine-grained temporal information. Consistent with these arguments is the fact that some of the words showing greatest classification accuracy improvement at 7T were frequently occurring words that would have more potential for non-linear, temporal interactions (Figures 5D–5F).

While the BOLD non-linearity concerns intuitively apply to quick successive repetitions of the same stimulus or word, they also apply to different but similar stimuli, which elicit similar spatial patterns of brain activity (Grill-Spector & Malach, 2001). For example, the improved linearity found at 7T could explain why similar concepts were found to have improved classification (Table 3; e.g., "priest" and "church"; "worker", "editor", and "lawyer"). Though some of the improvements found, for example, in emotional words (e.g., "angry", "lonely", "tired", "peaceful"), while arguably of similar concept dimension, could potentially be explained by improved CNR at 7T in emotion-processing regions such as the amygdala.

Conclusions

In our multi-modal sentence reading experiment, substantial word timing information was recoverable from the BOLD signal. Accurate and precise modelling of temporal events is key to optimally probing brain function as well as maximizing explained variance and decoding accuracy. Future studies with additional subjects and more advanced modelling of word semantics as well as BOLD nonlinearities will be important to furthering the results presented here. With proper modelling of BOLD non-linearities and the shorter TRs achievable with slice-accelerated fMRI, it may be possible to uncover additional temporal information at even finer temporal resolutions (Ogawa et al., 2000).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Air Force Research Laboratory [grant number FA8650-13-C-7356]; Biotechnology Research Center (BTRC) [grant number P41 EB015894] from NIBIB.

References

Berken, J. A., Gracco, V. L., Chen, J. K., Watkins, K. E., Baum, S., Callahan, M., & Klein, D. (2015). Neural activation in speech production and reading aloud in native and non-native languages. *Neuroimage*, *112*, 208–217. doi:10.1016/j.neuroimage.2015.03.016

Chang, W. T., Nummenmaa, A., Witzel, T., Ahveninen, J., Huang, S., Tsai, K. W., … Lin, F. H. (2013). Whole-head rapid fMRI acquisition using echo-shifted magnetic resonance inverse imaging. *Neuroimage*, *78*, 325–338. doi:10.1016/j.neuroimage.2013.03.040

Chen, L., Vu, A. T., Xu, J., Moeller, S., Ugurbil, K., Yacoub, E., & Feinberg, D. A. (2015). Evaluation of highly accelerated simultaneous multi-slice EPI for fMRI. *Neuroimage*, *104*, 452–459. doi:10.1016/j.neuroimage.2014.10.027

Constable, R. T., & Spencer, D. D. (2001). Repetition time in echo planar functional MRI. *Magnetic Resonance in Medicine*, *46*(4), 748–755.

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience*, *34*(1), 332–338. doi:10.1523/JNEUROSCI.1302-13.2014

Dapretto, M., and Bookheimer, S. Y. (1999). Form and content: Dissociating syntax and semantics in sentence comprehension. *Neuron*, *24*, 427–432.

Duong, T. Q., Yacoub, E., Adriany, G., Hu, X., Ugurbil, K., & Kim, S. G. (2003). Microvascular BOLD contribution at 4 and 7 T in the human brain: Gradient-echo and spin-echo fMRI with suppression of blood effects. *Magnetic Resonance in Medicine*, *49*(6), 1019–1027. doi:10.1002/mrm.10472

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Gunther, M., … Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. *PLoS One*, *5*(12), e15710. doi:10.1371/journal.pone.0015710

Flegal, K. E., Marin-Gutierrez, A., Ragland, J. D., & Ranganath, C. (2014). Brain mechanisms of successful recognition through retrieval of semantic context. *Journal of Cognitive Neuroscience*, *26*(8), 1694–1704. doi:10.1162/jocn_a_00587

Freedman, D. (2009). *Statistical models: Theory and practice*. Cambridge, NY: Cambridge University Press.

Glasgow, K., Roos, M., Haufler, A., Chevillet, M., Wolmetz, M. (2016). Evaluating semantic models with word-sentence relatedness. arXiv:1603.07253 [cs.CL].

Griffanti, L., Salimi-Khorshidi, G., Bechmann, C. F., Aurbach, E. J., Douaud, G., Sexton, C. E., … Smith, S. M. (2014). ICA-based artifact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage*, *95*, 232–247.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, *107*(1–3), 293–321.

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42. doi:10.1016/j.neuropsychologia.2015.04.018

Hulvershorn, J., Bloy, L., Gualtieri, E. E., Leigh, J. S., & Elliott, M. A. (2005). Spatial sensitivity and temporal response of spin echo and gradient echo bold contrast at 3 T using peak hemodynamic activation time. *Neuroimage*, *24*(1), 216–223. doi:10.1016/j.neuroimage.2004.09.033

Kay, K. N., David, S. V., Prenger, R. J., Hansen, K. A., & Gallant, J. L. (2008). Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Human Brain Mapping*, *29*(2), 142–156. doi:10.1002/hbm.20379

Kay, K. (2014). Getcanonicalhrf. Retrieved from http://kendrickkay.net/analyzePRF/doc/analyzePRF/utilities/getcanonicalhrf.html

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi:10.1038/nature06713

Lee, A. T., Glover, G. H., & Meyer, C. H. (1995). Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging. *Magnetic Resonance in Medicine*, *33*(6), 745–754.

Meltzer-Asscher, A., Mack, J. E., Barbieri, E., & Thompson, C. K. (2015). How the brain processes different dimensions of argument structure complexity: Evidence from fMRI. *Brain and Language*, *142*, 65–75. doi:10.1016/j.bandl.2014.12.005

Menon, R. S., Luknowsky, D. C., & Gati, J. S. (1998). Mental chronometry using latency-resolved functional MRI. *Proceedings of the National Academy of Sciences*, *95*(18), 10902–10907.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. doi:10.1126/science.1152876

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Ugurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*(5), 1144–1153. doi:10.1002/mrm.22361

Moisala, M., Salmela, V., Salo, E., Carlson, S., Vuontela, V., Salonen, O., & Alho, K. (2015). Brain activity during divided and selective attention to auditory and visual sentence comprehension tasks. *Frontiers in Human Neuroscience*, *9*, 3567. doi:10.3389/fnhum.2015.00086

Mukamel, R., Harel, M., Hendler, T., & Malach, R. (2004). Enhanced temporal non-linearities in human object-related occipito-temporal cortex. *Cerebral Cortex*, *14*(5), 575–585. doi:10.1093/cercor/bhh019

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*, *21*(19), 1641–1646. doi:10.1016/j.cub.2011.08.031

Ogawa, S., Lee, T. M., Stepnoski, R., Chen, W., Zhu, X. H., & Ugurbil, K. (2000). An approach to probe some neural systems interaction by functional MRI at neural time scale down to milliseconds. *Proceedings of the National Academy of Sciences*, *97*(20), 11026–11031.

Pfeuffer, J., McCullough, J. C., Van de Moortele, P. F., Ugurbil, K., & Hu, X. (2003). Spatial dependence of the nonlinear BOLD response at short stimulus duration. *Neuroimage*, *18*(4), 990–1000.

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., & Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, *67*(5), 1210–1224. doi:10.1002/mrm.23097

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., … Consortium, W. U.-M. H. (2013). Resting-state fMRI in the Human Connectome Project. *Neuroimage*, *80*, 144–168. doi:10.1016/j.neuroimage.2013.05.039

Tong, Y., Frederick, B. D. (2014). Studying the spatial distribution of physiological effects on BOLD signals using ultrafast fMRI. *Frontiers in Human Neuroscience*, *196*(8). doi:10.3389/fnhum.2014.00196

Triantafyllou, C., Polimeni, J. R., & Wald, L. L. (2011). Physiological noise and signal-to-noise ratio in fMRI with multi-channel array coils. *Neuroimage*, *55*(2), 597–606. doi:10.1016/j.neuroimage.2010.11.084

Ugurbil, K. (2014). Magnetic resonance imaging at ultrahigh fields. *IEEE Transactions on Biomedical Engineering*, *61*(5), 1364–1379. doi:10.1109/TBME.2014.2313619

Ugurbil, K., Xu, J., Auerbach, E. J., Moeller, S., Vu, A. T., Duarte-Carvajalino, J. M., … Consortium, W. U.-M. H. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage*, *80*, 80–104. doi:10.1016/j.neuroimage.2013.05.012S1053-8119(13)00506-5[pii]

Vaughan, J. T., Garwood, M., Collins, C. M., Liu, W., DelaBarre, L., Adriany, G., … Ugurbil, K. (2001). 7T vs. 4T: RF power, homogeneity, and signal-to-noise comparison in head images. *Magnetic Resonance in Medicine*, *46*(1), 24–30. doi:10.1002/mrm.1156[pii]

Vitello, S., Warren, J. E., Devlin, J. T., & Rodd, J. M. (2014). Roles of frontal and temporal regions in reinterpreting semantically ambiguous sentences. *Frontiers in Human Neuroscience*, *8*, 1124. doi:10.3389/fnhum.2014.00530

Wu, C. Y., Vissiennon, K., Friederici, A. D., and Brauer, J. (2016). Preschoolers' brains rely on semantic cues prior to the mastery of syntax during sentence comprehension. *NeuroImage*, 126, 256–66. doi:10.1016/j.neuroimage.2015.10.036

Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. A., … Ugurbil, K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3T. *Neuroimage*, *83*, 991–1001. doi:10.1016/j.neuroimage.2013.07.055

Yacoub, E., Shmuel, A., Pfeuffer, J., Van De Moortele, P. F., Adriany, G., Andersen, P., … Hu, X. (2001). Imaging brain function in humans at 7 Tesla. *Magnetic Resonance in Medicine*, *45*, 588–594.

Zhang, N., Yacoub, E., Zhu, X. H., Ugurbil, K., & Chen, W. (2009). Linearity of blood-oxygenation-level dependent signal at microvasculature. *Neuroimage*, *48*(2), 313–318. doi:10.1016/j.neuroimage.2009.06.071

Zhang, N., Zhu, X. H., & Chen, W. (2008). Investigating the source of BOLD nonlinearity in human visual cortex in response to paired visual stimuli. *Neuroimage*, *43*(2), 204–212. doi:10.1016/j.neuroimage.2008.06.033