Published in final edited form as:

Neuroimage. 2009 October 1; 47(4): 1649–1658. doi:10.1016/j.neuroimage.2009.06.001.

Hemodynamic Nonlinearities Affect BOLD fMRI Response Timing and Amplitude

Jacco A de Zwart, **Peter van Gelderen**, **J Martijn Jansma**, **Masaki Fukunaga**, **Marta Bianciardi**, and **Jeff H Duyn**

Advanced MRI section, LFMI, NINDS, National Institutes of Health, Bethesda, MD, USA

Abstract

The interpretation of functional Magnetic Resonance Imaging (fMRI) studies based on Blood Oxygen-Level Dependent (BOLD) contrast generally relies on the assumption of a linear relationship between evoked neuronal activity and fMRI response. While nonlinearities in this relationship have been suggested by a number of studies, it remains unclear to what extent they relate to the neurovascular response and are therefore inherent to BOLD-fMRI. Full characterization of potential vascular nonlinearities is required for accurate inferences about the neuronal system under study. To investigate the extent of vascular nonlinearities, evoked activity was studied in humans with BOLD-fMRI (n=28) and Magnetoencephalography (MEG) (n=5). Brief (600-800 ms) rapidly repeated (1 Hz) visual stimuli were delivered using a stimulation paradigm that minimized neuronal nonlinearities. Nevertheless, BOLD-fMRI experiments showed substantial remaining nonlinearities. The smallest stimulus separation (200-400 ms) resulted in significant response broadening (15-20% amplitude decrease; 10-12% latency increase; 6-14% duration increase) with respect to a linear prediction. The substantial slowing and widening of the response in the presence of preceding stimuli suggests a vascular rather than neuronal origin to the observed non-linearity. This was confirmed by the MEG data, which showed no significant neuro-electric nonlinear interactions between stimuli as little as 200 ms apart. The presence of substantial vascular nonlinearities has important implications for rapid event-related studies by fMRI and other imaging modalities that infer neuronal activity from hemodynamic parameters.

Introduction

The interpretation of blood oxygen-level dependent (BOLD) (Ogawa et al., 1990) and perfusion-based (Kim, 1995; Kwong et al., 1992) functional Magnetic Resonance Imaging (fMRI) data is dependent on the relationship between neuronal activity and the elicited hemodynamic response. Although this relationship is commonly implied to be linear, this may not be generally true. Significant deviations from linearity may occur between the BOLD response and both the level and duration of neuronal activity. For example, several studies have demonstrated a significant nonlinearity between the stimulus duration and the BOLD response integral, in particular for brief (1-4 s) stimuli (Birn and Bandettini, 2005; Birn et al., 2001; Boynton et al., 1996; Dale and Buckner, 1997; de Zwart et al., 2006; Friston et al., 1998; Glover, 1999; Ogawa et al., 2000; Vazquez and Noll, 1998; Zhang et al., 2008a; Zhang et al.,

Correspondence: Jacco de Zwart LFMI/NINDS/NIH Bldg 10, Rm B1D-728 9000 Rockville Pk - MSC 1065 Bethesda, MD 20892-1065 USA Phone: +1-301-594-7315 Fax: +1-301-480-2558 Email: Jacco.deZwart@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2008b). In these studies, repeated stimuli led to successively smaller responses. The quantitative findings of these studies vary substantially, in part because some of the reported nonlinearities can be attributed to neuronal effects (e.g. repetition effects (Grill-Spector et al., 2006)), either introduced by the stimulus design or inherent to the brain region under study. In order to truly investigate linearity of the BOLD response one has to assure linearity of the underlying neuronal response. Uncertainty about the origin of nonlinearities prohibits generalization of the findings and the characterization of the BOLD impulse response (IR), both of which are important for interpretation of BOLD fMRI data.

In the human visual system, apparent BOLD-fMRI response nonlinearities may arise from a transient neuronal response that generally occurs at the onset and ending of a stimulus and may lead to overestimation of hemodynamic nonlinearities. Initial studies that minimized this confounding effect found a small but persistent remaining BOLD nonlinearity in the human visual system (Boynton et al., 1996; Dale and Buckner, 1997; Kellman et al., 2003), albeit without establishing its origin.

The ability to distinguish between neuronal and vascular nonlinearities is crucial for fMRI studies (in particular event-related), and has implications for other imaging modalities that infer neuronal activity from the hemodynamic response, such as Positron Emission Tomography (PET) or Near-Infrared Spectroscopy (NIRS). While neuronal nonlinearities convey information about stimulus interactions in the brain region being studied, vascular nonlinearities are a byproduct of the BOLD contrast mechanism and thus an artifact of the measurement method. If not properly taken into account such vascular nonlinearities confound the interpretation of underlying neuronal events.

The purpose of this study was therefore to investigate the extent of hemodynamic nonlinearities in the BOLD signal and to characterize them based on their temporal evolution. We focused on nonlinearities related to stimulus duration, although substantial BOLD nonlinearities introduced by the strength of the stimulus may exist as well (e.g. (Boynton et al., 1996; Vazquez and Noll, 1998)). In order to accomplish this we delivered brief visual stimuli in rapid succession using an event-related paradigm based on the m-sequence probe method (Benardete and Victor, 1994; Sutter, 1987), which allows nonlinear system identification and characterization with high efficiency and temporal resolution (Buracas and Boynton, 2002). Minimization of confounding neuronal repetition effects was achieved by inserting a brief gap between the stimuli (See *Rationale*, below) (Kellman et al., 2003). Magnetoencephalography (MEG) experiments with these paradigms were performed to verify the absence of neuronal interactions between the individual stimuli in the paradigm.

Materials and Methods

Supplement A contains a limited amount of additional Materials and Methods.

Rationale

The rationale behind the experimental design of this study is that the various processes which contribute to nonlinearities in the BOLD response have distinctly different timescales, facilitating their separation and identification. The use of event-related fMRI with inter-stimulus gaps (stimulus separation) of appropriate duration allows reduction of short-lived neuronal nonlinearities, while minimally affecting nonlinearities originating from the hemodynamic response. The remaining nonlinearities observed in such gapped experiments can be characterized by studying the temporal dynamics of their effect on the IR, which can provide further clues about their origin.

Neuronal effects in the early areas of the visual system are relatively fast, having electrical imprints with timescales in the 10- to 100-ms range (Baseler et al., 1994; Schmolesky et al., 1998). Neuronal interaction effects between stimuli can therefore be minimized by the introduction of an inter-stimulus gap of a few hundred milliseconds (Kellman et al., 2003). On the other hand, nonlinearities originating from hemodynamic effects and the temporal evolution of vascular deoxyhemoglobin changes are less affected by short inter-stimulus gaps as they are substantially slower, with time-constants on the scale of seconds (Berwick et al., 2002; Rudin et al., 1997). Since the neuronal IR function is short, neuronal nonlinearities are not likely to engender changes in BOLD IR width or latency on a scale of seconds.

Vascular nonlinearities, on the other hand, are likely to affect BOLD IR shape since they affect the transit of (de)oxygenated blood through the vasculature, and possibly the temporal evolution of the deoxyhemoglobin concentration. In earlier work we demonstrated that the draining vasculature substantially affects latency and duration of BOLD IR (de Zwart et al., 2005). These results are supported by monocrystalline iron oxide nanocolloid (MION) based fMRI data (Silva et al., 2007). The fast component of the biphasic MION-response has been attributed to the arteriolar compartment (Lee et al., 2001) and shows a much more rapid response than BOLD fMRI data, which predominantly reflects the capillary plus venous domain.

The neurovascular control mechanism is assumed to operate on an intermediate timescale (hundreds of ms to 1-2 s). The upper estimate for this timescale is derived from the CBV response measured in rats using MION based fMRI (Silva et al., 2007) and optical imaging methods (Martindale et al., 2003).

In order to study BOLD nonlinearities with high sensitivity we used the m-sequence probe method (Benardete and Victor, 1994; Sutter, 1987), which allows accurate characterization of linear and nonlinear response components in a single experiment (see Appendix A). In addition, actual response time-courses for various stimulus conditions can be derived from these data. Apart from being more efficient than repeat experiments employing varying inter-stimulus intervals, the m-sequence method thus also eliminates experimental confounds such as attention effects.

Strong, visual stimuli of less than 1-s duration with a minimal stimulus separation (gap) of 200 ms were used to obtain a robust BOLD IR estimate. Both gap duration and stimulus intensity (luminosity) were varied to provide additional indication of the origin of observed nonlinearities. IR estimates were derived from the acquired fMRI data on a voxel-by-voxel basis (see Appendix A and (Kellman et al., 2003)). The temporal characteristics of the observed response to these stimuli in BOLD fMRI, and their dependence on stimulus separation (lag), allow characterization of the longer-timescale nonlinearities observed in BOLD fMRI. The substantial effect of lag on BOLD IR timing (stimulus width and latency) provides evidence of the non-neuronal (non-electrical) origin of these nonlinearities. As supporting evidence, MEG data were acquired to investigate the extent of residual neuronal interactions between stimuli during employment of the proposed stimulus paradigms.

Stimulus Design

The stimulation paradigm was based on a 255-bin (255-trial) m-sequence (Benardete and Victor, 1994; Sutter, 1987). The duration of each bin was 1 s, identical to the acquisition interval between MRI volumes. The first 45 m-sequence bins were repeated to fill 300 bins, and an inverse-repeat (repeat with polarity inverse of the bins) of the m-sequence was used (Kellman et al., 2003), resulting in an overall paradigm duration of 600 s. ‘On’ trials (stimuli) consisted of the display of a full-field radial checkerboard for 800 ms, contrast reversing every 66.7 ms (corresponding to a 7.5 Hz stimulus frequency). During the remaining 200 ms of the 1-s ‘on’ trial (the ‘gap’) a uniform grey disk was shown to assure a minimal stimulus separation of 200 ms. Throughout ‘off’ trials the same uniform grey disk was shown as during the gap. The average luminance of this field was equal to that of the checkerboard stimulus so that nonlinearities related to pupil dilation and flicker were avoided (Kellman et al., 2003). A small fixation dot in the center of the images was used to focus the volunteers' attention. The dot alternated between light and dark grey approximately once every 40 s. Volunteers were asked to indicate these changes with a button press, which was recorded to monitor attention.

Two alternative m-sequence stimulus paradigms were also used, one in which the stimulus duration was reduced to 600 ms, in combination with 400 ms gap, and another in which the 200-ms gap paradigm was used with 50%-reduced stimulus contrast. These three experiments were performed because one would expect the responses to be substantially different if there were (residual) neuronal nonlinearities at play. Since the gap increase from 200 to 400 ms is substantial on a neuronal time scale, one would expect a reduced contribution of nonlinearities in the 400-ms gap experiment compared to the 200-ms gap experiment if the nonlinearities were neuronal in origin. Similarly, if saturation effects would contribute substantially to the data, one would expect differences between the low-contrast and the full-contrast experiments. If the relative contribution of nonlinearities, as well as their temporal extent, is similar in the three experiment types, this is strong evidence of a non-neuronal origin of these effects.

Each volunteer was scanned with 2 out of 3 paradigms (with the exception of the first volunteer, which was scanned twice with the 200-ms gap, full-contrast paradigm, with only the first run being used for further analysis). In total, twenty full-contrast 200-ms gap datasets (referred to as *‘200’*), nineteen 400-ms gap datasets (called *‘400’*) and sixteen 200-ms low-contrast datasets (*‘200lc”*) were acquired.

A 5-min block paradigm with identical scan parameters was acquired after the m-sequence runs and used to select a functional region of interest (ROI). The block paradigm consisted of 5 blocks, each comprised of 30 s grey-disk rest stimulus followed by 30 s of the full-contrast checkerboard stimulus used in the m-sequence paradigm. A center dot fixation task similar to the one used during m-sequence scans was used.

In order to check if the observed effects were not affected by, or a result of, the specific m-sequence used, data from a previous very similar 200-ms gap study (Supplement B), employing a different m-sequence in an otherwise very similar set of fMRI experiments, were reanalyzed (referred to as *‘200old’*). The various interaction terms manifest themselves at completely different positions in the correlograms, both with respect to the primary (linear) response as well as with respect to each other (see *MRI Data Analysis* below, and Supplement B).

MRI Data Acquisition and Image Reconstruction

Twenty-eight studies were performed on 17 volunteers (8 m/9 f, average age 32.9 y), who underwent fMRI of the visual system on a General Electric 3 T scanner (GE Healthcare, Waukesha, WI, USA) equipped with a 16-channel head coil array (Nova Medical, Wilmington, MA, USA) (Bodurka et al., 2004; de Zwart et al., 2004). Volunteers gave informed consent to an IRB-approved protocol. Some m-sequence datasets were excluded on the basis of low SNR in the observed primary response amplitude (see Supplement A). As a result, eighteen *200* datasets, thirteen *400* and thirteen *200lc* datasets were used for further analysis.

A gradient-echo echo-planar imaging (EPI) sequence was employed. Ten slices were acquired parallel to and enveloping the calcarine fissure. Nominal spatial resolution was 1.6×1.6×2.0 mm^3, echo time 44 ms, and repetition time 1 s. MRI image reconstruction was performed as described earlier (de Zwart et al., 2002). Magnitude images from all scans were registered to the last image of the block paradigm scan for that volunteer using C-code based on software developed by Thévenaz et al. (Thévenaz et al., 1995).

MRI Data Analysis

Analysis was geared towards the extraction of the primary (first-order) and several second-order (nonlinear) response kernels from the data. Actual responses to an isolated stimulus or a stimulus preceded by other stimuli can subsequently be derived by combining these kernels (see Appendix A). This way, the response to a stimulus in isolation was obtained, as well as responses to the second of a pair of stimuli for two stimulus separation intervals. Shape analysis of these responses informed about the origin of observed nonlinearities.

All processing with exception of image registration was done in IDL (ITT Visual Information Solutions, Boulder, CO, USA). The block paradigm data were analyzed as described earlier (Waldvogel et al., 2000), assuming a hemodynamic response function with a time-to-peak (TTP) and full-width at half-maximum (FWHM) of 3.5 s. Voxels exceeding a threshold (t=5) in the resulting tmap were selected to derive a functional ROI for the analysis of m-sequence data. Temporal signal stability, expressed as the relative temporal SD (SD_t) was computed based on the residual signal in the block paradigm data after model fitting. SD_t was computed on a voxel-by-voxel basis and averaged for all voxels in the functional ROI. A block of 255 volumes was taken from each half of the m-sequence data (volumes 40-294 and 340-594, respectively).

M-sequence correlograms (covariance between signal timecourse and m-sequence stimulus) were computed on a voxel-by-voxel basis. The correlograms for the first m-sequence and its inverse-repeat were added to determine the primary response (referred to as ‘*pri*’), and subtracted to obtain the second-order responses (see (Kellman et al., 2003)). The first three second-order kernels (responses) were investigated, namely for lags 1 (‘*sec1*’), lag 2 (‘*sec2*’) and lag 3 (‘*sec3*’), as well as the first third order interaction for lag 1+2 (‘*tri12*’). These lags refer to the degree of separation of the current stimulus from the preceding stimulus, which causes the interaction. E.g., *sec1* describes the nonlinear component of the response that occurs when a stimulus is present in the bin directly preceding the current stimulus (having an onset time 1 second before the current stimulus and thus a stimulus separation of 200 ms or 400 ms depending on the experiment type). *Tri12* describes to what extent the nonlinear effect from the two preceding pulses (*sec1* and *sec2*) changes when both are present. (If the response to three consecutive pulses is different from the cumulative effect of *pri+sec1+sec2* then *tri12* is nonzero.) Characteristics of the m-sequence lead to a separation of these interactions in the correlogram, which permitted deriving them simultaneously from a single experiment (Appendix A). For each volunteer, the m-sequence responses were averaged within the functional ROI.

If the system is nonlinear, the IR to a stimulus depends on the events preceding it (presence or absence of a preceding stimulus; the time elapsed since the preceding stimulus occurred). Weighted combinations of the first- and higher-order kernels can be used to derive these different IRs. Here, several different IRs are expected, since significantly non-zero higher-order kernels were found. These various IRs can all be computed using the correct combination of the kernels (see Appendix A). The response to an isolated single-bin stimulus derived this way is referred to as ‘*respS*’. The responses to a single-bin stimulus that was preceded by an identical stimulus one or two bins earlier were also computed (referred to as ‘*resp1*’ and ‘*resp2*’, respectively). In the case of *resp1*, the preceding stimulus was separated from the current by a single gap, 0.2 s or 0.4 s in our experiments. Accordingly, the stimulus separation for *resp2* was 1.2 s for the 200-ms gap paradigms and 1.4 s for the 400-ms gap paradigm. Only the second-order kernels (responses) for lags 1 (‘*sec1*’) and lag 2 (‘*sec2*’) were taken into account when deriving *respS*, *resp1* and *resp2* since all other higher-order responses, including sec3 and tri12 were found to be insignificant compared to baseline fluctuations in the correlograms and explained less than 1% of the signal variance (with the only exception being *sec3* for the low-contrast paradigm, which explained 1.3% of the variance).

To quantitatively characterize the nonlinearity of these responses, both TTP (as a measure for latency) and FWHM for the different IRs were computed. Also, amplitude and surface area (over the first 10 s following stimulus onset) of *resp1* and *resp2* relative to *respS* were calculated. In order to evaluate whether the changes in the BOLD impulse response function in the presence of preceding stimuli can be explained by a dispersive mechanism, the computed average *respS* curves were stretched (in time) and scaled (amplitude) to determine the best possible fit to the *resp1* and *resp2* responses that were derived from the same scan on the same volunteer. The difference between the first 10 s of the *resp1* and *resp2* response and the deformed *respS* response was then computed. Of this difference, referred to as the residual, the root-of-summed-squares (RSS) was computed as a fraction of the maximum amplitude and used as a measure of goodness-of-fit.

MEG Experiments

Five normal volunteers (4 m, 1 f, average age 33.8 y, two of which also participated in the fMRI experiments) underwent MEG under the same protocol and equal stimulus paradigms as MRI. All but one volunteer successfully completed the MEG exam (see Supplement A). MEG detects the magnetic field associated with the current flow that results from axonal depolarization on neuronal activation (Hamalainen et al., 1993). Since the phase of the detected signal depends on the orientation of the current dipole with respect to the SQUID detector, the power and not the sign of the detected signal change is a measure of the underlying level of activation.

Experiments were performed on a 275-channel CTF (Coquitlam, BC, Canada) MEG scanner running software release 5.4.0. Volunteers were in seated position. Data were acquired at 600 Hz with continuous head position monitoring (3 channels, at nasion and two preauricular points). Two 10-min m-sequence runs (200- and 400-ms gap) were performed in random order, followed by a 5-min block paradigm scan. Signal from an optical sensor placed in the projector beam was sampled using a supplementary ADC channel of the MEG. Since it was not found to have a significant effect on the observed nonlinearities (see *BOLD IR nonlinearities* below), the stimulus luminance level was not calibrated to that during fMRI, only full-contrast stimuli were used.

MEG Data Analysis

MEG data were processed in IDL on a channel-by-channel basis. No source localization was performed since our primary interest concerned temporal signal characteristics. In the block paradigm scan, the 10 channels that correlated most strongly with the paradigm were selected. After band-pass filtering (2-30 Hz), m-sequence analysis was performed on sets of samples with an identical acquisition time relative to the start of each bin (e.g. m-sequence analysis on every first sample since bin onset), thus retaining the high temporal resolution that MEG provides. The phase-sensitive average of the 10 channels selected using the block paradigm scan was then computed. From these averages, the *pri*, *sec1*, *sec2*, *sec3* and *tri12* kernels for each volunteer were extracted. Similarly to fMRI analysis, the response to an isolated stimulus (*respS*) and responses in the presence of preceding stimuli (*resp1* and *resp2*) were subsequently computed by only taking into account *sec1* and *sec2*.

Although the m-sequence response-kernels for the different volunteers looked very similar, the response was bipolar and the response timing relative to stimulus onset was found to be volunteer-dependent. Straightforward averaging of the observed m-sequence response kernels or the derived response functions was therefore not feasible. The bipolarity of the measured response also complicated the assessment of the amount of nonlinear contribution. Straightforward comparison of the mean in the pri with the mean in the sec kernels would lead to a severe underestimation of the signal in pri. Computing the power of the signal based on the magnitude of the response is more appropriate, however the small amount of signal (if any) in the higher-order kernels would lead to rectified noise related artifacts. To overcome this, the power in resp1 and respS were computed. The difference in power between resp1 and respS was subsequently compared to the power in respS to assess the extent of nonlinear contribution in the MEG data.

Results

MEG confirms absence of neuronal nonlinearities

MEG experiments confirmed that contribution of neuronal nonlinearities was insignificant when the proposed stimulus paradigm was employed. An example of the primary and second order response kernels for a 200-ms gap experiment, averaged over 10 MEG channels for one of the volunteers, is shown in Fig. 1a. In all volunteers the 10 channels selected based on the block paradigm were located in the lateral occipital and posterior temporal cortices. Selected channels were grouped in either one or two (one in each hemisphere) clusters for all volunteers, albeit that clusters for different volunteers only partially overlapped. A robust primary response kernel was observed, showing strong correlation with the visual stimulus (cf. optical sensor output). On the other hand, no significant interaction between stimuli (nonlinearity) can be distinguished in any of the second-order kernels. The response to an isolated stimulus (*respS*) and responses for a stimulus directly preceded by another (*resp1*) were derived from these response-kernels (Fig. 1b). For the first 1-s interval of the response, the magnitude of *respS* was subtracted from the magnitude of *resp1*. The mean of this difference was subsequently divided by the mean magnitude of *respS* in the same 1-s interval to get a measure of the change in energy in the response. The result indicated that the differences between these responses were non-significant: For the 200-ms gap experiment a change of 7.5±8.9% (mean ± standard error over volunteers) was found, whereas 3.2±5.9% was found for the 400-ms gap experiment. As a second measure of a nonlinear effect, the *resp1* response was fitted with the *respS* response to determine the response amplitude change due to the presence of a preceding stimulus. Non-significant amplitude changes of −1.7±6.6% (mean ± standard error over volunteers) and −3.5±4.9% were found for 200-ms gap and 400-ms gap experiments, respectively. This confirmed the effectiveness of the stimulation protocol in minimizing neuronal nonlinearities and in facilitating the study of the effects of other nonlinear contributions.

fMRI data quality

All subjects completed the exam(s). Behavioral data (button presses in response to perceived changes in center dot brightness) were consistent throughout the runs; no volunteers were excluded based on behavioral data. Averaged over volunteers, the relative temporal standard deviation (SD) of residual signals in the functional ROI was 1.9±0.3% of the baseline signal. The average image signal-to-noise ratio (SNR) in the functional ROI was 64, corresponding to a relative SD of 1.6±0.3%. The difference between these values reflects physiological noise contributions, including those from the cardiac and respiratory cycles.

Good quality fMRI responses were obtained in visual cortex areas of all subjects for the stimulus contrasts and gap durations studied. Analysis of m-sequence data resulted in response kernels from which the response to a single stimulus was derived and nonlinear interactions were quantified. Fig. 2 shows an example of the first order response kernel, which is the average response to an 800-ms long stimulus in the experiment. The BOLD response can be distinguished as early as 2 s after stimulus onset. Consistent with earlier work (Birn and Bandettini, 2005;Birn et al., 2001;Glover, 1999;Vazquez and Noll, 1998), it peaks at approximately 4 s and is no longer distinguishable 9 s after stimulus onset.

BOLD IR nonlinearities

In addition to the first order response kernel, two second order response kernels (*sec1* and *sec2*) with progressively lower amplitudes were observed (Fig. 3). A third second order kernel (*sec3*) and a third order kernel (*tri12*) reached the limit of detectability for certain lags for some of the experiment types and are also shown in Fig. 3. No other kernels were significant. The presence of higher-order kernels indicates that there are significant nonlinearities (interactions between individual sub-second stimuli in the paradigm). From the observed first- and second order response kernels, we were able to estimate the response to single events in isolation and in the presence of preceding stimuli, and to quantify their differences. Results (Fig. 4 and Tab. 1) suggest that preceding stimuli have an effect that lowers, delays, and broadens the IR, and that this effect increases with decreasing stimulus separation. These substantial effects were similar for all stimulus types studied and resulted in significant response dispersion, which was strongest for the smallest stimulus separation (200-400 ms): Response amplitude decreased 15-20%, latency increased 7-12% and response duration increased 6-15%. Preceding stimuli did not significantly affect response integral. Furthermore, no significant effect of stimulus gap duration or luminance level was observed. The dispersive character (increased response delay and broadening) of the observed interaction was confirmed by fitting the estimated response of stimuli affected by a preceding stimulus with scaled (in time and amplitude) versions of the response to an isolated stimulus (Tab 2). This led to a residue that was not significantly above noise level. If *respS* was not stretched and scaled but directly subtracted from *resp1*, the observed residue as a fraction of the maximum response amplitude was 0.23 (on average for the three stimulus types), substantially higher than the expected minimal residual fraction of 0.09 (on average over the three stimulus types) resulting from noise in the experiment as determined from the fluctuation level over lags 20-39. When fitting with a stretched and scaled version of *respS*, the observed residual fractions were not significantly different from this expected value (Tab. 2). These results indicate that the responses in the presence of a preceding stimulus could be well described by a stretched and scaled version of the isolated stimulus response.

Differences between the responses (Fig. 4) were used to compute the maximal amplitude of the nonlinear contribution as a fraction of the amplitude of the single stimulus response. For the comparison of *respS* and *resp1*, this yielded fractions of 0.29±0.01 (29%) for the *200* experiment, 0.23±0.03 (23%) for *400*, and 0.31±0.02 (31%) for *200lc* (average ± standard error over volunteers). For the difference between *respS* and *resp2*, the nonlinear contributions were 17%, 21% and 14%, respectively, for *200*, *400* and *200lc* (all with 2% standard error). For the 200-ms gap data from the previous study, *200old* (Supplement B), 0.29±0.03 (29%) and 0.16 ±0.03 (16%) were found for comparison of *respS* with *resp1* and *resp2*, respectively.

Discussion

Nonlinearities of BOLD IR were investigated using a method that minimized confounding neuronal effects by focusing gaze and by introducing a minimal stimulus separation of at least 200 ms (Kellman et al., 2003), randomizing the presentation of these stimuli (Clark et al., 1998), ensuring their equiluminance (Boynton et al., 1996; Vazquez and Noll, 1998), thus minimizing attention effects.

This allowed investigation of the source of remaining nonlinearities by analyzing their temporal behavior. MEG experiments confirmed that electrical-neuronal nonlinear interactions between stimuli in these paradigms were indeed minimal. On the other hand, functional MRI results showed that:

1) A significant nonlinearity remains in BOLD data;

2) This residual nonlinearity causes substantial delay and broadening of the observed hemodynamic response when a closely preceding stimulus is present;

3) It reduces response amplitude without significantly affecting response integral;

4) The effect diminishes with increasing stimulus separation, dissipating when stimuli are more than 2-3 s apart. This can be seen from the decreased difference between *respS* and *resp2* when compared to the difference between *respS* and *resp1* (Fig. 4 and Table 1), as well as from the decreasing amplitude of the sec kernels (*sec1*>*sec2*>*sec3*, see Fig. 3).

These findings suggest that the residual BOLD-fMRI nonlinearities are vascular in origin, and that their characteristics are inconsistent with a neuronal origin, as we will elaborate on below.

The small magnitude of the BOLD IR nonlinearities found in this study (e.g. see Figure 3 and Table 1) is in line with previous work with comparable stimulation protocols (Dale and Buckner, 1997;Kellman et al., 2003). For example, the extent of nonlinearity in the data shown in Figure 4b in Dale and Buckner (Dale and Buckner, 1997), which used 2-s stimulus separation, resembles the nonlinear contributions found this work (c.f. Figure 4). Furthermore, the nonlinearities found here were much smaller than in previous work in which no efforts were made to ensure suppression of neuronal nonlinearities (Birn and Bandettini, 2005;Birn et al., 2001;Glover, 1999;Ogawa et al., 2000;Vazquez and Noll, 1998). E.g., Birn and Bandettini found that the observed BOLD responses to a 1-s stimulus were often 2-3 times larger than a linear prediction (Birn and Bandettini, 2005). The substantially reduced nonlinear contribution by the mere introduction of a small, sub-second stimulus separation is further evidence that a large fraction of the nonlinearities generally observed in BOLD fMRI of the visual cortex is of neuronal origin.

The cause of the dispersive character of the residual nonlinearities is not immediately obvious and deserves further discussion. Although increased response latency (TTP increase) caused by a preceding stimulus has been previously observed in BOLD fMRI (McClure et al., 2005), the finding of a response broadening that wanes with stimulus separation on a timescale of a few seconds is novel. These dispersive temporal characteristics are indicative of a slow, presumably vascular cause, since neuronal interactions would affect the response amplitude but are not expected to substantially alter the shape of the observed hemodynamic response. In other words, if the nonlinearities would be neuronal in origin, second order kernels with a shape similar to the observed primary response would be expected. This was indeed the case in experiments in which nonlinear neuronal contributions were enhanced (Kellman et al., 2003). It is plausible that temporal characteristics of the BOLD response are dependent on the baseline state of the vasculature, which is altered by neuronal activity associated with previous stimuli or by physiologic challenges. The changes in response amplitude, width and latency observed in BOLD fMRI experiments are similar to those found by Cohen and colleagues during hypercapnia-induced systemic vasodilatation (Cohen et al., 2002). In addition, the finding of a delayed IR is consistent with an earlier finding of slowed response to stimulation following a long period of elevated activity (McClure et al., 2005).

One possible mechanism for the observed vascular nonlinearities is an arteriolar compliance model proposed by Behzadi et al. (Behzadi and Liu, 2005), which suggests a reduced resistance to vasodilatation with increasing vessel diameter, causing the response to slow at higher dilation levels. This could explain response slowing observed in this study. An alternative mechanism responsible for our observations is the delayed vascular compliance model (Mandeville et al., 1999; Mandeville et al., 1998), in which the timing mismatch between cerebral blood volume (CBV) and cerebral blood flow (CBF) changes can lead to BOLD response slowing.

Some of the observed temporal nonlinearities may relate to mechanisms that support neuro-electrical activity, such as glutamate cycling and metabolic processes (Bak et al., 2006; Lu et al., 2004). These could significantly outlast neuronal signaling and affect the BOLD response to succeeding stimuli either by altering the hemodynamic coupling or by altering vascular oxygen extraction. Our methodology does not allow discrimination between such effects and the purely vascular effects described above.

In the MEG data shown in Figure 1 it can be seen that the response to the first event (checkerboard display onset) within the 800-ms duration stimulus is larger than for the subsequent events (checkerboard contrast reversals) within the same 800-ms duration stimulus. The 600-ms duration stimuli (400-ms gap) show the same effect (data not shown). It is a well documented phenomenon that the neuronal response to the checkerboard stimulus onset and its contrast reversal are different, e.g. in the visual evoked potential (VEP) literature (Estevez and Spekreijse, 1974). This stimulus onset nonlinearity affects each individual stimulus within the experiment equally. This should be distinguished from interactions between subsequent stimuli, investigation of which was the purpose of this study. The fact that this onset-nonlinearity is not significantly different when comparing the *respS* and *resp1* responses (Figure 1b) is actually further evidence of the lack of neuronal interaction between subsequent stimuli in these gapped experiments. If such an inter-stimulus interaction existed, one would expect decreased stimulus onset nonlinearity in the presence of a preceding stimulus, so a reduction of the exaggerated response to the initial event within a stimulus when another stimulus closely preceded the current one. This is not observed here, there is no significant difference between *respS* and *resp1* (see Figure 1b).

Although data presented here (e.g see Table 1) suggest a negligible nonlinearity between stimulus duration and response integral, they do indicate the presence of a nonlinear process in BOLD fMRI that affects hemodynamic response function (HRF) shape. This has important implications for the interpretation of (most notably event-related) experiments by fMRI as well as by other imaging modalities that infer neuronal activation from changes in cerebral hemodynamics (e.g. PET and NIRS). The observed effects should be accounted for in optimization of these techniques with regards to detection and estimation efficiency and need to be incorporated in data analysis models. Given that no significant effect was found on response integral in these experiments, the impact of the observed nonlinear effects on commonly-used block paradigm fMRI experiments is expected to be minimal: the steady-state response amplitude in block paradigms solely depends on this response integral.

The influence of prior events, on the scale of a few seconds, on the BOLD IR shape in response to a given stimulus can lead to misinterpretation of data when a canonical hemodynamic response function (HRF) is used in general linear model (GLM) analysis of such data. If, due to the effects observed here, temporal characteristics of the actual IR to one stimulus are substantially different from the response to another, this will adversely affect such a GLM analysis, especially in an event-related design: If the GLM design matrix used describes the response to one stimulus type less accurately than the response to another, the residual after fit will be larger. Since the standard deviation of this residual is used to compute significance of the fitted amplitude, the computed t-score will be reduced. This not only reduces detection power, but will often be erroneously interpreted as decreased activation. Secondly, the accuracy of the fitted amplitude is reduced in this scenario. These matters should be taken into account when comparing inferred changes in activation under different stimulus conditions, e.g. by using an analysis method that is not based on a canonical HRF, such as FIR (finite impulse response) in SPM.

Although the findings are expected to hold for other neocortical brain areas due to strong similarities in vascular architecture and neurovascular control mechanisms throughout these regions, this cannot currently be confirmed since only the human visual system was investigated here. Repeating these experiments in other brain areas is not straightforward. For example, paradigms that require more active participation of the subject (e.g. finger tapping to study this effect in primary motor areas) are more sensitive to attention effects and performance fluctuations. Furthermore, substantial long-lasting neuronal nonlinearities are well documented for other areas (e.g. somatosensory (Nangini et al., 2006) and olfactory (Freeman, 1979) cortices).

In conclusion, we demonstrated that vascular effects affect the BOLD response on a time scale of several seconds. This finding has implications for the interpretation of fMRI- and other functional data based on hemodynamic changes, especially for rapid event-related experiments in which stimuli are separated by less than a few seconds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Peter Kellman (NHLBI, NIH) for discussions and Karin Shmueli (AMRI) for proofreading the manuscript. This research was supported by the Intramural Research Program of the National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Appendix

Appendix A: Derivation of the various response kernels from m-sequence data

Two distinct classes of m-sequence stimuli can be identified, namely balanced and unbalanced designs. In a balanced design, as is simulated in Figure 2 of Kellman et al. (Kellman, Neuroimage 2003, 19:190-199), there are two distinct stimuli, which we will represent here by ‘1’ and ‘−1’ m-sequence bins. In a balanced stimulus paradigm it is assumed that the effect of a change from ‘1’ to ‘−1’ is equivalent to the transition from ‘−1’ to ‘1’. Also, two subsequent ‘1’ stimuli are assumed to yield similar activation as two subsequent ‘−1’ stimuli.

The other case, exemplified by a stimulus-versus-rest m-sequence as is shown in Figure 3 in Kellman et al., is unbalanced and can be described by ‘1’ and ‘0’ bins (stimulus ‘on’ and ‘off’). Two subsequent ‘on’ or ‘1’ stimuli will yield a different (in all likelihood more intense) activation than two subsequent ‘off’ or ‘0’ stimuli.

As a result, the primary and secondary response kernels observed in the correlograms yielded by balanced versus unbalanced m-sequence paradigms are different. In the balanced case, the observed primary response directly corresponds to the response that would result from a single, isolated stimulus (*h1* in Equation 3 in Kellman et al.). The secondary kernels derived from balanced m-sequence data correspond to the matching nonlinearities (*h2* in Equation 3 in Kellman et al.). However, in the case of an unbalanced paradigm like the one employed here, interactions affect only half of the stimuli, since in 50 % of the cases bins from which a given interaction would originate do not contain stimuli (are in the ‘off’ or ‘0’ state). As a result, the primary kernel (here also referred to as *pri*) yielded by covariance analysis (as is shown in Fig.

3 in the paper) is actually a mix of the true first-order response to an isolated stimulus with the responses to stimuli affected by the various interactions. A second effect of this reduced number of interactions is that the higher-order responses (here also referred to as *sec* and *tri*) in data derived from an unbalanced m-sequence design, as are shown in Fig. 3 in the paper, are underestimated compared to the observed primary response.

In order to determine the true first-order response to an isolated stimulus for an unbalanced m-sequence experiment, one has to correct for this effect. As will be set out mathematically below, this can be done by subtracting the measured higher-order responses from the measured *pri* response. Similarly, the various responses in the presence of preceding stimuli can be derived by adding the correct higher-order kernels.

The observed response, $r(t)$, to input $s(t)$ for a system with up to 2nd-order interactions is:

$$r(t)=h_0+\sum_k h_1(k)\,s(t-k)+\sum_{k_1}\sum_{k_2} h_2(k_1,k_2)\,s(t-k_1)\,s(t-k_2) \tag{1}$$

In an unbalanced design switching between 0 and 1, the input can be represented as a function of the m-sequence, which switches between -1 and 1, as follows:

$$s(t)=\frac{m(t)+1}{2} \tag{2}$$

Combination of Eq. [1] and [2] thus yields:

$$\begin{aligned} r(t) &=h_0+\sum_k h_1(k)\,\frac{m(t-k)+1}{2}+\sum_{k_1}\sum_{k_2} h_2(k_1,k_2)\,\frac{m(t-k_1)+1}{2}\cdot\frac{m(t-k_2)+1}{2}\\ &=h_0+\sum_k\frac{h_1(k)m(t-k)}{2}+\sum_k\frac{h_1(k)}{2}+\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)m(t-k_1)m(t-k_2)}{4}+\\ ]\quad &\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)m(t-k_1)}{4}+\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)m(t-k_2)}{4}+\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)}{4} \end{aligned} \tag{3}$$

Since terms without $m(t-k_n)$ are constants, they can be combined into one constant term A. Secondly, $m(t-k_1)\cdot m(t-k_2)$ results in another m-sequence, which will be referred to as $m(t-x_{k1,k2})$. Therefore, the above can be reduced to:

$$\begin{aligned} r(t) &=A+\sum_k\frac{h_1(k)m(t-k)}{2}+\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)m\left(t-x_{k1,k2}\right)}{4}+\sum_{k_1}\sum_{k_2}\frac{h_2(k_1,k_2)m(t-k_1)}{4}+\\ &\sum_{k_1}\sum_{K_2}\frac{h_2(k_1,k_2)m(t-k_2)}{4} \end{aligned} \tag{4}$$

Here, we only take the $h_2(k,k-1)$ and $h_2(k,k-2)$ interactions into account. Therefore:

$$\begin{aligned} r(t) &=A+\sum_k\frac{h_1(k)m(t-k)}{2}+\sum_k\frac{h_2(k,k-1)m\left(t-x_{k,k-1}\right)}{4}+\sum_k\frac{h_2(k,k-1)m(t-k)}{4}+\\ &\sum_k\frac{h_2(k,k-1)m(t-k-1)}{4}+\sum_k\frac{h_2(k,k-2)m\left(t-x_{k,k-2}\right)}{4}+\sum_k\frac{h_2(k,k-2)m(t-k)}{4}+\\ &\sum_k\frac{h_2(k,k-2)m(t-k-2)}{4} \end{aligned} \tag{5}$$

In the above, $m(t-x_{k,k-p})$ is a shifted version of the original m-sequence. The shift depends on the difference between k and $(k-p)$, and will be referred to as Δ_p. Therefore,

$$m\left(t-x_{k,k-p}\right)=m\left(t-\left(\Delta_{p}+k\right)\right)=m\left(t-\Delta_{p}-k\right) \tag*{[6a]}$$

and also

$$m\left(t-x_{k-1,k-1-p}\right)=m\left(t-\left(\Delta_{p}+1+k\right)\right)=m\left(t-\Delta_{p}-k-1\right) \tag*{[6b]}$$

During analysis, the measured signal $r(t)$ is correlated with m-sequence $m(t)$. This results in the correlogram $c(q)$:

$$c\left(q\right)=\sum_{t}m\left(t-q\right)s\left(t\right) \tag*{[7]}$$

Furthermore,

$$\sum_{t}m\left(t-q\right)m\left(t-k\right)=\delta\left(q-k\right) \tag*{[8]}$$

Due to the m-sequence properties, the above is only non-zero for q=k. (In truth all non-zero offsets in the m-sequence correlogram are a small constant negative value, which would only contribute constant terms to the equations that follow and are therefore ignored here.) Therefore:

$$\sum_{t}m\left(t-q\right)\sum_{k}h_{1}\left(k\right)m\left(t-k\right)=\sum_{k}h_{1}\left(k\right)\sum_{t}m\left(t-q\right)m\left(t-k\right)=N\sum_{k}h_{1}\left(k\right)\delta\left(q-k\right)=N\cdot h_{1}\left(q\right) \tag*{[9]}$$

In turn, this results in:

$$\begin{aligned}c\left(q\right)\quad&=\tfrac{N}{2}h_{1}\left(q\right)+\tfrac{N}{4}h_{2}\left(q-\Delta_{1},q-\Delta_{1}-1\right)+\tfrac{N}{4}h_{2}\left(q,q-1\right)+\tfrac{N}{4}h_{2}\left(q-1,q-2\right)+\\&\tfrac{N}{4}h_{2}\left(q-\Delta_{2},q-\Delta_{2}-2\right)+\tfrac{N}{4}h_{2}\left(q,q-2\right)+\tfrac{N}{4}h_{2}\left(q-2,q-4\right)\end{aligned} \tag*{[10]}$$

Here, the terms without Δ_p together form the observed pri kernel, terms with Δ_1 the *sec1* kernel and terms with Δ_2 the *sec2* kernel. Therefore, the h_1 kernel and slices of the h_2 kernel can be derived from the measured signal $c(q)$:

$$h_{2}\left(i,i-2\right)=\frac{4}{N}c\left(i+\Delta_{2}\right) \tag*{[11a]}$$

$$h_{2}\left(i,i-1\right)=\frac{4}{N}c\left(i+\Delta_{1}\right) \tag*{[11b]}$$

$$h_{1}\left(i\right)=\frac{2}{N}c\left(i\right)-\frac{1}{2}h_{2}\left(i,i-1\right)-\frac{1}{2}h_{2}\left(i-1,i-2\right)-\frac{1}{2}h_{2}\left(i,i-2\right)-\frac{1}{2}h_{2}\left(i-2,i-4\right) \tag*{[11c]}$$

It can be derived that for a balanced m-sequence design, the equivalent of Eq. [3] becomes:

$$\begin{aligned} r(t) &= h_0 + \sum_k h_1(k)\, m(t-k) + \sum_{k_1}\sum_{k_2} h_2(k_1,k_2)\, m(t-k_1)\, m(t-k_2) \\ & h_0 + \sum_k h_1(k)\, m(t-k) + \sum_{k_1}\sum_{k_2} h_2(k_1,k_2)\, m(t-x_{k1,k2}) \end{aligned} \quad [12]$$

Again, when only take the $h_2(k,k\text{-}1)$ and $h_2(k,k\text{-}2)$ interactions into account, this results in:

$$\begin{aligned} r(t) \quad &= h_0 + \sum_k h_1(k)\, m(t-k) + \sum_k h_2(k,k-1)\, m(t-x_{k,k-1}) + \sum_k h_2(k,k-2)\, m(t-x_{k,k-2}) + \\ & \sum_k h_2(k,k-3)\, m(t-x_{k,k-3}) + \sum_k h_3(k,k-1,k-2)\,(t-m_{k,k-1,k-2}) \end{aligned} \quad [13]$$

Covariance analysis of this measured signal $r(t)$ with m-sequence $m(t)$ yields:

$$c(q) = Nh_1(q) + Nh_2(q-\Delta) + Nh_2(q-\Delta_2) \quad [14]$$

This demonstrates that the various interactions are properly separated in the balanced m-sequence, and can therefore be directly measured at the various offsets $\triangle_p$:

$$h_1(i) = \frac{1}{N} c(i) \quad [15a]$$

$$h_2(i,i-1) = \frac{1}{N} c(i+\delta_1) \quad [15b]$$

$$h_2(i,i-2) = \frac{1}{N} c(i+\Delta_2) \quad [15c]$$

References

Bak LK, Schousboe A, Waagepetersen HS. The glutamate/GABA-glutamine cycle: aspects of transport, neurotransmitter homeostasis and ammonia transfer. J Neurochem 2006;98:641–653. [PubMed: 16787421]

Baseler HA, Sutter EE, Klein SA, Carney T. The topography of visual evoked response properties across the visual field. Electroencephalogr Clin Neurophysiol 1994;90:65–81. [PubMed: 7509275]

Behzadi Y, Liu TT. An arteriolar compliance model of the cerebral blood flow response to neural stimulus. Neuroimage 2005;25:1100–1111. [PubMed: 15850728]

Benardete, EA.; Victor, JD. An extension of the m-sequence technique for the analysis of multi-input nonlinear systems. In: Marmarelis, VZ., editor. Advanced methods of physiological system modeling. Plenum; New York: 1994. p. 87-110.

Berwick J, Martin C, Martindale J, Jones M, Johnston D, Zheng Y, Redgrave P, Mayhew J. Hemodynamic response in the unanesthetized rat: intrinsic optical imaging and spectroscopy of the barrel cortex. J Cereb Blood Flow Metab 2002;22:670–679. [PubMed: 12045665]

Birn RM, Bandettini PA. The effect of stimulus duty cycle and “off” duration on BOLD response linearity. Neuroimage 2005;27:70–82. [PubMed: 15914032]

Birn RM, Saad ZS, Bandettini PA. Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. Neuroimage 2001;14:817–826. [PubMed: 11554800]

Bodurka J, Ledden PJ, van Gelderen P, Chu R, de Zwart JA, Morris D, Duyn JH. Scalable multichannel MRI data acquisition system. Magn Reson Med 2004;51:165–171. [PubMed: 14705057]

Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci 1996;16:4207–4221. [PubMed: 8753882]

Buracas GT, Boynton GM. Efficient design of event-related fMRI experiments using M-sequences. Neuroimage 2002;16:801–813. [PubMed: 12169264]

Clark VP, Maisog JM, Haxby JV. fMRI study of face perception and memory using random stimulus sequences. J Neurophysiol 1998;79:3257–3265. [PubMed: 9636124]

Cohen ER, Ugurbil K, Kim SG. Effect of basal conditions on the magnitude and dynamics of the blood oxygenation level-dependent fMRI response. J Cereb Blood Flow Metab 2002;22:1042–1053. [PubMed: 12218410]

Dale AM, Buckner RL. Selective averaging of rapidly presented individual trials using fMRI. Hum Brain Mapp 1997;5:329–340.

de Zwart JA, Ledden PJ, van Gelderen P, Bodurka J, Chu R, Duyn JH. Signal-to-noise ratio and parallel imaging performance of a 16-channel receive-only brain coil array at 3.0 Tesla. Magn Reson Med 2004;51:22–26. [PubMed: 14705041]

de Zwart JA, Silva AC, van Gelderen P, Kellman P, Fukunaga M, Chu R, Koretsky AP, Frank JA, Duyn JH. Temporal dynamics of the BOLD fMRI impulse response. Neuroimage 2005;24:667–677. [PubMed: 15652302]

de Zwart, JA.; van Gelderen, P.; Jansma, JM.; Duyn, JH. Quantitative characterization of vascular non-linearities in BOLD fMRI; Proc. ISMRM 14th Scientific Meeting & Exhibition; Seattle, WA, USA. 2006. p. 2763

de Zwart JA, van Gelderen P, Kellman P, Duyn JH. Application of sensitivity-encoded echo-planar imaging for blood oxygen level-dependent functional brain imaging. Magn Reson Med 2002;48:1011–1020. [PubMed: 12465111]

Estevez O, Spekreijse H. Relationship between pattern appearance-disappearance and pattern reversal responses. Exp Brain Res 1974;19:233–238. [PubMed: 4819839]

Freeman WJ. EEG analysis gives model of neuronal template-matching mechanism for sensory search with olfactory bulb. Biol Cybern 1979;35:221–234. [PubMed: 526484]

Friston KJ, Josephs O, Rees G, Turner R. Nonlinear event-related responses in fMRI. Magn Reson Med 1998;39:41–52. [PubMed: 9438436]

Glover GH. Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 1999;9:416–429. [PubMed: 10191170]

Grill-Spector K, Henson R, Martin A. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 2006;10:14–23. [PubMed: 16321563]

Hamalainen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV. Magnetoencephalography - Theory, Instrumentation, and Applications to Noninvasive Studies of the Working Human Brain. Reviews of Modern Physics 1993;65:413–497.

Kellman P, Gelderen P, de Zwart JA, Duyn JH. Method for functional MRI mapping of nonlinear response. Neuroimage 2003;19:190–199. [PubMed: 12781738]

Kim SG. Quantification of relative cerebral blood flow change by flow-sensitive alternating inversion recovery (FAIR) technique: application to functional mapping. Magn Reson Med 1995;34:293–301. [PubMed: 7500865]

Kwong KK, Belliveau JW, Chesler DA, Goldberg IE, Weisskoff RM, Poncelet BP, Kennedy DN, Hoppel BE, Cohen MS, Turner R, et al. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proc Natl Acad Sci U S A 1992;89:5675–5679. [PubMed: 1608978]

Lee SP, Duong TQ, Yang G, Iadecola C, Kim SG. Relative changes of cerebral arterial and venous blood volumes during increased cerebral blood flow: implications for BOLD fMRI. Magn Reson Med 2001;45:791–800. [PubMed: 11323805]

Lu H, Golay X, Pekar JJ, Van Zijl PC. Sustained poststimulus elevation in cerebral oxygen utilization after vascular recovery. J Cereb Blood Flow Metab 2004;24:764–770. [PubMed: 15241184]

Mandeville JB, Marota JJ, Ayata C, Zaharchuk G, Moskowitz MA, Rosen BR, Weisskoff RM. Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. J Cereb Blood Flow Metab 1999;19:679–689. [PubMed: 10366199]

Mandeville JB, Marota JJ, Kosofsky BE, Keltner JR, Weissleder R, Rosen BR, Weisskoff RM. Dynamic functional imaging of relative cerebral blood volume during rat forepaw stimulation. Magn Reson Med 1998;39:615–624. [PubMed: 9543424]

Martindale J, Mayhew J, Berwick J, Jones M, Martin C, Johnston D, Redgrave P, Zheng Y. The hemodynamic impulse response to a single neural event. J Cereb Blood Flow Metab 2003;23:546–555. [PubMed: 12771569]

McClure KD, McClure SM, Richter MC, Richter W. The kinetics of the BOLD response depend on inter-stimulus time. Neuroimage 2005;27:817–823. [PubMed: 16043369]

Nangini C, Ross B, Tam F, Graham SJ. Magnetoencephalographic study of vibrotactile evoked transient and steady-state responses in human somatosensory cortex. Neuroimage 2006;33:252–262. [PubMed: 16884928]

Ogawa S, Lee TM, Kay AR, Tank DW. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A 1990;87:9868–9872. [PubMed: 2124706]

Ogawa S, Lee TM, Stepnoski R, Chen W, Zhu XH, Ugurbil K. An approach to probe some neural systems interaction by functional MRI at neural time scale down to milliseconds. Proc Natl Acad Sci U S A 2000;97:11026–11031. [PubMed: 11005873]

Rudin M, Beckmann N, Sauter A. Analysis of tracer transit in rat brain after carotid artery and femoral vein administrations using linear system theory. Magn Reson Imaging 1997;15:551–558. [PubMed: 9253999]

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, Leventhal AG. Signal timing across the macaque visual system. J Neurophysiol 1998;79:3272–3278. [PubMed: 9636126]

Silva AC, Koretsky AP, Duyn JH. Functional MRI impulse response for BOLD and CBV contrast in rat somatosensory cortex. Magn Reson Med 2007;57:1110–1118. [PubMed: 17534912]

Sutter, EE. A practical nonstochastic approach to nonlinear time-domain analysis. In: Marmarelis, VZ., editor. Advanced methods of physiological system modeling. University of Southern California; Los Angeles: 1987. p. 303-315.

Thévenaz, P.; Ruttiman, UE.; Unser, M. Iterative multi-scale registration without landmarks; IEEE international conference on image processing, IEEE international conference on image processing; Washington, DC, USA. 1995. p. 228-231.

Vazquez AL, Noll DC. Nonlinear aspects of the BOLD response in functional MRI. Neuroimage 1998;7:108–118. [PubMed: 9558643]

Waldvogel D, van Gelderen P, Muellbacher W, Ziemann U, Immisch I, Hallett M. The relative metabolic demand of inhibition and excitation. Nature 2000;406:995–998. [PubMed: 10984053]

Zhang N, Liu Z, He B, Chen W. Noninvasive study of neurovascular coupling during graded neuronal suppression. J Cereb Blood Flow Metab 2008a;28:280–290. [PubMed: 17700632]

Zhang N, Zhu XH, Chen W. Investigating the source of BOLD nonlinearity in human visual cortex in response to paired visual stimuli. Neuroimage 2008b;43:204–212. [PubMed: 18657623]

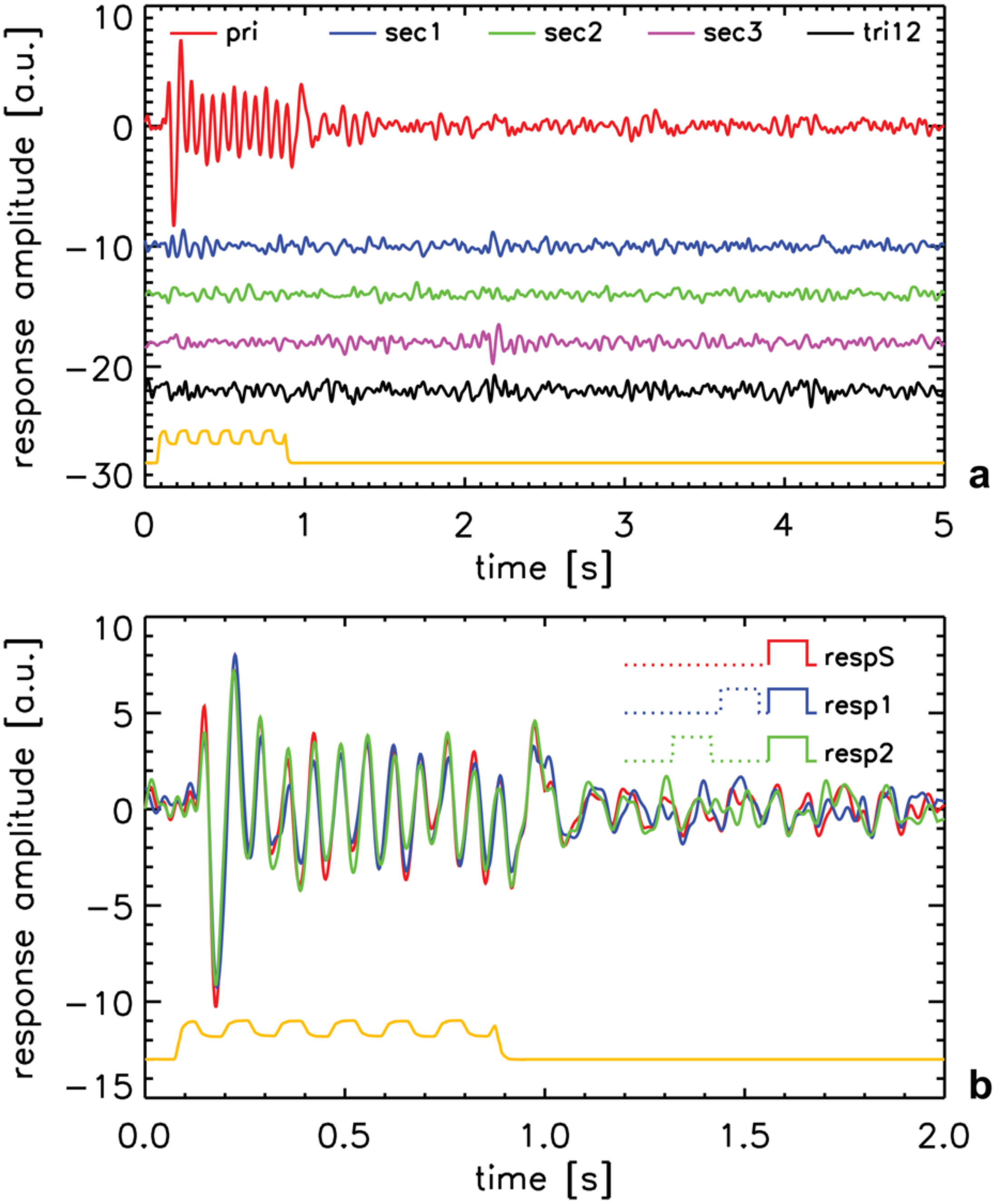

Figure 1.

(a) M-sequence response kernels (cf. Fig. 3) for the 200-ms gap MEG experiment for 1 volunteer, derived from the average signal over 10 detectors. The primary kernel is shown in red, the first three second-order kernels in blue, green and pink, respectively, and the first third-order kernel in black. The yellow line shows the stimulus timing and is derived from the output of an optical sensor in the projector beam. For clarity an arbitrary baseline offset is used for all but the primary response kernel. (**b**) The response to an isolated stimulus (*respS*) and the responses to a stimulus that closely follows on a preceding identical stimulus (*resp1* and *resp2*, corresponding to inter-stimulus intervals of 0.2 s and 1.2 s, respectively). These responses were derived using the *pri*, *sec1* and *sec2* kernels shown in (**a)** (see text and Appendix A for details). The yellow line again shows the stimulus timing.

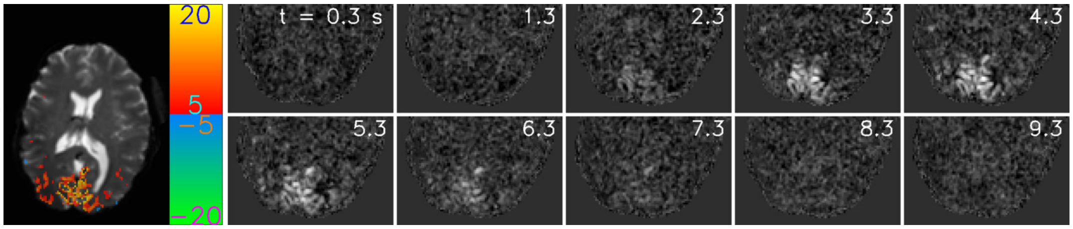

Figure 2.

Example of the observed average BOLD-fMRI hemodynamic response (first order kernel *pri*) in a representative slice in one of the volunteers during a full-contrast experiment with 200-ms gap. In the left-most pane, a t-score map (derived from the same experiment) is superimposed on an anatomical image (first image of the EPI time-series data) of the slice. The first 10 s of the response in the bottom-half of the slice are shown in the remaining 10 images. The time in s relative to stimulus onset is indicated in the top right-hand corner of each image. Since this specific slice was number 7 out of 10 its acquisition timing was delayed 0.3 s relative to stimulus onset.

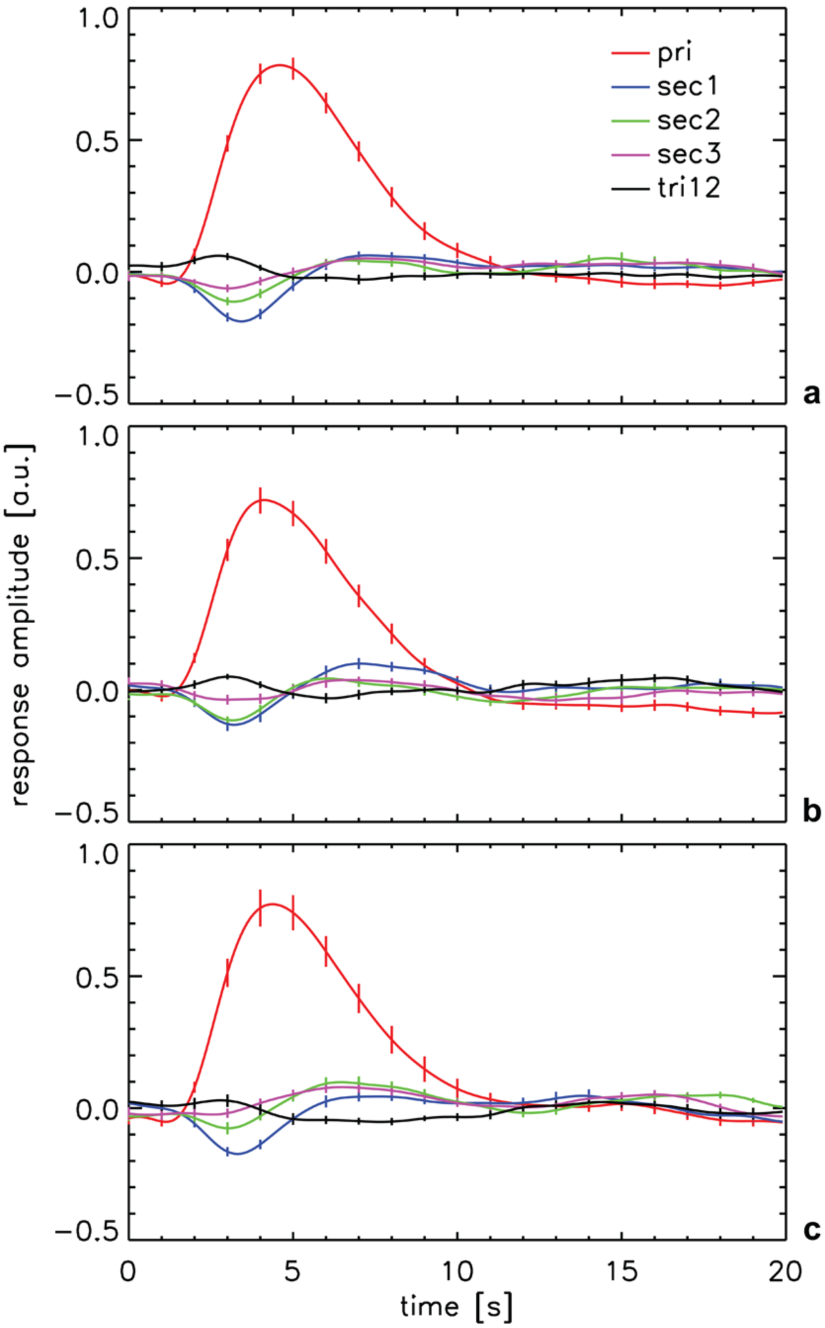

Figure 3.

ROI and volunteer-averaged response time-courses (m-sequence response kernels) obtained from BOLD fMRI experiments. **(a)** full-contrast 200 ms gap (n=18); **(b)** 400-ms gap (n=13); and **(c)** low-contrast 200-ms gap (n=13). The mean first-order kernel (red) and the first 3 second-order kernels (respectively blue, green and pink) are shown, as well as the first third-order kernel (black). Error bars indicate inter-subject standard error.

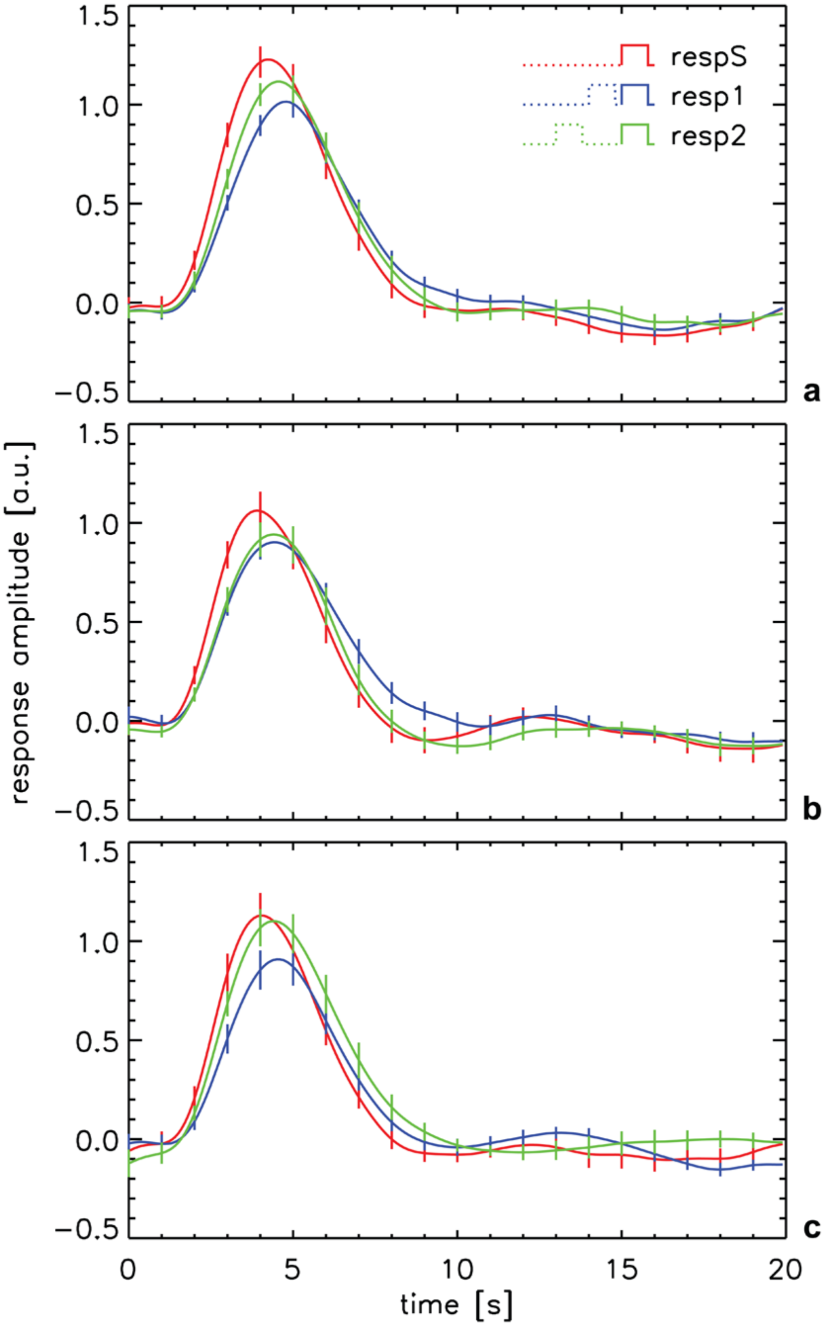

Figure 4.

The *pri*, *sec1* and *sec2* kernels (see Fig. 3) were used to compute the hemodynamic response to a single stimulus (*respS*) on a volunteer-by-volunteer basis. Also computed were the response that occurred when the current stimulus was closely preceded by another, identical stimulus, either in the bin directly preceding it (*resp1*, 0.2-0.4 s stimulus separation) or in the bin before that (*resp2*, 1.2-1.4 s stimulus separation). **(a)** full-contrast 200 ms gap (n=18); **(b)** 400-ms gap (n=13); and **(c)** low-contrast 200-ms gap (n=13). Error bars indicate inter-subject standard error.

Table 1

Parameters of the BOLD-fMRI responses for the different stimuli under various conditions. For the time-to-peak (TTP) and full width at half-maximum (FWHM), the absolute as well as the relative values are given for all three paradigms in the current and an earlier (*200old*, see Supplement B) studies. For amplitude and surface area data, only the relative values are supplied. Here, '*200*' and '*200old*' refer to full-contrast stimuli with 200-ms gap (800-ms duration stimuli), '*400*' to 600-ms duration full-contrast stimuli with 400-ms gap and '*200lc*' to low-contrast 200-ms gap stimuli.

	time to peak [s]				**full width at half max. [s]**				**relative amplitude**				**relative surface area**			
	200	**400**	**200lc**	**200old**	**200**	**400**	**200lc**	**200old**	**200**	**400**	**200lc**	**200old**	**200**	**400**	**200lc**	**200old**
respS	4.3 (0.1)	4.0 (0.1)	4.1 (0.1)	4.4 (0.2)	3.6 (0.1)	3.6 (0.2)	3.5 (0.1)	3.4 (0.3)	1.00 (0.07)	1.00 (0.09)	1.00 (0.11)	1.00 (0.11)	1.00 (0.11)	1.00 (0.13)	1.00 (0.13)	1.00 (0.17)
resp2	4.6 (0.1)	4.4 (0.1)	4.5 (0.1)	4.6 (0.2)	3.8 (0.2)	3.5 (0.2)	3.8 (0.2)	3.8 (0.2)	0.91 (0.05)	0.90 (0.09)	0.98 (0.09)	0.93 (0.11)	0.93 (0.09)	0.90 (0.11)	1.10 (0.12)	0.91 (0.14)
resp1	4.8 (0.1)	4.4 (0.1)	4.6 (0.1)	4.7 (0.2)	3.9 (0.3)	4.1 (0.3)	3.7 (0.2)	3.9 (0.2)	0.82 (0.05)	0.85 (0.06)	0.80 (0.09)	0.81 (0.06)	0.89 (0.07)	1.05 (0.10)	0.87 (0.12)	0.99 (0.09)

	relative time to peak				**Relative FWHM**			
	200	**400**	**200lc**	**200old**	**200**	**400**	**200lc**	**200old**
respS	1.00 (0.02)	1.00 (0.03)	1.00 (0.02)	1.00 (0.04)	1.00 (0.03)	1.00 (0.06)	1.00 (0.03)	1.00 (0.07)
resp2	1.07 (0.02)	1.10 (0.04)	1.10 (0.03)	1.05 (0.04)	1.04 (0.04)	0.98 (0.05)	1.09 (0.05)	1.11 (0.05)
resp1	1.12 (0.02)	1.11 (0.03)	1.13 (0.02)	1.08 (0.04)	1.07 (0.04)	1.14 (0.07)	1.07 (0.07)	1.13 (0.06)

Table 2

The results of fitting *resp1* and *resp2* with a stretched and scaled version of *respS*. Data show that the stretch factor increases with decreasing stimulus separation, confirming that the presence of a preceding stimulus increases the TTP and FWHM of the response. Furthermore, the fitted response amplitude decreases with decreasing stimulus separation, also in agreement with our experimental findings (Tab. 1). A residual fraction of 0.23 (on average for the three stimulus types) was observed if *respS* was directly subtracted from *resp1*, without stretching or scaling.

	stretch factor			**amplitude factor**			**residual fraction**		
	200	**400**	**200lc**	**200**	**400**	**200lc**	**200**	**400**	**200lc**
resp2	1.07 (0.01)	1.07 (0.01)	1.09 (0.02)	0.90 (0.02)	0.85 (0.02)	1.01 (0.05)	0.08 (0.01)	0.10 (0.02)	0.08 (0.01)
resp1	1.12 (0.01)	1.13 (0.01)	1.10 (0.02)	0.79 (0.01)	0.84 (0.03)	0.77 (0.03)	0.08 (0.01)	0.10 (0.02)	0.10 (0.02)