# Decoding of semantic categories of imagined concepts of animals and tools in fNIRS

**Abstract.** *Objective.* Semantic decoding refers to the identification of semantic concepts from recordings of an individual's brain activity. It has been previously reported in fMRI and EEG. We investigate whether semantic decoding is possible with functional near-infrared spectroscopy (fNIRS). Specifically, we attempt to differentiate between the semantic categories of animals and tools. We also identify suitable mental tasks for potential brain-computer interface (BCI) applications. *Approach.* We explore the feasibility of a silent naming task, for the first time in fNIRS, and propose three novel intuitive mental tasks based on imagining concepts using three sensory modalities: visual, auditory, and tactile. Participants are asked to visualize an object in their minds, imagine the sounds made by the object, and imagine the feeling of touching the object. A general linear model is used to extract hemodynamic responses that are then classified via logistic regression in a univariate and multivariate manner. *Main results.* We successfully classify all tasks with mean accuracies of 76.2% for the silent naming task, 80.9% for the visual imagery task, 72.8% for the auditory imagery task, and 70.4% for the tactile imagery task. Furthermore, we show that consistent neural representations of semantic categories exist by applying classifiers across tasks. *Significance.* These findings show that semantic decoding is possible in fNIRS. The study is the first step toward the use of semantic decoding for intuitive BCI applications for communication.

*Keywords*: semantic decoding, mental imagery, functional near-infrared spectroscopy (fNIRS), machine learning, general linear model (GLM), brain-computer interface (BCI), concepts

## 1. Introduction

Neuroimaging studies have shown the possibility of semantic decoding, defined as the identification of specific semantic concepts an individual is focusing on, or thinking of, from a recording of their brain activity [1]. Work by several groups has shown that several pairs of semantic categories can be differentiated, for instance, animals and tools [2, 3, 4, 5, 6, 7], tools and dwellings [8], animals and body parts [9], animals and faces [10]. Recent work has also shown that it is possible to differentiate more than two categories at a time [11, 12, 13].

While the most promising results to date have been reported using functional magnetic resonance imaging (fMRI) [14, 15, 8, 2], semantic decoding using neural signals recorded from the scalp, such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS), is of particular interest for potential applications such as brain-computer interfaces (BCIs) [16, 17]. EEG and fNIRS are portable, relatively

cheap, and provide better ecological validity in comparison with fMRI. EEG provides a good temporal resolution but has a poor spatial resolution. This could be potentially improved by combining it with fNIRS, as these two techniques have complementary strengths.

BCIs provide an alternative pathway between a human brain and external devices. They have been successfully used as a technological solution to aid communication for people who experience difficulties communicating via other means [18, 19, 20]. However, current BCI communication speeds and accuracies are relatively low in comparison with other communication platforms [21, 22, 23]. New BCIs based on semantic decoding could have the potential to be highly intuitive but it is currently unknown if they might allow comparable levels of accuracy and communication speed to the other state-of-the-art BCIs.

In this paper, we investigate whether semantic decoding is possible with fNIRS. To this end, we must carefully choose mental tasks to elicit clear patterns of brain activity that allow for the differentiation between semantic concepts or categories of interest. We attempt to identify mental tasks suitable for potential BCI applications, particularly for communication.

Semantic neuroimaging studies to date have used a variety of different mental tasks. For instance, in a silent naming task [6, 7], participants are asked to silently name, in their minds, an object displayed on the screen. The silent naming task has not been tested in fNIRS yet. We test its feasibility for semantic decoding.

Several studies [8, 3, 4, 15] asked participants to think of the same object properties each time the object was displayed. Each participant was asked to come up with a set of properties to think about for each object before the start of the experiment. Another study [9] removed this constraint and let participants think freely about the meaning of the given object or any memory it evoked. In general, many studies have shown that similar patterns of brain activity arise when perceiving and imagining objects [24, 25, 26, 27]. We follow their lead and propose three novel intuitive imagery sensory-based tasks using visual, auditory, and tactile perception. Participants are asked to visualize an object in their minds, imagine the sounds made by the object, and imagine the feeling of touching the object.

In [2, 5, 14], participants were presented with target and non-target semantic categories and asked to respond upon the appearance of objects from the non-target category, for instance, by pressing a mouse button. Similar tasks have been reported to be feasible for BCI applications. For instance, Geuze and colleagues [28] presented a BCI based on semantic relations in an association network. The BCI determined which prime word a participant had in mind by presenting probe words. The BCI detected whether the presented probe word was related to the prime word by the detecting semantic priming response, which can be detected at the single-trial level [29]. The BCI continued to choose new probe words from connections in the association network until a pre-determined confidence threshold was reached.

To determine the feasibility of our mental tasks, we attempt to differentiate between

semantic categories of animals and tools, that have mostly been used in existing semantic studies based on EEG [6, 5, 30] and fMRI [2]. A variety of external stimulus modalities have been used to cue participants to focus on an object, such as visual (photographs [2, 6], line drawings [5, 8]), written words [2, 5, 14], spoken words [2, 5], or combinations [9, 15]. We use the visual modality because it has frequently been used in semantic EEG studies [6, 31, 10, 30, 5] and it is appropriate for all our mental tasks.

In contrast with most semantic neuroimaging studies, we attempt to ensure a separation of input stimulus modality and task. Many studies focused on a comparison of different input modalities (images, written words, spoken words, different languages, etc.) while we focus on the comparison between different mental tasks when the concept to focus on has already been communicated to the participants. This separation avoids the problem of potential processing-related confounds in the classification process that is present in many studies. For instance, focusing on a concept while seeing its image raises the question of what is used for the differentiation between different concepts: the visual processing of the image or the imagination of the concept. This is a pertinent question as it has been shown that the viewed object can be identified from passive viewing of images [24, 25]. Some studies explicitly analyzed the influence of some possible confounds. For instance, the study by Murphy and colleagues [6] examined brightness, mean spatial frequency, and visual complexity of the stimuli images. However, the set of potential confounds and methods (for instance, how to measure image complexity) has not been comprehensively studied. Our study design mitigates this problem. We also use a mask after image presentation to reduce visual persistence [32].

Our objectives are: 1) to determine whether it is possible to differentiate between the semantic categories of animals and tools using the silent naming task and our three novel imagery sensory-based tasks in fNIRS, and 2) to explore whether consistent neural representations of animals and tools exist across tasks.

## 2. Methods

In our experiments, we simultaneously acquired EEG and fNIRS. However, we focus solely on fNIRS in this study. The fNIRS signals were modeled via a general linear model [33, 34] to remove noise and influences from previous tasks and to extract evoked hemodynamic responses that were then classified by regularized logistic regression in both a univariate and multivariate manner.

### *2.1. Participants*

Twelve right-handed native English speakers (3 males and 9 females, age range 20-57, mean 32.75, standard deviation 11.55) from the student and staff population of the University of Essex participated in the study. We recruited only native English speakers to avoid a potential problem of differences in neural representations of semantic concepts by individuals who speak different languages [35, 36], which could be particularly

important for the silent naming task. All participants had normal or corrected-to-normal vision. Participants received compensation of £16 for their time. They all read, understood, and signed a consent form. The study was approved by the Ethics Committee of the University of Essex on 25th October 2018.

### *2.2. Tasks*

Participants were presented with images of concepts from two semantic categories (animals and tools) to focus on. Each image presentation was followed by a set of four individual tasks: silent naming, visual, auditory, and tactile imagery. The order of tasks was randomized across blocks.

In the silent naming task, participants were asked to name the presented object in their minds, in their mother tongue (English). In the visual imagery task, participants were asked to visualize the presented object. They were instructed to try not to visualize the particular image they had seen but their own representation of the concept. Similarly in the auditory imagery task, they were asked to imagine sounds made by the presented object when they interact with it. For instance, the sounds made by an animal (such as the mewing of a cat) or the sounds produced when using a tool (such as the banging of a hammer). Lastly, in the tactile imagery task, participants were asked to imagine the feeling of touching the presented object. For instance, when petting an animal or touching different parts of a tool.

Participants were presented with the above-mentioned descriptions of the tasks (including the examples reported) but instructed to use the imagery strategy that came most naturally to them. For all imagery tasks, they were instructed to be engaged for the whole task duration (3 seconds). They were also instructed to try to avoid eye movements, facial muscle contractions, and head or body movements during the tasks. (See supplementary materials for the specific instructions provided to participants to read while setting up fNIRS and EEG.)

Additionally, we did try not to influence the participants' interpretation of images. They were shown all images before the experiment. We named a particular image only if they could not recognize it. Otherwise, we let them use their own interpretations.

### *2.3. Stimuli*

A set of 18 animals and 18 tools were selected. We selected as many concepts from previous studies [5, 6] as possible. We used concepts that are suitable for all tasks and recognizable by most people. For instance, many animals were excluded from our initial set as they were unsuitable for the auditory imagery task because we judged that a considerable number of people would not be able to recall, and then imagine, the sound such animals make. Images (with a license allowing non-commercial reuse with modifications) were sourced from the Internet. They were converted to gray-scale, cropped, resized to $400 \times 400$ pixels, and contrast stretched. In all images, the object was presented on a white background. We used photos instead of line drawings because

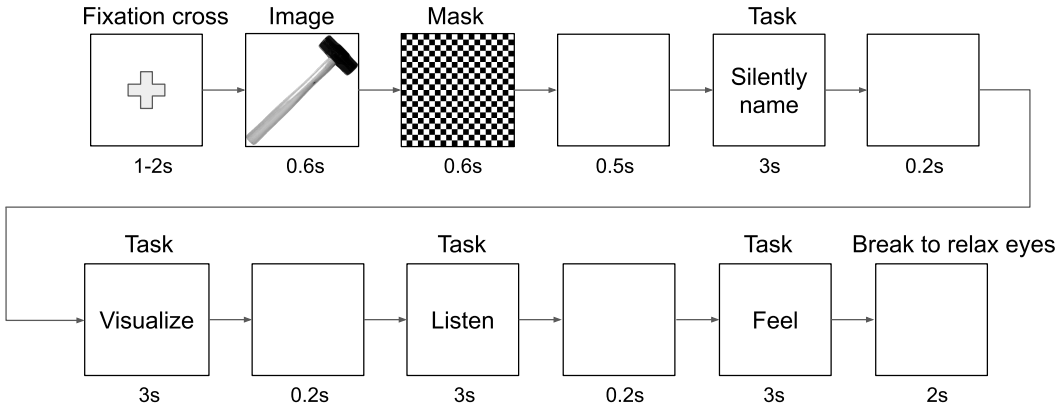


**Figure 1.** Illustration of one trial of concept presentation. The order of tasks is randomized across blocks.

we did not want to be constrained by their limited variety. The selected concepts are listed below (see supplementary materials for the corresponding images).

*Animals:* bear, cat, cock, cow, crab, crow, dog, donkey, duck, elephant, frog, lion, monkey, owl, pig, sheep, snake, and tiger.

*Tools:* axe, bottle-opener, broom, chain saw, computer keyboard, computer mouse, corkscrew, hammer, hand saw, hoover, kettle, knife, microwave, pen, phone, scissors, shovel, and toothbrush.

### *2.4. Experimental design*

As shown in figure 1, a trial started with the presentation of a black fixation cross on a white background for 1-2 seconds (uniformly distributed), followed by the image of a concept for 0.6 seconds. A mask (the image of a checkerboard) followed for 0.6 seconds to reduce visual persistence and thus to get rid of potential visual confounds [32]. Then, a blank white screen was presented for 0.5 seconds followed by a sequence of all four tasks, each task lasting for 3 seconds and was separated from the following one by a blank white screen lasting for 0.2 seconds. Participants were informed as to which task to do by a textual command appearing on the screen for the whole duration of the task: Silently name, Visualize, Listen, or Feel. The last task was followed by a short break for 2 seconds indicated by a blank screen, which changed color over time from white to black and back. One trial of the concept presentation took 17.3-18.3 seconds depending on the duration of the fixation cross.

Each concept was presented five times, for a total of 90 trials per category (18 concepts, 5 repetitions each). The experiment was split into 15 blocks with 12 concepts per block (207.6-219.6 seconds). Blocks were separated by breaks of at least 30 seconds. There was also a longer break of at least 3 minutes in the middle of the experiment.

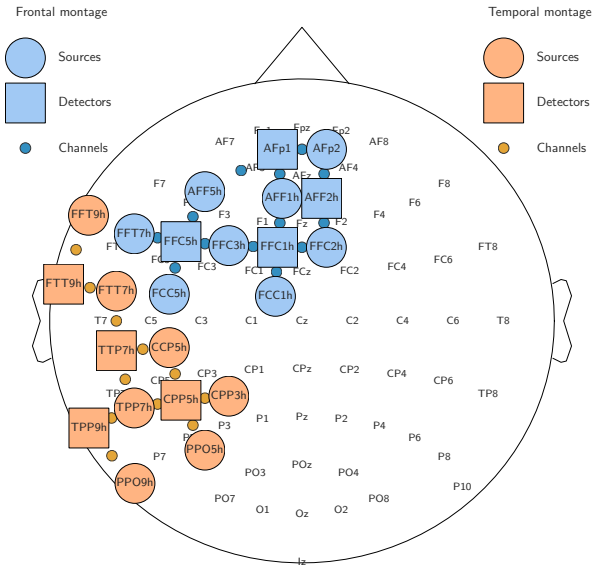


**Figure 2.** The frontal and temporal montage used for fNIRS data acquisition with the joint EEG with 64 electrodes following the international 10-20 system. fNIRS sources (circles) and detectors (squares) positioned in the 10-5 system form channels (small circles) shown between them.

The experiment started with two additional short blocks (86.5-91.5 seconds, containing a random subset of 5 concepts, each repeated two times) for familiarization with the experiment.

The order of concepts and tasks was pseudo-randomized for each participant with the following constraints. No concept was repeated twice in succession. All tasks in one block had the same order. Different blocks had different orders of tasks and no order in a given block was repeated in the following block.

### *2.5. Data acquisition*

fNIRS data were acquired with a NIRx NIRScoutXP continuous wave imaging system equipped with 4 light detectors, 8 light emitters (sources), and low-profile fNIRS optodes. EEG data were acquired with a BioSemi ActiveTwo system with 64 electrodes positioned according to the international 10-20 system, plus one electrode on each earlobe as references. Additionally, 2 electrodes placed on the left hand measured galvanic skin response (GSR) and a respiration belt around the waist measured respiration. The sampling rate was 2048 Hz. Both electrodes and optodes were placed in a NIRx NIRScap for integrated fNIRS-EEG layouts.

The brain regions selected for fNIRS channel placement were based on a review by Binder and colleagues [37] which analyzed 120 functional neuroimaging studies to locate the semantic system for semantic processing, which refers to the cognitive act of accessing stored knowledge about the world. Their analysis showed a left-lateralized network comprised of 7 regions: posterior inferior parietal lobe (angular gyrus and

adjacent supramarginal gyrus), middle temporal gyrus (and posterior portions of the inferior temporal gyrus), fusiform and parahippocampal gyri, dorsomedial prefrontal cortex, inferior frontal gyrus (especially pars orbitalis), ventromedial prefrontal cortex, and posterior cingulate gyrus (and adjacent ventral precuneus). We focused on regions close to the scalp which could potentially be studied by fNIRS, namely left-lateralized: posterior inferior parietal lobe, middle temporal gyrus, and dorsomedial prefrontal cortex.

We designed two fNIRS montages because our equipment did not include enough channels to record from all regions of interest (see figure 2). The first montage focused on the left temporal lobe and posterior inferior parietal lobe. The montage was selected based on the fOLD toolbox [38], which allows for probe arrangement guided by brain regions of interest, with regions of interest in the left-lateralized inferior parietal lobe, angular gyrus, middle and inferior temporal gyrus. We selectively eliminated optodes with the lowest specificity for our regions of interest until the number of optodes satisfied our equipment limitations. This montage used 4 detectors and only 7 sources creating 11 channels with a sampling rate of $f_s = 8.92$ Hz. The second montage focused on the left frontal cortex. The montage was selected using the same method based on the fOLD toolbox with regions of interest in Brodmann areas 9 and 46. This montage used 4 detectors and 8 sources creating 14 channels with a sampling rate of $f_s = 7.81$ Hz. The inter-optode separations were approximately 3 cm. For the first six participants, we used the prefrontal montage and for the final six, we used the temporal montage.

### *2.6. fNIRS data preprocessing*

Bad channels were identified and removed based on the scalp coupling index (SCI) [39], which assesses the correlation between two wavelength channels in the cardiac band. The main visible artifact in the fNIRS signal is due to the cardiac cycle which is approximately 1 Hz. Its presence is related to intracranial physiological parameters and indicates a good contact between the optical probe and the scalp [40, 41]. To compute an SCI, raw fNIRS data were band-pass filtered between 0.5 and 2.5 Hz (4th-order Butterworth filter, forward-backward filtering by the `sosfiltfilt` method from SciPy [42]) and z-scored. For each channel, the SCI was computed as the Pearson correlation coefficient between the fNIRS signals of two wavelengths of 785 and 830 nm defining the channel. In-phase and counter-phase identical waveforms have an SCI of 1 and -1, respectively, whereas uncorrelated signals have an SCI of 0. Channels with a mean SCI across all blocks below 0.75 were removed. This threshold was suggested in [39]. Two participants (participants 5 and 12) were excluded from further fNIRS analysis because they had all channels removed.

Motion artifacts caused by relative motion between fNIRS optical fibers and the scalp were corrected with a wavelet transform [43, 44, 45] using the `hmrMotionCorrectWavelet` method from the Homer2 software package [46] with an interquartile range of 1.5, as suggested in the function's documentation. Motion-

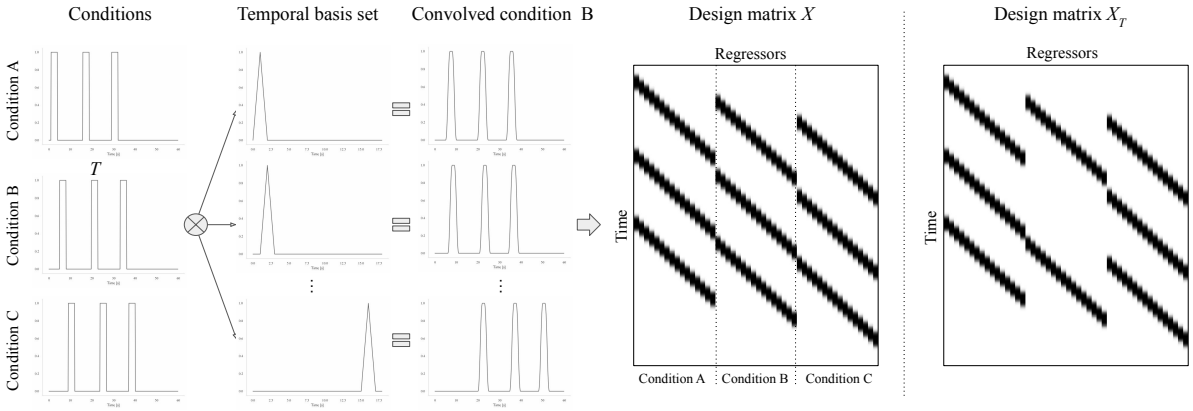


**Figure 3.** Illustration of creating the design matrix $X$ with three conditions. Condition B (e.g., visual imagery task) is separately convolved by each function from the temporal basis set. The results are regressors placed in the corresponding columns of $X$. $T$ represents one task trial from condition B. The design matrix $X_T$ is used to extract the evoked hemodynamic response for $T$. $X_T$ is created the same way as $X$ but excluding $T$ from the condition signal before the convolution.

corrected fNIRS signals with a wavelength of 785 and 830 nm were converted into changes of concentration in oxygenated and deoxygenated hemoglobin (in M) using the modified Beer-Lambert law [47] with differential path length factors (unitless) $DPF^{785} = 7.25$, and $DPF^{830} = 6.38$, and molar absorption coefficients (in $\mathrm{M}^{-1}\mathrm{cm}^{-1}$) $\alpha^{785}_{oxy} = 1798.643$, $\alpha^{785}_{deoxy} = 2295.285$, $\alpha^{830}_{oxy} = 2321.424$, and $\alpha^{830}_{deoxy} = 1791.734$.

### *2.7. fNIRS analysis*

Each fNIRS channel (i.e., changes of concentration in oxygenated or deoxygenated hemoglobin) was modeled by a general linear model (GLM) [33, 34] to remove noise and influences from preceding tasks. This is important due to the short gap (200 ms) between tasks. Since fNIRS records brain activity in a similar way to fMRI, the GLM approach adopted for fMRI [33] can be applied to fNIRS [48, 49, 50, 51].

A single fNIRS channel can be modeled by the GLM as follows:

$$y = X\beta + \epsilon \tag{1a}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I_N) \tag{1b}$$

where $y \in \mathbb{R}^N$ represents $N$ time samples of fNIRS signal, $X \in \mathbb{R}^{N \times P}$ is a design matrix, $\beta \in \mathbb{R}^P$ are unknown parameters, $\epsilon$ are normally distributed error terms with zero mean and variance $\sigma^2$, and $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix. The ordinary least squares (OLS) estimates of $\beta$ are given by

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2}$$

The design matrix $X$ models the experiment from the beginning of the first block until the end of the last block. Each row represents a different time step in the experiment. Columns represent regressors modeling the experiment. The first 160

regressors are the result of the convolution of 10 conditions modeled in the experiment and a set of 16 temporal basis functions.

We modeled the following conditions: four conditions represent the execution of 4 tasks for the animal category, four conditions represent the execution of 4 tasks for the tools category, one condition represents the presentation of an image, and one condition represents the mask. An element of a particular condition is 1 if the condition was taking place at the time represented by the element. All other elements are 0. This is illustrated in figure 3.

We used a temporal basis set based on the B-spline functions (see figure 3). It is an alternative to the finite impulse response (FIR) basis set that uses contiguous boxcar functions. The first function is non-zero from 0 to 2 seconds with a peak at 1 second. The following function is shifted by 1 second and so on. Our basis set is flexible enough to model different hemodynamic responses, with different timing between oxygenated and deoxygenated hemoglobin [52]. The set contains 16 functions with the last peak at the 16th second and the overall maximum length is 17 seconds. The number of functions was chosen based on the canonical hemodynamic response function to include its most informative segment.

All conditions and temporal basis functions were sampled at a higher temporal resolution ($16f_s$) and then downsampled (decimated) back to $f_s$ after the convolution, similarly to the default setting in the fMRI SPM toolbox implementation (`defaults.stats.fmri.t` = 16) [33]. The last 2 regressors are one column representing the respiration signal (resampled to the sample rate of the fNIRS signal, $f_s$), and one column of ones representing an offset. Thus, the model contains $P = 162$ parameters to be estimated.

Before the GLM analysis, fNIRS and the respiration signal were filtered between 0.01 and 0.7 Hz by two filters; first by an IIR high-pass 9th-order Butterworth filter with 0.01 Hz cutoff (forward-backward filtering, `sosfiltfilt` method) and then convolved by a FIR low-pass filter of length $\lfloor 42f_s \rfloor$ (374 for the temporal montage, 328 for the frontal montage; 21 seconds is the shortest length after the last block before the end of the experiment from all participants) with 0.7 Hz cutoff designed by the `firwin` method from SciPy [42].

The error term $\epsilon_t$ at time point $t$ in the fNIRS signal is correlated with its temporal neighbors due to several slow physiological processes, such as respiration, heartbeat, and blood pressure changes. This temporal correlation has to be modeled. Most of the fNIRS-GLM analysis studies have used pre-whitening based on an autoregressive model (AR) of the error terms [49, 53, 54]. A first-order autoregressive model (AR(1)) of the error terms is

$$\epsilon_t = a_1 \epsilon_{t-1} + \upsilon_t \tag{3a}$$

$$\upsilon_t \sim \mathcal{N}(0, \sigma_w^2) \tag{3b}$$

where the value at the current time $t$ depends on the value from the previous time $t-1$ and the normally distributed error term $\upsilon_t$ with zero mean and variance $\sigma_w^2$. First, the

OLS estimates are calculated, as per equation (2). The parameter $a_1$ is estimated by fitting AR(1) to the resulting residuals. Then, the Cochrane-Orcutt method [55] to reduce temporal correlation is defined as

$$y_t - a_1 y_{t-1} = X_t \beta_w - a_1 X_{t-1} \beta_w + \epsilon_t - a_1 \epsilon_{t-1} \tag{4}$$

which can be rewritten as

$$y_t^W = X_t^W \beta_w + \upsilon_t \tag{5}$$

where $y_t^W = y_t - a_1 y_{t-1}$ and $X_t^W = X_t - a_1 X_{t-1}$ represent the pre-whitened data and design matrix respectively, and $\beta_w \in \mathbb{R}^P$ are new unknown parameters. We will use the superscript $W$ to denote the pre-whitened data $y^W$ and design matrix $X^W$. Then, the transformed GLM with reduced temporal correlation is given by

$$y^W = X^W \beta_w + \upsilon \tag{6}$$

and the OLS estimates of $\beta_w$ are given by

$$\hat{\beta}_w = (X^{W^T} X^W)^{-1} X^{W^T} y^W \tag{7}$$

### *2.8. Classification*

Our GLM approach for a single fNIRS channel was adapted for the classification of evoked hemodynamic responses of a particular task (e.g., the visual imagery task) to differentiate between the semantic categories of animals and tools in 15-block-wise cross-validation in which one block is used for testing, while other blocks are used for training.

In one step of the cross-validation, let $X_{train}$ be the full design matrix $X$ with all rows outside the training blocks set to zero, with data blocks extended 5 seconds beyond the end of the experimental blocks to compensate for the hemodynamic delay and before possible artifacts in the fNIRS signal at the beginning of the breaks. Similarly, let $y_{train}$ be the fNIRS signal zeroed outside the training blocks. Keeping the original dimensions (but zeroing rows) makes things easier for overall manipulations, fitting AR(1) and pre-whitening without any time discontinuities. The terms $a_1$ and $\hat{\beta}_w$ are then estimated from $X_{train}$ and $y_{train}$.

The next step is to extract the evoked hemodynamic response for each trial of the desired task. Let $T$ be one task trial for which we want to extract the evoked hemodynamic response. The estimated GLM is used to remove influence from preceding tasks and thus to isolate the evoked hemodynamic response for $T$. Let $X_T$ be the design matrix created in the same way as described for $X$. The only difference is that, before the convolution, the elements of the corresponding condition for $T$ are 0, instead of 1, when $T$ is taking place. In other words, the design matrix $X_T$ is excluding $T$ from the design matrix $X$ as if $T$ did not take place in the experiment. This is illustrated in figure 3. The evoked hemodynamic response $r$ for $T$, plus the normally distributed error, is then computed via

$$r = y^W - X_T^W \hat{\beta}_w \tag{8}$$

On the other hand, the estimated hemodynamic response for $T$ from the GLM model is $(X - X_T)^W \hat{\beta}_w$, which is potentially non-zero from the task onset for up to 20 seconds (given that the temporal basis set is up to 17 seconds long, plus 3 seconds of task duration). We will consider only this segment of 20 seconds in the estimated and extracted hemodynamic response from now on.

Finally, the extracted hemodynamic response $r$ was classified by regularized logistic regression with L2 norm (from scikit-learn [56], default settings but with a maximum number of iterations of 1000). The classifier input was the normalized response (z-scored, mean subtraction and division by standard deviation based on extracted responses from training blocks) for each time point separately.

We explored classification performance using time windows of different lengths for the hemodynamic responses starting at the task onset. The time lag between the onset of the first task and the end of the last task in a trial is 12.6 seconds. Therefore, any significant improvement in classification accuracies past these times is most probably an artifact of "overfitting" due to the short gap of 0.2 seconds between tasks. A similar argument can be made for classification accuracies starting at the task onset as this is most probably due to the previous task. Considering a hemodynamic delay of up to 6 seconds plus the task duration of up to 3 seconds, anything outside this region is most likely to be an artifact. Thus, we restrict our analysis to a maximal window length of 10 seconds. Most fMRI and fNIRS studies use a maximal temporal window of up to 9 seconds [8, 3, 4, 2, 9].

## 3. Results

### *3.1. Univariate classification*

As already mentioned, each fNIRS channel was separately modeled by the GLM to remove noise and influences from previous tasks and to extract evoked hemodynamic responses that were then classified by regularized logistic regression for univariate classification. The whole pipeline was applied within a 15-block-wise cross-validation. The classifier can be said to have a classification accuracy that is statistically significantly above chance level when the accuracy is above 56.11% for a significance level of $p < 0.05$ (one-sided Binomial test ($n = 180$)), 58.89% for a significance level of $p < 0.01$, and 61.67% for a significance level of $p < 0.001$.

Figure 4 shows classification accuracies of all channels for each participant using different window lengths of hemodynamic responses starting at the time of task onset. As the window length increases more and more channels can be individually used to classify the semantic categories at a significant level of accuracy ($p < 0.05$) and the classification accuracies increase for all tasks and both montages. Each participant may have a different internal neural representation of the concepts and, thus, a different response in each channel. Figure 5 shows mean classification accuracies across participants for all channels. It shows the same trend of increasing classification

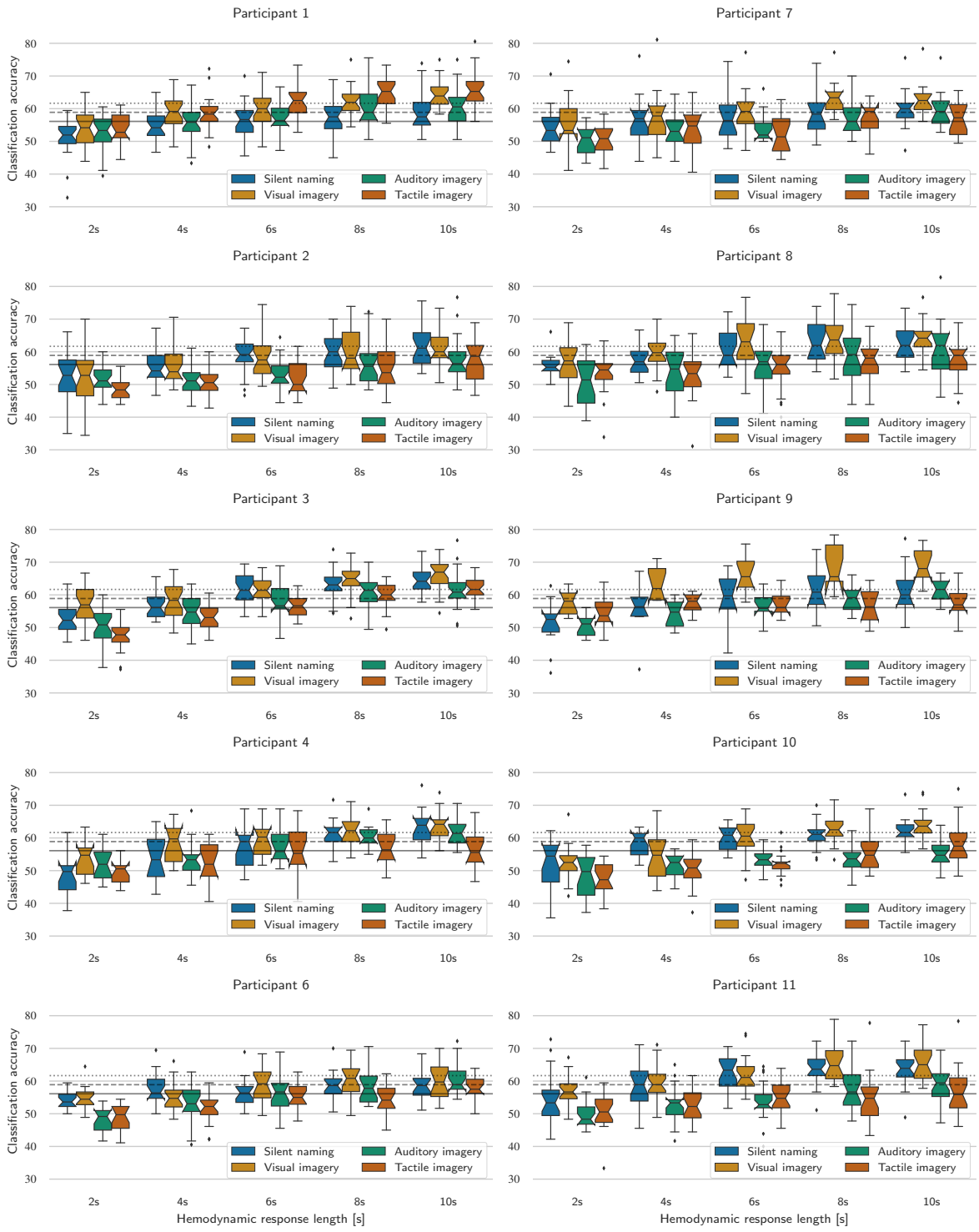


**Figure 4.** Univariate classification accuracies of all channels for each participant using different window lengths of hemodynamic responses. Horizontal lines represent a significance borderline for $p = 0.05$ (56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted). Participants with the frontal montage are in the left column and participants with the temporal montage are in the right column.

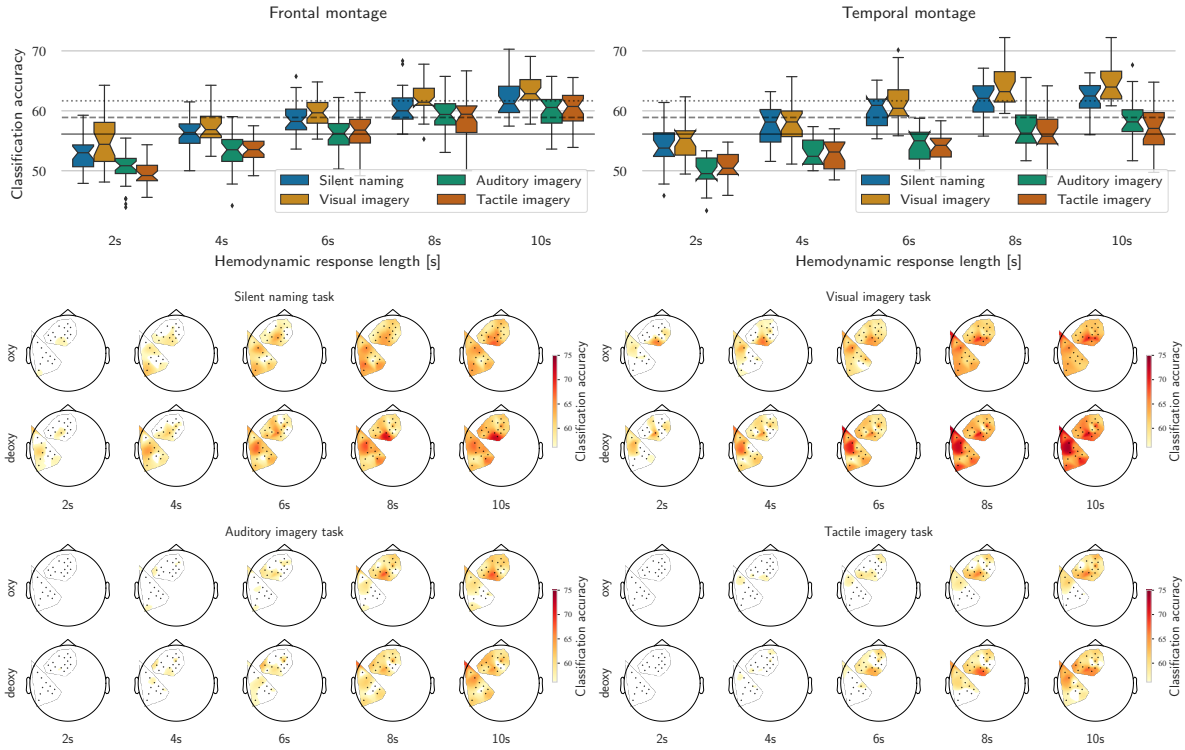


**Figure 5.** Univariate mean classification accuracies across participants for all channels using different window lengths of hemodynamic responses with corresponding scalp locations of the channels. Horizontal lines represent a significance borderline for $p = 0.05$ (56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted) for a single participant, similarly to figure 4. Scalp maps indicate performance above the significance borderline ($p = 0.05$, 56.11%).

accuracies with increasing window length as seen for single participant classification. This indicates that some channels give consistently good semantic category classification accuracies across different participants. These channels varied between tasks. Their spatial distribution over time is also illustrated in figure 5. The average accuracy is calculated only from two to five participants, depending on which particular channel is used, and thus any reasonable statistical testing is not possible. Thus we will use the same statistical thresholds as for a single participant in all the following analysis.

We inspect participant 6 in more detail as this is a typical participant. Figure 6 shows estimated and extracted hemodynamic responses for animals and tools for channel AFF5h-FFC5h when trained using all data (blocks). The generality of the temporal basis set adopted allows the timing of the main peak, if present, to vary for each channel type and task. Estimated and extracted hemodynamic responses follow the expected relationship between changes in oxygenated and deoxygenated hemoglobin. Dotted lines show classification accuracies from cross-validation using different window lengths of extracted responses. We can directly see which times are important for the differentiation between animals and tools. For instance, in the visual imagery task, classification accuracies are above chance level ($p = 0.05$, 56.11%) after 2.3 and 4.6

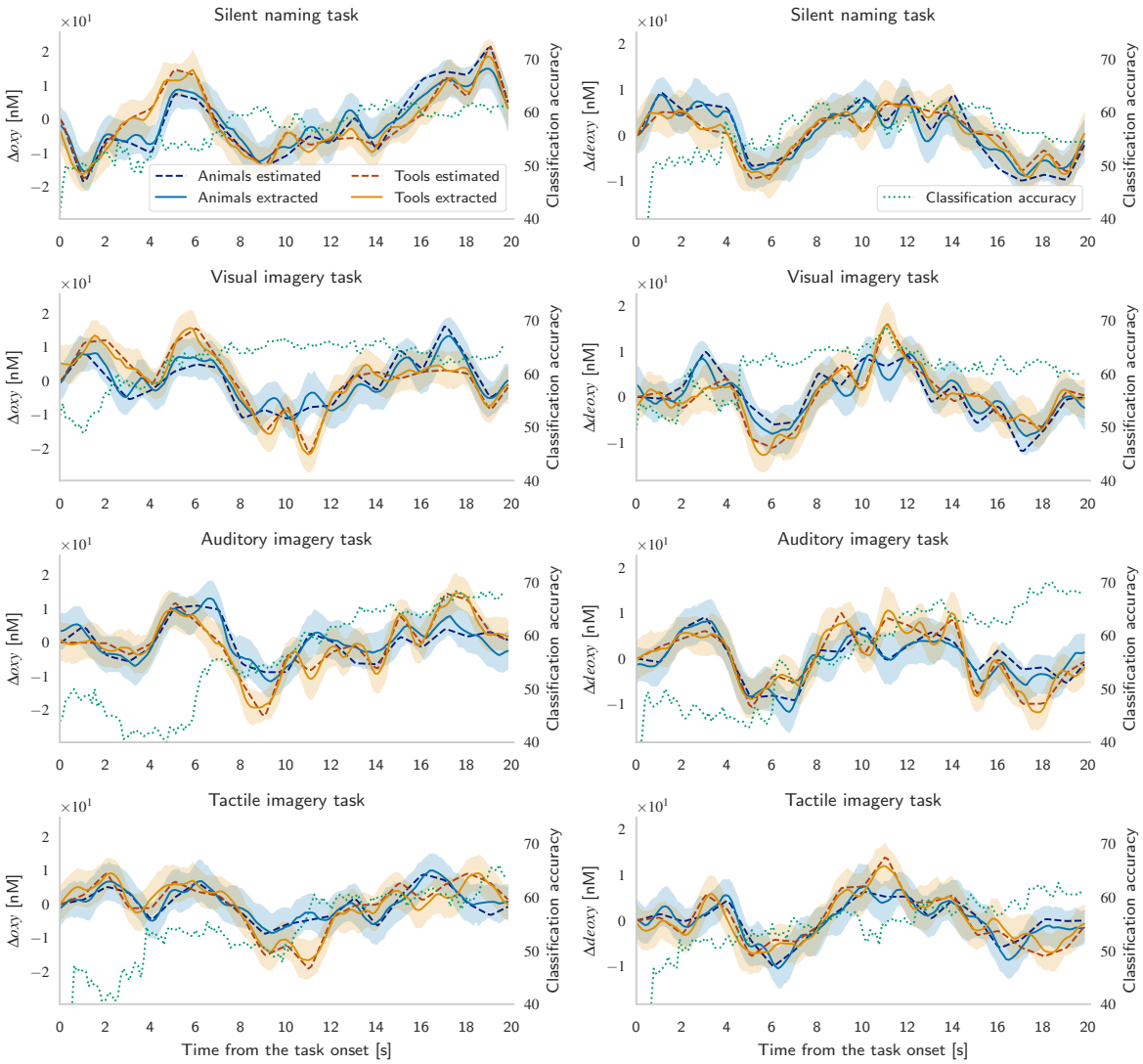


**Figure 6.** Estimated and extracted hemodynamic responses of animals and tools for channel AFF5h-FFC5h from participant 6 when trained using all data (blocks). Extracted responses are shown with mean and 95% confidence interval. Changes in oxygenated hemoglobin are in the left column and changes in deoxygenated hemoglobin are in the right column. Dotted lines represent classification accuracies from cross-validation using corresponding window lengths of extracted responses.

seconds for changes in oxygenated and deoxygenated hemoglobin respectively.

### *3.2. Multivariate classification with channel selection*

We explored the feasibility of aggregating multiple channels for classification. We used 15-block-wise nested-cross-validation for automatic channel selection. The $n$ channels with the highest classification accuracies from the inner cross-validation were selected. The selected channels were used for soft voting (sum of classifier probabilities) on the test block of the outer cross-validation.

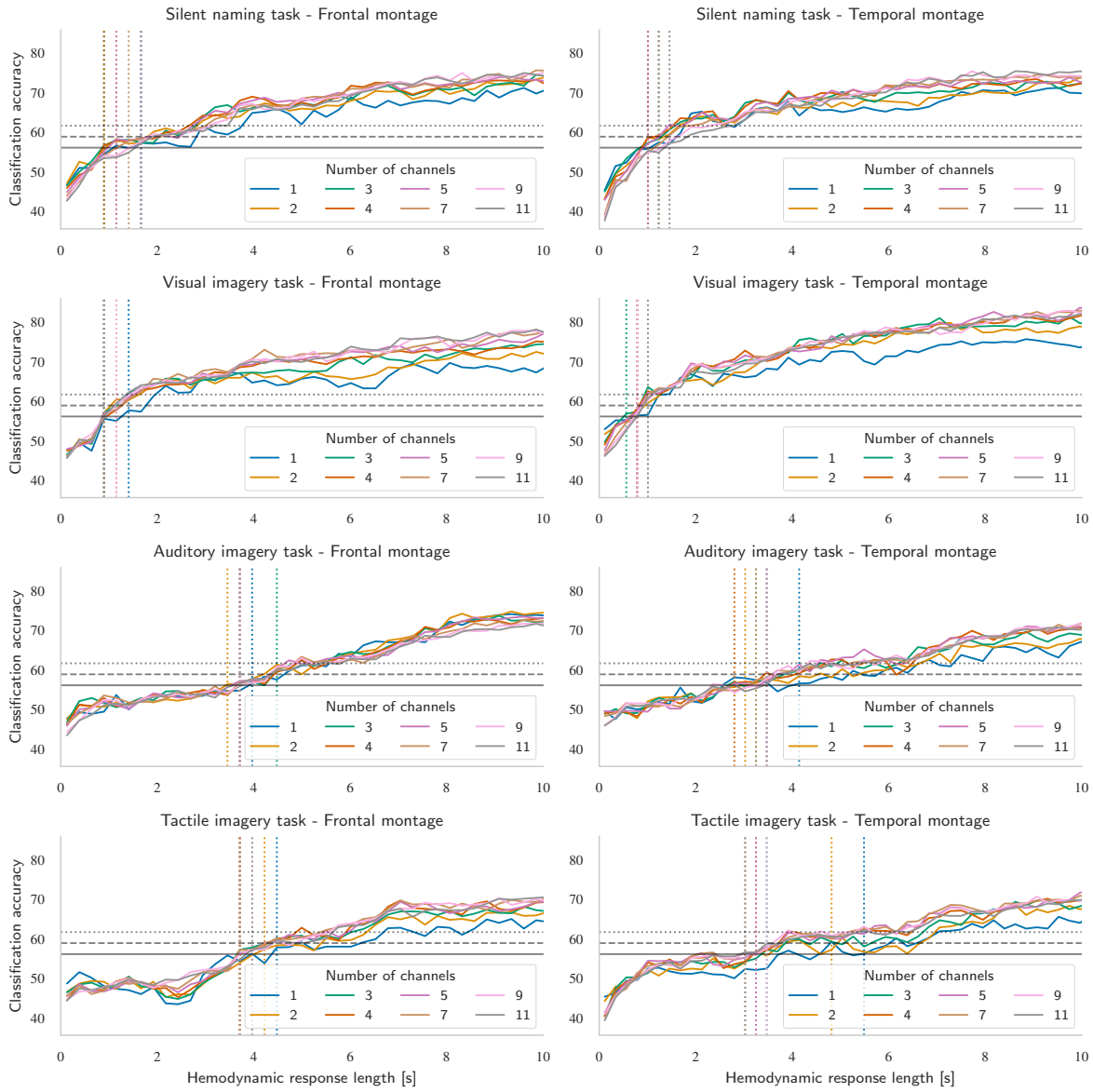


**Figure 7.** Mean classification accuracies across participants from nested-cross-validation for automatic channel selection using different window lengths of hemodynamic responses. Participants with the frontal montage are in the left column and participants with the temporal montage are in the right column. Horizontal lines represent a significance borderline for $p = 0.05$ (56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted) for a single participant, similarly to figure 4. Dotted vertical lines represent time points (i.e., lengths of responses) for each number of channels using the same color from when mean accuracies stayed above the significance borderline ($p = 0.05$, 56.11%).

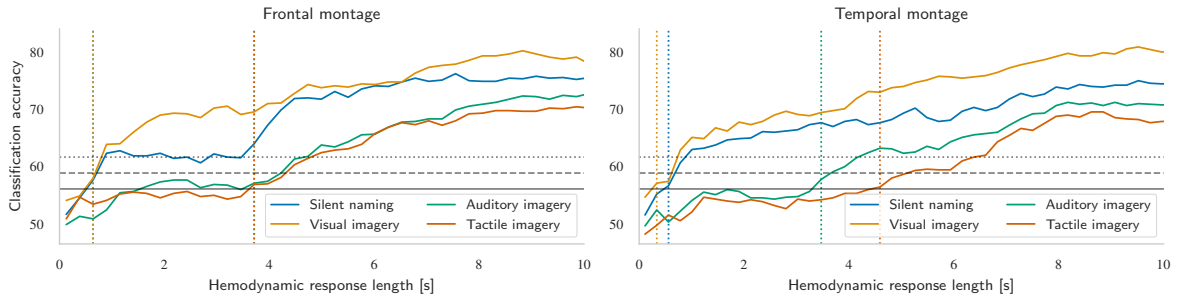


**Figure 8.** Mean multivariate classification accuracies across participants when the classifier is trained using all channels. Horizontal lines represent a significance borderline for $p = 0.05$ (56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted) for a single participant, similarly to figure 4. Dotted vertical lines represent time points (i.e., lengths of responses) for each task using the same color from when mean accuracies stayed above the significance borderline ($p = 0.05$, 56.11%).

Figure 7 shows mean classification accuracies across participants from this nested-cross-validation for $n \in \{1, 2, ..., 12\}$. The results for each task had a similar behavior for both montages. Mean classification accuracies stayed above the threshold of statistical significance ($p = 0.05$, 56.11%) from between 0.9 seconds (for the visual imagery task) to 3.8 seconds (for the tactile imagery task). In general, a shorter time window was needed to reach statistically better-than-chance classification for the silent naming task and the visual imagery task, while a longer window was needed for the auditory and tactile imagery tasks.

### *3.3. Multivariate classification of all channels*

Results from univariate classification showed that most channels are suitable for classification, especially when using the silent naming task and the visual imagery task. This suggests the possibility of using all channels for the classification rather than selecting a subset of channels. Instead of using one classifier for each channel, extracted responses from all channels were concatenated, normalized as before, and input into the regularized logistic regression during the cross-validation.

Figure 8 shows the mean classification accuracies across participants when classifiers were trained on all channels. Mean classification accuracies stayed above the threshold of statistical significance ($p = 0.05$, 56.11%) from between 0.4 seconds (for the visual imagery task) to 4.2 seconds (for the tactile imagery task). Similar to previous results, a shorter time window was needed for the silent naming and the visual imagery task, while a longer window was needed for the auditory and tactile imagery tasks.

### *3.4. Consistent representations across tasks*

To inspect whether the current results are task-specific or whether there is some "universal" neural representation of the semantic categories of animals and tools that

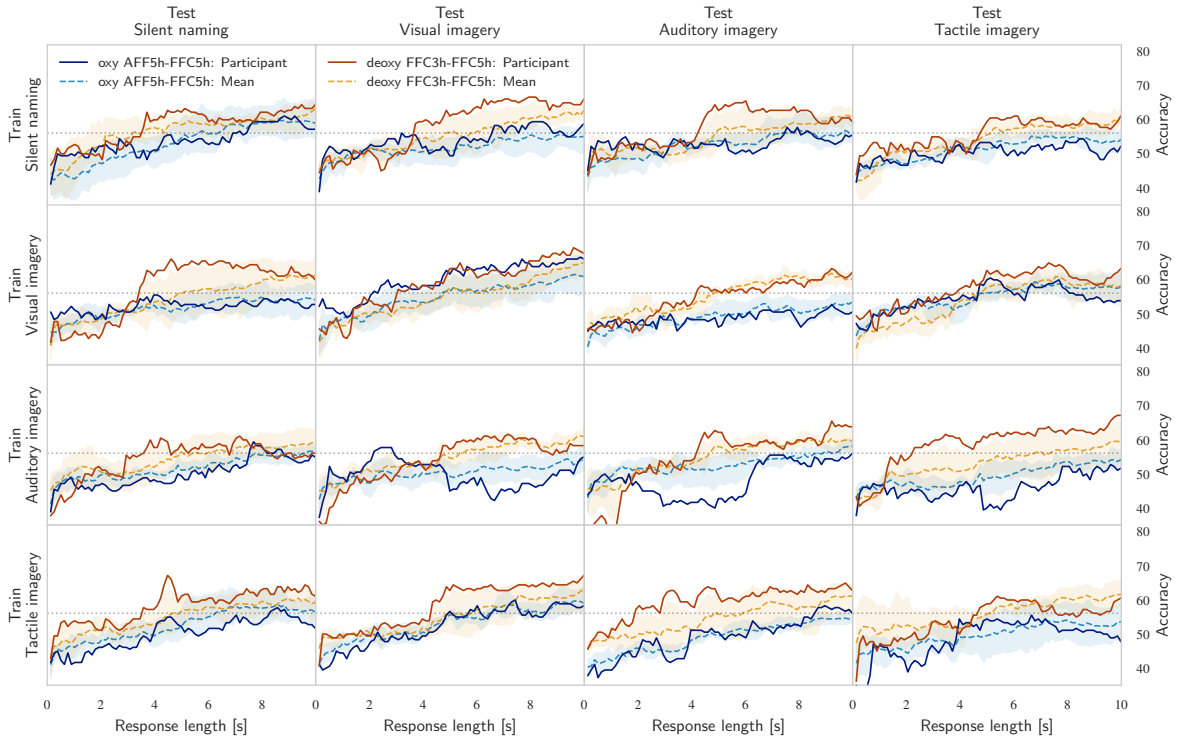


**Figure 9.** Univariate classification accuracies for participant 6 (solid lines) and mean classification accuracies with 95% confidence interval across all participants with the same montage (dashed lines) for two typical channels oxy AFF5h-FFC5h and deoxy FFC3h-FFC5h when the classifier was trained on one particular task and tested on a different task. Rows represent tasks used for training and columns represent tasks used for testing. Horizontal dotted lines represent a significance borderline for $p = 0.05$ (56.11%) for a single participant.

is consistent across tasks, we trained classifiers on one particular task and then tested them on a different task (from the test fold during the cross-validation).

First, we inspected the results from univariate classification. Figure 9 shows classification accuracies for two typical channels from participant 6 and the mean accuracies, for those same channels, across all the participants with the same montage. In channel oxy AFF5h-FFC5h, most classifiers for participant 6 across tasks struggled to stay above the chance level. In some instances, they crossed the chance level but then dropped below again when the window length was increased. Mean classification accuracies of oxy AFF5h-FFC5h showed the same struggle to generalize across tasks. On the other hand, channel deoxy FFC3h-FFC5h showed the opposite behavior. Most classifiers for participant 6 could be interchanged across tasks while staying above the chance level. This behavior is also observed when inspecting mean classification accuracies. This suggests the possibility to generalize across tasks if the correct fNIRS channels are chosen.

Furthermore, figure 10 shows results of the same analysis when using all channels for multivariate classification, as in section 3.3. Mean classification accuracies across

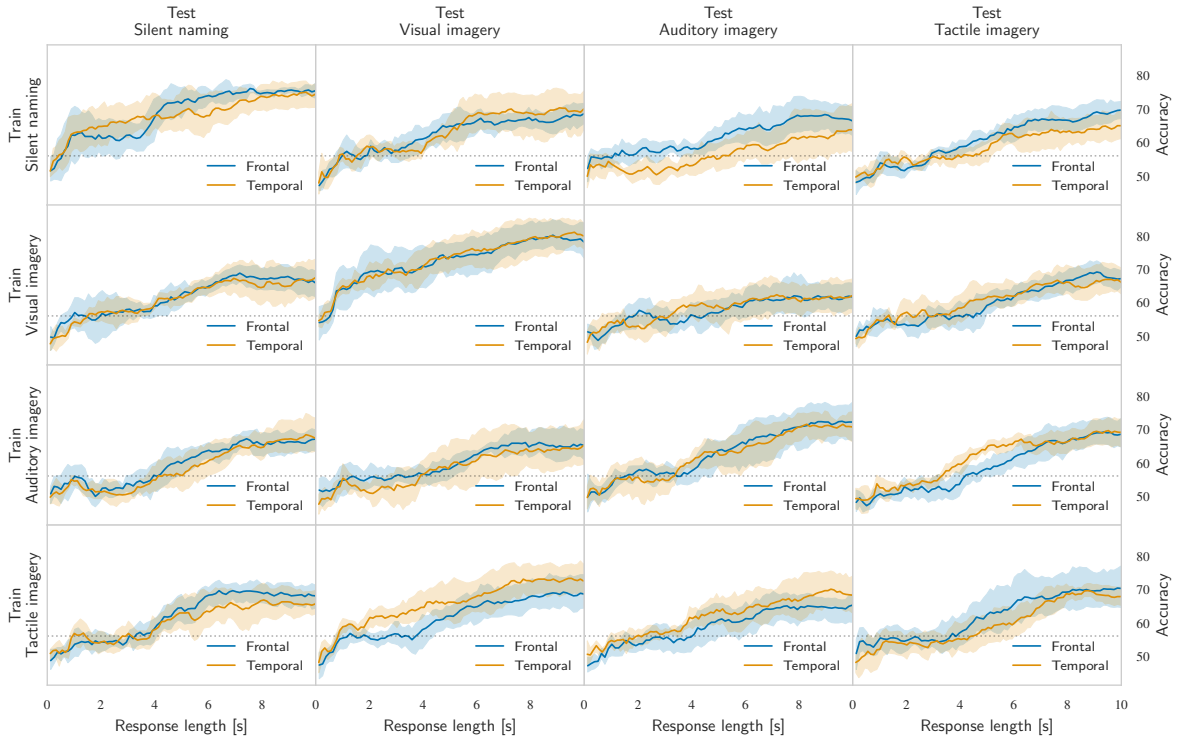


**Figure 10.** Mean multivariate classification accuracies across participants for each montage when the classifier was trained using all channels on one particular task and tested on a different task. Rows represent tasks used for training and columns represent tasks used for testing. Accuracies are shown with mean and 95% confidence interval. Horizontal dotted lines represent a significance borderline for $p = 0.05$ (56.11%) for a single participant.

participants show that all classifiers could be interchanged across different tasks with an increasing window length of responses. This strongly supports the idea that there are consistent neural representations of the semantic categories across different imagery tasks.

### *3.5. Order of tasks*

In the course of our analysis, we discovered a minor error in our experiment script that was used to present stimuli to participants. All participants but participant 1 had the same seed for the random number generator and they thus shared the same order of tasks and concepts. We will focus on these participants and ignore participant 1 in this section. Figure 11 shows how many times tasks appeared as the first, second, third, and fourth task in blocks. The distribution is clearly not uniform. It was skewed to present the auditory and tactile imagery tasks more frequently as the first task and the auditory imagery task as the second task. This might potentially explain why longer time windows were needed for the auditory and tactile imagery tasks. These tasks followed after the initial stimuli image and mask presentation more often and

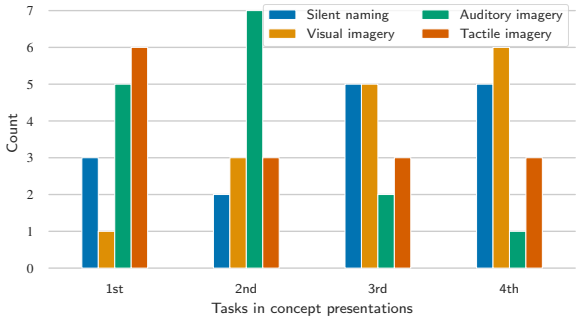


**Figure 11.** Number of task appearances as the first, second, third, and fourth task in the shared order of tasks in participants 2 to 12.

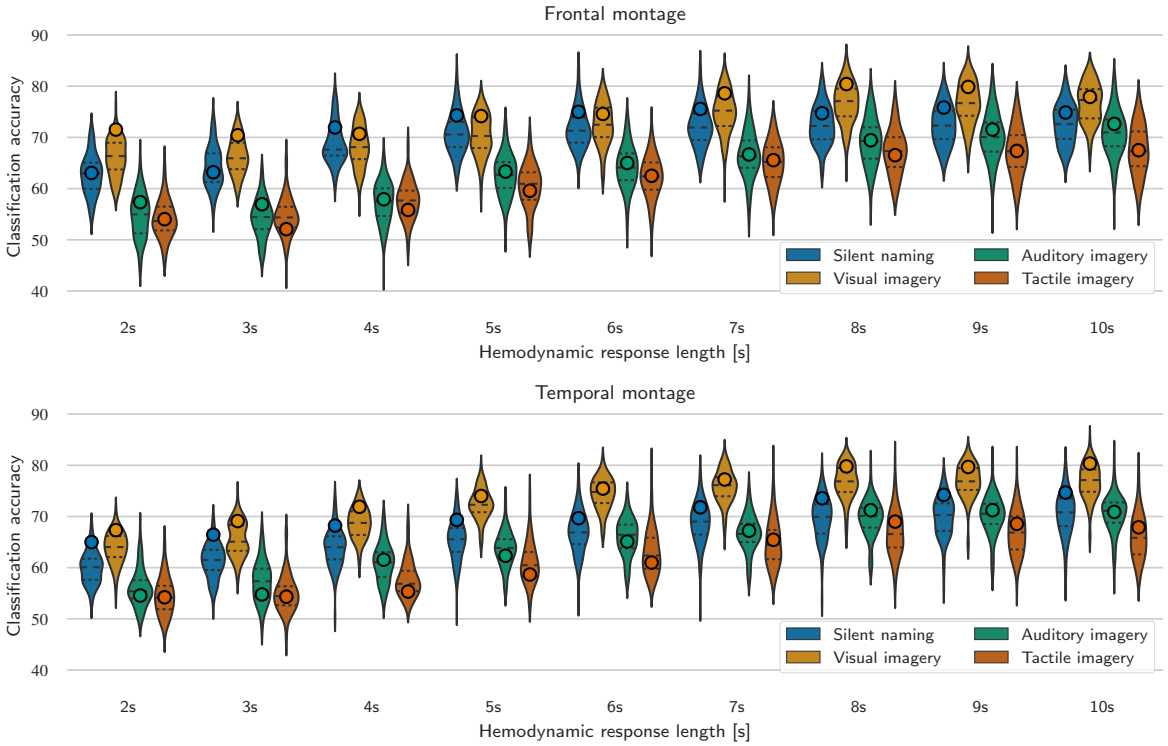


**Figure 12.** Mean multivariate classification accuracies using all channels across participants for each subset of blocks simulating different distributions of task appearance orders. Distributions of accuracies are shown with quartiles. Circles represent the results found when the complete set of data is used in the analysis.

participants then needed to switch to a new concept, which may explain why longer time windows were needed to classify these tasks.

We performed a post-hoc analysis to investigate how much our results depend on different distributions of the order of task appearances. We re-computed our analysis on subsets of blocks from the experiment by excluding individual blocks from the analysis to simulate different distributions. We used subsets of blocks with sizes from 7 to 14 blocks and only considered subsets in which all tasks appeared as the first, second, third, and

fourth task. From these 4023 subsets, 100 were randomly selected. Figure 12 shows the mean multivariate classification accuracies using all channels, as in section 3.3, across participants for each subset of blocks. Our results from the complete set of data visually appear to fall within the "middle" of the distribution. Mean classification accuracies across all subsets were above 60% for all tasks from 5 seconds onwards. Thus, our results would be significant even if the order of tasks had been correctly randomized across participants. In other words, the error we discovered in our experiment script did not adversely affect our results.

### *3.6. Possible confounds*

Components of fNIRS signals that are not driven by neuronal activity (i.e., neurovascular coupling) can act as possible confounds. These components include extracerebral hemodynamics and intracerebral hemodynamics caused by task-related systemic activity. They act as confounds as they may mimic the presence of, or attenuate, the neuronally induced hemodynamic response [57, 58, 59]. In our study, the inclusion of the respiration signal as a regressor in the GLM and autoregressive model reduced the influence of only some of these components. However, we also tested using two preprocessing methods for fNIRS signals before the model estimation to further mitigate this issue in the multivariate classification, as in section 3.3. The first method was based on the principal component analysis to remove global interference by removing an eigenvector with the highest eigenvalue [60, 61]. The second method was the common average reference [62]. Both methods were applied to channels of oxygenated and deoxygenated channels separately. Mean classification accuracies from both methods were not statistically different in comparison with the results in section 3.3 (two-sided Wilcoxon test corrected with the false discovery rate). Based on these results, we believe that there is not a significant interference from the aforementioned sources in the classification.

## 4. Discussion

We showed that it is possible to differentiate between the semantic categories of animals and tools in all mental tasks with mean classification accuracies of 76.2% for the silent naming task, 80.9% for the visual imagery task, 72.8% for the auditory imagery task, and 70.4% for the tactile imagery task using window lengths of hemodynamic responses of up to 10 seconds long in the multivariate classification approach using all channels.

Our results show that generalization across different tasks is possible to a certain degree (in certain channels), and thus it seems that the classifier exploits common features of the neural representations of animals and tools across different tasks. Another possible interpretation is that information from preceding and following task(s) could be exploited for classification due to the short gap of 0.2 seconds between tasks. This issue should be partly mitigated by the random order that the tasks were presented

across different blocks. However, we cannot be sure purely based on this experiment.

A presentation of stimulus image causes semantic priming. This form of visual priming could enhance performance of the visual imagery task in comparison with using non-visual stimulus [63, 24, 64]. This prior semantic processing can also enhance subsequent picture naming performance [65]. Additionally, participants could recall the object name during the presentation of the stimulus image and kept it in mind for all subsequent tasks. This could explain the difference in shorter time windows that were needed for the silent naming and visual imagery task in comparison with longer windows for the auditory and tactile imagery task.

Further research is needed to investigate which modality of mental imagery is the most suitable for semantic decoding and whether a single modality or a sequence of modalities performs better. Multimodal (multisensory) mental imagery could be especially promising [66, 67, 68]. In informal talks with participants during and after the experiment, some participants mentioned that they were unable to distinguish between different tasks. For instance, imagining touching an object included visualizing the object itself. This is not a problem for us because our goal is to choose mental tasks that elicit clear neural correlates for differentiation between the semantic categories. These questions are ultimately interconnected with imagery mental strategies and task instructions given to participants. Nevertheless, imagery vividness and strength varies between individuals [69, 70]. In potential BCI applications, the mental imagery strategy should be specifically selected for each individual according to their abilities and needs.

It is important to note that this study has been conducted offline. It is currently unknown whether similar results could be achieved in an online experimental paradigm for BCI applications. Online implementation entails several challenges. For instance, preprocessing, mainly motion correction and filtering, must be replaced with online alternatives. However, the method presented here, based on the GLM to extract the evoked hemodynamic response for a task trial, can be used in online analysis and classification by modifying the experimental design. Intervals between task trials must be appropriately chosen in order to avoid, or minimize, the influence between their hemodynamic responses.

In follow-up research, we will analyze recorded EEG signals and/or the combination of EEG and fNIRS for classification. We will make the recorded data publicly available afterwards. We intend to replicate the results presented in this paper in an online experimental paradigm. Also, future experiments should use additional short-distance detectors to detect and directly remove extracerebral signals close to the source [71, 72].

### *4.1. Silent naming task*

Murphy *et al.* [6] were able to differentiate between semantic categories of mammals and tools in EEG with a mean classification accuracy of 72% under 0.5 second after the stimulus onset, while participants viewed images and performed the silent naming task. Due to the hemodynamic delay, fNIRS signals cannot compete with a similar

short duration after the task onset for classification. Similar mean accuracies above 70% were achieved after 4.5 seconds in the frontal montage and after 6.5 seconds in the temporal montage while accuracies stayed above the chance level (56.1%, $p < 0.05$) from 0.6 seconds in both montages.

### *4.2. Imagery tasks*

The most similar mental tasks in comparison with our proposed imagery tasks are tasks based on imagining of object properties.

Shinkareva *et al.* [8] were able to differentiate between semantic categories of tools and dwellings in fMRI, while participants viewed line drawings for 3 seconds followed by a 7 seconds rest period. Participants were asked to think of the same object properties generated before the recording. A single trial-level data set for classification was computed by averaging between 4 and 7 seconds after the stimulus onset. Mean classification accuracies were 87%.

Akama *et al.* [3] were able to differentiate between semantic categories of mammals and tools in fMRI, while participants viewed photographs with either an accompanying caption or with a spoken word description of the object in different sessions. As in the study above, participants were asked to silently enumerate concept properties that were generated before the recording. Each stimulus was presented for 3 seconds followed by a rest period for 7 seconds. A single trial-level fMRI image for classification was computed by averaging consecutive fMRI volumes between 4 and 7 seconds after the stimulus onset. Uni-modal prediction achieved accuracies in the 80-90% range and cross-modal classification accuracies were in the 65-75% range. In the same experimental design but using only accompanying captions in 2 different languages for each participant in different sessions, Akama *et al.* [4] achieved mean uni-modal classification accuracies of 90.8-93.4% and cross-modal classification accuracies (different languages) of 82.9-86.7%.

In comparison with these fMRI studies, our mean classification accuracies using hemodynamic response lengths of up to 7 seconds were 77.2% for the visual imagery task, 67.8% for the auditory imagery task, and 66.7% for the tactile imagery task. The accuracies are lower, as expected, because fNIRS can only measure cortical brain activity up to about 2 cm deep and has a smaller spatial resolution. On the other hand, our experimental design did not have the canonical 7 seconds rest period in between tasks.

In fNIRS, Zinszer *et al.* [9] discriminated between semantic categories of animals and body parts while participants focused on audiovisual stimuli (photographs with a simultaneous auditory presentation of object name) and thought about the meaning of that stimulus or any memory it evoked. Each stimulus was presented for 3 seconds and followed by an interstimulus interval of 6-9 seconds composed of fireworks and a short musical clip. Mean accuracies were 66%. Data were epoched in 6.5 to 9.0 seconds after the stimulus onset.

In comparison, our mean classification accuracies using hemodynamic response lengths of up to 9 seconds were 80.1% for the visual imagery task, 71.8% for the auditory

imagery task, and 69.7% for the tactile imagery task.

## 5. Conclusion

We demonstrated that semantic decoding is possible in fNIRS by differentiating between the semantic categories of animals and tools in all mental tasks with mean classification accuracies above 70%. Selection of suitable mental tasks is the first step towards intuitive BCI applications for communication. In follow-up research, we will include recorded EEG signals into classification. Also, we intend to replicate the results presented in this paper in an online experimental paradigm.

## References

[1] Hiroyuki Akama and Brian Murphy. Emerging methods for conceptual modelling in neuroimaging. *Behaviormetrika*, 44(1):117–133, jan 2017.

[2] Irina Simanova, Peter Hagoort, Robert Oostenveld, and Marcel A J van Gerven. Modality-Independent Decoding of Semantic Information from the Human Brain. *Cerebral Cortex*, 24(2):426–434, feb 2014.

[3] Hiroyuki Akama, Brian Murphy, Li Na, Yumiko Shimizu, and Massimo Poesio. Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. *Frontiers in Neuroinformatics*, 6:24, aug 2012.

[4] Hiroyuki Akama, Brian Murphy, Miao Mei Lei, and Massimo Poesio. Cross-participant modelling based on joint or disjoint feature selection: an fMRI conceptual decoding study. *Applied Informatics*, 1(1):1, dec 2014.

[5] Irina Simanova, Marcel van Gerven, Robert Oostenveld, and Peter Hagoort. Identifying Object Categories from Event-Related EEG: Toward Decoding of Conceptual Representations. *PLoS ONE*, 5(12):e14465, dec 2010.

[6] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, 117(1):12–22, apr 2011.

[7] Brian Murphy, Marco Baroni, and Massimo Poesio. EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, volume 2, pages 619–627, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[8] Svetlana V. Shinkareva, Robert A. Mason, Vicente L. Malave, Wei Wang, Tom M. Mitchell, and Marcel Adam Just. Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. *PLoS ONE*, 3(1):e1394, jan 2008.

[9] Benjamin D. Zinszer, Laurie Bayet, Lauren L. Emberson, Rajeev D. S. Raizada, and Richard N. Aslin. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003, aug 2017.

[10] Guillaume A. Rousselet, M. J.-M. Mace, and Michèle Fabre-Thorpe. Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component? *Journal of Vision*, 4(1):13–21, jan 2004.

[11] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M. Norcia, and Patrick Suppes. A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *PLOS ONE*, 10(8):e0135697, aug 2015.

[12] Greta Tuckute, Sofie Therese Hansen, Nicolai Pedersen, Dea Steenstrup, and Lars Kai Hansen. Single-Trial Decoding of Scalp EEG under Natural Conditions. *Computational Intelligence and Neuroscience*, 2019:1–11, apr 2019.

[13] Alex Clarke, Barry J Devereux, Billi Randall, and Lorraine K Tyler. Predicting the Time Course of Individual Objects with MEG. *Cerebral Cortex*, 25(10):3602–3612, oct 2015.

[14] Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to Decode Cognitive States from Brain Images. *Machine Learning*, 57(1):145–175, oct 2004.

[15] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, K.-M. Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195, may 2008.

[16] Bernhard Graimann, Brendan Allison, and Gert Pfurtscheller. *Brain–Computer Interfaces: A Gentle Introduction*, pages 1–27. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[17] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain Computer Interfaces, a Review. *Sensors*, 12(2):1211–1279, jan 2012.

[18] Andrea Kübler, Adrian Furdea, Sebastian Halder, Eva Maria Hammer, Femke Nijboer, and Boris Kotchoubey. A Brain-Computer Interface Controlled Auditory Event-Related Potential (P300) Spelling System for Locked-In Patients. *Annals of the New York Academy of Sciences*, 1157(1):90–100, mar 2009.

[19] N. Jeremy Hill, Erin Ricci, Sameah Haider, Lynn M. McCane, Susan Heckman, Jonathan R. Wolpaw, and Theresa M. Vaughan. A practical, intuitive brain–computer interface for communicating 'yes' or 'no' by listening. *Journal of Neural Engineering*, 11(3):035003, jun 2014.

[20] Sonja C. Kleih, Andreas Herweg, Tobias Kaufmann, Pit Staiger-Sälzer, Natascha Gerstner, and Andrea Kübler. The WIN-speller: a new intuitive auditory brain-computer interface spelling application. *Frontiers in Neuroscience*, 9:346, oct 2015.

[21] Eman Albilali, Hatim Aboalsamh, and Areej Al-Wabil. Comparing brain-computer interaction and eye tracking as input modalities: An exploratory study. In *2013 International Conference on Current Trends in Information Technology (CTIT)*, pages 232–236. IEEE, dec 2013.

[22] Kaori Suefusa and Toshihisa Tanaka. A comparison study of visually stimulated brain–computer and eye-tracking interfaces. *Journal of Neural Engineering*, 14(3):036009, jun 2017.

[23] Emanuele Pasqualotto, Tamara Matuz, Stefano Federici, Carolin A. Ruf, Mathias Bartl, Marta Olivetti Belardinelli, Niels Birbaumer, and Sebastian Halder. Usability and Workload of Access Technology for People With Severe Motor Impairment. *Neurorehabilitation and Neural Repair*, 29(10):950–957, 2015.

[24] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634, oct 2019.

[25] Radoslaw M. Cichy, Jakob Heinzle, and John-Dylan Haynes. Imagery and Perception Share Cortical Representations of Content and Location. *Cerebral Cortex*, 22(2):372–380, feb 2012.

[26] Sue-Hyun Lee, Dwight J. Kravitz, and Chris I. Baker. Disentangling visual imagery and perception of real-world objects. *NeuroImage*, 59(4):4064–4073, feb 2012.

[27] Leila Reddy, Naotsugu Tsuchiya, and Thomas Serre. Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2):818–825, apr 2010.

[28] Jeroen Geuze, Jason Farquhar, and Peter Desain. Towards a Communication Brain Computer Interface Based on Semantic Relations. *PLoS ONE*, 9(2):e87511, feb 2014.

[29] Jeroen Geuze, Marcel A. J. van Gerven, Jason Farquhar, and Peter Desain. Detecting Semantic Priming at the Single-Trial Level. *PLoS ONE*, 8(4):e60377, apr 2013.

[30] Markus Kiefer. Perceptual and semantic sources of category-specific effects: Event-related potentials during picture and word categorization. *Memory & Cognition*, 29(1):100–116, jan 2001.

[31] Guillaume A. Rousselet, Jesse S. Husk, Patrick J. Bennett, and Allison B. Sekuler. Single-trial EEG dynamics of object and face visual processing. *NeuroImage*, 36(3):843–862, jul 2007.

[32] E. T. Rolls and M. J. Tovee. Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London. Series B: Biological Sciences*,

257(1348):9–15, jul 1994.

[33] William Penny, Karl Friston, John Ashburner, Stefan Kiebel, and Thomas Nichols. *Statistical Parametric Mapping*. Elsevier, 2007.

[34] Ronald Christensen. *Plane Answers to Complex Questions*. Springer Texts in Statistics. Springer New York, New York, NY, 2011.

[35] Ying Yang, Jing Wang, Cyntia Bailer, Vladimir Cherkassky, and Marcel Adam Just. Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. *NeuroImage*, 146:658–666, feb 2017.

[36] Benjamin D. Zinszer, Andrew J. Anderson, Olivia Kang, Thalia Wheatley, and Rajeev D. S. Raizada. Semantic Structural Alignment of Neural Representational Spaces Enables Translation between English and Chinese Words. *Journal of Cognitive Neuroscience*, 28(11):1749–1759, nov 2016.

[37] Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12):2767–2796, dec 2009.

[38] Guilherme Augusto Zimeo Morais, Joana Bisol Balardin, and João Ricardo Sato. fNIRS Optodes' Location Decider (fOLD): a toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, 8(1):3341, dec 2018.

[39] Luca Pollonini, Cristen Olds, Homer Abaya, Heather Bortfeld, Michael S. Beauchamp, and John S. Oghalai. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hearing Research*, 309:84–93, mar 2014.

[40] George Themelis, Juliette Selb, Sonai Thaker, Jonathan J. Stott, Anna Custo, David Boas, and Maria Angela Franceschini. Depth of arterial oscillation resolved with NIRS time and frequency domain. In *Biomedical Topical Meeting*, page WF2, Washington, D.C., jul 2004. OSA.

[41] George Themelis, Helen D'Arceuil, Solomon G. Diamond, Sonal Thaker, Theodore J. Huppert, David A. Boas, and Maria Angela Franceschini. Near-infrared spectroscopy measurement of the pulsatile component of cerebral blood flow and volume from arterial oscillations. *Journal of Biomedical Optics*, 12(1):014033, 2007.

[42] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, mar 2020.

[43] Behnam Molavi and Guy A Dumont. Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological Measurement*, 33(2):259–270, feb 2012.

[44] Robert J. Cooper, Juliette Selb, Louis Gagnon, Dorte Phillip, Henrik W. Schytz, Helle K. Iversen, Messoud Ashina, and David A. Boas. A Systematic Comparison of Motion Artifact Correction Techniques for Functional Near-Infrared Spectroscopy. *Frontiers in Neuroscience*, 6(OCT):147, oct 2012.

[45] Sabrina Brigadoi, Lisa Ceccherini, Simone Cutini, Fabio Scarpa, Pietro Scatturin, Juliette Selb, Louis Gagnon, David A. Boas, and Robert J. Cooper. Motion artifacts in functional near-infrared spectroscopy: A comparison of motion correction techniques applied to real cognitive data. *NeuroImage*, 85:181–191, jan 2014.

[46] Theodore J. Huppert, Solomon G. Diamond, Maria A. Franceschini, and David A. Boas. HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied Optics*, 48(10):D280–D298, apr 2009.

[47] D. T. Delpy, M. Cope, P van der Zee, S. Arridge, S. Wray, and J. Wyatt. Estimation of optical

pathlength through tissue from direct time of flight measurement. *Physics in Medicine and Biology*, 33(12):1433–1442, dec 1988.

[48] Matthias L. Schroeter, Markus M. Bücheler, Karsten Müller, Kâmil Uludağ, Hellmuth Obrig, Gabriele Lohmann, Marc Tittgemeyer, Arno Villringer, and D.Yves von Cramon. Towards a standard analysis for functional near-infrared imaging. *NeuroImage*, 21(1):283–290, jan 2004.

[49] M.M. Plichta, S. Heinzel, A.-C. Ehlis, P. Pauli, and A.J. Fallgatter. Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data: A parametric validation study. *NeuroImage*, 35(2):625–634, apr 2007.

[50] Peck H. Koh, Daniel E. Glaser, Guillaume Flandin, Stefan Kiebel, Brian Butterworth, Atsushi Maki, David T. Delpy, and Clare E. Elwell. Functional optical signal analysis: a software tool for near-infrared spectroscopy data processing incorporating statistical parametric mapping. *Journal of Biomedical Optics*, 12(6):064010, 2007.

[51] Jong Chul Ye, Sungho Tak, Kwang Eun Jang, Jinwook Jung, and Jaeduck Jang. NIRS-SPM: Statistical parametric mapping for near-infrared spectroscopy. *NeuroImage*, 44(2):428–447, jan 2009.

[52] T.J. Huppert, R.D. Hoge, S.G. Diamond, M.A. Franceschini, and D.A. Boas. A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *NeuroImage*, 29(2):368–382, jan 2006.

[53] Markus J. Hofmann, Martin J. Herrmann, Ippeita Dan, Hellmuth Obrig, Markus Conrad, Lars Kuchinke, Arthur M. Jacobs, and Andreas J. Fallgatter. Differential activation of frontal and parietal regions during visual word recognition: An optical topography study. *NeuroImage*, 40(3):1340–1349, apr 2008.

[54] Jeffrey W. Barker, Ardalan Aarabi, and Theodore J. Huppert. Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical Optics Express*, 4(8):1366–1379, aug 2013.

[55] D. Cochrane and G. H. Orcutt. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61, mar 1949.

[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[57] Felix Scholkmann, Stefan Kleiser, Andreas Jaakko Metz, Raphael Zimmermann, Juan Mata Pavia, Ursula Wolf, and Martin Wolf. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage*, 85:6–27, jan 2014.

[58] Ilias Tachtsidis and Felix Scholkmann. False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward. *Neurophotonics*, 3(3):031405, mar 2016.

[59] Toshimitsu Takahashi, Yoriko Takikawa, Reiko Kawagoe, Satoshi Shibuya, Takayuki Iwano, and Shigeru Kitazawa. Influence of skin blood flow on near-infrared spectroscopy signals measured on the forehead during a verbal fluency task. *NeuroImage*, 57(3):991–1002, aug 2011.

[60] Yiheng Zhang, Dana H. Brooks, Maria Angela Franceschini, and David A. Boas. Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of Biomedical Optics*, 10(1):011014, 2005.

[61] Maria Angela Franceschini, Danny K. Joseph, Theodore J. Huppert, Solomon G. Diamond, and David A. Boas. Diffuse optical imaging of the whole head. *Journal of Biomedical Optics*, 11(5):054007, 2006.

[62] G. Bauernfeind, S. C. Wriessnegger, I. Daly, and G. R. Müller-Putz. Separating heart and brain: on the reduction of physiological noise from multichannel functional near-infrared spectroscopy (fNIRS) signals. *Journal of Neural Engineering*, 11(5):056010, oct 2014.

[63] Daniel L Schacter, Ian G Dobbins, and David M Schnyer. Specificity of priming: a cognitive neuroscience perspective. *Nature Reviews Neuroscience*, 5(11):853–862, nov 2004.

[64] Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9):635–642, sep 2001.

[65] Shiree Heath, Katie McMahon, Lyndsey Nickels, Anthony Angwin, Anna MacDonald, Sophia van Hees, Kori Johnson, and David Copland. Priming Picture Naming with a Semantic Task: An fMRI Investigation. *PLoS ONE*, 7(3):e32809, mar 2012.

[66] Bence Nanay. Multimodal mental imagery. *Cortex*, 105:125–134, aug 2018.

[67] Chris McNorgan. A meta-analytic review of multisensory imagery identifies the neural correlates of modality-specific and modality-general imagery. *Frontiers in Human Neuroscience*, 6(September):285, sep 2012.

[68] Simon Lacey and K. Sathian. Multisensory object representation. In *Progress in Brain Research*, volume 191, pages 165–176. Elsevier B.V., jan 2011.

[69] Joel Pearson, Thomas Naselaris, Emily A. Holmes, and Stephen M. Kosslyn. Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences*, 19(10):590–602, oct 2015.

[70] Xu Cui, Cameron B. Jeter, Dongni Yang, P. Read Montague, and David M. Eagleman. Vividness of mental imagery: Individual variability can be measured objectively. *Vision Research*, 47(4):474–478, feb 2007.

[71] Takanori Sato, Isao Nambu, Kotaro Takeda, Takatsugu Aihara, Okito Yamashita, Yuko Isogaya, Yoshihiro Inoue, Yohei Otaka, Yasuhiro Wada, Mitsuo Kawato, Masa-aki Sato, and Rieko Osu. Reduction of global interference of scalp-hemodynamics in functional near-infrared spectroscopy using short distance probes. *NeuroImage*, 141:120–132, nov 2016.

[72] Meryem A. Yücel, Juliette Selb, Christopher M. Aasted, Mike P. Petkov, Lino Becerra, David Borsook, and David A. Boas. Short separation regression improves statistical significance and better localizes the hemodynamic response obtained by near-infrared spectroscopy for tasks with differing autonomic responses. *Neurophotonics*, 2(3):035005, sep 2015.