Corrigendum: Decoding of semantic categories of imagined concepts of animals and tools in fNIRS (2021 J. Neural Eng. 18 046035)

Milan Rybář, Riccardo Poli, and Ian Daly

Brain-Computer Interfacing and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, UK

E-mail: {milan.rybar, rpoli, i.daly}@essex.ac.uk

In the published article, the evoked hemodynamic responses for each mental task were extracted using a general linear model (GLM) approach in Equation 8 (2.8 Classification section in the Methods section). However, all four mental tasks in one concept trial share the same underlying hidden variable about the semantic category. Thus, it is not possible to extract the evoked hemodynamic response for an individual mental task. Instead, the evoked hemodynamic response must be extracted from the sequence of four mental tasks. Hemodynamic response for each mental task must then be extracted from this sequence.

We fixed this issue and further validated our results by not using the GLM to extract the evoked hemodynamic responses. With our new results, we still demonstrated that semantic decoding is possible in fNIRS by differentiating between the semantic categories of animals and tools.

In this corrigendum, we describe the revised analyses, present the new results, and correct the parts influenced by the new results.

Corrections to Abstract

The new results changed the achieved classification accuracies thus the Main result part of the Abstract should read:

- It is possible to successfully classify the semantic category in each mental task for several participants with classification accuracies up to the range of 60–65%.

Corrections to 2. Methods

In this section, we correct the extraction of hemodynamic responses and present revised analyses.

Corrigendum: Decoding of semantic categories of imagined concepts of animals and tools in fNIRS (2021 J.

Single trial analysis

We performed a single trial analysis of two types of signals. First, the preprocessed data were used. We will refer to this data as 'no-GLM'. Second, the preprocessed fNIRS signals were modeled via a general linear model (GLM) [1, 2] to remove noise and influences from previous concept presentations. We will refer to this data as 'GLM'. The GLM approach is described in detail in Section 2.7 of the published article. Epochs of mental task trials were then extracted and further processed in the same way for both approaches.

We first extracted concept trial periods that start from the image presentation onsets and include the sequence of four mental tasks. Epochs of concept trial periods were extracted from 1 second before the image presentation (i.e., the last 1 second of the fixation cross) until 26 seconds after the image presentation. The last mental task starts 11.3 seconds after the image onset. So, this 27 seconds period thus contains 14.7 seconds of the hemodynamic response of the last mental task.

Epochs of concept trial periods were further preprocessed before any analysis by using: (i) linear detrending by subtracting each epoch's least squares fit and (ii) baseline correction by subtracting the mean of the 1 second period before the image presentation. All the following analyses were tested: (1) no detrending and no baseline correction, (2) detrending but no baseline correction, (3) no detrending but baseline correction, and (4) detrending and baseline correction. Although, there were differences between each setting, the overall message was the same for all of them. Only one setting is thus reported here to simplify the presentation in which only no detrending and no baseline correction (option (1) above) is used.

Finally, epochs of mental tasks were extracted from the preprocessed epochs of concept trial periods above. Mental task epochs start from the mental task onsets until 13.5 seconds after the mental task onsets. Lastly, data were downsampled by a factor of 2 that is from 7.81 to 3.905 Hz for the frontal montage and from 8.92 to 4.46 Hz for the temporal montage.

Corrections to 2.8 Classification

All four mental tasks in one concept trial share the same underlying hidden variable about the semantic category. Thus, it is not possible to extract the evoked hemodynamic response for an individual mental task. Instead, the evoked hemodynamic response must be extracted for the sequence of four mental tasks.

Let T be one such sequence of four mental tasks in one concept trial for which we want to extract the evoked hemodynamic response. The estimated GLM is used to remove influence from preceding trials and thus to isolate the evoked hemodynamic response for T. Let X_T be another modification of the design matrix X with the difference that, before the convolution, the elements of the corresponding conditions for T are set to 0, instead of 1, when T is taking place. In other words, the design matrix X_T is a version of the design matrix X with T excluded as if T did not take place in the experiment. The evoked hemodynamic response r for T, plus the normally distributed error, is then computed via

$$r = y^W - X_T^W \hat{\beta}_w \ . \tag{1}$$

In order to use the same analysis pipeline as for the non-GLM approach, the extracted hemodynamic responses were de-whitened. This approach brings the signal back to the original fNIRS space. To undo the effect of pre-whitening from equation 5 in the published article, the de-whitened response r^D is estimated from r, which is in the whitened space, as

$$r_t^D = r_t + a_1 r_{t-1} + a_1^2 r_{t-2}^D \tag{2}$$

with starting conditions $r_0^D = 0$ and $r_1^D = 0$.

Single-channel classification

We tested two approaches for semantic decoding in a single channel: (1) a classification of the whole mental task period of 13.5 seconds from the mental task onset (53 samples for the frontal montage and 61 samples for the temporal montage), and (2) a sliding window approach to investigate the temporal evolution of semantic decoding. In the latter approach, a window size was 16 samples, which correspond to about 4 seconds (4.1 seconds for the frontal montage and 3.59 seconds for the temporal montage). This temporal window was shifted in steps of half the window size.

In both approaches, the data in each temporal window and each channel were classified separately in 15-block-wise cross-validation. We used 15 blocks that correspond to blocks defined in the experimental design to use the training blocks and a testing block from the extraction of hemodynamic responses in the GLM data. In each temporal window, the data were normalized (z-scored) and classified by a classifier. We tested the following classifiers (all from scikit-learn [3]): support vector machine (SVM) with a radial basis function kernel ($C = 1$), SVM with a nested-cross-validation (with inner stratified 10-fold cross-validation) to choose an appropriate parameter C, logistic regression (LR) with L2 norm, and linear discriminant analysis (LDA).

Multi-channel 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 $n \in \{1, 2, \ldots, 11\}$ channels with the highest classification accuracies from the inner stratified 10-fold 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.

Multi-channel classification with PCA features

An approach based on a principal component analysis (PCA) was employed to decrease data dimensionality in the channel space while allowing the classifier to use information from all channels. In each temporal window, each channel was normalized (z-scored) separately. The spatial PCA then projected the data onto a smaller subspace by keeping only the $N \in \{1, 2, \ldots, 10\}$ PCA components which explain most of the data variance. The selected components were normalized (z-scored) separately before being passed as features to classifiers mentioned above. Additionally, we tested using only oxygenated or deoxygenated channels, instead of using both channel types, with $N \in \{1, 2, \ldots, 5\}$.

Corrections to 3. Results

In this section, we present the new results from the revised analyses.

Figure 1 shows hemodynamic responses of animals and tools for channel AFF5h-AFp3h from participant 1 in both the GLM and no-GLM data. The difference in scale between the GLM and no-GLM data is due to the de-whitening step in the GLM data pre-processing to bring the cleaned signal back to the original fNIRS space (see equation (2)). Nevertheless, this difference in scale has no influence on the decoding process because both data will be normalized (z-scored, channel-wise). Hemodynamic responses follow the expected relationship between changes in oxygenated and deoxygenated hemoglobin. Due to our experimental design, it was not possible to remove the influence of preceding mental tasks in the GLM data in the same concept trial because all mental tasks share the same underlying information about the semantic category. For this reason, hemodynamic responses in Figure 1 have visually similar trends between the GLM and no-GLM data.

Single-channel classification

The mental task period (13.5 seconds from the mental task onset) was classified in the single-channel classification approach to differentiate between the semantic categories of animals and tools using different classifiers and two types of preprocessed data. Figure 2 shows numbers of channels with statistically significant classification accuracies (that is above 56.11%, corresponding to $p < 0.05$ in a one-sided Binomial test with $n = 180$) for each participant. To combat a multiple comparison problem between channels, we used a (conservative) threshold ($p < 0.05$) for the number of statistically significant channels for each participant using a bootstrapping simulation. This bootstrapping simulation was based on 10^6 simulations by sampling from the binomial distribution for each channel, counting the number of significant channels (that are above 56.11%, corresponding to $p < 0.05$ in the one-sided Binomial test), and computing the 95 percentile of this distribution. The number of channels with significant classification accuracies is considered statistically significant when it is above this 95 percentile. Note that the 95 percentile is different for each participant due to the different number of excluded channels.

In the silent naming task, it was possible to differentiate between the semantic categories of animals and tools in three participants with the frontal montage (1, 2,

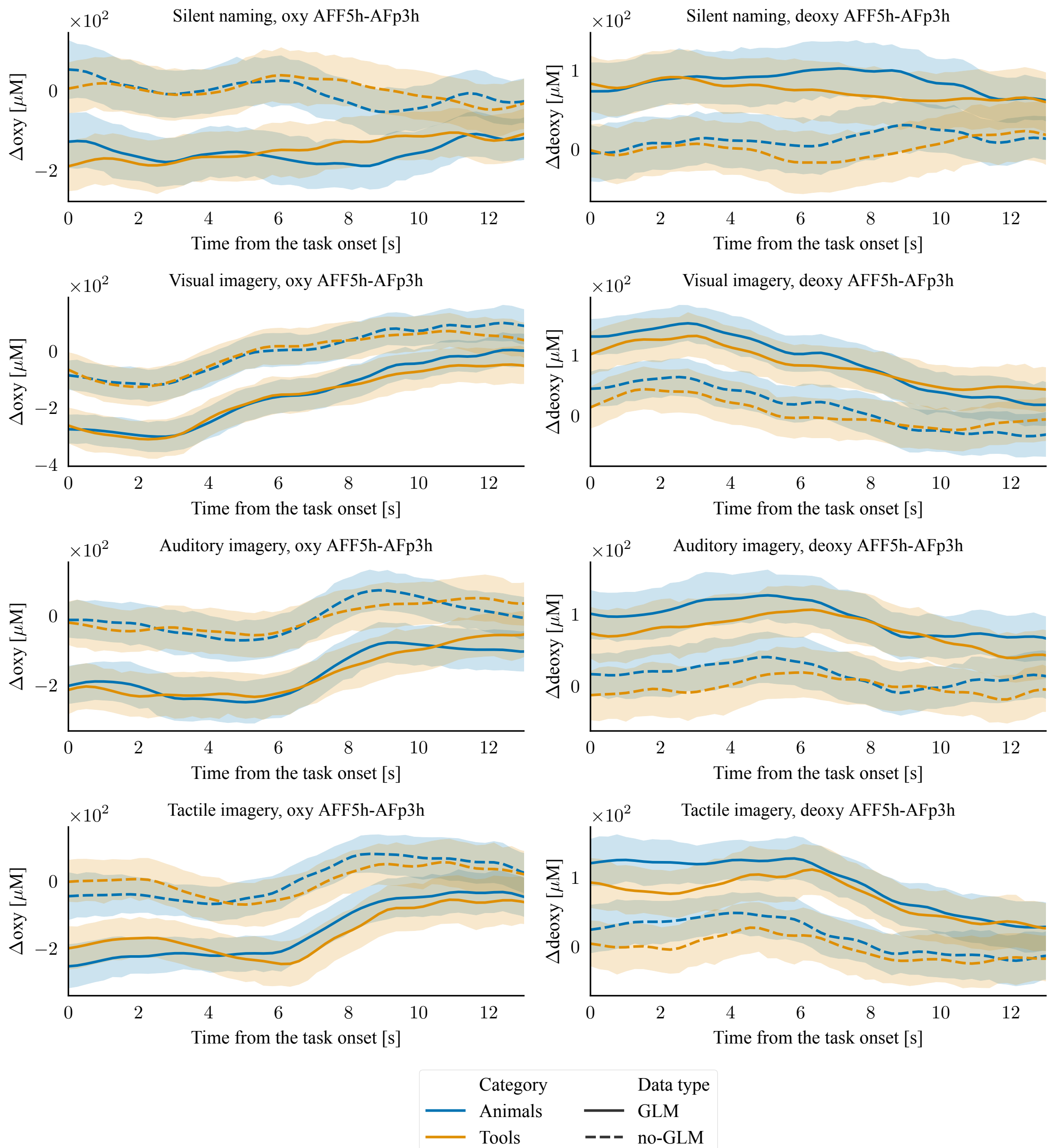

Figure 1. Hemodynamic responses of animals and tools for channel AFF5h-AFp3h from participant 1 in the GLM (solid) and no-GLM data (dashed). Hemodynamic responses are shown with mean and 95% confidence interval. Changes in oxygenated hemoglobin are in the left column and changes in deoxygenated hemoglobin are in the right column. The difference in scale between the GLM and no-GLM data is due to the de-whitening step in the GLM data to bring the cleaned signal back to the original fNIRS space.

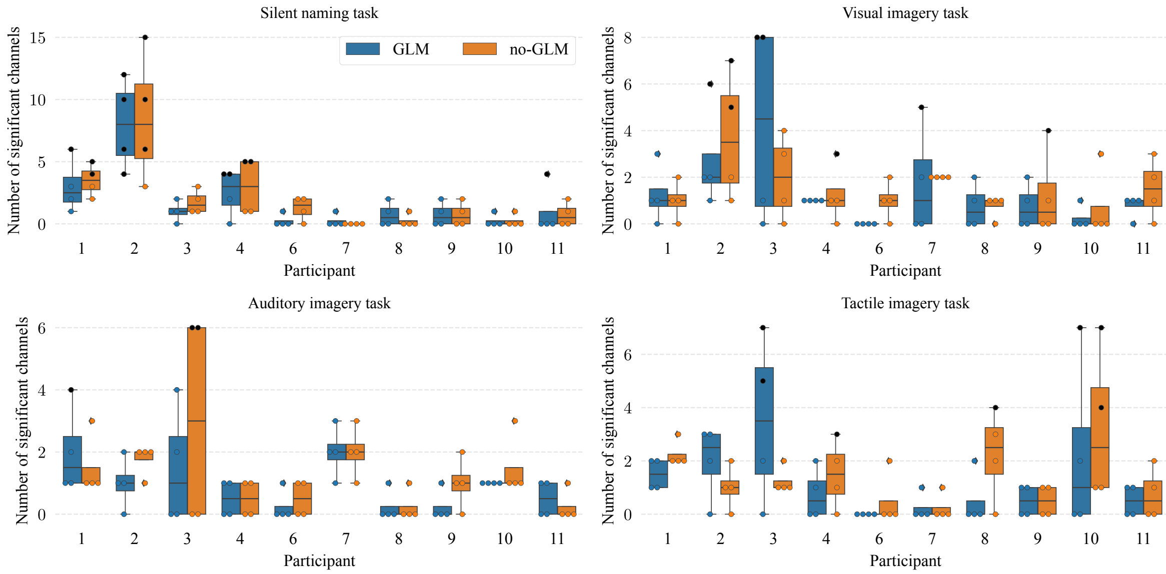

Figure 2. Numbers of channels with statistically significant classification accuracies (56.11%, $p < 0.05$, one-sided Binomial test) in the single channel classification approach. Dots correspond to four used classifiers. Black dots indicate statistical significant numbers of channels ($p < 0.05$, bootstrapping simulation, see text).

and 4) with both the GLM and no-GLM data. While, only one participant with the temporal montage (11) allowed significant semantic decoding. However, this semantic decoding was possible only with one classifier on the GLM data.

In the visual imagery task, semantic decoding was possible in three participants with the frontal montage (2, 3, and 4). However, participant 4 allowed significant semantic decoding only with one classifier on the no-GLM data and participant 3 with two classifiers only on the GLM data. In the temporal montage, two participants (7 and 9) allowed significant semantic decoding but with only one classifier on one type of data (the GLM and no-GLM, respectively).

In the auditory imagery task, two participants with the frontal montage (1 and 3) allowed significant semantic decoding. However, this semantic decoding was possible only with one classifier on the GLM data for participant 1 and for two classifiers only on the no-GLM data for participant 3. On the other hand, it was not possible to differentiate between the semantic categories in any participant with the temporal montage.

In the tactile imagery task, semantic decoding was possible in two participants with the frontal montage (3 and 4). But this was possible only on one type of data (two classifiers on the GLM data, and one classifier on the no-GLM data, respectively). For the temporal montage, semantic decoding was possible in two participants (8 and 10). However, participant 8 allowed significant semantic decoding only with one classifier on the no-GLM data.

In the sliding window approach, we observed a similar trend with only a few

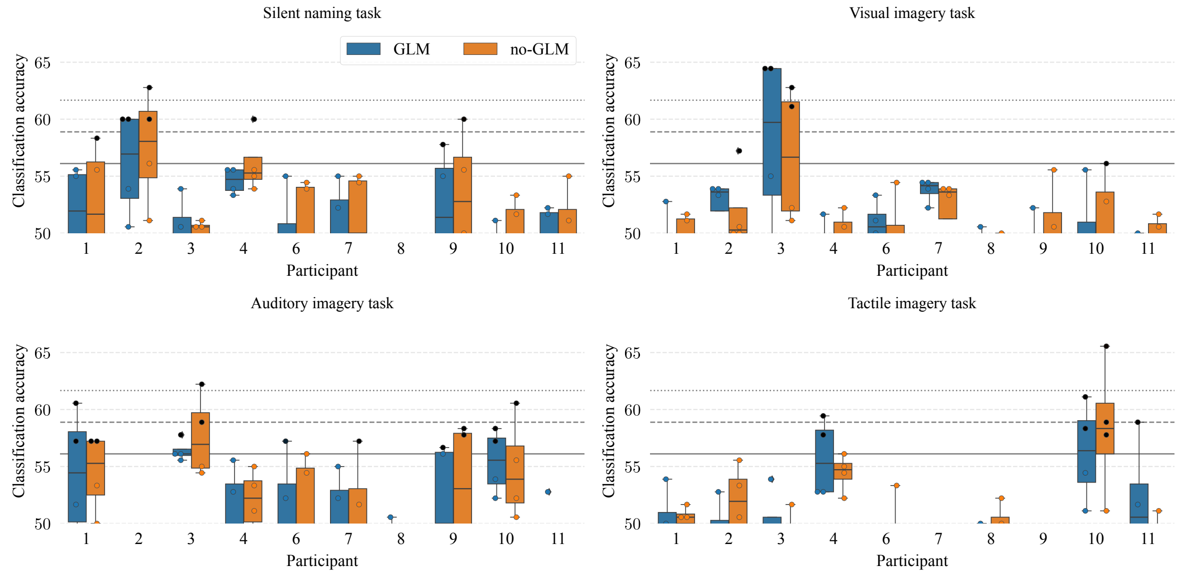

Figure 3. Classification accuracies using information from all channels. Dots correspond to four used classifiers. Horizontal lines indicate significance borderlines for $p = 0.05$ (one-sided Binomial test, 56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted). Black dots indicate statistical significant classification accuracies ($p < 0.05$).

differences. We thus decided not to report these results in the rest of this paper to simplify the presentation with the overall same message.

All-channels classification

We tested using information from all channels for the semantic decoding of the whole mental task period, see Figure 3.

In the silent naming task, it was possible to differentiate between the semantic categories ($p < 0.05$, one-sided Binomial test) in three participants with the frontal montage (1, 2, and 4). However, participants 1 and 4 had statistically significant classification accuracies only with one classifier on the no-GLM data. While, only participant 9 with the temporal montage had significant accuracies with both the GLM and no-GLM data.

In the visual imagery task, semantic decoding was possible in two participants with the frontal montage (2 and 3). However, participant 2 had significant accuracies only with one classifier on the no-GLM data. In the temporal montage, only participant 10 had significant accuracies with one classifier on the no-GLM data.

In the auditory imagery task, semantic decoding was possible in three participants with the frontal montage (1, 3, and 6). However, participant 6 had significant accuracies with only one classifier on the GLM data. In the temporal montage, three participants (7, 9, and 10) had significant accuracies but participant 7 achieved this only with one classifier on the no-GLM data.

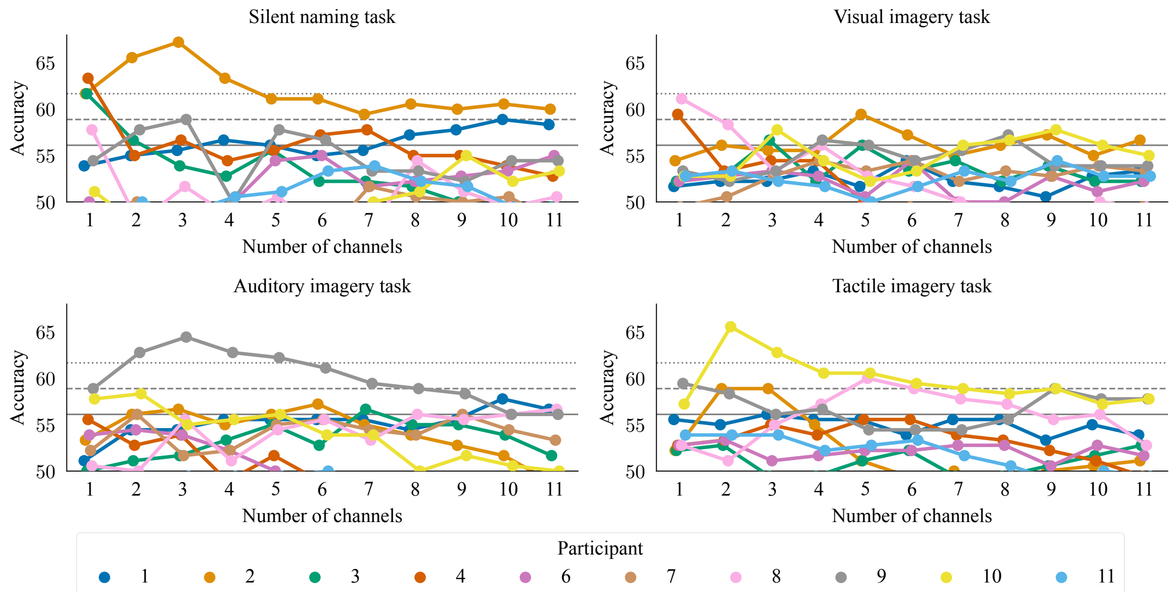

Figure 4. Maximal classification accuracies across different classifiers for the channel selection approach with the no-GLM data. Horizontal lines indicate significance borderlines for $p = 0.05$ (one-sided Binomial test, 56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted).

In the tactile imagery task, only participant 4 with the frontal montage had significant accuracies with two classifiers only on the GLM data. On the other hand, two participants with the temporal montage (10 and 11) had significant accuracies. However, participant 11 had significant accuracies only with one classifier on the GLM data.

Multi-channel classification with channel selection

Figure 4 shows maximal classification accuracies achieved across all tested classifiers with the no-GLM data in the channel selection approach. The results generally followed a trend from the single-channel classification results. The higher number of channels with significant accuracies from the single-channel analysis provided a higher chance for them to be used in automatic channel selection. However, the conservative threshold for statistical significance in the single-channel classification approach excluded participants with only several channels suitable for semantic decoding. For instance, participant 8 had significant accuracies using one and two selected channels in the visual imagery task. While, the same participant needed to have more than three channels with significant accuracies in the single-channel classification approach to be considered to have statistically significant semantic decoding.

In the silent naming task, semantic decoding was possible (one-sided Binomial test corrected with the false discovery rate) in three participants with the frontal montage (2, 3, and 4) and in no participant with the temporal montage. In the visual imagery

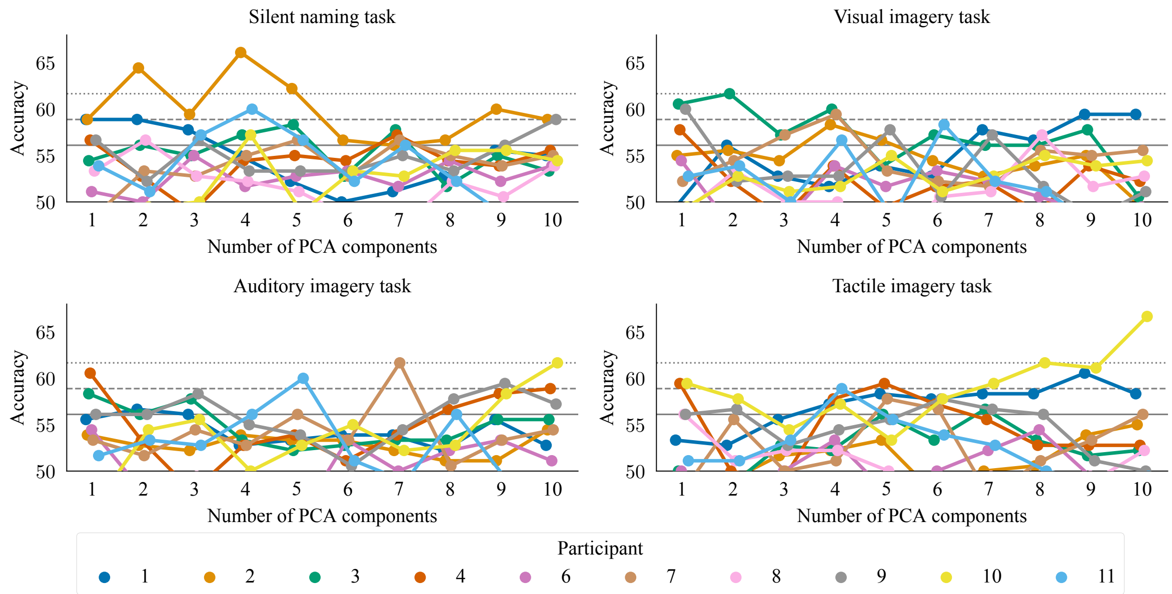

Figure 5. Maximal classification accuracies across different classifiers using a number of (spatial) PCA components as classifier features with the no-GLM data. Horizontal lines indicate significance borderlines for $p = 0.05$ (one-sided Binomial test, 56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted).

task, semantic decoding was possible only in one participant with the temporal montage (8). In the auditory imagery task, only one participant with the temporal montage (9) had significant accuracies. In the tactile imagery task, semantic decoding was possible only in three participants with the temporal montage (8, 9, and 10).

Multi-channel classification with PCA features

To decrease data dimensionality in the channel space while allowing classifiers to use information from all considered channels, the spatial PCA was computed on only oxygenated, only deoxygenated, or both channel types together. Figures 5, 6, and 7 show results with the no-GLM data for both channel types, for oxygenated channels, and for deoxygenated channels, respectively.

In the silent naming task, it was possible to differentiate between semantic categories (one-sided Binomial test corrected with the false discovery rate) in two participants (1 and 2) with both oxygenated and deoxygenated channels, separately, and in two participants (3 and 9) with only deoxygenated channels. When both channel types were used together, only two participants (2 and 11) had significant classification accuracies.

In the visual imagery task, semantic decoding was possible in one participant (9) with both channel types, in two participants (2 and 3) with only oxygenated channels, and in one participant (8) with only deoxygenated channels. On the other hand, three participants (1, 3, and 9) had significant accuracies when both channel types were used

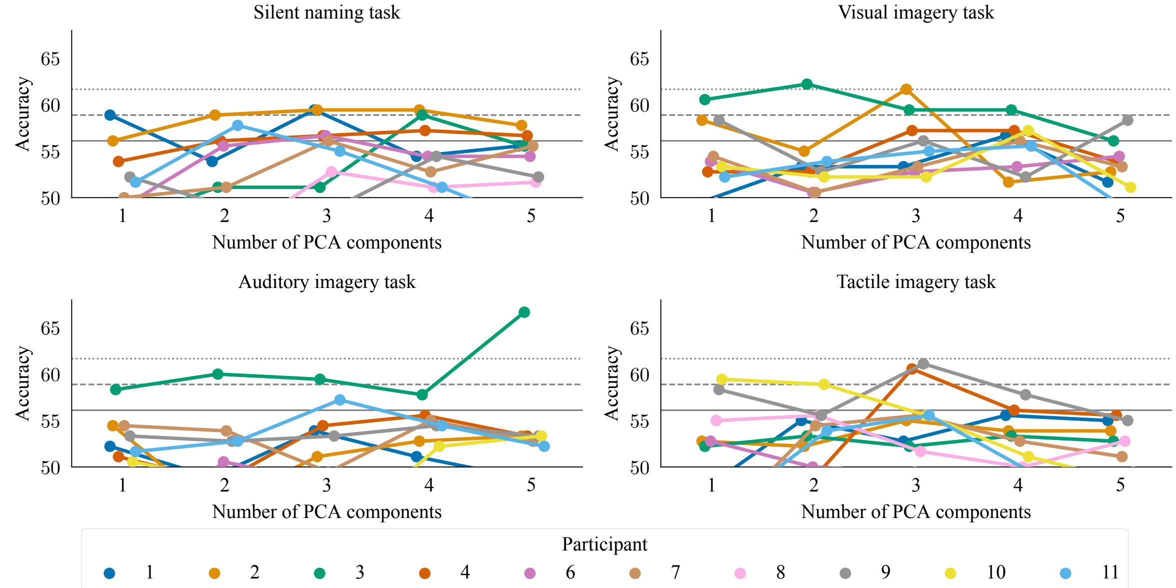

Figure 6. Maximal classification accuracies across different classifiers using a number of (spatial) PCA components from only oxygenated channels as classifier features with the no-GLM data. Horizontal lines indicate significance borderlines for $p = 0.05$ (one-sided Binomial test, 56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted).

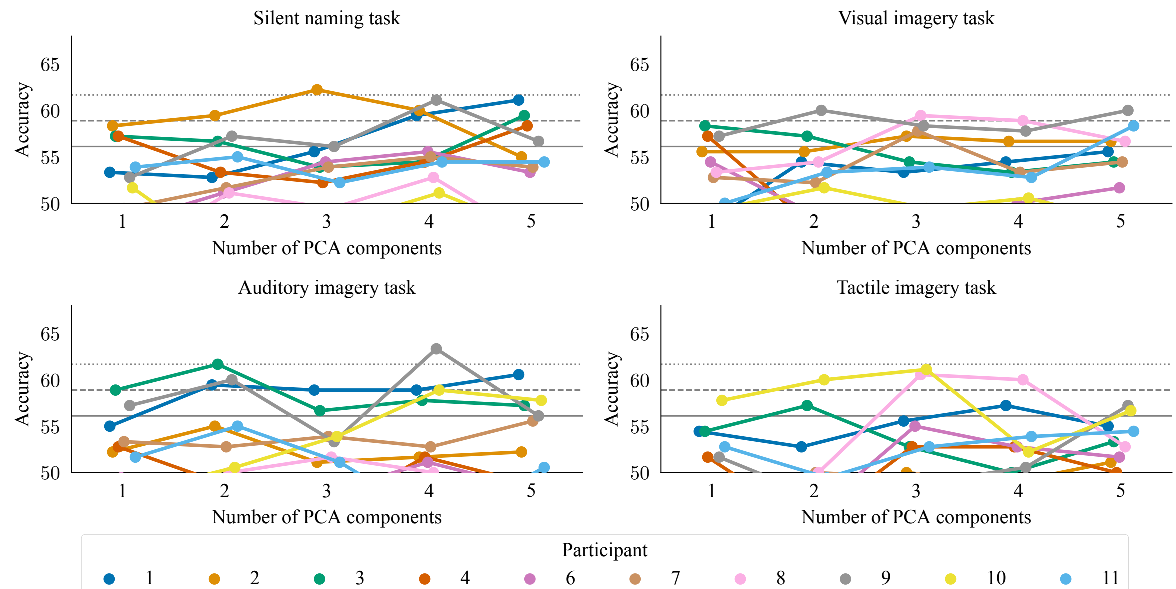

Figure 7. Maximal classification accuracies across different classifiers using a number of (spatial) PCA components from only deoxygenated channels as classifier features with the no-GLM data. Horizontal lines indicate significance borderlines for $p = 0.05$ (one-sided Binomial test, 56.11%, solid), $p = 0.01$ (58.89%, dashed), and $p = 0.001$ (61.67%, dotted).

together.

In the auditory imagery task, participant 3 had significant accuracies with both channel types and two participants (1 and 9) only with deoxygenated channels. When both channel types were used together, four participants (4, 7, 10, and 11) had significant classification accuracies.

In the tactile imagery task, participant 10 had significant accuracies with both channel types, while two participants (4 and 9) only with oxygenated channels and participant 8 only with deoxygenated channels. On the other hand, three participants (1, 4, and 10) had significant accuracies when both channel types were used together.

Corrections to 4. Discussion

We showed that it is possible to differentiate between the semantic categories of animals and tools in fNIRS in each mental task in some participants. We explored this possibility of semantic decoding over many different options: different data preprocessing methods, classifiers, and analysis approaches. Although, results differed slightly between each tested option, the overall message of the possible semantic decoding was clear. We first explored semantic decoding in a single channel (either oxygenated or deoxygenated) and showed that some channels carry useful information for differentiating between the semantic categories (in some participants). We then allow classifiers to utilize information from all channels or multiple channels, either by the channel selection approach or by dimensionality reduction by the PCA. Overall, semantic decoding was possible for up to 5 participants (3 with the frontal and 2 with the temporal montage) in the silent naming and the visual imagery task, up to 7 participants (3 with the frontal and 4 with the temporal montage) in the auditory imagery task, and up to 5 participants (2 with the frontal and 3 with the temporal montage) in the tactile imagery task.

While our experimental design of the sequence of four different mental tasks and short gaps of only 200 ms between mental tasks is appropriate for EEG recordings, this experimental design was not the most optimal for fNIRS recordings in terms of the conclusiveness of the results. Due to the short gap between mental tasks, classifiers could exploit information from preceding and following mental task(s). While the GLM approach can model different mental tasks, it is not possible to remove the influence of preceding mental tasks in the same concept trial because all mental tasks share the same underlying information about the semantic category. This issue should be partly mitigated by the random order in which the mental tasks were presented across different blocks. In future experiments to further investigate the neural correlates of semantic decoding in fNIRS, we would modify our experimental design in the following ways. First, we would use only a single mental task after the image presentation. Second, a longer interstimulus interval should be employed, such as a canonical 6–9 seconds interval from fMRI research, to properly identify evoked hemodynamic responses. Lastly, to suppress the influence of image presentation, a longer interval between the image presentation and the mental task could also be used.

There has been only one other fNIRS-based semantic neural decoding study to date. A study by Zinszer and colleagues [4] discriminated between semantic categories of animals and body parts while participants focused on audiovisual stimuli (photographs with a simultaneous auditory presentation of the object names) 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 from 6.5 to 9.0 seconds after the stimulus onset. We were not able to achieve similar mean classification accuracies. However, we were able to achieve significant classification accuracies for several participants in the range of 60–65%.

Corrections to 5. Conclusion

We demonstrated that semantic decoding is possible in fNIRS by differentiating between the semantic categories of animals and tools.

References

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

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

[3] 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.

[4] 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.