Brain Decoding Using fNIRS

Lu Cao[1]*, Dandan Huang[2,3]*, Yue Zhang[2,3]†, Xiaowei Jiang[4], Yanan Chen[4]
[1]Singapore University of Technology and Design, Singapore
[2]School of Engineering, Westlake University, China,
[3] Institute of Advanced Technology, Westlake Institute for Advanced Study, China
[4]Henan University, China
lu_cao@mymail.sutd.edu.sg, {huangdandan, zhangyue}@westlake.edu.cn, jiangxiaowei@aipsycho.com, chenyn@henu.edu.cn

Abstract

Brain activation can reflect semantic information elicited by natural words and concepts. Increasing research has been conducted on decoding such neural activation patterns using representational semantic models. However, prior work decoding semantic meaning from neurophysiological responses has been largely limited to ECoG, fMRI, MEG, and EEG techniques, each having its own advantages and limitations. More recently, the functional near infrared spectroscopy (fNIRS) has emerged as an alternative hemodynamic-based approach and possesses a number of strengths. We investigate brain decoding tasks under the help of fNIRS and empirically compare fNIRS with fMRI. Primarily, we find that: 1) like fMRI scans, activation patterns recorded from fNIRS encode rich information for discriminating concepts, but show limits on the possibility of decoding fine-grained semantic clues; 2) fNIRS decoding shows robustness across different brain regions, semantic categories and even subjects; 3) fNIRS has higher accuracy being decoded based on multi-channel patterns as compared to single-channel ones, which is in line with our intuition of the working mechanism of human brain. Our findings prove that fNIRS has the potential to promote a deep integration of NLP and cognitive neuroscience from the perspective of language understanding. We release the largest fNIRS dataset by far to facilitate future research.

Introduction

The increasing development and use of neuroimaging techniques represent a significant advance in the field of brain decoding, in which computational scientists interpret implicit brain activities based on explicit linguistic representations and deep-learning algorithms (Mitchell et al. 2008; Wehbe et al. 2014b; Hale et al. 2018; Gauthier and Levy 2019; Cao and Zhang 2019). The main research task is to establish a mapping between the concepts and neural activation patterns through neuroimaging experiments. As shown in the Equation 1, given brain images which imply mental content for words or text snippets, the task of brain decoding is to predict

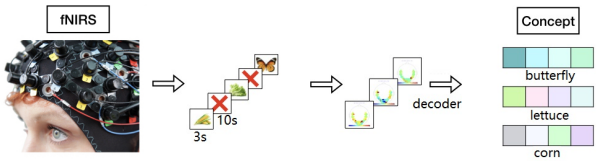


Figure 1: Schematic representation of fNIRS-based brain decoding: (a) emitter and detector probes are positioned over the scalp surface; (b) stimuli are presented to participants; (c) participants' neural activation are recorded by fNIRS; (d) a decoder is trained to take brain images as input and output the corresponding text semantic vectors.

the stimulus word evoking such a neural pattern,

$$v = f(i), \tag{1}$$

where i, v is the brain activation image and the stimulus word, respectively. The target is to learn the mapping function $f(\cdot)$, which in most cases is a linear model (Mitchell et al. 2008).

Brain decoding is a significant fundamental research topic both for cognitive science and artificial intelligence. For neuroimaging research, brain decoding methods are appreciated tools for localizing and distinguishing intricate brain response patterns and making predictions about undetectable neural states (Shinkareva et al. 2008; Just et al. 2010). And for artificial intelligence research, in particular the domain of natural language processing (NLP), the appealing properties of brain decoding technology are increasingly being used to explore to what extent the intelligent language understanding behaviors captured by those artificial models are consistent with human reading physiology (Gauthier and Levy 2019; Abnar et al. 2019). The underlying theory is that the brain neural basis and the corpus distributional properties of the same word are highly correlated (Mitchell et al. 2008).

However, as an important prerequisite of brain decoding, the measurement of brain activation remains a challenge due to limitations on neuroimaging technology. Existing methods include electrocorticography (ECoG) (Kuruvilla and Flink 2003), electroencephalogram (EEG) (Murphy, Baroni, and Poesio 2009), functional magnetic resonance imag-

*The first two authors have equal contribution.
†Yue Zhang is the corresponding author.

ing (fMRI) (Pereira, Just, and Mitchell 2001; Wehbe et al. 2014a; Gauthier and Levy 2019), and magnetoencephalography (MEG) (Wehbe et al. 2014b; Fyshe et al. 2014), each having its own relative advantages and weaknesses. One comparatively less studied tool is the functional near-infrared spectroscopy, shortly fNIRS, which can interpret the brain through cerebral hemodynamic responses associated with neuron behaviors. In practice, as shown in Figure 1a, fNIRS recording requires participants to wear a cap over the scalp surface, which embeds with emitters and detectors of near-infrared light. The emitters emit light, and the detectors receive light that passes through the tissue. A detector-emitter pair forms a *channel*, within which hemodynamic responses can be recorded and active cortex regions can be detected based on the absorption and scattering of near-infrared light by hemoglobin. At the end, the detected emerging fNIRS signal comes mainly from oxygenated hemoglobin and deoxygenated hemoglobin located in small vessels. Therefore, compared with other neuroimaging tools, fNIRS shows advantages of noninvasiveness, high temporal resolution, low experimental cost, full compatibility, multiple biomarkers and high tolerance for motion (Ferrari and Quaresima 2012).

Zinszer et al. (2017)0 make the first attempt to link fNIRS signals to representations of concrete words. The dataset and channel configuration they exploit are relatively small and limited, leaving questions open to comprehensively detect the decodability of fNIRS patterns. We try to better understand fNIRS-based brain language processing and draw more conclusions on the property of fNIRS. To this end, we conduct one pilot study and one large-scale decoding experiment using fNIRS technology. In the pilot study, we conduct a small scale preliminary test using materials from Zinszer et al. (2017), aiming to evaluate the feasibility, duration, channel configuration, electrode setting, interested brain region and decoding strategy of fNIRS operation. Then, we conduct a full-scale decoding experiment with hyperparameters decided by the pilot study, and verify whether fNIRS has the potential to promote a deep integration of NLP and cognitive neuroscience from the perspective of language understanding. Beyond findings of Zinszer et al. (2017), we find that:

1. fNIRS indeed encodes rich linguistic information into hemodynamic neurological signals, but shows limitations on the possibility of decoding fine-grained semantic clues.

2. fNIRS decoding has weaker accuracy compared with fMRI due to its limited penetration depth, but shows robustness across different brain regions, semantic categories and even subjects. This provides a basis for establishing a unified model across subjects and diverse semantic spaces.

3. fNIRS has higher accuracy decoded on multi-channel patterns as compared to single-channel ones, in line with our intuition of the working mechanism of the human brain.

In addition, to complement the extensive fMRI, MEG and EEG datasets published for brain decoding tasks (Wehbe et al. 2014a,b; Pereira, Just, and Mitchell 2001; Sudre et al. 2012; Pereira et al. 2018; Murphy, Baroni, and Poesio 2009), we release the largest fNIRS dataset by far for future research, which covers 50 objects from 10 semantic categories[1].

Related Work

The past decade has witnessed considerable progress in the field of brain decoding regarding language processing. Mitchell et al. (2008) used distributed word representation to decode neural activation associated with concrete nouns. Subsequent studies extended the research from simple word stimulus (Palatucci et al. 2009; Just et al. 2010; Murphy, Talukdar, and Mitchell 2012) to successive phrases (Wehbe et al. 2014a,b; Huth et al. 2016) and even sentences (Matsuo et al. 2016; Pereira et al. 2018; Sun et al. 2019), based on fMRI and MEG neuroimaging. Murphy, Baroni, and Poesio (2009) demonstrated that corpus-based semantic representations can predict neural activation recorded by EEG. Research based on EEG has further received much interests (Murphy and Poesio 2010; Murphy et al. 2011; Hale et al. 2018).

Compared with various brain decoding studies based on fMRI, MEG and EEG neural patterns, little attention has been paid to fNIRS decoding. As mentioned earlier, Zinszer et al. (2017) conducted the first study to explore fNIRS on language processing problems. In their studies, subjects passively viewed 8 stimuli (*bunny, bear, kitty, dog, mouth, foot, hand, nose*) and their blood oxygen levels were measured by fNIRS. The main conclusion was that fNIRS signals encoded information suitable for neural decoding via extrinsic representation models. However, their experiments were limited in using a small dataset and relatively constrained probe settings, leaving it an open question whether fNIRS can generalize to broader concepts and show robust performance.

In line with Zinszer et al. (2017), we also explore the decodability of fNIRS. Our contribution is four fold: First, we adopt more extensive experiment settings, which include 50 concepts covering 10 semantic categories and 46 channels covering 3 brain regions. In contrast, Zinszer et al. (2017) adopted 8 concepts covering 2 categories and 42 channels covering 2 brain regions. Second, we explore single and multi-channel decoding strategies respectively, and find that fNIRS signal is more stable and accurate under the multi-channel setting. In contrast, Zinszer et al. (2017) did not make a comparison in this aspect. Third, we compare the performance of fNIRS with fMRI under the same conditions, finding that fNIRS has generally weaker accuracy due to its limits in penetration depth, but shows robustness when being decoded across different brain regions, semantic categories and even subjects. Last, we also investigate the time window in which the signal points have the best decodability and the time extension for an acceptable experiment setting. To our knowledge, we are the first to explore these properties of fNIRS associated with language processing under the interdiscipline of natural language processing and neuroscience.

Task Specification

Decoding Model We learn linear regression models that map oxygenated hemoglobin (HbO) recorded by fNIRS into representations of words produced by natural language understanding (NLU) models. For HbO, when a brain area is

[1] https://github.com/caolusg/decoding_fnirs

involved in execution of a certain task, its metabolic demand for oxygen and glucose changes, leading to an increase in HbO concentrations. This is called hemodynamic response and can be measured through fNIRS at multiple locations of cerebral cortex. For NLU models, Pereira et al. (2018) carried out a comparison of all types of semantic vectors available at that time with regard to how well they can predict human judgments on behavioral tasks. The word2vec (Mikolov et al. 2013) and GloVe embedding (Pennington, Socher, and Manning 2014) were superior to others. We adopt the GloVe embedding for its widespread application (Jat et al. 2019; Cao et al. 2020) and homogeneity of value ranges in different dimensions and vocabulary size. New semantic representations still have been put forward but we believe that improvements in brain decoding have been marginal at best. For GloVe, specifically, let $H_j \in \mathbb{R}^{N \times D}$ represent the $D-$dimentional HbO in response to the j^{th} stimulus, and V represents the GloVe word vector. For each subject and his/her stimuli-triggered HbO variation, we use the ridge regression to learn a linear map $w : H_j \rightarrow V$ by minimizing the function:

$$J = ||wH_j - V||_2^2 + \alpha||w||^2, \tag{2}$$

where α is a regularization hyperparameter. Note that HbO is not used as input for the regression models. Given an arbitrary stimulus word j, we first collect its brain activity data H, which is HbO concentration transferred from near-infrared light wavelength. Then we encode the meaning of j via GloVe, obtaining a word vector V. Lastly, we use ridge regression to learn a linear map from H to V. GloVe embedding produces a dimensional vector representation of each word. While these representations are unique to model, we apply representational similarity methods to abstract GloVe and fNIRS data from respective sources into a shared similarity space. The linear regression model is trained and tested by *leave-two-out* and *leave-one-out* cross validation. In the leave-two-out approach, the model is trained repeatedly using $C(_N^{N-2})$ stimuli and tested using the two stimuli left out. In the leave-one-out approach, the model is trained using $C(_N^{N-1})$ stimuli and tested using the one left out. The procedure repeats until all stimuli have been trained and tested.

We explore two decoding strategies, namely *single-channel decoding* (SCD) and *multi-channel decoding* (MCD). The former decodes each channel separately and gives a result on average. The latter considers signals from multiple channels as a whole at once. For neuroimaging studies, conventional analyses treat each voxel independently to localize brain regions activated by a specific condition (Friston et al. 1994). While the popular multivariate statistical analyses use multiple voxels simultaneously in a multivariate fashion. The single/multivariate statistical method is a standard setting in the literature (Mitchell et al. 2008; Mahmoudi et al. 2012).

Baseline The implicit assumption for the model is that it would perform at chance level. For the necessity to show that the mapping from brain activity to word vector is robust, reporting chance performance is not enough, thus we adopt the *random scrambled pairs* (RSP) as one of our baselines. The stimuli and corresponding fNIRS signals are randomly shuffled in this setting. The comparison of model performance with RSP baseline is to determine that the mapping is reliable and not a result of noise.

Category	Exemplar
animal	bunny, bear, kitty, dog
body-part	mouth, foot, hand, nose

Table 1: Exemplars used in the pilot study.

Additionally, we also include a fMRI concept decoding task (Mitchell et al. 2008) for comparison. Although fNIRS has high time resolution like EEG does, we do not take EEG into comparison because the working scheme of fNIRS is quite different from that of EEG. EEG obtains neuronal activity through bioelectrical activities, while fNIRS (and fMRI) from hemodynamic aspects. fNIRS measures similar physiological signal to fMRI as both of them obtain neuronal activity by measuring changes in blood oxygen. The difference is that fMRI is measured by magnetism which covers the whole brain, while fNIRS is measured by lights and covers only the brain surface. Several studies have been conducted to validate and compare the metabolic correlates of neural activity as measured by fNIRS (i.e., increase in HbO and decrease in HbR) with the gold standard measured by fMRI (i.e., the blood oxygenation level-dependent response). Positive results have been established that the hemodynamic responses as measured by fNIRS and fMRI are spatially and temporally correlated (Strangman et al. 2002). Thus it is of interest to investigate the difference of the brain decoding ability between fNIRS and fMRI in the setting of brain decoding.

Evaluation Metric The decoding performance of the leave-two-out approach is evaluated by the matching score (MS) metric, and the leave-one-out approach is evaluated by the mean squared error (MSE) metric.

Matching score was first used in the brain decoding studies by Mitchell et al. (2008). Given a trained model, two test stimuli (w_1, w_2) and ground truth word vectors (v_1, v_2), the model predicts the word vector p_1 for w_1 and p_2 for w_2. It then decides which one is a better match: $(p_1 = v_1, p_2 = v_2)$ or $(p_1 = v_2, p_2 = v_1)$. The matching score is assigned as:

$$\begin{aligned} MS(p_1 = v_1, p_2 = v_2) = cosine(p_1, v_1) \\ + cosine(p_2, v_2). \end{aligned} \tag{3}$$

Similarity between the predicted and ground-truth vectors is measured by the cosine function, and the decoding accuracy for each subject is the fraction of correct pairs.

Mean squared error is also commonly used (Gauthier and Levy 2019), which measures average squares of errors between the predicted and ground-truth word vectors.

The two evaluation metrics serve complementary roles: the MS metric simply requires that fNIRS signals are semantically distinguishable, while the MSE metric strictly evaluates the ability of fNIRS signals to match the representational geometry of model activation. We use these two metrics to fully understand the decoding performance of fNIRS signals.

Pilot Study

We first conduct a pilot study to evaluate the feasibility and adjust experimental settings and hyperparameters. This study is conducted preliminarily before large-scale experiments due to expenses and difficulty in managing human experiments. Figure 1 depicts the high-level design of the process: a sequence of concepts is presented to subjects, and their cortical activity is recorded during the processing of each instant. We then decode fNIRS signals into semantic representations.

Participant Four right-handed undergraduates (two males and two females, mean age 22) participate in the task. [2] The tendency to include only right-handed people in neuroimaging research stems from the finding that certain processes are different in the brains of adextrals. While the left hemisphere dominates language processing in almost all right-handed people, in about 30% of adextrals this processing occurs predominantly in the right or both hemisphere. Handedness also influences how the brain represents sensation and movement of hands. Neuropsychologists have therefore avoided recruiting adextrals for fear of affecting data.

Procedure Following the settings of Zinszer et al. (2017), the stimuli are pictures and audios of 8 common objects from 2 semantic categories (Table 1). Each stimulus is presented 12 times, with a random permutation of item sequence in each presentation. Visual presentation lasts for 3 seconds, with audio presented immediately at the onset, followed by a 10-second rest period. During rest period, participants are instructed to fixate on an X displayed in the center of the screen (Figure 1b). The task for participants is to passively view and listen, simply focusing on each stimulus and thinking about properties of the object freely when it is presented. Participants' blood oxygen levels are measured by the NIRx NIRScout fNIRS system throughout the exposure. Note that the paradigm of viewing and listening simultaneously is effectively used in language-processing-related neural encoding and decoding tasks (Huth et al. 2016; Jat et al. 2019), thus there is no doubt that the results can reflect semantic processing rather than low-level sensory differences.

fNIRS Measurement and Preprocessing

The probes are arranged in three arrays: 26 channels in posterior, approximately covering the occipital lobe; 10 channels in left temporal lobe; and 10 channels in right temporal lobe. Detailed arrangement of probes and channels is demonstrated in Figure 2a. The posterior array is centered on the back of the head, with the most inferior row of channels just over the inion, and the two lateral arrays are positioned directly above the ears. Compared with Zinszer et al. (2017) who arrange probes in two arrays on the left and posterior of the head only, we adopt three arrays and take the right brain regions into account, aiming to achieve a more elaborate channel configuration. By this setting, we try to verify whether the cerebral hemodynamic responses from fNIRS varies across recording regions, and if so, which decoding region is the most responsible for language and semantic processing.

Brain activity data collected by NIRScout are detector readings for near-infrared light wavelength, with sampling rate being set as 3.9 Hz. Preprocessing of fNIRS data is performed using nirsLab (Xu, Graber, and Barbour 2014). We first remove discontinuities and spike artifacts, then bandpass filter the data (high pass: 0.01 Hz, low pass: 0.1 Hz), following the common way in this field (Zinszer et al. 2017; Pereira et al. 2018; Blankertz, Curio, and Müller 2002). Finally, we convert the wavelength data to oxygenated and deoxygenated hemoglobin concentration according to the modified Beer-Lambert law (Kocsis, Herman, and Eke 2006). We use the resulting HbO concentration data for subsequent analysis.

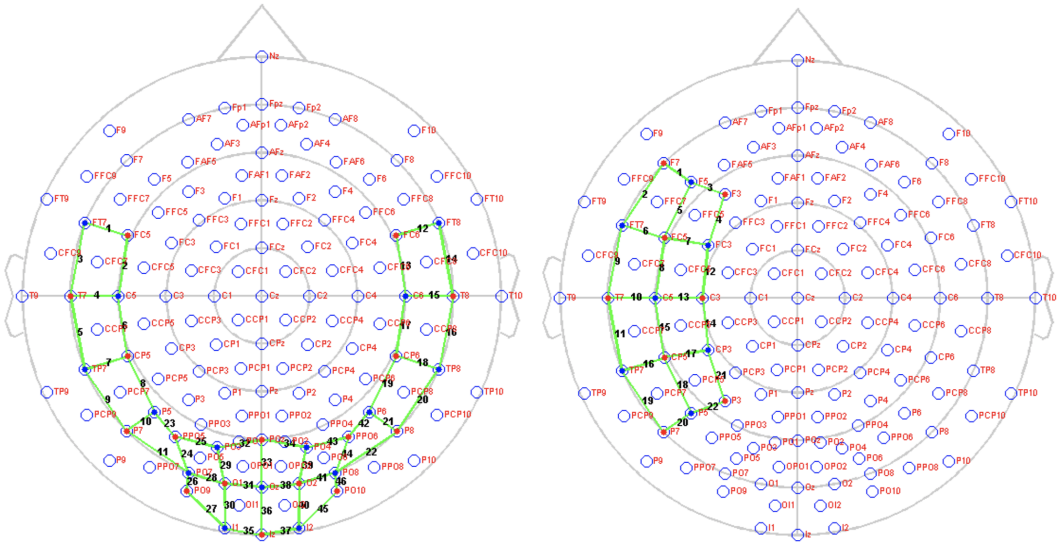
Results and Analysis

We determine: (a). Which part of the brain area is the most informative for language-related neural decoding? (b). What time period after hemodynamic response has the most abundant information? (c). Will word embedding parameters influence the decoding performance? (d). What is a better decoding strategy? We discuss the results of the pilot study from the perspective of decoding strategy, word embedding dimension, decoding time window, channel configuration and interested brain region, respectively, with the aim to determine the feasibility of the following large-scale experiment.

Decoding Strategy Conventional analysis for neuroimaging treats each voxel independently of any other to localize brain regions activated by a specific condition (Friston et al. 1994). The recently popular approach, termed multi-voxel pattern analysis (MVPA), on the other hand, utilizes multiple voxels simultaneously in a multivariate fashion, and is commonly exploited in fMRI decoding (Mitchell et al. 2008; Mahmoudi et al. 2012). The fact is that one fMRI voxel covers $3 \times 3 \times 3mm^3$, while one fNIRS channel covers $2.5cm - 3cm$ distance, larger than area covered by a single fMRI voxel. Thus, we are interested in verifying whether signals from such a distance are sufficient for encoding brain information. We compare the performance of single-channel decoding (SCD) and multi-channel decoding (MCD). Our null hypothesis is that SCD or MCD does not differ from a randomly scrambled pair baseline. We set the significance level to 0.05, where the test statistic is the difference between SCD, MCD and the RSP baseline. The results are validated by permutation test, with statistic distribution created by permutating test statistic 1000 times.

As presented in Table 2, the SCD performance is not significantly better than RSP under both metrics (MS: $p > 0.16$, MSE: $p > 0.92$), while the MCD is significantly better than RSP under the MS metric ($p < 0.03$). This demonstrates that for fNIRS, multi-channel decoding is a better choice compared to the single-channel scheme, which is in accordance with fMRI studies (We further verify this point on the large-scale experiment with more data, and the results are consistent and significant under both MS and MSE metrics. We list the results in Table 8). Thus the subsequent analysis is mainly based on multi-channel decoding.

[2] Informed consent procedures and experimental methods are approved by the institutional review board in advance.



(a) Pilot study.

(b) Full-scale experiment.

Figure 2: Probe and channel arrangement for two experiments. Red circles, blue circles, green lines indicate lasers, detectors and channels, respectively. The goal for the pilot study is to cover a broad brain area to determine which part is the most informative for language-related neural decoding. Thus we focus on the number of brain regions covered rather than the number of channels in each area. In order to cover more areas, there will naturally be fewer probes arranged in each area. The goal for the full-scale study is to focus on one area and explore the potentiality of fNIRS patterns as much as possible in the decoding. With permission of experimental equipment, we arrange more probes in the left hemisphere than the pilot study.

Subjects	1	2	3	4
RSP	0.50	0.49	0.51	0.50
SCD	0.55	0.46	0.56	0.55
MCD	0.57	0.50	0.71	0.71

(a) Matching score.

Subjects	1	2	3	4
RSP	0.2517	0.2632	0.2621	0.2524
SCD	0.2469	0.3103	0.2549	0.2479
MCD	0.2927	0.2652	0.2532	0.2395

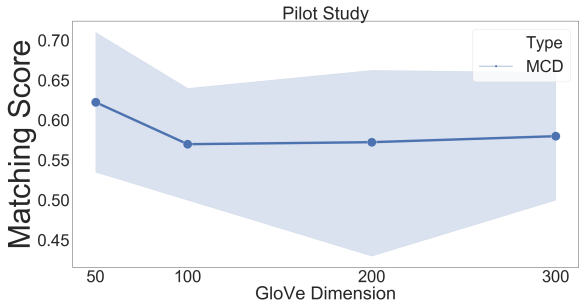
(b) Mean square error.

Table 2: The results of SCD and MCD in the pilot study, compared with the RSP baseline.

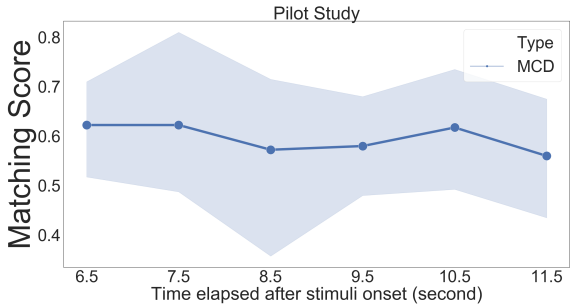
Word Embedding Dimension As shown in Figure 3a, we test the influence of word embedding dimension size on the decoding performance, with GloVE embedding sizes of 50, 100, 200 and 300. A relatively stable matching score is observed in the figure, with lower dimensional GloVe word embeddings achieving slightly better decoding performance for fNIRS patterns. This phenomenon may suggest that limitations exist on the possibility of decoding *fine-grained* semantic information encoded in *high-dimensional* embeddings from fNIRS human neuroimaging, which is also observed under fMRI operation (Gauthier and Levy 2019). Hence, we fix the dimension of word vector as 50 in the following analysis.

Decoding Time Window When the word vector is set to 50 dimensions, we test model performance under various decoding time window sizes. It has been reported that the hemodynamic response peaks 6 seconds after the neurons' immediate activation in a region (Kohl et al. 2000; Devor et al. 2008). Thus we decode the signals from various time windows to determine the time point of the most informative signal after the 6th second. A noticeable trend is demonstrated in Figure 3b: decoding signals between 6.5s to 7.5s result in the best matching score, and the performance decreases as further time elapses. This is in line with our intuition that the intensity of physiological signals gradually weakens after stimuli. Hence, we fix the decoding time window as 6.5-7.5 seconds in the following experiments.

Brain Region Previous studies have shown that decoding fMRI data in different brain regions yields significant variations (Mitchell et al. 2008; Pereira et al. 2018). We test whether fNIRS patterns are very different across recording sessions and which part is the most informative for neural decoding. We comprehensively collect fNIRS data from the left, right and occipital side of the brain, with a null hypothesis that its decoding performance does not differ among brain regions. We set the significance level as 0.05, and test the significance by permutation test. As illustrated in Table 3, the p-values under both metrics are all above the significance level, not rejecting the null hypothesis. Thus this suggests that in contrast to fMRI, the fNIRS performances in different regions are not significantly different.



(a) Performance changes over word embedding dimensions.



(b) Performance changes over time.

Figure 3: The influence of word embedding dimension and time window selection on decoding performance, evaluated by matching score on validation test. Shaded regions represent 95% confidence intervals, pooling across all subjects.

	L-R [1]	L-O [2]	R-O [3]
Matching score	0.12	0.21	0.68
MSE	0.14	0.89	0.26

[1] left vs. right temporal lobe [2] left vs. occipital lobe
[3] right vs. occipital lobe

Table 3: The p-values of decoding performance cross brain regions. The test statistic is the decoding difference between each two brain regions. The test statistic distribution is created by randomly permutating 1000 times.

Category	Exemplar
tool	pliers, saw, screwdriver, scissor, hammer
vegetable	celery, corn, carrot, tomato, lettuce
building	bird's nest, tiananmen, oriental pearl TV tower, pyramid, water cube
insect	bee, butterfly, dragonfly, ant, fly
transportation	car, train, truck, airplane, bicycle
furniture	sofa, chair, desk, bed, bookshelf
cloth	sweater, jeans, shirt, skirt, dress
animal	panda, cat, dog, horse, cow
body-part	arm, eye, foot, palm, leg
kitchen	knife, pan, spoon, glass, chopsticks

Table 4: Exemplars used in the full-scale experiment. Though some test words are compounds in form, what they express is a unified concept and we examine brain activity associated with the meanings of concepts.

Full-scale Experiment

We conduct a large-scale experiment based on hyperparameters selected from the pilot study. The schematic procedure is the same as the pilot study.

Participant Seven additional right-handed undergraduates (four males and three females, mean age 22) participate in this task. The number of participants in our study is comparable with existing research (Pereira, Just, and Mitchell 2001; Mitchell et al. 2008; Cox and Savoy 2003).

Subjects	1	2	3	4	5	6	7
Between-category	0.58	0.57	0.55	0.58	0.61	0.49	0.48
Within-category	0.55	0.48	0.57	0.51	0.50	0.50	0.45
Leave-one-category	0.57	0.47	0.52	0.55	0.55	0.49	0.48

Table 5: Matching score of cross-category decoding.

#subjects	1	2	3	4	5	6
MSE	0.393	0.374	0.372	0.372	0.372	0.372

Table 6: Mean squared error of cross-subject decoding.

Procedure To extend the pilot study, we adopt 50 frequently-used concepts from 10 broader semantic categories, with 5 exemplars per category (Table 4). Each stimulus is presented 7 times randomly, which is comparable to the fMRI literature on this topic (Mitchell et al. 2008). To avoid fatigue, we divide the experiment into two sessions, each presenting 25 words. There is a ten-minute break between two sessions. Other settings follow the same protocols as the pilot study strictly. The amount of data we adopted is comparable to previous literature using fMRI and EEG. For example, Mitchell et al. (2008) used 60 concrete nouns to decode fMRI activation associated with the meaning of nouns; Jat et al. (2019) used 32 sentences to understand simple sentence processing in deep neural networks and the brain. This reflects cost and difficulty in obtaining human data.

fNIRS Measurement and Preprocessing

As shown in Figure 2b, fNIRS probes are arranged to cover the left temporal, parietal and prefrontal lobes, with a total of 22 channels. As the performance across regions is not significantly different under fNIRS recording, we decrease fNIRS probes from 46 channels in the pilot study to 22 channels and mainly focus on the left hemisphere. The left brain has two cortical areas known to be involved in language processing, namely the Left Inferior Frontal Gyrus (LIFG), also known as Broca's area (Dronkers et al. 2007), and the Left Superior Temporal Gyrus (LSTG) / Left Posterior Middle and Superior Temporal Gyrus (LMTG), also known as Wernicke's area (Bogen and Bogen 1976). The left hemisphere

henceforth becomes the main region of interest in our study.

The sampling rate is set as 7.8 Hz in this task. The sampling rate of fNIRS depends on the number of emitters and detectors of near-infrared light. Since the emitters emit infrared light in turn, the sampling rate becomes higher with less detector-emitter pairs. Compared to the pilot study, this task has less detector-emitter pairs and thus a higher sampling rate. The collection and preprocessing of fNIRS data are all conducted according to the same guidelines as the pilot study.

Results and Analysis

Across Semantic Category Distinguishing within-category differences (e.g., skirt, dress) is more challenging than between-category differences (e.g., skirt, celery). Mitchell et al. (2008) demonstrated that predicting words from the same category yields lower accuracy than from different categories by fMRI. It is also interesting to determine the category distinguishing ability of fNIRS signals. Besides, we are also interested in predicting words from a new category by excluding all words in the same category from the training set (e.g., for the test words ant and $scissor$, we exclude all $insects$ and $tools$ from the training set). Our null hypothesis is that the performance of within-category and leave-one-category decoding will differ from the between-category decoding for fNIRS, like what has been observed from fMRI scans (Mitchell et al. 2008).

The results are shown in Table 5. We set the significance level of 0.05. The test statistic distribution is created by permutating test statistic 1000 times. The matching score of within-category ($p > 0.09$) and leave-one-category ($p > 0.21$) decoding is not significantly inferior to that of between-category decoding. We conclude that fNIRS demonstrates robust differentiation power, and the decoded concept is still distinguishable even under the leave-out condition. Palatucci et al. (2009) demonstrated that it can predict words that people were thinking about from fMRI of their neural activity, even without training examples for those words. Our results suggest that fNIRS also has the potential to be expanded into diverse semantic spaces. While most subject show results significantly above the random baseline ($p < 0.05$), we find that subjects 6 and 7 give low results. Possible reason can be poor optical contact and light obstruction by their dense hair.

Across Subjects Following most previous work on fMRI (Mitchell et al. 2004; Pereira et al. 2018), our models analyzed above are trained and tested separately for each subject. However, training subject-specific models is not feasible for real life applications since users do not participate in product-turing work. Thus we further investigate whether fNIRS signals preserve commonalities among different subjects. We divide subjects into training and testing groups, conducting training with 1 subject and testing with the remaining 6 subjects in turn, and averaging the 6 performances as the overall performance in this iteration. Then, we train with 2 subjects and test with the remaining 5 subjects in turn. The experiments are repeated until the last group is trained with 6 subjects and tested with the remaining one.

We iterate all possible combinations and demonstrate the results in Table 6. In contrast to fMRI scans which are very different across subjects (Mitchell et al. 2008), we find that for fNIRS, the performance is stronger with more training subjects as MSE decreased, indicating that the model has learned more commonalities among subjects. This provides a basis for establishing a unified model across subjects.

Figure 4: MSE of cross-subjects decoding.

Stimuli Influence The cross-subject robustness of fNIRS can be exploited in different applications. Theoretically, we can build a large dictionary for the human brain by means of fNIRS. The keys are brain activity images, and the values are corresponding concepts evoking such a neural pattern (Poldrack 2018). However, this is not easy in practice. One of critical factors of brain decoding is that, as the experiment stimuli expends, the experiment time becomes longer, thus subjects may distract more or less and infect the signal quality. And individual differences can also affect. We study whether fNIRS decoding will be affected as the concepts increase and experiment length increases. We compare the decoding performance of the pilot study, which consists of 8 concepts and lasts for 25 minutes, and the full-scale experiment, which consists of 50 concepts and lasts for 50 minutes. The results are presented in Table 7. Significance test is performed by the permutation test with the same setting mentioned above. Our results show that with the extension of experiment time and the increase of concepts, the decoding performance does not decrease significantly ($p > 0.12$). This means that it is acceptable to increase the number of stimuli to a certain degree, and it is an interesting question to record more concepts and cover more semantic categories in future study.

fMRI vs. fNIRS

fMRI has become a de-facto standard for in vivo imaging of the human brain in recent years. Compared to fMRI, fNIRS has its own advantages and limitations. fNIRS is critical for extending cognitive neuroscience beyond MRI scanners and enabling research with partici-pants who are not well-suited for fMRI studies (e.g., children, clinical populations). The optical nature of fNIRS stands out for its low cost, portability and robustness to motion noise, bringing functional imaging into a more realistic environment. It also has a higher temporal resolution than fMRI, allowing measurements of concentration changes in both oxygenated and deoxygenated hemoglobin. But fNIRS is limited by its inferior spatial resolution. It is therefore of interest how fNIRS compares to fMRI in studies of brain decoding. We adopt a fMRI decod-

	Pilot Study				Full-scale Experiment						
Subjects	1	2	3	4	1	2	3	4	5	6	7
Matching Score	0.57	0.50	0.71	0.71	0.58	0.56	0.55	0.58	0.60	0.49	0.47

Table 7: Stimuli Influence.

	Matching Score							MSE						
Subjects	1	2	3	4	5	6	7	1	2	3	4	5	6	7
RSP	0.50	0.52	0.49	0.49	0.49	0.51	0.49	0.3830	0.3821	0.3839	0.3809	0.3812	0.3851	0.3804
SCD	0.49	0.46	0.50	0.51	0.49	0.51	0.47	0.3803	0.3804	0.3805	0.3802	0.3887	0.4121	0.3805
MCD	0.58	0.56	0.55	0.58	0.60	0.49	0.47	0.3803	0.3809	0.3806	0.3801	0.3800	0.3805	0.3803

Table 8: The results of SCD and MCD in the full-scale experiment, compared with the RSP baseline. We find that the SCD does not reject the null hypothesis (Matching Score: $p > 0.32$, MSE: $p > 0.62$), while the MCD rejects the null hypothesis under both matching score and MSE metrics (Matching Score: $p < 0.03$, MSE: $p < 0.003$).

Categories	Exemplar
animal	bear, cat, dog, horse, cow
vegetable	lettuce, carrot, corn, tomato, celery
body part	eye, arm, foot, leg, hand
man-made	telephone, key, bell, watch, refrigerator
building	igloo, barn, house, apartment, church
kitchen	spoon, bottle, cup, knife, glass
vehicle	truck, car, train, bicycle, airplane
clothing	dress, skirt, coat, pants, shirt
furniture	chair, dresser, desk, bed, table
build part	door, chimney, closet, arch, window
insect	fly, bee, butterfly, ant, beetle
tool	hammer, chisel, screwdriver, saw, pliers

Table 9: Exemplars used in the Mitchell et al. (2008).

	between-category	within-category	leave-one-category
fNIRS	0.58	0.52	0.53
fMRI	0.85	0.59	0.77

Table 10: Mean matching score of cross-category decoding. fNIRS exlcudes the subject 6 and 7 due to poor signal quality. The results of fMRI are better than those reported in the original paper (0.77, 0.62, 0.70) in between-category and leave-one-category conditions. One possible reason is that GloVe is better than frequency based semantic model.

ing dataset from Mitchell et al. (2008) to compare the fNIRS to fMRI decoding. Their task is to decode 60 concrete nouns (Table 9) based on fMRI activation. During the fMRI data collection, 60 words are presented to 9 subjects with each stimulus exhibited 6 times randomly. The stimuli setting, experiment procedure and subject size are all comparable to our study, which laid a foundation for comparability of two datasets. Therefore, we decode the fMRI scans in the same way as we decode fNIRS, and make comparison from the perspective of cross category and cross subject, respectively.

The performance of cross-category decoding is summarized in Table 10. The evaluation metric is the matching score. The results show that generally fMRI decoding has a higher accuracy than fNIRS decoding, in terms of between-category, leave-one-category and within-category aspects. One likely reason is that fMRI scan covers the whole-brain (lateral surface and depth) and can be $50,000$ to $200,000$ dimensions, whereas fNIRS is limited to brain surface and is usually less than $10,000$ dimensions. This shows fNIRS's limits in penetration depth. However, the results also show that fNIRS has a robust decodibility across categories, since the matching score under three conditions does not vary significantly. In contrast, for fMRI signals, the accuracy drops dramatically from between category decoding to leave-one category decoding, and declines more sharply for within category decoding.

The performance of cross-subject decoding is summarized in Figure 4. The evaluation metric is the mean square error. As discussed in the earlier section, fNIRS signal can preserve commonalities among different subjects. In contrast, the mean square error of fMRI increases when more subjects are involved. We conclude that fMRI has higher subject-specific and category-specific recording accuracy than fNIRS, and fNIRS has higher robustness across different conditions. This may offer new possibilities for a combination of fMRI and fNIRS technologies. The two methods can complement each other and allow for more complex research paradigms that are unfeasible with either technique alone.

Conclusion

We presented a set of experimental results for decoding different mental states evoked by language processing, on the basis of the underlying brain activation measured with fNIRS. A large-scale fNIRS study is conducted among 50 frequently-used concepts across 10 semantic categories. Through an empirical comparison between semantic vectors generated by neural networks and brain activities observed by fNIRS, we explored the decodebility and robustness of fNIRS signals. Results show that 1) fNIRS can encode rich linguistic information into neurological signals, but show limits on the possibility of decoding fine-grained semantic information; 2) fNIRS decoding shows robustness across different brain regions, semantic categories and even subjects; 3) in line with the expectation that the brain uses multiple parts simultaneously to comprehend concepts, multi-channel fNIRS signals demonstrate stable precision compared with single-channel ones. We made our effect on validating fNIRS as a well-suited technology to promote a deep integrate of natural language processing and cognitive neuroscience.

References

Abnar, S.; Beinborn, L.; Choenni, R.; and Zuidema, W. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539* .

Blankertz, B.; Curio, G.; and Müller, K.-R. 2002. Classifying single trial EEG: Towards brain computer interfacing. In *Advances in neural information processing systems*, 157–164.

Bogen, J. E.; and Bogen, G. 1976. Wernicke's region–Where is it. *Annals of the New York Academy of Sciences* (280): 834–843.

Cao, L.; Chen, Y.; Huang, D.; and Zhang, Y. 2020. Investigating Rich Feature Sources for Conceptual Representation Encoding. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, 12–22. Online: Association for Computational Linguistics.

Cao, L.; and Zhang, Y. 2019. Investigating Lexical and Semantic Cognition by Using Neural Network to Encode and Decode Brain Imaging. In *International Workshop on Human Brain and Artificial Intelligence*, 84–100. Springer.

Cox, D. D.; and Savoy, R. 2003. fMRI Brain Reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19(2): 261–270.

Devor, A.; Hillman, E.; Tian, P.; Waeber, C.; Teng, I.; Ruvinskaya, L.; Shalinsky, M.; Zhu, H.; Haslinger, R.; Narayanan, S.; Ulbert, I.; Dunn, A.; Lo, E.; Rosen, B.; Dale, A.; Kleinfeld, D.; and Boas, D. 2008. Stimulus-induced changes in blood flow and 2-deoxyglucose uptake dissociate in ipsilateral somatosensory cortex. *Journal of Neuroscience* 28(53): 14347–14357. ISSN 0270-6474. doi:10.1523/JNEUROSCI.4307-08.2008.

Dronkers, N. F.; Plaisant, O.; Iba-Zizen, M. T.; and Cabanis, E. A. 2007. Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. *Brain* 130(5): 1432–1441.

Ferrari, M.; and Quaresima, V. 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63(2): 921–935.

Friston, K. J.; Holmes, A. P.; Worsley, K. J.; Poline, J.-P.; Frith, C. D.; and Frackowiak, R. S. 1994. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping* 2(4): 189–210.

Fyshe, A.; Talukdar, P. P.; Murphy, B.; and Mitchell, T. M. 2014. Interpretable Semantic Vectors from a Joint Model of Brain- and Text- Based Meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 489–499. Baltimore, Maryland: Association for Computational Linguistics. doi:10.3115/v1/P14-1046.

Gauthier, J.; and Levy, R. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 529–539. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1050.

Hale, J.; Dyer, C.; Kuncoro, A.; and Brennan, J. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2727–2736. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1254.

Huth, A. G.; De Heer, W. A.; Griffiths, T. L.; Theunissen, F. E.; and Gallant, J. L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532(7600): 453–458.

Jat, S.; Tang, H.; Talukdar, P.; and Mitchell, T. 2019. Relating Simple Sentence Representations in Deep Neural Networks and the Brain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5137–5154. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1507.

Just, M. A.; Cherkassky, V. L.; Aryal, S.; and Mitchell, T. M. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one* 5(1).

Kocsis, L.; Herman, P.; and Eke, A. 2006. The modified Beer–Lambert law revisited. *Physics in Medicine & Biology* 51(5): N91.

Kohl, M.; Lindauer, U.; Royl, G.; Kühl, M.; Gold, L.; Villringer, A.; and Dirnagl, U. 2000. Physical model for the spectroscopic analysis of cortical intrinsic optical signals. *Physics in Medicine and Biology* 45(12): 3749–3764. doi:10.1088/0031-9155/45/12/317.

Kuruvilla, A.; and Flink, R. 2003. Intraoperative electrocorticography in epilepsy surgery: useful or not? *Seizure* 12(8): 577–584.

Mahmoudi, A.; Takerkart, S.; Regragui, F.; Boussaoud, D.; and Brovelli, A. 2012. Multivoxel pattern analysis for FMRI data: a review. *Computational and mathematical methods in medicine* 2012.

Matsuo, E.; Kobayashi, I.; Nishimoto, S.; Nishida, S.; and Asoh, H. 2016. Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity. In *Proceedings of the ACL 2016 Student Research Workshop*, 22–29. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-3004.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mitchell, T. M.; Hutchinson, R.; Niculescu, R. S.; Pereira, F.; Wang, X.; Just, M.; and Newman, S. 2004. Learning to decode cognitive states from brain images. *Machine learning* 57(1-2): 145–175.

Mitchell, T. M.; Shinkareva, S. V.; Carlson, A.; Chang, K.-M.; Malave, V. L.; Mason, R. A.; and Just, M. A. 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science* 320(5880): 1191–1195. ISSN 0036-8075. doi:10.1126/science.1152876.

Murphy, B.; Baroni, M.; and Poesio, M. 2009. EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 619–627. Singapore: Association for Computational Linguistics.

Murphy, B.; and Poesio, M. 2010. Detecting Semantic Category in Simultaneous EEG/MEG Recordings. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, 36–44. Los Angeles, USA: Association for Computational Linguistics.

Murphy, B.; Poesio, M.; Bovolo, F.; Bruzzone, L.; Dalponte, M.; and Lakany, H. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and language* 117: 12–22. doi:10.1016/j.bandl.2010.09.013.

Murphy, B.; Talukdar, P.; and Mitchell, T. 2012. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of COLING 2012*, 1933–1950. Mumbai, India: The COLING 2012 Organizing Committee.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP (EMNLP)*, 1532–1543.

Pereira, F.; Just, M.; and Mitchell, T. 2001. Distinguishing natural language processes on the basis of fMRI-measured brain activation. In *European Conference on Principles of Data Mining and Knowledge Discovery*, 374–385. Springer.

Pereira, F.; Lou, B.; Pritchett, B.; Ritter, S.; Gershman, S. J.; Kanwisher, N.; Botvinick, M.; and Fedorenko, E. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications* 9(1): 1–13.

Poldrack, R. A. 2018. *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts*. Princeton University Press. ISBN 9780691178615.

Shinkareva, S. V.; Mason, R. A.; Malave, V. L.; Wang, W.; Mitchell, T. M.; and Just, M. A. 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One* 3(1).

Strangman, G.; Culver, J. P.; Thompson, J. H.; and Boas, D. A. 2002. A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *Neuroimage* 17(2): 719–731.

Sudre, G.; Pomerleau, D.; Palatucci, M.; Wehbe, L.; Fyshe, A.; Salmelin, R.; and Mitchell, T. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage* 62(1): 451–463.

Sun, J.; Wang, S.; Zhang, J.; and Zong, C. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7047–7054.

Wehbe, L.; Murphy, B.; Talukdar, P.; Fyshe, A.; Ramdas, A.; and Mitchell, T. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* 9(11).

Wehbe, L.; Vaswani, A.; Knight, K.; and Mitchell, T. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 233–243.

Xu, Y.; Graber, H. L.; and Barbour, R. L. 2014. nirsLAB: a computing environment for fNIRS neuroimaging data analysis. In *Biomedical optics*, BM3A–1. Optical Society of America.

Zinszer, B. D.; Bayet, L.; Emberson, L. L.; Raizada, R. D. S.; and Aslin, R. N. 2017. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics* 5(1): 1 – 8. doi:10.1117/1.NPh.5.1.011003.