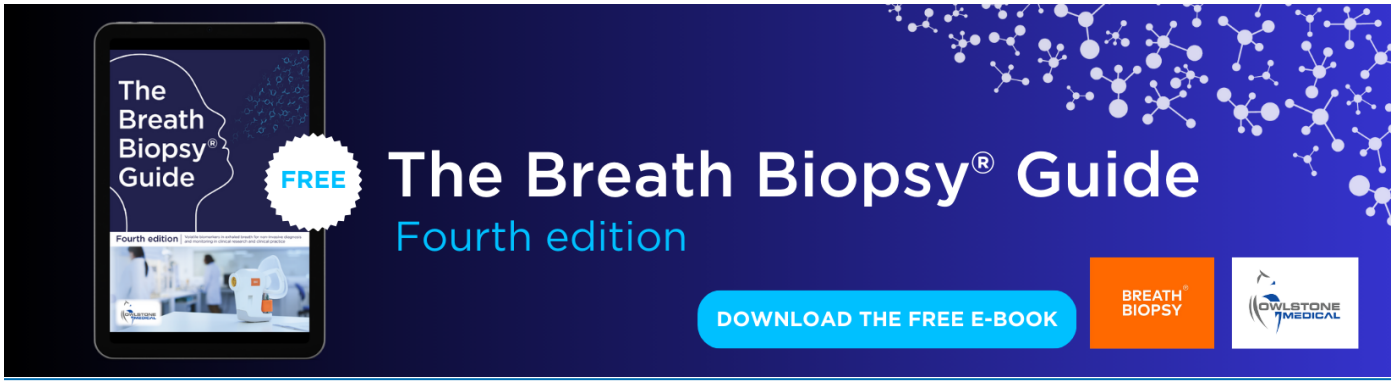
PAPER

Online classification of imagined speech using functional near-infrared spectroscopy signals



IOP Publishing Journal of Neural Engineering
J. Neural Eng. **16** (2019) 016005 (12pp) https://doi.org/10.1088/1741-2552/aae4b9

Online classification of imagined speech using functional near-infrared spectroscopy signals

Alborz Rezazadeh Sereshkeh[1,2], **Rozhin Yousefi**[1,2], **Andrew T Wong**[1,2] **and Tom Chau**[1,2]

[1] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Canada
[2] Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada

E-mail: tom.chau@utoronto.ca

Received 3 June 2018, revised 25 September 2018
Accepted for publication 27 September 2018
Published 16 November 2018

Abstract

Objective. Most brain–computer interfaces (BCIs) based on functional near-infrared spectroscopy (fNIRS) require that users perform mental tasks such as motor imagery, mental arithmetic, or music imagery to convey a message or to answer simple yes or no questions. These cognitive tasks usually have no direct association with the communicative intent, which makes them difficult for users to perform. *Approach*. In this paper, a 3-class intuitive BCI is presented which enables users to directly answer yes or no questions by covertly rehearsing the word 'yes' or 'no' for 15 s. The BCI also admits an equivalent duration of unconstrained rest which constitutes the third discernable task. Twelve participants each completed one offline block and six online blocks over the course of two sessions. The mean value of the change in oxygenated hemoglobin concentration during a trial was calculated for each channel and used to train a regularized linear discriminant analysis (RLDA) classifier. *Main results*. By the final online block, nine out of 12 participants were performing above chance ($p < 0.001$ using the binomial cumulative distribution), with a 3-class accuracy of 83.8% $\pm$ 9.4%. Even when considering all participants, the average online 3-class accuracy over the last three blocks was 64.1 % $\pm$ 20.6%, with only three participants scoring below chance ($p < 0.001$). For most participants, channels in the left temporal and temporoparietal cortex provided the most discriminative information. *Significance*. To our knowledge, this is the first report of an online 3-class imagined speech BCI. Our findings suggest that imagined speech can be used as a reliable activation task for selected users for development of more intuitive BCIs for communication.

Keywords: brain–computer interfaces, functional near-infrared spectroscopy, imagined speech, regularized linear discriminant analysis

(Some figures may appear in colour only in the online journal)

1. Introduction

Brain–computer interfaces (BCIs) can be used to provide a communication channel for individuals with severe motor impairments who are unable to communicate independently [1]. Since the emergence of BCIs, various activation protocols have been suggested and tested. A subset of these protocols are known as reactive BCIs [2], which require the user to attend to external stimuli. Examples include P300 spellers [3] and BCIs based on steady-state visually evoked potentials [4]. BCI protocols that do not require an external stimulus give rise to active BCIs [2], where instead, users perform a mental task. Some of the most common examples of these mental tasks are motor imagery [5], mental arithmetic [6] and word generation [7]. Given an adequate classification accuracy, a BCI user can perform each of these mental tasks to convey a different message, e.g. to answer yes or no questions. However, these mental tasks are usually difficult to perform

by the target population since the tasks are non-intuitive and unrelated to the actual intended message.

An intuitive mental task for BCIs which has attracted attention during the last decade is imagined speech—also known as covert speech [8]. A review of reported BCIs based on imagined speech and their performances are provided in [8, 9]. According to these reviews, invasive measurement techniques such as electrocorticography (ECoG) have been required in most cases where accuracies of classifying electrophysiological brain signals during imagined speech have exceed 70% (the touted threshold for practical BCI application [10]) [11–13]. In contrast, most BCIs based on non-invasive electrophysiological measurements, including electroencephalography (EEG) and magnetoencephalography (MEG), have yielded accuracies less than 70% when discriminating between two different imagined speech tasks [14–16]. Moreover, only one study used a real-time paradigm which reported an average classification accuracy of ~69% using EEG signals recorded during covert repetition of 'yes' and 'no' [17].

Another brainwave response which has been investigated during speech related tasks is the hemodynamic response [18]. Initial studies on the hemodynamic response related to speech generation and comprehension deployed positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) to study activated brain areas [19].

Initial attempts to use the hemodynamic response to decode different speech tasks focused on the averaged hemodynamic response over many repetitions of a speech task [8]. However, a successful imagined speech BCI should be able to decode speech in a single trial [8]. In [20], covert repetition of a nursery rhyme was used as an activation task in a 4-class BCI based on fMRI and yielded an average classification accuracy greater than 90%. However, due to the limitations of fMRI, the duration of each trial was relatively long (~2 min). More importantly, fMRI cannot be used in the development of a portable BCI.

Another modality to measure the hemodynamic response is functional near-infrared spectroscopy (fNIRS). An fNIRS device can be portable, and the duration of each trial can be as short as 10–15 s [21]. Early applications of fNIRS in speech recognition focused on distinguishing among different speech modes: overt, silent and imagined speech, and trials without any speech activity [22, 23]. In [22], each speech task included a whole sentence, and different speech modes were successfully discriminated using fNIRS data. In another fNIRS study, different patterns of hemodynamic responses were reported during trials of inner recitation of hexameter or prose verses [24].

Due to the slow nature of the hemodynamic response [25], decoding small units of language, such as nouns, is more difficult compared to full sentences or different speech modes [8]. Gallegos-Ayala *et al* reported an fNIRS-BCI for answering 'yes' or 'no' questions [26]. This BCI was tested on a patient with amyotrophic lateral sclerosis (ALS) who answered different questions by simply thinking 'yes' or 'no'. The duration of each trial was 25 s and an online classification accuracy of 71.7% was reached for this patient.

Hwang *et al* tested a similar 'yes' or 'no' paradigm on eight able-bodied participants using fNIRS [27]. The duration of each trial was reduced to 10 s. Different types of hemodynamic features, feature numbers and time window sizes were tested and their accuracies were compared. An offline average accuracy of ~75% was reported when the best feature set was employed for each participant. Surprisingly, the location of the fNIRS channels did not cover any of the temporal regions which are some of the most important speech-related brain areas.

In [28], Chaudhary *et al* expanded the work presented in [26]. Four ALS patients used the same fNIRS-BCI to answer yes or no questions by thinking 'yes' or 'no'. An average online classification accuracy of more than 70% (above the chance-level) was reported across participants.

As summarized, none of the previous online non-invasive, portable neuroimaging studies (EEG and fNIRS) have investigated classification of more than two classes. The classification was either limited to imagined speech versus a control condition (e.g. rest) or between two imagined speech tasks. In this study, we developed an fNIRS-BCI for online 3-class classification of the following three tasks: thinking 'yes' while mentally rehearsing the phrase 'yes', thinking 'no' while mentally rehearsing the phrase 'no', and unconditional rest.

The contributions of this work are threefold. Firstly, to the best of our knowledge, this is the first 3-class online BCI based on imagined speech using a portable and non-invasive neuroimaging technique, i.e. fNIRS. Secondly, the impact of using a regularization parameter in the classification model and optimizing and updating its value during the study is investigated. Finally, the role of different channel locations in providing discriminative information between the mental tasks are explored.

2. Methods

2.1. Participants

Twelve able-bodied participants (seven males) between the ages of 23 and 33 (mean age: 28.4 ± 2.9 years) participated in this study. Participants were fluent in English, had normal or corrected-to-normal vision, and had no health issues that could adversely affect the measurements or their ability to follow the experimental protocol. These issues included neurological, cardiovascular, respiratory, psychiatric, metabolic, degenerative, or alcohol-related conditions. Participants were asked to refrain from drinking alcoholic or caffeinated beverages at least 3 h prior to each session. This study was approved by the research ethics boards of the Holland Bloorview Kids Rehabilitation Hospital and the University of Toronto. Written consent was obtained from all participants prior to study participation.

2.2. Instrumentation

fNIRS measurements were collected from the frontal, parietal and temporal cortices using a continuous-wave near-infrared

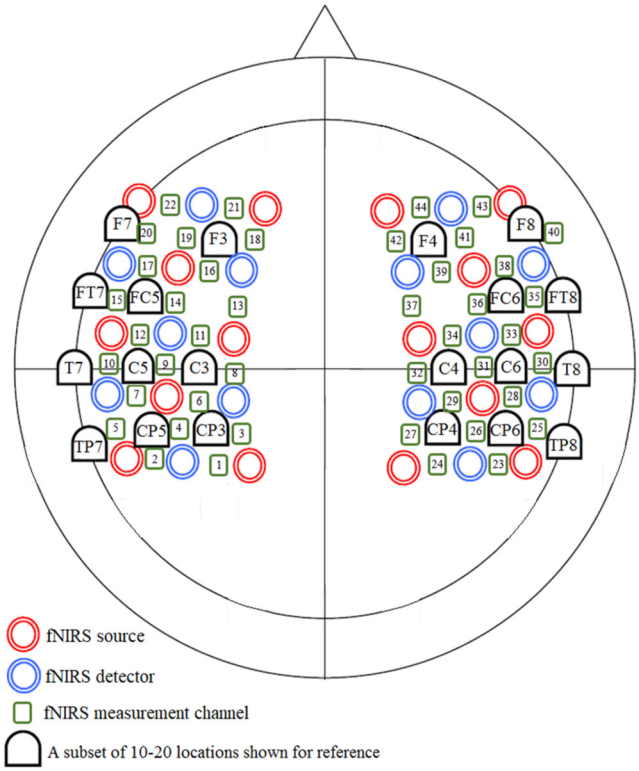


Figure 1. The placement of fNIRS sources and detectors. A subset of 10–20 locations are also shown for reference.

spectrometer (ETG-4000 Optical Topography System, Hitachi Medical Co., Japan). As shown in figure 1, 16 NIR emitters and 14 photodetectors were integrated in two 3 × 5 rectangular grids of optical fibers in a standard EEG cap (EasyCap, Germany). Each NIR emitter contained two laser diodes that simultaneously emitted NIR light at wavelengths of 695 nm and 830 nm. The optical signals were sampled at 10 Hz.

Adjacent positions in each of the two 3 × 5 grids, were 3 cm apart. Only optical signals arising from source-detector pairs (or 'channels') separated by 3 cm were acquired for analysis. This separation distance yielded a depth penetration of light between 2 and 3 cm [29, 30], which surpasses the average scalp-to-cortex depth within the brain areas monitored [31]. Using this configuration, optical signals were acquired from a total of 44 measurement sites on the cerebral cortex, 22 on each hemisphere (see figure 1). In addition to fNIRS measurements, EEG signals were recorded from 32 locations using BrainAmp DC amplifier (Brain Products GmbH, Germany). These data are not analyzed herein.

2.3. Experimental protocol

Participants attended two sessions on two separate days, within a span of 6–21 d. The first session consisted of three blocks, starting with an offline block and followed by two online blocks. In the offline block, participants performed 36 trials, including 12 'yes' imagined speech trials, 12 'no' imagined speech trials and 12 unconstrained rest trials. The trials were presented in a pseudorandom order. At the end of the offline block, a 3-class classifier was trained using the data from the offline block. Each online block consisted of 24 trials, eight trials per class, presented in a pseudorandom order. Participants were presented with the classifier decision subsequent to each trial. The 3-class classifier was re-trained after each block using the data from all previous blocks.

The second session consisted of four online blocks, each with 24 trials equally distributed among the three classes presented in pseudorandom order. Similar to the first session, the 3-class classifier was retrained after each block. The timing diagram is depicted in figure 2.

A fixation cross appeared at the center of a blank screen at the beginning of each trial and persisted throughout the trial. Each trial started with a 14 s baseline period which allowed the hemodynamic signal to return to a basal level [32, 33]. Participants were asked to refrain from performing any of the imagined speech tasks during this period. They had no knowledge of the type of the next trial at the time of baseline collection.

In the imagined speech trials, a question appeared on the screen (below the fixation cross) after the baseline period for 3 s. Then it was replaced by the instruction 'start', which disappeared after 1 s. The question was always the same: 'Is this word in uppercase letters? WORD'. For the yes trials, the word was written in uppercase letters. For the no trials, the word was written in lowercase letters. The words were different in each question and were selected at random from a list of emotionally neutral words suggested by [34]. In the unconstrained rest trials, the phrase 'rest' appeared on the screen (again, below the fixation cross) for 3 s, which was then replaced by the instruction, 'start', for 1 s.

Participants were instructed to commence the mental task as soon as the 'start' instruction disappeared. For the imagined speech trials, participants were instructed to think 'yes' or 'no' while iteratively repeating the word 'yes' or 'no' mentally. They were explicitly instructed to perform the task without any vocalization or motor movement, especially of the lips, tongue or jaw. In the unconstrained 'rest' trials, participants allowed normal thought processes to occur without restriction. The participant was asked to perform the mental task for 15 s for all trial types. This duration was determined based on previous similar fNIRS studies and the suggested minimum measurement time for a hemodynamic response in literature [21].

At the end of each session, the participants were asked to rate from 1 to 5 (where 1 was the lowest and 5 the highest) their perceived ability to perform the task (data not shown).

2.4. Data analysis

2.4.1. Signal processing. First, using the modified Beer–Lambert law [35], we converted optical intensities to relative oxygenated and deoxygenated hemoglobin concentration changes, denoted as [HbO] and [Hb], respectively. The signals were then filtered using a using a 3rd order Chebyshev type II low-pass filter with a passband cutoff frequency of 0.1 Hz, passband ripple of 0.1 dB, stopband cut off frequency of 0.5 Hz and minimum stopband attenuation of 50 dB. This

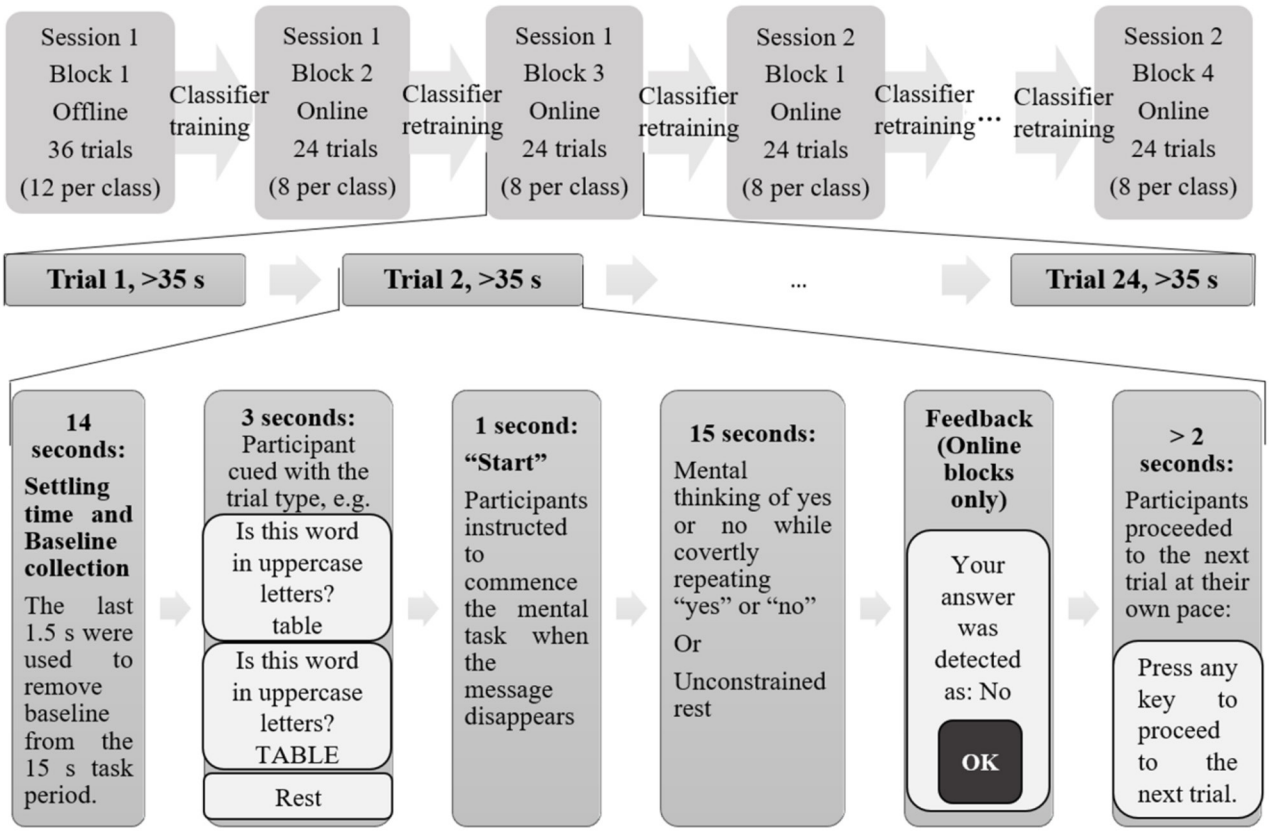


Figure 2. The timing diagram of the experiment.

filter removed any high frequency physiological noise, including Mayer waves at 0.1 Hz, respiration at ~0.3 Hz and cardiac activity at 0.8–1.2 Hz [36–38]. For participants 1, 2, 10 and 12, after the initial setup and visual evaluation of the fNIRS data, a number of fNIRS channels (five, secen, six and six channels, respectively), were discarded from the training data due to excessive noise. The data collected from these channels were omitted from further analysis.

2.4.2. Baseline removal. The baseline value of HbO can change from 1 d to another or even from the beginning to the end of a session [39]. Hence, some BCI studies have added baseline collection periods to the beginning of each session or block to adjust for this natural fluctuation [39, 40].

In this study, baseline data were collected prior to each trial to calculate a more precise and trial-specific mean baseline value. From the 14 s settling time and baseline period, we calculated the mean of [HbO] during the last 1500 ms for each fNIRS channel and subtracted this value from the subsequent trial on a per-channel basis. The last 1.5 s was chosen instead of the entire 14 s; given the duration of the mental task in this experiment, the hemodynamic signal required approximately 12 s to return to its baseline value [32, 33, 41].

2.4.3. Feature extraction. The mean value of [HbO] for each channel during the entire length of each mental task (15 s) were used as features for classification. Hence, each trial was represented as a 1×44 vector of features (44 channels $\times$ 1 feature).

Other common types of fNIRS features are variance, slope, skewness and kurtosis of [HbO], [Hb], and changes in total hemoglobin concentrations. These features were examined during pilot sessions, but the mean of [HbO] led to the highest classification accuracy and therefore was selected to provide real-time feedback during the online trials. This feature has been previously used in a similar 'yes' versus 'no' fNIRS study on ALS patients [28]. Furthermore, it has been shown in another 'yes' versus 'no' study on healthy participants [27] that features extracted from [HbO] provide more discriminative information than features derived from [Hb].

It should be noted that given enough training data, other types of features, such as slope and variance, as well as features derived from deoxyhemoglobin concentrations may provide additional discriminatory information. During pilot sessions, the addition of these features resulted in lower classification accuracy, which may be attributed to the small number of trials in this study. In other words, considering the number of trials (52 trials per class prior to the last online block and even fewer trials prior to earlier blocks), extracting only one feature from each channel (in this case, mean [HbO]) provided the best performance during pilot sessions. However, a longer feature vector with the inclusion of different feature types generated from oxygenated and deoxygenated hemoglobin concentrations, may have resulted in the best performance if more trials were available. In that case, the classifier could be retrained to take advantage of the expanded feature set.

2.4.4. Classification. For classification, a regularized linear discriminant analysis (RLDA) algorithm was used [42]. This

method was chosen as it led to the highest average accuracy during the pilot sessions compared to support vector machines (linear, polynomial, radial basis function and sigmoid kernels), neural networks (multilayer perceptron with one hidden layer) and naïve Bayes classifiers.

To discriminate between the three classes, a multiclass LDA was used for classification. In contrast with other types of discriminant analysis, e.g. quadratic discriminant analysis, LDA assumes that all classes have the same covariance. This common pooled covariance matrix is defined as:

$$\widehat{\Sigma} = \sum_{k=1}^{K} \sum_{i \in I_k} (X_i - \mu_k)(X_i - \mu_k)^T / (N - K) \tag{1}$$

where K is the number of classes, X_i is the feature vector for the ith example, $I_k = \{i \mid y_i = k\}$ is the subset of indices identifying the examples of the kth class, y_i is the class label of the ith example, μ_k is the mean of all examples of the kth class, and N is the total number of examples.

LDA classification is done based on the analysis of the following two scatter matrices: the within-class scatter matrix and the between-class scatter matrix. The within-class scatter matrix can be expressed in terms of the common covariance matrix defined in equation (1):

$$S_{\mathrm{w}} = (N - K) \times \widehat{\Sigma}. \tag{2}$$

The between-class scatter matrix is defined as:

$$S_{\mathrm{b}} = \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{3}$$

where μ is the overall mean of all examples and N_k is the number of examples in the kth class, or $N_k = |I_k|$ where $|.|$ denotes cardinality.

The main goal of LDA is to find a set of coefficients, $\mathbf{W}$, that maximizes the following ratio:

$$W_{\mathrm{LDA}} = \underset{W}{\operatorname{argmax}} \frac{W^T S_{\mathrm{b}} W}{W^T S_{\mathrm{w}} W}. \tag{4}$$

This ratio is called the Fisher criterion.

In regularized LDA, the common pooled covariance matrix is replaced with the following covariance matrix for each class:

$$\widehat{\Sigma}_\gamma = (1 - \gamma)\,\widehat{\Sigma} + \gamma \cdot \mathrm{diag}\left(\widehat{\Sigma}\right) \tag{5}$$

where $\mathrm{diag}\left(\widehat{\Sigma}\right)$ are the diagonal elements of $\widehat{\Sigma}$ and γ is the regularization parameter. It can be seen that when γ is equal to zero, $\widehat{\Sigma}_\gamma$ is equal to $\widehat{\Sigma}$, and the optimization equation reduces to that of non-regularized LDA.

2.4.5. Optimization of the regularization parameter. The only hyper-parameter in an RLDA classifier is the regularization parameter, gamma, which can be any value between 0 and 1. This parameter was optimized every time the classifier was trained using a leave-one-out cross-validation (LOOCV) method. Prior to each online block, we calculated the LOOCV accuracy on the data from all previous blocks for different gamma values in the range of 0–1 in 0.05 increments, with two exceptions which are explained below. The gamma which resulted in the highest LOOCV accuracy was selected for subsequent classifier training. In case of a tie, the largest gamma was selected to obtain a more generalized classifier. The classifier was then trained on the entire training set using that gamma.

Two restrictions were applied to the gamma range. Firstly, during each session, the maximum value of gamma in the gamma-optimization step was set to the gamma used in the previous block. Since more same-day data was acquired as the session progressed, the need for generalizing the classifier was reduced and the classifier could be further optimized. Secondly, at the beginning of the second session, a minimum of 0.3 was used for gamma to prevent overfitting, i.e. overemphasis on data from the first session and thereby preserve generalizability. The value of 0.3 was optimized and selected during pilot sessions. The data analysis steps are summarized in figure 3.

3. Results

3.1. Online 3-class accuracies

Table 1 provides the online 3-class classification accuracies obtained during the six online blocks performed by each participant. Nine out of twelve participants reached above-chance online classification accuracy in their final three blocks ($p < 0.001$ using the binomial cumulative distribution [43]), achieving an average online accuracy above the 70% threshold (minimum acceptable threshold for practical BCI applications [10]).

Excluding three participants, P5, P7 and P11, whose second sessions were interrupted upon participants' request to remove the cap due to discomfort and fatigue.

For participants P5, P7 and P11, the second session was interrupted as these participants asked to have the cap removed due to discomfort. After the removal of the cap, these participants took a short break and continued the experiment. In the post-session questionnaire, in response to the question of 'How hard was it to perform the mental tasks?', all three of these participants chose 5 (on a scale of 1–5, 5 being the hardest) and stated that that the task was difficult (5 on a scale of 1–5) to perform given the discomfort of the cap. Due to this reason, as well as potential variations in fNIRS cap positioning between successive donning of the cap, the mean accuracy is also reported without these three participants in table 1. Note that calculating the mean accuracy without these three participants is done only as a secondary analysis and is not part of the primary results and discussion.

3.2. The role of different fNIRS channels in providing discriminative information

In order to determine the role of each fNIRS channels in providing the discriminative information, we used the value of the Fisher criterion calculated for each feature. Since RLDA was used for classification (which works based on

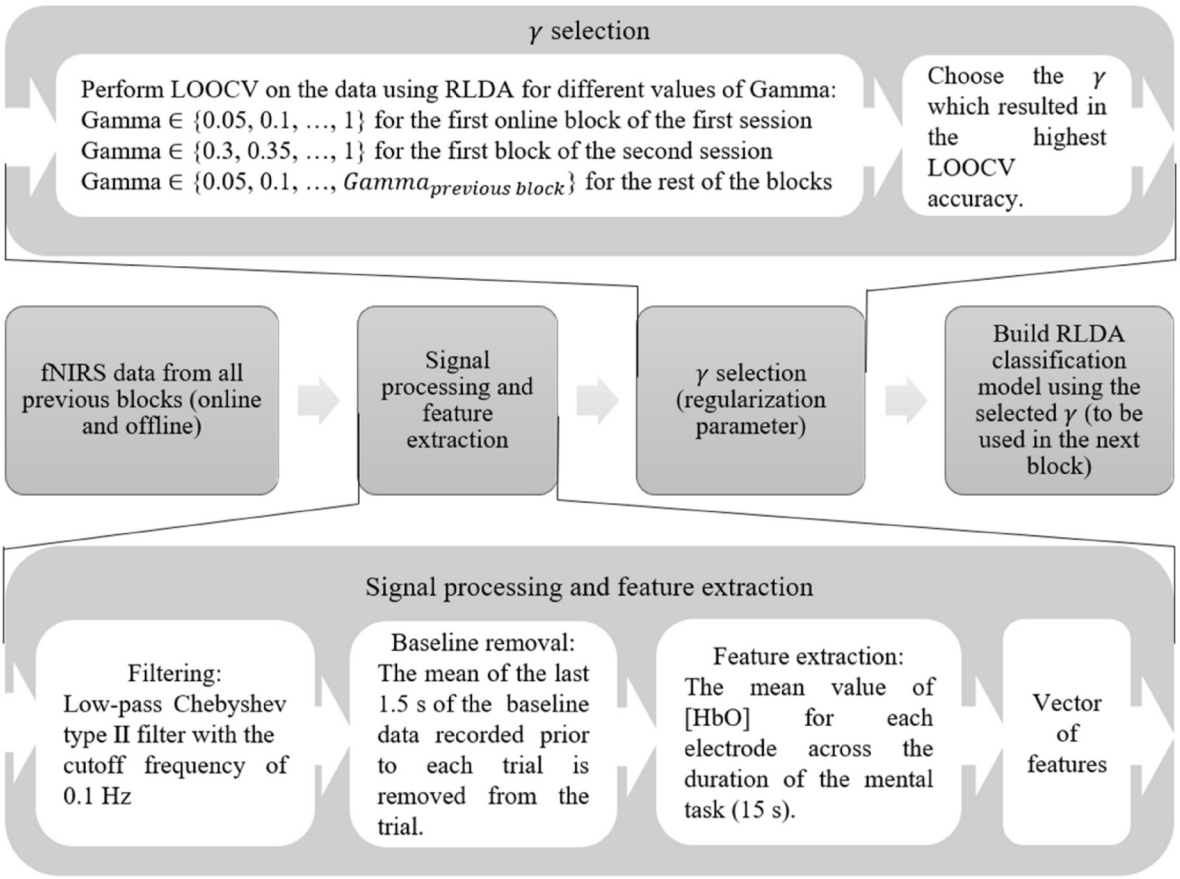


Figure 3. The mathematical steps for building a classifier prior to each online block (LOOCV = leave one out cross-validation, RLDA = regularized linear discriminant analysis).

Table 1. Online 3-class accuracies (%) for each participant for all online blocks. Average accuracies exceeding the upper limit of the 95%, 99% and 99.9% confidence interval of chance are marked with *, ** and ***, respectively. These limits were calculated using the binomial cumulative distribution and based on the number of trials [43]. For individual test blocks (24 trials), these limits were 50.0%, 58.3% and 62.5%, respectively. For the combination of all blocks (72 trials), these limits were 43.1%, 45.8% and 51.4%, respectively.

Participant	Session one—block two	Session one—block three	Session two—block one	Session two—block two	Session two—block three	Session two—block four	Average of last three blocks
P1	50.0*	87.5***	75.0***	75.0***	62.5***	79.2***	72.2***
P2	37.5	45.8	37.5	83.3***	83.3***	95.8***	87.5***
P3	75.0***	66.7***	58.3**	66.7***	75.0***	79.2***	73.6***
P4	75.0***	66.7***	41.7	66.7***	70.8***	83.3***	73.6***
P5	41.7	54.2*	37.5	33.3	45.8	25.0	34.7
P6	95.8***	100***	87.5***	83.3***	100***	100***	94.4***
P7	62.5***	58.3**	62.5***	37.5	45.8	75.0***	52.8***
P8	62.5***	41.7	41.7	58.3**	50.0*	83.3***	63.9***
P9	45.8	50.0*	41.7	54.2*	50.0*	70.8***	58.3***
P10	83.3***	79.2***	62.5***	70.8***	66.7***	87.5***	75.0***
P11	29.2	33.3	41.7	45.8	45.8	25.0	38.9
P12	37.5	58.3**	58.3**	33.3	45.8	54.2*	44.4*
Mean (all participants)	58.0 ± 21.0	61.8 ± 19.4	53.8 ± 16.2	59.0 ± 18.3	61.8 ± 17.8	71.5 ± 24.7	64.1 ± 20.6
Mean (P1–P4, P6, P8–P10, P12)	62.5 ± 21.0	66.2 ± 19.7	56.0 ± 17.2	65.7 ± 15.7	67.1 ± 17.6	81.5 ± 13.5	71.5 ± 16.7

maximizing the Fisher criterion) and each fNIRS channel produced only one feature, the calculated Fisher criterion for that feature represented the level of discriminative information that fNIRS channel provided. Figure 4(a) depicts the brain map of the calculated Fisher criterion for each channel averaged across participants. Figure 4(b) provides the brain map of the standard deviation of the calculated Fisher criterion across participants.

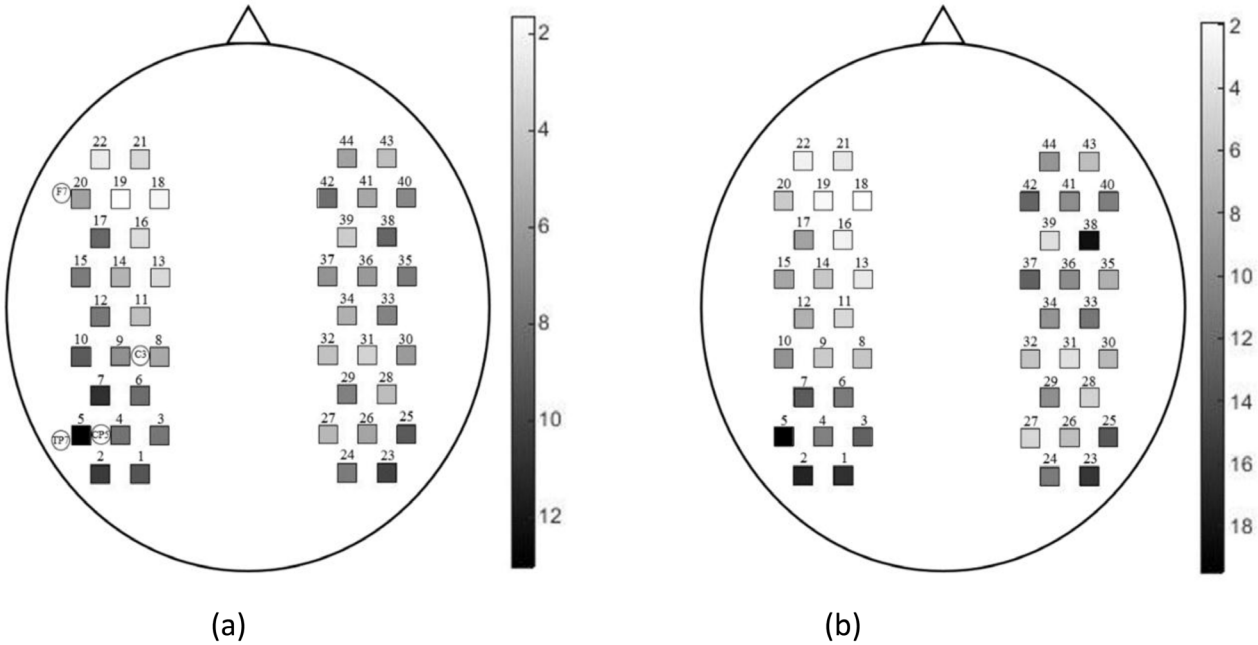


Figure 4. The brain map of (a) the average of the Fisher criterion value across participants and (b) the standard deviation of the Fisher criterion value across participants.

4. Discussion

4.1. Comparison with previous multiclass fNIRS-BCIs and previous imagined speech BCIs

In this paper, we proposed an online 3-class BCI based on imagined speech. An average ternary classification accuracy of 71.5% ± 24.7% was reached across all participants in their last block, with nine out of 12 participants surpassing the chance level ($p < 0.001$ using the binomial cumulative distribution [43]).

Some studies have explored the possibility of developing a multiclass (>2 classes) BCI using fNIRS [44]. Power *et al* [6] developed an fNIRS-BCI to classify between mental singing, mental arithmetic and unconstrained rest and reported an offline ternary classification accuracy of 56.2% ± 8.7% across seven participants. Herff *et al* [45] used fNIRS to classify between three levels of the *n*-back task (where participants were instructed to continuously remember the last *n* letters of a series of rapidly flashing letters) and the rest state, reporting an average offline accuracy of 44.5% ± 10.0% across ten participants. Weyand *et al* [46] investigated different combinations of six cognitive tasks and reported an offline ternary accuracy of 60.5% ± 6.0% across ten participants. Recently, Schudlo *et al* [44] reported one of the first online 3-class fNIRS-BCIs. The three tasks included verbal fluency, Stroop task and rest and were differentiated online with an accuracy of 74.2% ± 14.8% across 11 participants.

Using a BCI for online classification of brainwaves when participants mentally think yes or no has been limited to two studies, one with EEG [17] and one with fNIRS [28]. Both studies reported ~70% average binary classification accuracies (see section 1 for a more extensive summary of these BCIs).

In terms of the average classification accuracy across participants, our results surpassed the outcome of all previous mentioned BCIs except [44]. However, the Stroop task used in [44] required users to attend to a screen which is not practical for individuals with visual impairments.

Our results exhibited higher standard deviation across participants, namely, 18.3%, 17.8% and 24.7% in the last three online blocks, compared to previously mentioned BCIs. The standard deviation increased in the last block since most participants obtained higher accuracies (e.g. 100% for the last online block of P6) as the session progressed while a few hovered at chance level accuracies across all blocks (e.g. 25% in the last online block for P5 and P11). If we were to exclude the participants who were unable to complete the session without interruptions, the accuracy in the last block would jump to 81.5% ± 13.5%. This new standard deviation is in the same range as those previously reported for multiclass BCIs. Note that P5 and P11 had difficulties obtaining accuracies above chance in both sessions, and P7 experienced difficulties at the beginning of session two. This is not unusual as certain individuals may have difficulties performing certain tasks or using certain BCI modalities (sometimes referred to 'BCI illiteracy' [47]).

4.2. The role of different brain regions

In figure 4(a), we see that the channels in the left temporal and left temporoparietal regions yielded the highest Fisher criterion value, and therefore provided the most discriminative information. Channels five, seven and two provided the three highest average Fisher criterion values (13.02, 10.80 and 10.33, respectively). Channel five is located between CP5 and TP7, while channel seven is positioned between CP5 and

C5, and channel two is situated close to CP5. Although the exact Brodmann areas of these channels cannot be determined without an fMRI scan, previous concurrent EEG-fMRI studies can provide an estimation of the associated Brodmann regions of these channels. Based on the channel maps from [48], these three channels (five, seven and two) approximately cover parts of Brodmann areas 21, 22, 39, 40 and 42. These Brodmann areas represent in part, Wernicke's area, the left angular gyrus, the left supramarginal gyrus and the left auditory cortex. All these areas are belong to the speech network of the brain [49] and have been previously identified in other imagined speech studies as yielding discriminative information [19].

Another channel of note is channel 20, as it is close to F7 and Broca's area (see figure 1) [48]. Although Broca's area is known to play an important role in speech production, the average Fisher criterion of this channel was 5.80, ranking it 20th out of 44 channels. This finding is in line with several previous studies on the classification of imagined speech, where greater discriminative information was found in the temporal and temporoparietal regions close to Wernicke's area, compared to Broca's area, especially when the imagined speech task did not involve the production of complicated phrases [17, 19].

Figure 4(b) depicts the brain map of the standard deviation of the calculated Fisher criterion across all participants. Again, channel five provided the highest standard deviation of the Fisher criterion. The value of the Fisher criterion varied from 0.31 (P12) to 64.33 (P6) for this channel. The large variations in the Fisher criterion values of the speech-related regions may be attributable to inter-individual performance variations of the imagined speech task. As mentioned, all participants were instructed to think yes or no while covertly repeating the phrase without any motor movements. Since they were given the freedom to 'think' yes or no in their own way, some individuals may have focused on the meaning of the phrases (affirmative versus negative response), the articulation of the phrases (with or without motor imagery of the articulation), or on imagining hearing the phrases, while covertly repeating them. For example, the highest Fisher criterion for participant 12 was obtained on channel eight (located approximately between C3 and C1) with the value of 7.09, while channel five produced the lowest Fisher criterion at 0.31. The area covered by channel eight is known to be activated during motor imagery tasks, which may indicate that this participant mainly focused on imagining motor movements required for speech production. The low accuracy of P12 compared to other participants seems to support the hypothesis that P12 may have utilized a unique approach to covert speech.

The variation across participants in the location of the channels which provided the maximum discriminative information for classification has been frequently reported in previous imagined speech studies [8, 19], and more generally, in most active BCI tasks [2]. Other than participant-specific performance of active mental tasks, this inconsistency could also, in part, be attributed to inter-individual variation in the shape and size of various brain regions. fMRI or similar imaging techniques could be used to confirm the brain regions interrogated at each of the 10–20 locations. Therefore, without the use of fMRI and structural data for each individual, it is not possible to assign a 10–20 location to a specific brain region and make a claim about the performance of a specific brain region [48].

In order to illustrate how [HbO] changed during a trial in channel five, which provided the highest average and standard deviation of Fisher criterion and was approximately the closest channel to Wernicke's area [48], a graph illustrating [HbO] versus time, averaged over all trials of each participant is shown in figure 5. Individualized activation patterns were elicited in 'yes', 'no' and rest trials, which may be attributable to participant-specific performance of imagined speech. Unsurprisingly, the difference among the three trial types was generally more visually discernable in the data of participants with the highest classification accuracies.

4.3. Accuracies across different blocks

As seen in table 1, the last online block yielded the highest average accuracy across participants, which was significantly higher than the first online block ($p = 0.022$; as determined by a Friedman test on the accuracies from different blocks and post-hoc Wilcoxon signed rank tests on the accuracies of every pair of blocks with a Holm–Bonferroni correction for multiple comparisons). This increase in average accuracy is likely the combined effect of two factors: improved classifier robustness due to the accumulation of training data and more consistent task performance (and hence brain signals) by the user upon receiving real-time feedback [1]. The important role of feedback in improving the performance in BCI training has been reported in literature [50, 51]. Moreover, Schultz *et al* emphasized the importance of real-time feedback in BCIs based on speech [8].

At the beginning of the second session, there was a drop in the average accuracy. As the classifier was trained using data from a different day, this decline in accuracy may be attributable to slight variations in fNIRS cap positioning, changes in mental states between sessions (e.g. fatigue or attention [52]), or variations in metabolic states [53].

4.4. The role of regularization

Regularization can be necessary in classification models to preserve generalizability, especially when the number of samples are of the same order of magnitude as the number of features [54]. To determine whether using a regularization parameter was helpful for online classification, we retrospectively calculated the accuracies for all online blocks without any regularization. As evident in figure 6, regularization improved the average accuracies across all blocks across participants ($p < 0.01$), as suggested by a two-way repeated measures ANOVA with accuracy as the dependent variable and regularization (i.e. regularized versus not regularized) and block number as independent factors. However, when the classification accuracies with and without regularization were compared separately for each block, only the first three online blocks exhibited a significant difference ($p = 0.003$, 0.032 and 0.044 for the first three blocks, respectively, using

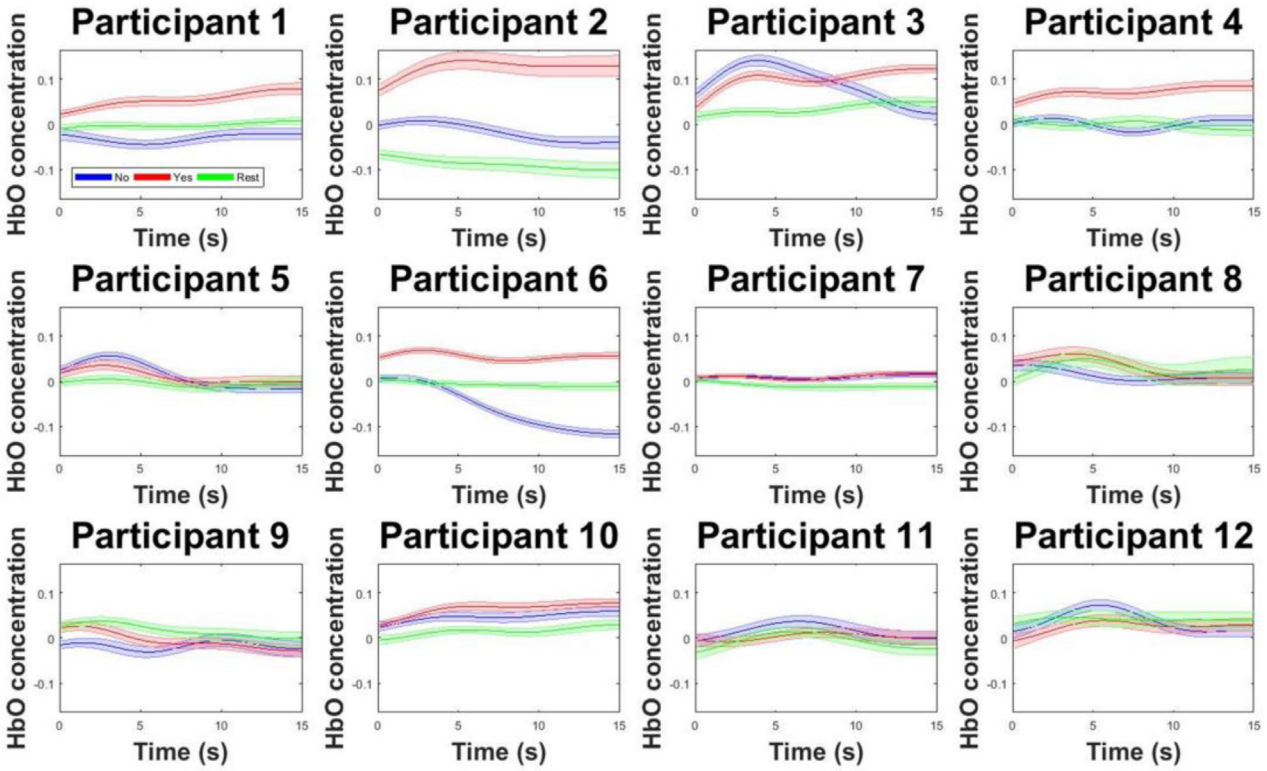


Figure 5. Averaged [HbO] responses recorded during the three trial types for channel five. The shaded regions indicate the standard errors computed across all trials of the same class.

Wilcoxon signed rank test and Holm–Bonferroni correction for multiple comparisons). The significant difference in early blocks confirmed the importance of regularization when the training dataset is relatively small, as well as when a BCI is trained only on the data from a previous day. The difference was not significant in the remaining blocks, the last three online blocks, due to the inclusion of more same-day data in the classifier training.

The selected gamma values for all participants in different blocks as well as the changes across blocks can be found in figure 7. Note that, as stated in the section 2.4.5, during each session, we restricted the upper limit of the search range for the optimized gamma to the selected gamma in the previous block. So, across each session, the selected value of gamma could either remain unchanged or decrease. As seen in figure 7, for all participants except one, the chosen gamma value decreased throughout each session most probably due to the increase in the number of training examples. Having more training examples decreases the risk of overfitting, and hence smaller regularization parameters can yield higher cross-validation accuracies.

4.5. Toward an asynchronous 2-class BCI

The control task in this study was an 'unconstrained rest'—participants were only asked to refrain from performing the other two imagined speech tasks during these trials. The admission of an idle state facilitates potential asynchronous implementation. As such, for users who achieved reliable 3-class accuracy (such as P2 and P6), imagined speech might eventually be exploited in a 2-class asynchronous BCI, where the BCI can be activated by mentally repeating the phrases 'yes' or 'no'. Such an asynchronous BCI could be used as a binary switch for an assistive device (i.e. with two activation modes) where the user might could call his/her caregiver, or start a music player, for example, without additional prompts. Using 'yes' or 'no' mental tasks to activate an assistive device may not be intuitive, but these tasks can be easier to perform than those commonly invoked in fNIRS-BCIs (e.g. mental arithmetic) [17].

Depending on the application and preference of each user, the sensitivity and specificity levels for each activation task can be tuned. For example, if the task is of high importance, such as activating a call bell for assistance, the user may prefer a high sensitivity setting to err on the side of caution. On the other hand, if accidental activations are unwanted, such as switching on and off a music player, a higher specificity may be warranted.

It should be noted that even for participants with highly reliable performance with the synchronous BCI, further training sessions within an asynchronous paradigm are required to gauge the feasibility of a truly user-paced imagined speech BCI. The development of such a BCI might proceed with only one imagined speech task to activate the BCI (like an on/off switch) and if deemed reliable after a few sessions, a second imagined speech task could be added.

4.6. Limitations and future directions

For future studies, the authors suggest the use of additional sessions, as the increased number of trials may

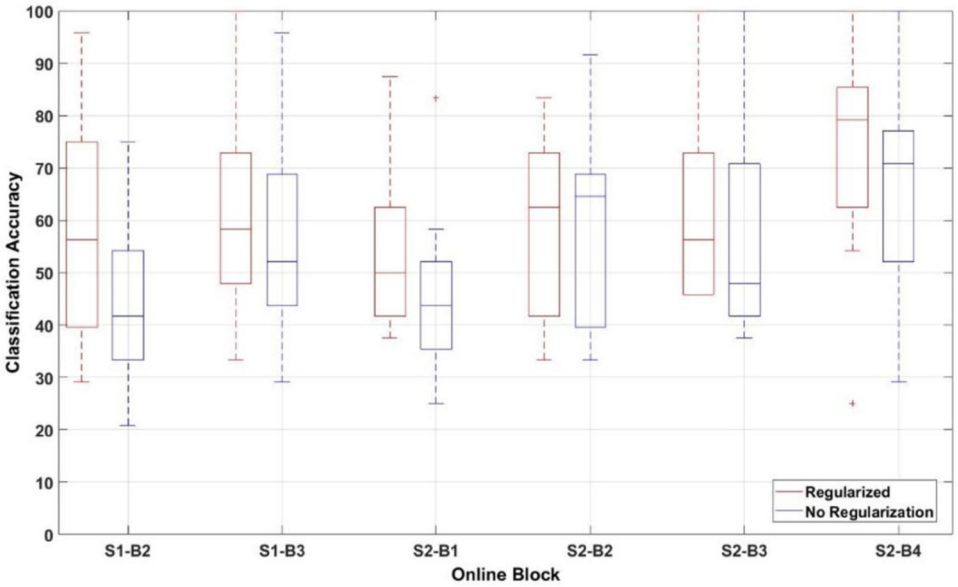


Figure 6. The classification accuracies in different online blocks averaged over all participants with (red) and without (blue) regularization. The notation Sn—Bm identifies the *m*th block (B) of the *n*th session (S).

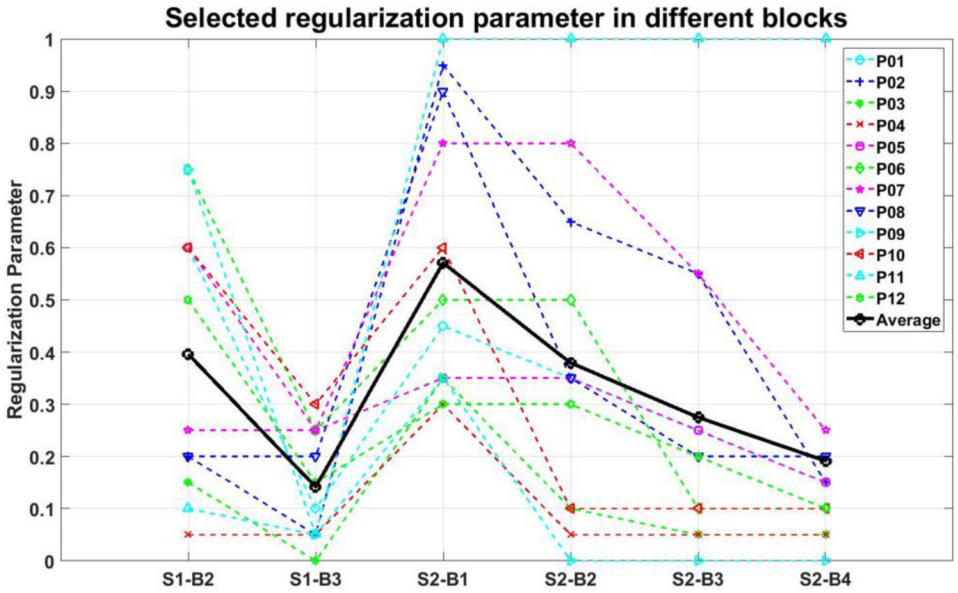


Figure 7. Changes in the selected regularization parameter, γ, for different participants across different blocks. The selected γ is the value which provided the highest leave-one-out cross validation accuracy on the data from all previous blocks. The notation Sn—Bm identifies the *m*th block (B) of the *n*th session (S).

enhance classifier performance. This study demonstrated that during the second session, the average accuracies during the last online block were significantly higher than those of the first online block, which is possibly due to the increased training data. Additional sessions would shed further insight on the achievable classifier performance and robustness. Also, prior to clinical translation, the findings herein must be replicated with individuals who present as locked-in.

Future research could also explore additional BCI-specific of additional BCI-specific intuitive commands, such as 'left', 'right', 'stop' and 'go' for navigation. Using words other than 'yes' and 'no' will also reveal if the classification results obtained herein were mainly due to users' intention to provide an affirmative versus a negative response, or due to the difference between covert articulation of 'yes' and 'no'. Also, for future imagined speech BCI studies on able-bodied participants, an ultrasound system could be used to detect and discard trials which may contain possible motor confounds associated with subvocalization.

Finally, as each modality has been individually applied to the classification of imagined speech, a combination of EEG and fNIRS may exploit the advantages of each modality, potentially leading to improved BCI performance.

5. Conclusion

This study investigated an intuitive 3-class BCI based on imagined speech. Our findings suggest that fNIRS is a suitable modality for reliably differentiating affirmative and negative responses from unconstrained rest for selected BCI users. An average online classification accuracy of 64.1% ± 20.6% was reached across all participants in the last three online blocks with nine participants exceeding the chance level ($p < 0.001$). Task-related differences in the left temporal and left temporo-parietal regions tended to provide discriminatory information. The proposed BCI could eventually empower individuals with severe disabilities with an intuitive means of interacting with their environment. To our knowledge, this is the first report of an online fNIRS 3-class classification of imagined speech.

Acknowledgments

The authors would like to thank Dr Frank Rudzicz, Dr Stephen Strother and Dr Silvia Orlandi for their guidance throughout this project. Special thanks also go to all the members of the PRISM Lab. This work was supported by the Natural Science and Engineering Research Council of Canada. Funding bodies had no involvement in the study, nor in the decision to publish.

ORCID iDs

Alborz Rezazadeh Sereshkeh https://orcid.org/0000-0003-2680-646X
Rozhin Yousefi https://orcid.org/0000-0003-2402-5467
Tom Chau https://orcid.org/0000-0002-7486-0316

References

[1] van Gerven M *et al* 2009 The brain–computer interface cycle *J. Neural Eng.* **6** 041001
[2] Nicolas-Alonso L F and Gomez-Gil J 2012 Brain computer interfaces, a review *Sensors* **12** 1211–79
[3] Krusienski D J, Sellers E W, McFarland D J, Vaughan T M and Wolpaw J R 2008 Toward enhanced P300 speller performance *J. Neurosci. Methods* **167** 15–21
[4] Nezamfar H, Mohseni Salehi S S, Moghadamfalahi M and Erdogmus D 2016 FlashTypeTM: a context-aware c-VEP-based BCI typing interface using EEG signals *IEEE J. Sel. Top. Signal Process.* **10** 932–41
[5] Pfurtscheller G, Brunner C, Schlögl A and Lopes da Silva F H 2006 Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks *NeuroImage* **31** 153–9
[6] Power S D, Kushki A and Chau T 2012 Automatic single-trial discrimination of mental arithmetic, mental singing and the no-control state from prefrontal activity: Toward a three-state NIRS-BCI *BMC Res. Notes* **5** 141
[7] Weyand S, Schudlo L, Takehara-Nishiuchi K and Chau T 2015 Usability and performance-informed selection of personalized mental tasks for an online near-infrared spectroscopy brain-computer interface *Neurophotonics* **2** 025001
[8] Schultz T, Wand M, Hueber T, Krusienski D J, Herff C and Brumberg J S 2017 Biosignal-based spoken communication: a survey *IEEE/ACM Trans. Audio Speech Lang. Process.* **25** 2257–71
[9] Rezazadeh Sereshkeh A, Trott R, Bricout A and Chau T 2017 EEG classification of covert speech using regularized neural networks *IEEE/ACM Trans. Audio Speech Lang. Process.* **25** 2292–300 (https://ieeexplore.ieee.org/document/8114360)
[10] Kübler A, Mushahwar V K, Hochberg L R and Donoghue J P 2006 BCI meeting 2005—workshop on clinical issues and applications *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 131–4
[11] Brumberg J S, Krusienski D J, Chakrabarti S, Gunduz A, Brunner P, Ritaccio A L and Schalk G 2016 Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task *PLoS One* **11** 1–21
[12] Mugler E M, Patton J L, Flint R D, Wright Z A, Schuele S U, Rosenow J, Shih J J, Krusienski D J and Slutzky M W 2014 Direct classification of all American English phonemes using signals from functional speech motor cortex *J. Neural Eng.* **11** 035015
[13] Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone N E, Rieger J, Schalk G, Knight R T and Pasley B N 2014 Decoding spectrotemporal features of overt and covert speech from the human cortex *Frontiers Neuroeng.* **7** 1–15
[14] DaSalla C S, Kambara H, Sato M and Koike Y 2009 Single-trial classification of vowel speech imagery using common spatial patterns *Neural Netw.* **22** 1334–9
[15] Brigham K and Kumar B V K V 2010 Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy *2010 4th Int. Conf. Bioinformatics and Biomedical Engineering iCBBE 2010* pp 1–4
[16] Zhao S and Rudzicz F 2015 Classifying phonological categories in imagined and articulated speech *ICASSP, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Proc.* vol 2015 pp 992–6
[17] Sereshkeh A R, Trott R, Bricout A and Chau T 2017 Online EEG classification of covert speech for brain–computer interfacing *Int. J. Neural Syst.* **27** 1750033
[18] Weiskopf N, Mathiak K, Bock S W, Scharnowski F, Veit R, Grodd W, Goebel R and Birbaumer N 2004 Principles of a brain-computer interface (BCI) based on real-time functional magnetic resonance imaging (fMRI) *IEEE Trans. Biomed. Eng.* **51** 966–70
[19] Price C J 2012 A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading *NeuroImage* **62** 816–47
[20] Yoo S-S, Fairneny T, Chen N-K, Choo S-E, Panych L P, Park H, Lee S-Y and Jolesz F A 2004 Brain–computer interface using fMRI: spatial navigation by thoughts *Neuroreport* **15** 1591–5
[21] Naseer N and Hong K-S 2015 fNIRS-based brain–computer interfaces: a review *Frontiers Hum. Neurosci.* **9** 1–15
[22] Herff C, Putze F, Heger D, Guan C and Schultz T 2012 Speaking mode recognition from functional near infrared spectroscopy *Proc. Annual Int. Conf. IEEE Eng. Med. Biol. Soc., EMBS*
[23] Herff C, Heger D, Putze F, Guan C and Schultz T 2012 Cross-subject classification of speaking modes using fNIRS *Int. Conf. on Neural Information Processing* (Berlin, Heidelberg: Springer) pp 417–24
[24] Scholkmann F, Wolf M and Wolf U 2013 The effect of inner speech on arterial CO_2 and cerebral hemodynamics and oxygenation: a functional NIRS study *Oxygen Transport to Tissue XXXV* (*Advances in Experimental Medicine*

and Biology vol 789) ed S Van Huffel *et al* (New York: Springer) pp 81–7
[25] Miezin F M, Maccotta L, Ollinger J M, Petersen S E and Buckner R L 2000 Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing *NeuroImage* **11** 735–59
[26] Gallegos-Ayala G, Furdea A, Takano K, Ruf C A, Flor H and Birbaumer N 2014 Brain communication in a completely locked-in patient using bedside near-infrared spectroscopy *Neurology* **82** 1930–2
[27] Hwang H-J, Choi H, Kim J-Y, Chang W-D, Kim D-W, Kim K, Jo S and Im C-H 2016 Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy *J. Biomed. Opt.* **21** 091303
[28] Chaudhary U, Xia B, Silvoni S, Cohen L G and Birbaumer N 2017 Brain–computer interface–based communication in the completely locked-in state *PLoS Biol.* **15** 1–26
[29] Cui X, Bray S, Bryant D M, Glover G H and Reiss A L 2011 A quantitative comparison of NIRS and fMRI across multiple cognitive tasks *NeuroImage* **54** 2808–21
[30] Haeussinger F B, Heinzel S, Hahn T, Schecklmann M, Ehlis A C and Fallgatter A J 2011 Simulation of near-infrared light absorption considering individual head and prefrontal cortex anatomy: Implications for optical neuroimaging *PLoS One* **6** e26377
[31] Okamoto M *et al* 2004 Three-dimensional probabilistic anatomical cranio-cerebral correlation via the international 10–20 system oriented for transcranial functional brain mapping *NeuroImage* **21** 99–111
[32] Schudlo L C and Chau T 2018 Development and testing an online near-infrared spectroscopy brain–computer interface tailored to an individual with severe congenital motor impairments *Disabil. Rehabil. Assist. Technol.* **13** 581–91
[33] Thompson D E *et al* 2014 Performance measurement for brain-computer or brain-machine interfaces: a tutorial *J. Neural Eng.* **11** 035001
[34] Bradley M M and Lang P P J 1999 Affective norms for english words (ANEW): instruction manual and affective ratings *Technical Report* C-1 (Florida: The Center for Research in Psychophysiology, University of Florida) (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.306.3881&rep=rep1&type=pdf)
[35] Yamashita Y, Maki A and Koizumi H 2001 Wavelength dependence of the precision of noninvasive optical measurement of oxy-, deoxy-, and total-hemoglobin concentration *Med. Phys.* **28** 1108–14
[36] Boas D A, Dale A M and Franceschini M A 2004 Diffuse optical imaging of brain activation: Approaches to optimizing image sensitivity, resolution, and accuracy *NeuroImage* **23** S275–S288
[37] Zhang Y, Brooks D H, Franceschini M A and Boas D A 2005 Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging *J. Biomed. Opt.* **10** 011014
[38] Franceschini M A, Joseph D K, Huppert T J, Diamond S G and Boas D A 2006 Diffuse optical imaging of the whole head *J. Biomed. Opt.* **11** 054007
[39] Power S D, Kushki A and Chau T 2012 Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by NIRS *PLoS One* **7** e37791
[40] Moghimi S, Kushki A, Power S, Guerguerian A M and Chau T 2012 Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy *J. Neural Eng.* **9** 026022
[41] Yin X, Xu B, Jiang C, Fu Y, Wang Z, Li H and Shi G 2015 A hybrid BCI based on EEG and fNIRS signals improves the performance of decoding motor imagery of both force and speed of hand clenching *J. Neural Eng.* **12** 36004
[42] Guo Y, Hastie T and Tibshirani R 2007 Regularized linear discriminant analysis and its application in microarrays *Biostatistics* **8** 86–100
[43] Combrisson E and Jerbi K 2015 Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy *J. Neurosci. Methods* **250** 126–36
[44] Schudlo L C and Chau T 2018 Development of a ternary near-infrared spectroscopy brain–computer interface: online classification of verbal fluency task, stroop task and rest *Int. J. Neural Syst.* **28** 1750052
[45] Herff C, Heger D, Fortmann O, Hennrich J, Putze F and Schultz T 2014 Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS *Front. Hum. Neurosci.* **7** 935
[46] Weyand S and Chau T 2015 Correlates of near-infrared spectroscopy brain–computer interface accuracy in a multi-class personalization framework *Front. Hum. Neurosci.* **9** 536
[47] Nijholt A, Tan D, Pfurtscheller G, Brunner C, Millón J D R, Allison B, Graimann B, Popescu F, Blankertz B and Müller K R 2008 Brain-computer interfacing for intelligent systems *IEEE Intell. Syst.* **23** 72–9
[48] Koessler L, Maillard L, Benhadid A, Vignal J P, Felblinger J, Vespignani H and Braun M 2009 Automated cortical projection of EEG sensors: anatomical correlation via the international 10–10 system *NeuroImage* **46** 64–72
[49] Demonet J F, Chollet F, Ramsay S, Cardebat D, Nespoulous J L, Wise R, Rascol A, Frackowialk R and Frackowiak R 1992 The anatomy of phonological and semantic processing in normal subjects *Brain* **115** 1753–68 (PMID:1486459)
[50] Nyberg L, Eriksson J, Larsson A and Marklund P 2006 Learning by doing versus learning by thinking: an fMRI study of motor and mental training *Neuropsychologia* **44** 711–17
[51] Gentili R, Han C E, Schweighofer N and Papaxanthis C 2010 Motor learning without doing: trial-by-trial improvement in motor performance during mental training *J. Neurophysiol.* **104** 774–83
[52] Myrden A and Chau T 2015 Effects of user mental state on EEG-BCI performance *Frontiers Hum. Neurosci.* **9** 1–11
[53] Merzagora A C, Foffani G, Panyavin I, Mordillo-Mateos L, Aguilar J, Onaral B and Oliviero A 2010 Prefrontal hemodynamic changes produced by anodal direct current stimulation *NeuroImage* **49** 2304–10
[54] Lotte F, Congedo M, Lécuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1–13