A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech

Ciaran Cooney, Raffaella Folli, and Damien Coyle

Abstract*—*Objective: **Brain-computer interfaces (BCI) studies are increasingly leveraging different attributes of multiple signal modalities simultaneously. Bimodal data acquisition protocols combining the temporal resolution of electroencephalography (EEG) with the spatial resolution of functional near-infrared spectroscopy (fNIRS) require novel approaches to decoding.** ***Methods:*** **We present an EEG-fNIRS Hybrid BCI that employs a new bimodal deep neural network architecture consisting of two convolutional sub-networks (subnets) to decode overt and imagined speech. Features from each subnet are *fused* before further feature extraction and classification. Nineteen participants performed overt and imagined speech in a novel cue-based paradigm enabling investigation of stimulus and linguistic effects on decoding.** ***Results:*** **Using the hybrid approach, classification accuracies (46.31% and 34.29% for overt and imagined speech, respectively (chance: 25%)) indicated a significant improvement on EEG used independently for imagined speech (p=0.020) while tending towards significance for overt speech (p=0.098). In comparison with fNIRS, significant improvements for both speech-types were achieved with bimodal decoding (p<0.001). There was a mean difference of ~12.02% between overt and imagined speech with accuracies as high as 87.18% and 53%. Deeper subnets enhanced performance while stimulus effected overt and imagined speech in significantly different ways.** ***Conclusion:*** **The bimodal approach was a significant improvement on unimodal results for several tasks. Results indicate the potential of multi-modal deep learning for enhancing neural signal decoding.** ***Significance:*** **This novel architecture can be used to enhance speech decoding from bimodal neural signals.**

***Index Terms*— electroencephalography, EEG, functional near-infrared spectroscopy, fNIRS, brain-computer interfaces, imagined speech, deep learning, bimodal deep learning**

I. Introduction

COMBINING electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) acquisition protocols has become a popular approach in brain-computer interface (BCI) research [1]–[3]. This is due to the potential offered by merging the temporal resolution of the brain's electrical signals (EEG) with the spatial resolution of the hemodynamic response acquired from fNIRS [3]. Integration of modalities for concurrent data acquisition can mitigate the weaknesses of unimodal protocols [4], and the complimentary characteristics of EEG and fNIRS, as well as their shared portability and low cost, have made them a strong candidate for the development of multimodal BCIs [5], [6].

Research into methods for decoding EEG-fNIRS has advanced as neural signal acquisition protocols improve [7]. Most studies have used standard features such as band power for EEG and oxy-hemoglobin (HbO) for fNIRS [1], [8], with common machine learning methods such as linear discriminant analysis (LDA) [1], [2] and support vector machines (SVM) [9], [10]. As in other fields, deep learning (DL) offers an important avenue for decoding neural signals [11]–[15]. However, few studies have investigated multimodal DL with EEG-fNIRS data [3], [16], [17]. Difficulties associated with combining multiple modalities, for example asymmetric predictive capacity [1], [18] and varying noise topology (muscular and eyeblink artefacts in EEG [19], heartbeat and Mayer Waves in fNIRS [20]), partially account for the sparsity of published research. In addition, the temporal alignment of EEG and fNIRS presents challenges which must be addressed [7], [9], [21]. With respect to EEG-fNIRS DL methods, the most important studies have used artificial neural networks (ANN) [3], recurrent neural networks (RNN) [16] and a combined recurrent-convolutional neural network (RCNN) [17]. Multimodal convolutional neural networks (CNN) have been applied to EEG, electro-oculogram and electromyogram for sleep stage classification [22] but despite being used for mental workload classification [23] CNNs have not been widely employed for EEG-fNIRS. Here, to the best of our knowledge, we present the first study using EEG-fNIRS with a bimodal CNN method for speech decoding.

Research into BCI systems for decoding speech-related processes from neural activity have gained prominence recently [11], [12], [24]–[26]. Implanted electrodes are often used in speech decoding studies which evaluate overt speech [11], [25], [26] or response to auditory speech stimuli as the mode of communication [12], [24]. Imagined speech decoding poses a number of additional challenges [6], [15], [27], and results are typically lower than overt speech, yet there is limited consensus in the literature on the relationship between the two speech modalities with respect to BCI development [28], [29]. Additionally, paradigms vary widely, with studies predominantly using audio [30], [31] or text-based [6], [27] stimuli. However, spontaneous speech [32] and question-and-answer [6], [11] paradigms have also been researched. A related issue is the different units of language participants are asked to speak, ranging from phonemes [33] and syllables [30] to words [15] and sentences [25]. Few studies have examined the impact of linguistic properties such as semantics or syntax on decoding words or sentences [34]. The difficulty of decoding speech from

This work is supported by a Northern Ireland Department for the Economy studentship.

C. Cooney and Damien Coyle are with the Intelligent Systems Research Centre, Faculty of Computing, Engineering and the Built Environment, Ulster University, Londonderry, Northern Ireland, United Kingdom.

R. Folli is with Institute for Research in Social Sciences, Ulster University, Jordanstown, Northern Ireland, United Kingdom

non-invasive recordings has been demonstrated by studies reporting no better than chance accuracy with a binary classifier [35] and only 9 of 12 participants exceeding chance in a 3-class classification task [6]. However, others have indicated the potential of non-invasive speech decoding with one study reporting 38.5% accuracy when decoding three imagined speech envelopes [36] and another reporting 64.1% accuracy for 3-class classification of 15s repetitions of *yes vs no vs rest* [37]. Our recent research achieved 24.90% and 30.25% for decoding 6 words and 5 vowels from a 4s task period [15].

In a previous study, we recorded EEG and fNIRS as participants undertook trials designed to examine the relative decoding potential of overt and imagined speech from EEG [38]. Here, we present a new deep neural network architecture for decoding bimodal neural signals (EEG-fNIRS) in a single training procedure. This network consists of two sub-networks (subnets) which act as data-specific feature extractors before *fusion* [39] is used to form a combined featureset for further processing and classification. The experiment was designed to examine the effects of three stimulus types, and two linguistic properties of speech on decoding accuracy (section II.A.) [38]. This facilitated six classification tasks, one for each stimulus/ word-type combination. The bimodal network was trained and tested on each task for both overt and imagined speech and compared with unimodal EEG and fNIRS approaches.

The bimodal network achieved higher decoding accuracies than both unimodal EEG and fNIRS methods for overt and imagined speech. These results were statistically significant for all but overt speech EEG (p=0.098). The impact of fNIRS due to the constrained duration of our task execution period was identified as a limiting factor in enhancing the performance of the bimodal approach. We also found that deeper subnets for feature extraction were conducive to enhanced decoding accuracy. Results confirmed previous findings that overt speech decoding consistently outperformed imagined speech while also indicating that stimulus significantly impacted decoding performance and that this effect differed between types of speech. The effect of linguistic properties was not significant. Finally, we discuss ways in which performance may be improved by tailoring the network to different data types and extending the time-period of fNIRS signals used.

II. Methods

A. Experimental Paradigm

To investigate effects of stimuli used to cue imagined speech on BCI decoding our experiments employed three modalities to cue participants: *text*, *image*, and *audio* (Fig. 1(a, b)). Motivation for selection of these modalities is discussed in detail in [38], and briefly here. With *text* stimuli, participants can read directly from prompts but risk bypassing initial stages of speech production i.e., conceptual preparation and lexical selection [40], [41], and there is thus an important difference between *text*-prompted speech and spontaneous speech. Through indirect presentation of words as *images* in a picture-naming task, participants are engaged in the earlier stages of speech production [42], [43] and it has been hypothesized that increased cognitive load with picture-naming in comparison with word repetition can improve signal-to-noise ratio in speech decoding tasks [44]. Auditory stimuli have potentially confounding effects as they present participants with the words they are expected to speak in another person's voice. Previous studies have demonstrated neural decoding of response to auditory stimuli [12], [26] but the challenge of fully disentangling speech listening from production of speech is extremely difficult. Use of all three modalities enabled comparison of different effects. In addition, two categories were used to select words for the study: *action words* and *combinations* (Fig. 1(c)). Word groups were selected to examine whether linguistic properties of semantics and syntax impact speech decoding. The first group was predicated on the theory of linguistic embodiment which posits that *action words* (e.g. *kick, lick, pick*) associated with different body parts elicit activity in cortical regions associated with muscle groups used to perform that action (e.g. *foot, tongue, hand*) [45]. Here, two concrete examples of embodiment were used to select *action words* (Fig. 1(c)). The words "*squeeze*" and "*jump*" correspond to actions associated with bodily limbs, whereas the words "*kiss*" and "*smile*" are associated with the face and, more specifically, the lips. The second word group was chosen to examine effects that presence or absence of syntactic modification has on decoding. These *combinations* were selected on the basis of an observation that lists of words lack the critical computation to combine them into a single concept [46]. Therefore, two phrases and two lists were chosen (Fig. 1(c)). They were "*red ball*" and "*green hat*" (phrases) and "*red green*" and "*ball hat*" (lists).

Several common methods were considered in designing the experimental procedure. One requires participants to begin speaking immediately in response to stimulus [27], [33], [47]. Another partitions the two component parts i.e., stimulus/cuing and task execution, with stimulus directly preceding execution [31], [48]. The final approach considered requires a defined interlude between stimulus and task production periods, with participants holding the target word or phrase in memory before task execution [49], [50]. Despite each method having associated pros and cons, we selected the first, a dual stimulus and task execution period (Fig. 1(a) - green) to limit cognitive load associated with working memory, decrease total time per trial and to avoid disruption of speech production processes described by common models of production [29].

At -500ms, each trial began with a fixation cross presented on screen. Following this, one of the three stimulus-types were presented at time 0s to prompt participants to produce a certain word(s). That is, for any given trial a word would be prompted by presenting the participant with either a *text*, *image* or *audio* representation of that word. *Text* and *image* stimuli were displayed on-screen for 1s, before being replaced by a blank grey background for a further 1s (Fig. 1(a)). *Images* selected to represent words are presented in supplementary Fig 1 and all images were resized to standard dimensions of 325 × 325 pixels, except for the "*ball hat*" *image* which was resized to 488 × 325 (to enable clear visual display of both objects). *Audio* began playing at 500ms with all *audio* clips played for less than 1s. During *audio* presentation the monitor displayed a recognisable symbol indicating that this was the stimulus presentation period (Fig. 1(b)). The 2s period represented by

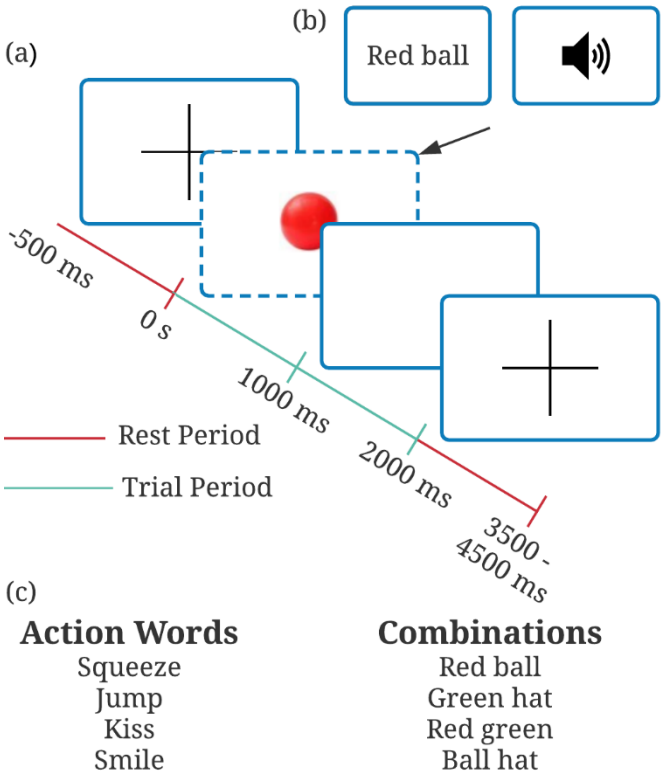

Fig. 1. (a) Trial periods began at time 0s, with a fixation cross presented for 500ms. Stimuli were then presented on-screen for 1s, followed by a blank screen for 1s. This 2s period (green) was considered the trial-period for experiments. (b) Three types of stimuli were used to present words: *text*, *image* and *audio*. (c) Words used for experiments, broadly categorised as action words and combinations.

green shading (1s stimulus + 1s blank; Fig. 1(a)) was the task execution period. This was considered the classification period for EEG with required adjustments made for fNIRS (see section II.D). A post-task production period, during which a fixation cross was displayed on-screen, was randomized between 1.5 and 2.5s. All participants were provided with identical written directions on how to produce imagined speech (supplementary material) and, given the integrated experimental protocol, instructed to begin producing speech immediately upon perceiving each word. Participants were explicitly instructed to say each word or pair of words only once during each trial. Each possible combination of stimulus and word were presented to participants 50 times each. Sessions were split into 6 blocks with 2 runs each per block and 100 trials per run, therefore, 1200 trials per session. Participants were permitted to take short breaks between runs. Trials were randomized across blocks and runs. In total, experiments lasted approximately 2 hours. For full details of the experimental protocol, see [38].

B. Participants

Nineteen participants undertook experiments (10 female; mean age 26.63 ± 2.13). Each participant was scheduled to complete 4 sessions: 2 overt speech and 2 imagined speech. However, due to Covid19 restrictions, not all sessions were completed. Eight of the 19 completed all 4 planned sessions, 5 completed 3 sessions, 2 completed 2 sessions and 4 completed 1 session. All participants completed at least one overt speech session and 15 completed at least one imagined speech session. All were native English speakers, had normal or corrected-to-normal vision and reported no history of neurological disorders. Participants provided written informed consent prior to experiments. Ethical approval was granted by Ulster University's research ethics committee. Participants were remunerated for involvement in the study.

C. Data Acquisition

EEG and fNIRS data were recorded concurrently using the g.Nautilus fNIRS-8 (g.tech medical engineering GmbH Austria), a fully integrated EEG and fNIRS recording device. The g.Nautilus fNIRS-8 facilitates wireless digital transmission of acquired signals at a distance of 10 meters. Synchronous signal recording is achieved using the MATLAB-Simulink platform with bespoke Simulink blocks for EEG and fNIRS.

A 64-channel EEG montage (Fig. 2) was configured using g.SCARABEO active wet electrodes. Electrodes were positioned according to the unified standard montage10-5 system to enable even distribution across scalp locations and to facilitate positioning of fNIRS optodes across bihemispheric motor regions. A sampling rate of 250 Hz was used for EEG recordings. A 0.1Hz high-pass filter was used to remove slow drifts during recordings and a 48-52Hz notch filter used to remove 50Hz line noise. fNIRS data were recorded at 10Hz and upsampled to 250Hz during acquisition. Data were acquired using 8 LED based transmitters, each of which emit light at wavelengths of 760 and 850 nm. Two receivers, each associated with 4 transmitter channels, produce 2×4 fNIRS channels. Each fNIRS channel recorded optical densities at both wavelengths, resulting in a total of 16 channels containing optical densities for each recording. Additionally, the g.Nautilus fNIRS-8 facilitates online conversion of optical densities into concentration changes of HbO and deoxy-hemoglobin (HbR), using the Modified Beer-Lambert law [51], [52]:

$$A(t;\lambda) = ln\frac{I_{in}(\lambda)}{I_{out}(t;\lambda)} = \alpha(\lambda) \times c(\lambda) \times d(\lambda) + \eta, \quad (1)$$

$$\begin{bmatrix}\Delta c_{HbO}(t)\\ \Delta c_{HbR}(t)\end{bmatrix} = \begin{bmatrix}\alpha_{HbO}(\lambda_1) & \alpha_{HbR}(\lambda_1)\\ \alpha_{HbO}(\lambda_2) & \alpha_{HbR}(\lambda_2)\end{bmatrix}^{-1} \times \begin{bmatrix}\Delta A(t;\lambda_1)\\ \Delta A(t;\lambda_2)\end{bmatrix}\frac{1}{l \times d(\lambda)} \quad (2)$$

where A is the optical density, t is time in seconds, λ_1 and λ_2 are the stated wavelengths, I_{in} is the incident intensity of light, I_{out} is the detected intensity of light, α is the extinction coefficient in $\mu M^{-1}cm^{-1}$, c is the absorber concentration in micromolars, l is the distance between source and detector optodes in centimeters, d is the differential path-length factor (6), and η is the loss of light due to scattering (here it is cancelled out on the assumption that it is neglible due to attenuation in continuous-wave fNIRS [53]). The incident intensity of light is the initial intensity of light emitted from the g.Nautilus Fnirs-8 and is a property of the device.

Receiver optodes were positioned at C3 and C4, with each transmitter positioned at 30 mm from the receivers. Transmitter optodes were placed at the same scalp locations and connected to the same channels for each session.

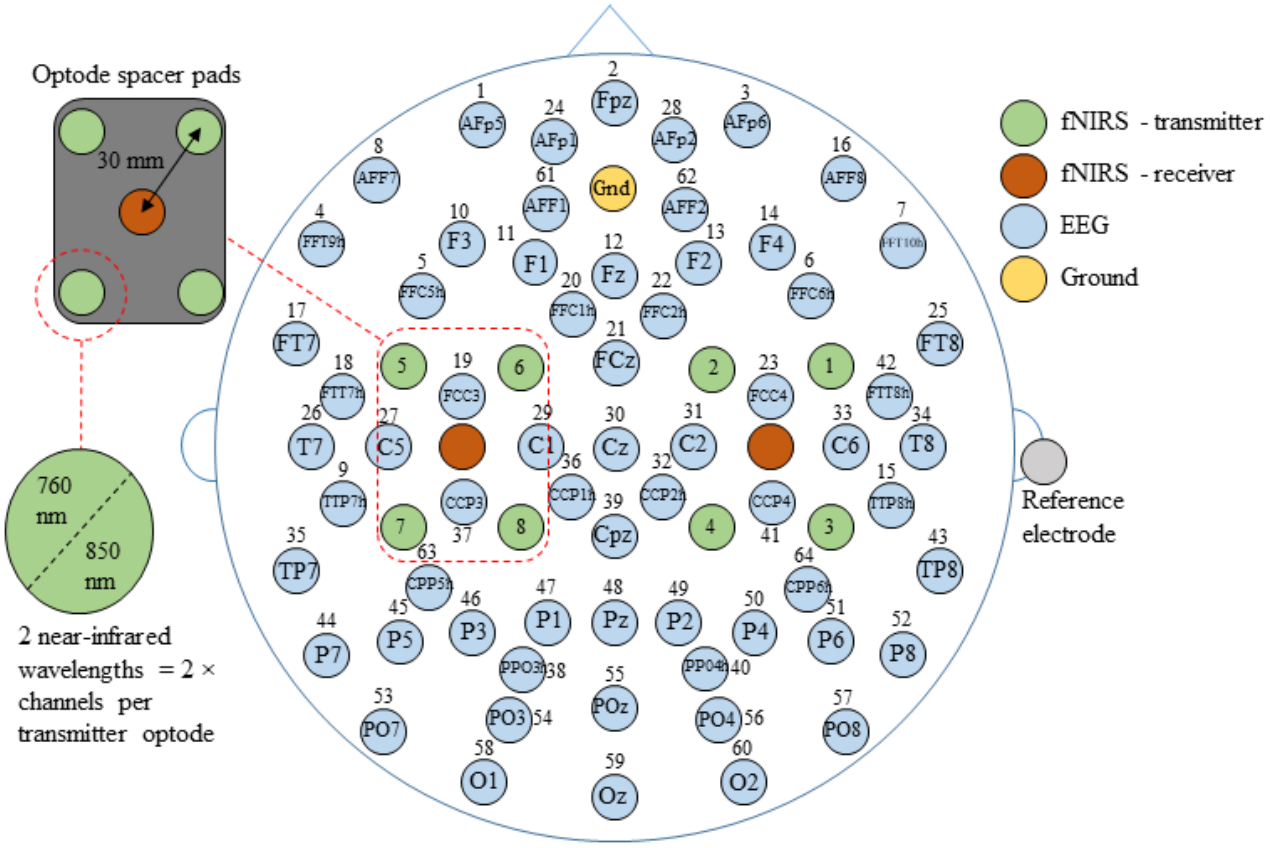

Fig. 2. The 64-channel EEG montage was configured using the international 10-5 system and designed to provide coverage across all scalp regions while also allowing placement of fNIRS optodes. A ground electrode was positioned at AFz and a reference electrode attached to the right earlobe. fNIRS optodes were positioned bihemispherically over central-motor regions. Receiver optodes (orange) were positioned at C3 and C4 respectively, with each centrally located among four associated transmitter optodes (green). Transmitter optodes were precisely positioned at 30 mm from the receivers. Each transmitter optode consists of two channels which transmit light at wavelengths of 760 nm and 850 nm.

D. Signal Processing

EEG data were processed using EEGLAB [54] in MATLAB 2017a (Mathworks, Natick, MA, USA). Channel rejection due to excessive noise (±500μV max.) or signal loss was applied following visual inspection of the raw EEG. A Hamming windowed finite impulse response (FIR) filter, with EEGLAB's built-in heuristic automatically determining filter length, was implemented to bandpass raw continuous EEG between 0.5-40Hz, with all signals rereferenced using common average referencing [54]. Baseline removal was applied by computing the mean for each trial in the time period -500ms – 0s (Fig. 1(b)) and subtracting it from the task period. Trials containing muscular artefacts were rejected by visual inspection. Finally, independent components analysis (ICA) was performed on remaining preprocessed channels using the infomax algorithm to remove artefacts [55]. ICA components were visually assessed and those with clear frontal distribution of weights indicating ocular artefacts were removed. Between one and three components were removed per session data. EEG data were transformed back into channel space for further analysis.

fNIRS data were processed in Fieldtrip [56]. Due to poor fNIRS signal quality during setup (S5-Session 1, S6-Session 1 (Overt); S13-Session 1 (Imagined)) or signal dropout during experiments (S2-Session 2, S3-Session 1 (Overt); S2-Session 1 (Imagined)), several sessions reported in the original EEG study [38] were not used here. Channels with poor signal quality due to inadequate contact were eliminated from further analysis following visual inspection. Signals were bandpass filtered from 0.1-0.8Hz to reduce artefacts from physiological signals such as cardiac interference (0.8Hz). Data were epoched into periods of –500ms-3.5s (longer than EEG to account for slower fNIRS time courses) and baseline corrected. Trial rejection due to movement artefacts was applied through visual inspection.

The 2s task execution period (Fig. 1(a)) was used for classification. Due to differential time courses of EEG and fNIRS, a temporal offset was applied to fNIRS for all classification tasks. Hybrid EEG-fNIRS studies have used a variety of windows for extracting features from fNIRS, including one 4s post-cue onset for a 10s trial [9] and a 2-7s post-cue window for a 10s task [7]. A recent study reported peak correlation between EEG and fNIRS signals occurred when the fNIRS lagged the EEG signal by approximately 1.7s during a 3.5s trial period [21]. Due to the relatively short task execution period (2s), we applied a 1.5s offset to fNIRS data i.e., a 0-800ms classification window corresponds to fNIRS data recorded 1.5-2.3s post cue onset.

Training a bimodal classifier requires data samples from the different modalities to be perfectly class-aligned. As we applied trial-rejection to EEG and fNIRS independently, we ensured that trials for bimodal classification were aligned by rejecting all independently rejected trials from both data types prior to training. Finally, data were split into the six different 4-class decoding tasks facilitated by the experimental design. These

were: *action-text* (*AT*), *action-image* (*AI*), *action-audio* (*AA*), *combinations-text* (*CT*), *combinations-image* (*CI*) and *combinations-audio* (*CA*).

E. Bimodal DL Architecture

The bimodal architecture (Fig. 3) consists of two subnets, each associated with a specific data type, and a wider network architecture in which they are contained. The two subnets consist of an initial convolutional block combining temporal and spatial convolutions [13]. Filters in the first layer (number of filters = 40; filter size = 1×5) are convolved with the input data along the time dimension. The resulting weights are then spatially filtered (number of filters = 40; filter size = N channels × 1) with weights for all possible pairs of electrodes. Batch normalization [57] adds regularization and an activation function adds non-linearity (section II.F). This is followed by dropout (p=0.1). To avoid diminishing spatial information in the data, no pooling operations were used. During hyperparameter (HP) optimization, an extension of this design, with convolution, batch normalization, activation function and dropout layers (Fig. 3(b)), was evaluated. The output of each subnet is a FC layer with 500 hidden units. Outputs of the subnets feed into the remaining layers of the network where they are combined in a process described as *late fusion*, where features are extracted separately and merged at later layers [39]. Here, outputs of the subnets are concatenated and passed to a GRU layer [58] (250 hidden layer units). This is followed by an activation function, a dropout layer (p=0.2), a FC layer and a final activation function. The output layer of the bimodal network is a log softmax classifier (section II.E).

Dimensions of data as it progressed through the basic and extended versions of the network are reported in supplementary TABLE I, with additional text indicating how the dimensionality of the feature maps differ as the number of layers vary. Due to windowing, input tensor dimensions were *32×64×200* for EEG and *32×16×200* for fNIRS (*batch×channels×samples*). The output of the GRU was a *32×250* tensor which fed into the next FC layer with an output shape of *32×4* to be applied to the log softmax classifier. The network was built using PyTorch [59] with the braindecode [13] software package (https://github.com/braindecode/braindecode). The bimodal network is available at: https://github.com/cfcooney/BiModNeuroCNN.

F. Unimodal DL Architecture

For comparison with EEG- and fNIRS-only decoding, we used a unimodal DL network similar to the subnets. The layout consists of an initial convolutional block combining temporal and spatial convolutions [13], with identical filter dimensions to the subnets. Batch normalization [57] added regularization and an activation function added non-linearity. These layers were followed by dropout (p=0.1) and a log softmax classifier.

G. Network Training

Training procedures were identical for the bimodal, unimodal EEG and unimodal fNIRS networks, respectively. Data which previously had trials removed or were imbalanced due to incomplete recordings, were oversampled using SMOTE [60] to ensure that all minority classes were balanced with the majority class. This step was applied to training data only. *Xavier* uniform initialization [61] was used to initialize weights in the temporo-spatial convolution layers. Later layers used *He* uniform initialization due to its utility for rectified linear units (ReLU) based activation functions [62]. Training was optimized using Adam [63], a popular approach to gradient-based optimization of stochastic objective functions. The method updates exponential moving averages of the gradient and the squared gradient with HPs $\beta_1, \beta_2 \in [0, 1]$ (here 0.9 and 0.999) controlling exponential rates of decay. Moving averages are estimates of the mean and uncentred variance of the gradient. The maximum number of training epochs was 50, and an early-stopping strategy was applied to all training instances. Training was stopped when validation accuracy stopped improving over a predefined number of epochs (patience=20). Training resumed with parameter values re-initialized to those that resulted in the best validation accuracies thus far. Training was terminated when validation loss dropped to the same value as the training loss achieved at the end of the first training phase [13]. A batch size of 32 was used for all experiments. The initial learning-rate was 0.001 and learning-rate decay was applied using multi-step scheduling. This method decays the learning-rate by a fixed value, *gamma*, at specified intervals during training. Here, *gamma* was set to 0.1 and learning-rate decay applied at epochs 20 and 25 due to the small number of training epochs. Categorical probability distributions were obtained by transforming the output from the final convolution layer using a log softmax function (see supplementary material). A negative log-likelihood loss was used to minimize the error between the ground truth and predictions obtained by the log softmax function. Loss was minimised, and network parameters updated, using the Adam optimizer with backpropagation.

H. Hyperparameter Optimization

Rather than manually selecting all HP values, we optimized a subset using the nested cross-validation strategy (nCV) described in [15]. This method consists of outer- and inner- fold protocols with data split into k=5 folds for both, resulting in a train/validation/test split of 128/32/40 for a 200 sample classification task. The inner-fold selects optimal HP values, with multiple inner validations used to train and validate a model for all possible HP combinations. Maximal mean inner-fold validation accuracy was used to select optimal HP values. The outer-fold procedure evaluated model performance given tuned HPs. As with the inner-fold, validation accuracy was the metric used to evaluate the model during training, with the final model evaluated on test data. Optimized HPs were categorised as feature-extraction and network parameters. Feature-extraction parameters relate directly to the data (frequency band, classification window) and network parameters are used to instantiate and train the network (number of layers, activation function). We evaluated different frequency bands for both EEG and fNIRS. For EEG, five bands were considered: delta (0.5-4Hz), theta (4-8Hz), alpha (8-12Hz), beta (12-28 Hz) and gamma (28-40Hz). Bands were iteratively filtered from EEG during nCV using a 5th order Butterworth filter. For fNIRS, four low-frequency bands were evaluated: 0.1-0.5Hz, 0.2-0.6Hz,

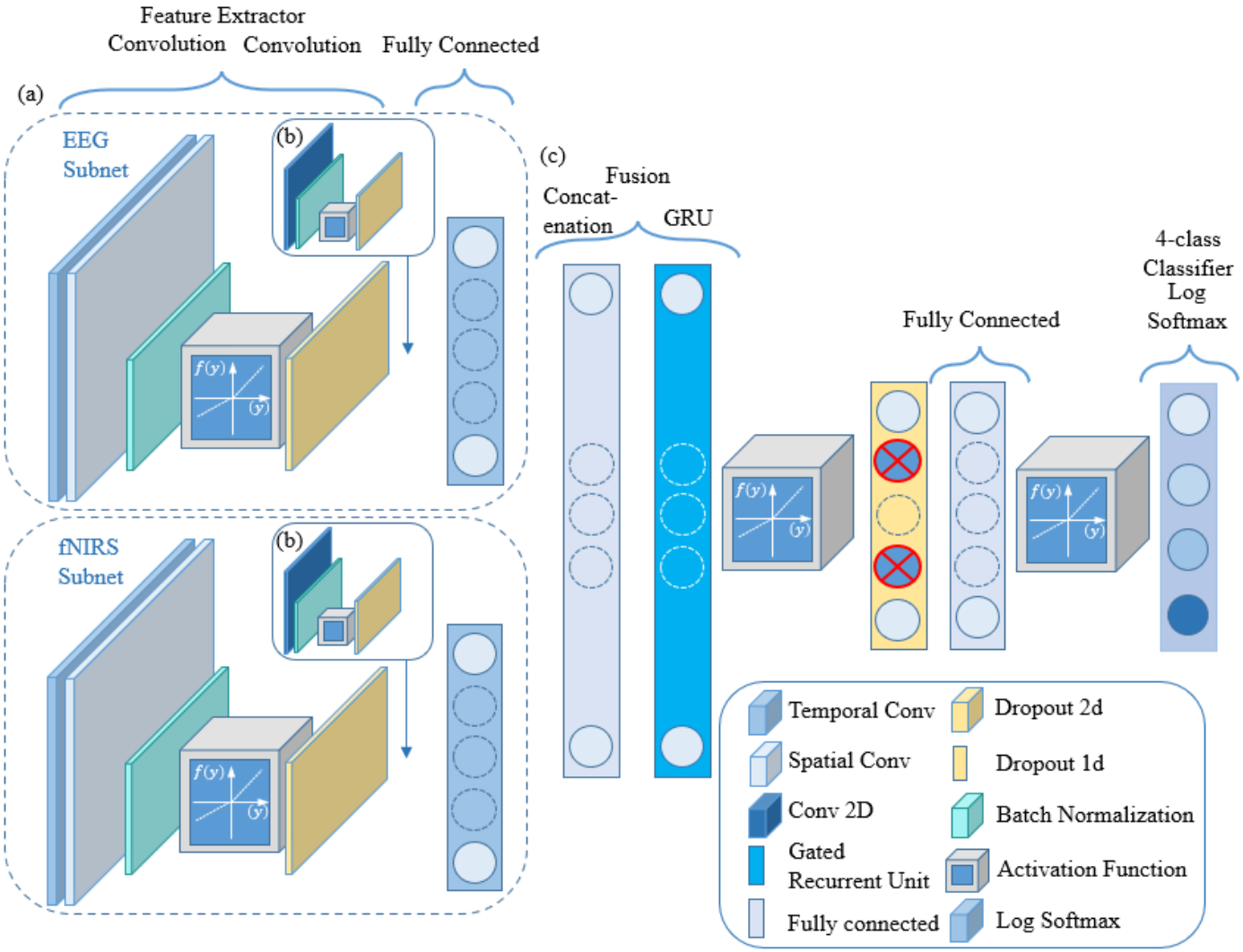

Fig. 3. Bimodal network for training with EEG and fNIRS. (a) Two identical CNNs (EEG and fNIRS subnets) form a dual feature extractor. The CNNs' initial layers consist of combined temporal and spatial convolution. Batch normalization is then applied, followed by one of two possible activation functions (ELU or Leaky ReLU). This is followed by dropout (p=0.1). The final layer of the subnets is a FC layer with 500 hidden units. (b) Parameters used to extend the depth of the CNN during HP optimization. This consists of 2d convolution, batch normalization, activation function and dropout. (c) *Fusion* and classification layers. Subnet outputs are concatenated in *late fusion* and fed to a GRU. This is followed by activation function, dropout, FC layer and another activation function. The final layer is a log softmax classifier used here for 4-class classification.

0.3-0.7Hz and 0.4-0.8Hz. Bands were filtered using a 2nd order Butterworth filter. With a task execution window of 2s, three overlapping 800ms classification windows were evaluated. That is, windows of 0-800ms, 600-1400ms and 1200-2000ms post cue-onset (+ 1.5s for fNIRS; see Signal Processing) were evaluated to determine optimal classification time-periods.

Optimized network parameters were the activation function and depth of subnet. Two non-linear activation functions were evaluated. The first of these was exponential linear units (ELU) [64], defined as $f(x) = x$ for $x \geq 0$ and $f(x) = e^x - 1$ for $x < 0$. The second was Leaky ReLU, defined as $f(x) = x$ for $x \geq 0$ and $f(x) = \alpha x$ for $x < 0$, where α defines the extent to which the function "leaks" i.e., the slope of the function for $x < 0$. Structural HPs were optimized by extending the depth of the initial subnet (Fig. 3(a)) with additional layers (Fig. 3(b)). This consisted of an additional convolution layer, batch normalization, activation function and dropout. Optimal network depth indicated by the validation set was then used to obtain results during testing stages.

HPs evaluated with the nCV scheme were coupled across the subnets. That is, for each HP value the entire network was instantiated with that value and thus both subnets were always paired in this manner i.e., at no point was one subnet using ELU as its activation function when the other was using Leaky ReLU. EEG and fNIRS frequency bands were not coupled in this way as each data type was associated with a single subnet.

I. Evaluation metrics and statistics

Classification accuracy was used for evaluating the performance of the trained models. Accuracies were obtained for each test fold of the outer nCV procedure, and a mean and variance calculation used for reporting results.

Here, we considered p<0.05 to indicate statistical significance. Statistical analyses were performed using all the data collected as per section II.B. We used Analysis of Variance (ANOVA) tests based on the assumptions that the sampling distribution of the mean of the population is normally

distributed and that all samples are drawn independently. For all ANOVAs, when statistical significance was indicated a *post hoc* analysis was performed using a Tukey Honest Significant Difference (HSD) multiple comparisons test [65] to evaluate pairwise differences between results. Further details on statistical analysis are available in the supplementary material.

III. Results

A. Bimodal network improves on unimodal performance

The bimodal method achieved higher overall decoding accuracies than both unimodal approaches with statistically significant improvements for all but overt speech EEG (TABLES I & II). Task specific scores for overt speech classified with the bimodal approach were *AT*=49.61%, *AI*=48.72%, *AA*=45.02%, *CT*=49.20%, *CI*=46.76% and *CA*=38.52% (TABLE I; Fig. 4). With mean EEG decoding accuracies of *AT*: 46.04%, *AI*: 46.66%, *AA*: 41.55%, *CT*: 46.90%, *CI*: 45.08%, *CA*: 36.72% (TABLE I), bimodal decoding outperformed unimodal EEG in all overt speech tasks. For imagined speech, the bimodal network also outperformed unimodal EEG in all tasks with mean decoding accuracies of *AT*=31.78%, *AI*=38.37%, *AA*=33.89%, *CT*=32.21%, *CI*=36.67% and *CA*=32.80% (bimodal; TABLE *CI*: 34.56%, *CA*: 30.60% (EEG; TABLE II; Fig. 4), respectively.

On average, bimodal decoding improved on unimodal EEG by 2.48% (overt) and 1.59% (imagined). This result hints at potential performance improvements from combining EEG and fNIRS. A 2-way ANOVA *network* × *classification task* indicated differences between the two methods were significant for imagined speech ($F(1, 5)=5.45$, $p=0.0203$) while tending towards significance for overt speech ($F(1, 5)=2.75$, $p=0.098$). Significance corresponding to enhanced imagined speech decoding results from the bimodal classifier being an improvement in 16 of the 21 sessions. Despite a p-value >0.05, 21 of the 28 overt speech sessions were improved upon with bimodal decoding. Further analysis of results indicated that the bimodal approach suffered from a degree of negative transfer associated with several subjects' fNIRS data. For example, the overt speech scores for Subject 14 - Session 2 achieved mean accuracy of 41.79% with EEG but dropped to 32.77% with hybrid decoding (supplementary TABLES II & VI). In addition, supplementary TABLES IV & V present instances of fNIRS data being classified at or below chance level (25%), indicating that a small portion of the fNIRS data was not likely to benefit bimodal decoding. Reasons for this negative transfer are suggested by comparison with fNIRS decoding, below.

The bimodal network was significantly better than unimodal fNIRS for each of the six classification tasks for both overt and imagined speech (overt: $F(1, 5)=131.13$, $p<0.001$; imagined: $F(1, 5)=69.11$, $p<0.001$). Mean fNIRS decoding accuracies for overt speech were *AT*: 32.46%, *AI*: 32.46%, *AA*: 33.66%, *CT*: 31.73%, *CI*: 31.91%, *CA*: 33.49% (TABLE I). fNIRS results for imagined speech were *AT*: 30.62%, *AI*: 28.61%, *AA*: 29.72%, *CT*: 31.32%, *CI*: 28.95%, *CA*: 28.64% (TABLE II). A possible cause for the fNIRS having relatively poor decoding performance and limited impact on bimodal decoding can be observed in Fig. 5 where the fNIRS signal does not exhibit the typical time course associated with longer trial-periods. Instead, the HbO signal only begins its expected rise associated with task production approximately 2 – 2.5s post cue. As stated in section II.D., task-related fNIRS is usually expressed over longer periods. However, the constraints of our experiment, i.e., the relatively short task period required to investigate the different stimuli and word groups, means that there are potential performance gains to be made from a longer fNIRS period.

TABLE I
BIMODAL COMPARISON WITH UNIMODAL FOR OVERT SPEECH

Word-type	Action Words			Combinations		
	Text	Image	Audio	Text	Image	Audio
Bimodal	49.61	48.72	45.02	49.20	46.76	38.52
EEG	46.04	46.66	41.55	46.90	45.08	36.72
fNIRS	32.46	32.46	33.66	31.73	31.91	33.49

TABLE II
BIMODAL COMPARISON WITH UNIMODAL FOR IMAGINED SPEECH

Word-type	Action Words			Combinations		
	Text	Image	Audio	Text	Image	Audio
Bimodal	31.78	38.37	33.89	32.21	36.67	32.80
EEG	30.25	36.21	32.11	31.55	34.56	31.60
fNIRS	30.62	28.61	29.72	31.32	28.95	28.64

B. Decoding performance of the bimodal network

Fig. 4(a, b) are scattered boxplots visualizing accuracies obtained using the bimodal network for overt and imagined speech. The boxplots highlight two results: 1) the bimodal network classifies overt and imagined speech with accuracy substantially greater than chance level while exhibiting significant variance between classification tasks. 2) there is a clear performance gap between the two speech types, with overt speech resulting in significantly better decoding accuracy ($F(1, 5)=3.06$, $p<0.05$; 2-way ANOVA). Mean decoding accuracy across all tasks was 46.31% for overt speech and 34.29% for imagined speech, resulting in a 12.02% difference. Maximum decoding performance also illustrated differences with overt speech achieving 87.18% for *AT* and imagined speech achieving a best score of 53% for *AI* (supplementary TABLE II & III). Statistical analysis of differences between overt and imagined speech was undertaken with different sample sizes for the two conditions, with statistical power consequently limited by the smaller set (imagined speech).

C. Effect of stimuli and word-type on decoding

Results indicated some variation in decoding performance dependent on the type of stimuli used to prompt tasks. In addition, trends across the different classification tasks were not common across speech types. A 2-way ANOVA *stimulus* × *word-type* indicated that the main effects of different stimuli were significant ($F(2,162)=4.59$, $p<0.05$) but that the effects of different word types were not ($F(1,162)=1.87$, $p=0.174$). *Post hoc* tests attributed significance to the inferior scores obtained from audio trials (AA, CA ($p<0.05$), with differences between text and images negligible ($p=0.80$).

A 2-way ANOVA *stimulus* × *word-type* indicated that the main effects of stimuli were highly significant for imagined speech ($F(2,120)=12.27$, $p=1.42\times10^{-5}$), although the main effect of words was not ($F(1,120)=1.22$, $p=0.272$). *Post hoc*

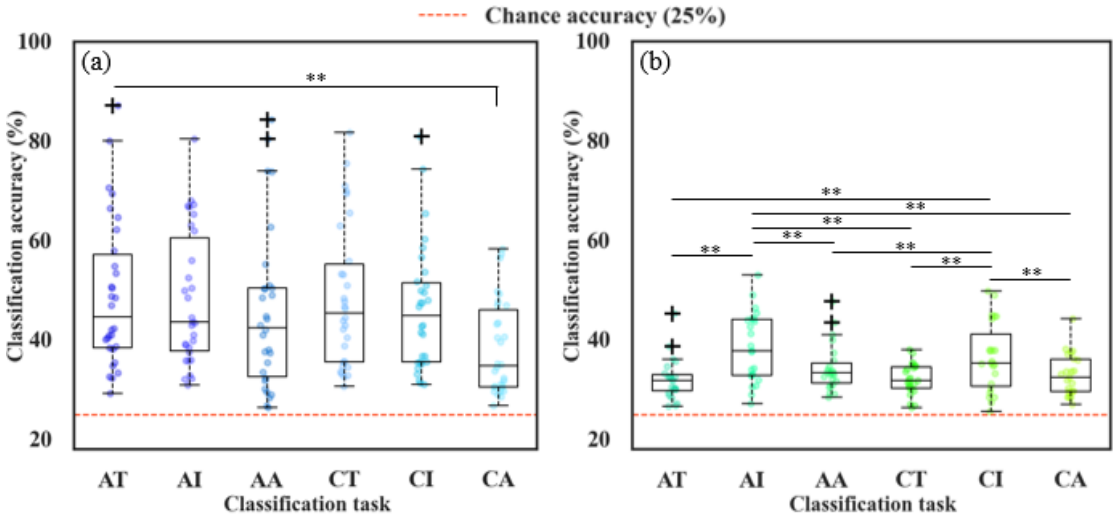

Fig. 4. Classification results across all classification tasks for both overt and imagined speech. Each data point corresponds to classification accuracy for one of six conditions, and for a single session (participants engaged in one or two sessions each). Boxplots visualize the distribution of results, indicating the median value (the point at which 50% of results are above and below), the interquartile range (box heights) and 1.5 times the interquartile range (whiskers extending beyond box edges). (a) Variability in performance across subjects and sessions for each classification task for overt speech. (b) Variability in performance across subjects and sessions for each classification task for imagined speech. **p<0.005.

tests revealed that superior accuracies obtained from trials using image stimuli were significant with respect to both *text* (p=1.44×10^{-5}) and *audio* (p<0.005) trials. Comparison of *text v audio* revealed no significance (p=0.312).

In section II.C, we reported that fNIRS optodes were placed above bihemispheric motor regions with the expectation that this may aid decoding of *action words*. However, the statistical analyses clearly indicate that there was no increase in performance for *action words* in relation to *combinations*.

D. Hyperparameter optimization of bimodal network

Results from HP optimization indicated the importance of EEG frequency band, classification window and depth of CNN subnet (Fig. 6). A 3-way ANOVA *frequency band* × *stimulus* × *word type* revealed that the main effects of frequency bands was highly significant for both overt (F(4,878)=17.273, p<1×10^{-6}) and imagined speech (F(4,668)=21.98, p<1×10^{-6}). For overt speech, *post-hoc* tests revealed that validation accuracies obtained with the delta band were significant with respect to all others (p<1×10^{-8}) (Fig. 6(a) - top). The gamma band was significantly poorer than all others (p<0.05). For imagined speech, delta was significantly greater than theta (p<0.005), alpha (p=1×10^{-6}) and gamma (p<1×10^{-5}), but not beta (Fig. 6(a) – bottom). In contrast with overt speech, beta band results were significantly greater with respect to theta (p<0.005), alpha (p<1×10^{-6}) and gamma (p<1×10^{-5}).

Main effects analysis showed that the impact of different windows was significant for both speech types (F(2,526)=90.8,p<1×10^{-6}; F(2,400)=7.25, p<0.001). For overt speech, *post hoc* tests revealed that the greater accuracies of both the second and third windows, compared to the first, were highly significant (p<1×10^{-6}) and that the difference between the second and third windows was significant (p<0.05). This translated to 67.8% selection of the third classification window and only 5.7% for the first (Fig. 6(b) –left). Similarly, for imagined speech significance resulted from the lower accuracies obtained from the first classification window in relation to the second (p<0.01) and third (p<0.005) windows. This resulted in 81.2% selection for windows two and three and only 18.8% for the first window (Fig. 6(b) – right). Greater inner-fold validation accuracies obtained with deeper subnets (overt: 43.60% vs 37.99%; imagined: 41.20% vs 37.49%) were significant (F(1,350)=90.8, p<1×10^{-6}; F(1,266)=43.45, p<1×10^{-6}). Neither fNIRS frequency bands nor activation functions had a significant impact on decoding performance.

IV. Discussion

Simultaneous recording of EEG and fNIRS can increase the volume of data and may be particularly useful for certain BCI applications as it facilitates acquisition of electrical and hemodynamic brain activity corresponding to a single task. Just as methods for recording brain activity continue to evolve, techniques for decoding multiple data streams must also be advanced. Here, we presented a bimodal deep learning architecture consisting of subnets previously developed for neural decoding applications [13], a *fusion* layer for combining features extracted by subnets and later GRU and FC layers feeding a log softmax classifier. Although multimodal deep learning has been applied elsewhere to neurological data streams [3], [22]this is the first instance of a bimodal architecture with convolutional subnets being applied to decoding overt and imagined speech from EEG-fNIRS data.

The bimodal approach demonstrated performance improvement upon unimodal approaches, with statistically significant improvement for all but overt speech EEG. While these results are suggestive of future use of this bimodal network, limiting factors such as the duration of the task execution period must be addressed to fully harness its potential. The design of the bimodal network includes some important conceptual differences from related methods for hybrid EEG-fNIRS decoding. Whereas other approaches have used distinct feature extraction algorithms for the two data

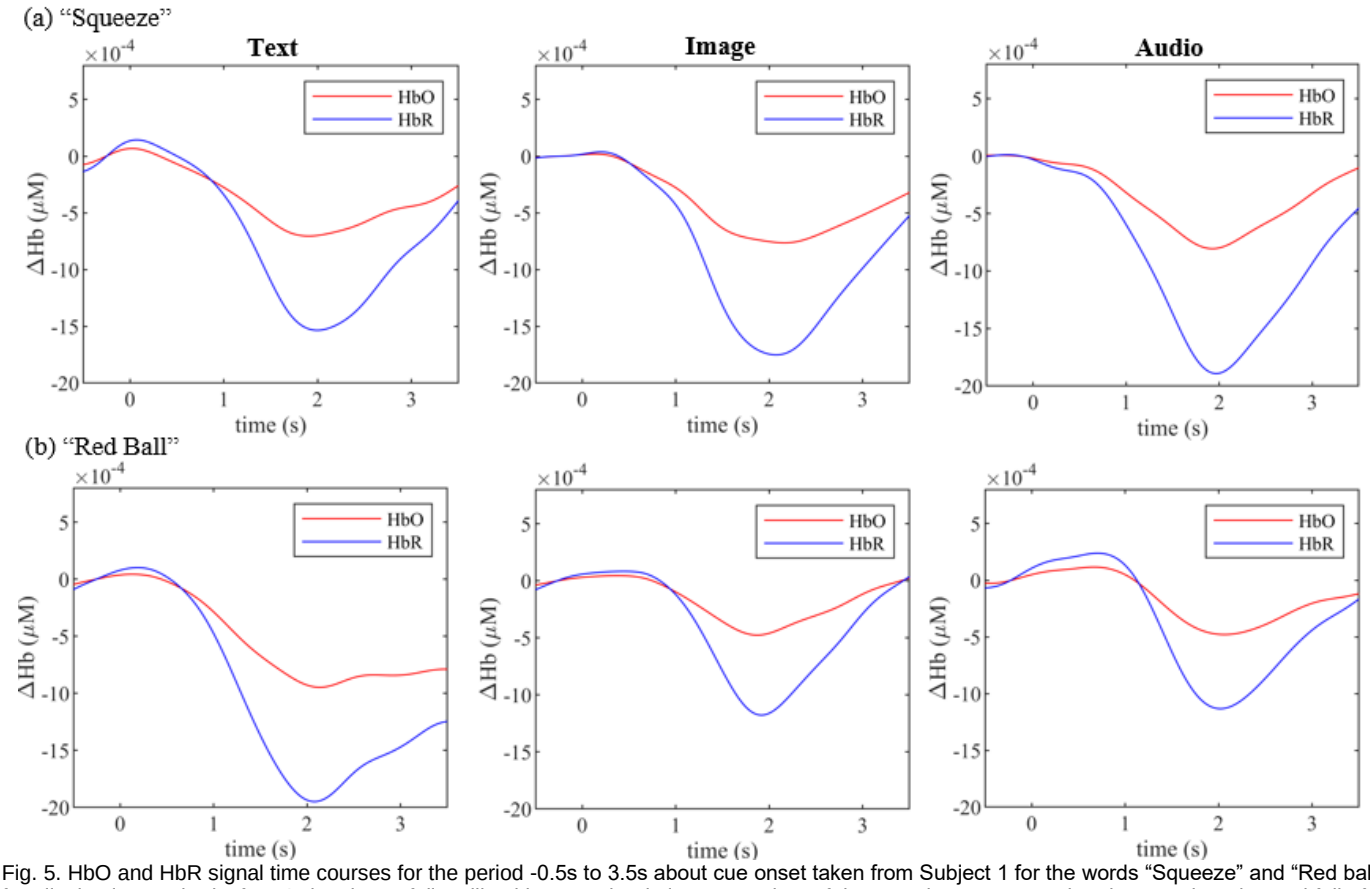

Fig. 5. HbO and HbR signal time courses for the period -0.5s to 3.5s about cue onset taken from Subject 1 for the words "Squeeze" and "Red ball" for all stimulus methods. fNIRS data is not fully utilized here as the timing constraints of the experiments meant that the complete rise and fall of a typical HbO signal associated with task production was not possible.

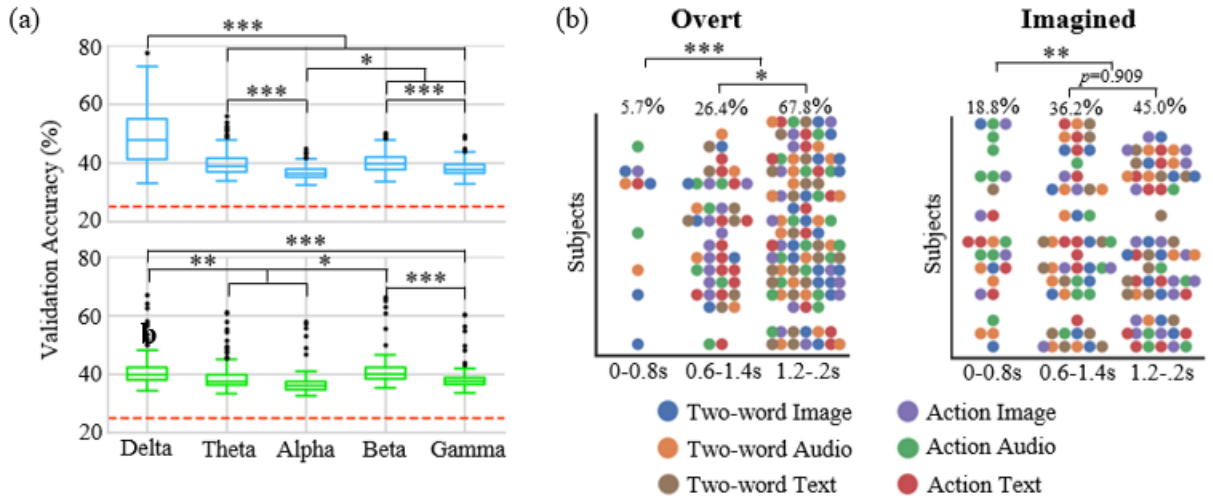

Fig. 6. Hyperparameter optimization for overt and imagined speech tasks. (a) Inner-fold validation accuracy for delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-28 Hz), gamma (28-40 Hz) (overt speech – top/blue; imagined speech bottom/green). (b) Optimal classification windows for overt (left) and imagined (right) speech. *p<0.05, **p<0.01, ***p<1×10^{-8}

types [2] or combined the data prior to its being fed into a neural network [3], our approach processed the data through individual subnets through which the network could co-adaptively learn features for both EEG and fNIRS. This approach allows the network to update parameters during training as it learns features from both data types simultaneously. Here, two identical subnets were used for extracting features from EEG and fNIRS. One of the reasons for this was to enable pairing of HPs across subnets, thus reducing the overall search space during optimization. However, it is possible that this is a sub-optimal solution and further research is required to ascertain whether bespoke subnets for each data type would yield significant performance improvement. A feature extraction approach specifically tailored to the characteristics of fNIRS is a potential improvement that should be investigated.

The use of subnets in the design also facilitated concatenation of the features they extracted in a *fusion* process. There are

several points at which parallel data streams can be fused in a bimodal classifier [39], [66]. They can be concatenated before being fed into a network or fused in a penultimate layer just prior to classification. Our network applied *fusion* immediately after the two convolutional subnets were used for feature extraction and then performed further feature extraction on this combined featureset with GRU and FC layers. The rationale for this is that there may be more information in one or other of the extracted featuresets which further deep learning could identify and exploit. Combining the use of subnets and *fusion* enabled optimization of feature extraction and classification in a single training procedure without the necessity of manual feature engineering processes for each data type.

As well as enhancing decoding in comparison with unimodal EEG and fNIRS, the bimodal network resulted in all classification tasks achieving above-chance decoding accuracy, with overt speech reaching as high as 87.18%. This is promising for the future potential of bimodal decoding of non-invasively acquired speech correlates. Peaking at 53%, imagined speech results also indicated potential, particularly when prompted by *images*. Results are significant despite not relying on word repetition to enhance performance as in other studies [6], [27], and using relatively few trials per class.

Although not the primary subject of the study, here we consider results obtained from a unique experimental procedure reported elsewhere [38]. Direct comparisons of results obtained from overt and imagined speech are sparse in non-invasive BCI literature [36], [50] as studies have focused on overt speech [11], [25], [26]. Here, we confirmed the results of our previous work [38], reporting a clear disparity between decoding potential for overt and imagined speech. This is to be expected, and results reported here even exhibit a narrowing gap of 12.02% in comparison with similar studies [36], [50]. Nevertheless, it is clear that imagined speech cannot currently be decoded with accuracy equivalent to that of overt speech.

The experimental procedure also enabled investigation of the effects of stimulus type and the semantics and syntax of different words on decoding performance. Replicating previous findings [38], statistical tests indicated that the effects of selecting words based on semantic and syntactic criteria did not significantly impact decoding performance. While the effects of using different categories of words was not significant, the impact of stimulus clearly was. Results from imagined speech trials prompted with *images* show a consistent and significant performance improvement over *text* and *audio*. *Image* presentation has some advantages over *text* and *audio* in that it does not directly present the word to be spoken and thus participants must engage in the word retrieval phase of speech production models. On the other hand, it has been shown that *images* evoke higher amplitude responses than *text* [67], and it is possible that this may impact decoding. This being the case, presentation modality must be carefully considered by researchers when designing experiments.

Limitations of this research accrued from constraints imposed by recording equipment, our investigation of both overt and imagined speech, and the effects of different stimulus and word groups. Despite our fNIRS montage consisting of a similar number of channels to others employing hybrid EEG-fNIRS for BCI applications [2], [3], [10], it is possible that higher-density fNIRS may have impacted this study. For example, [68] used fifty fNIRS channels to benefit from extensive scalp coverage when trialing a BCI for covert intention classification. Greater fNIRS coverage may have mitigated some of the imbalance between EEG and fNIRS results. However, the totality of difference is not likely due to coverage alone as studies have demonstrated the utility of few-channel fNIRS [69], [70]. Related to this is the likely impact of fNIRS optode placement at different functional regions across the cortex. We placed optodes over motor regions to coordinate with the selection of *action words* in our experimental procedure. However, studies have reported speech-related decoding from fNIRS with optodes over Broca's and Wernicke's areas [71] while others have demonstrated mental character writing [72] and visual stimuli [73] with optodes placed at the prefrontal cortex. The relatively short 2s trial period, and corresponding 800ms classification window, was a function of limiting session recording times to 2 hours. It is possible that the decoding performance of the bimodal approach would be improved with an extended trial period, particularly as the time courses presented in Fig. 5 indicate that the fNIRS data may not have been fully utilized. The signal did not exhibit the full characteristic curve demonstrated in studies with longer time-periods [2], [3], but nevertheless did show the process of increasing and decreasing HbO and HbR concentrations which may suggest the fNIRS in this time-period was representative of task execution. Further research is required to understand the extent to which, if at all, the time-period limited the benefit of using fNIRS. Clearly, there are trade-offs between the length of fNIRS time-period and the applicability of fNIRS to real-time speech decoding. Additionally, comparison of the effects of different wavelengths is a potential future research question.

There are downsides to extended trial periods, as a virtue of using EEG for communication is the high temporal resolution that facilitates real-time interaction. This would be lost in extending the trial period for fNIRS. Timing constraints also limited the number of trials per class to 50. Previous studies have recorded 100+ trials per class [27], [30], and it is highly probable that additional training data would improve the generalizability of the bimodal model and consequently overall performance. Finally, further research is required to validate the efficacy of this approach in online BCI experiments. The development of methods for speech decoding must be functional in real-time scenarios if they are to be a feasible mode of communication.

V. Conclusion

In this paper, we presented a bimodal deep neural network architecture for decoding neural signals from two data streams and showed it improved upon unimodal approaches. The design facilitates concurrent feature extraction by instantiating two convolutional subnets which are trained using a common loss function. Data-specific features are then combined in a *fusion* layer before further layers are used to extract features for classification. To test the network, we trained it on EEG and fNIRS data recorded while participants performed tasks using overt and imagined speech. These tasks also enabled an investigation into the effects of stimulus and linguistic properties on speech decoding.

Results demonstrated that the bimodal network significantly improved upon unimodal decoding for all imagined speech tasks. Most subjects' results improved with the bimodal network, despite subnets not being specifically tailored for different data types and the duration of fNIRS data not being optimal. These are areas in which future research and development is required. Overall accuracies hinted at the potential for decoding speech from non-invasive neural recordings despite a significant performance gap between overt and imagined speech. In addition, results indicated that deeper subnets improved performance. Our findings also support significant differences in the effect of stimulus on decoding performance, with *image* stimulus presentation resulting in the highest classification accuracies for imagined speech.

VI. Acknowledgment

We are grateful for access to the Tier 2 High Performance Computing resources provided by the Northern Ireland High Performance Computing (NI-HPC) facility, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/T022175/1. Damien Coyle is grateful for a UKRI Turing AI Fellowship 2021-2025, funded by the EPSRC, Grant No, EP/V025724/1. Ciaran Cooney is grateful for a PhD studentship funded by the Department for the Economy, Northern Ireland.

VII. References

[1] A. P. Buccino, H. O. Keles, and A. Omurtag, "Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks," *PLoS One*, vol. 11, no. 1, pp. 1–16, 2016.

[2] M. J. Khan and K. S. Hong, "Hybrid EEG-FNIRS-based eight-command decoding for BCI: Application to quadcopter control," *Front. Neurorobot.*, vol. 11, no. FEB, 2017.

[3] A. M. Chiarelli, P. Croce, A. Merla, and F. Zappasodi, "Deep learning for hybrid EEG-fNIRS brain-computer interface: Application to motor imagery classification," *J. Neural Eng.*, vol. 15, no. 3, p. 36028, 2018.

[4] S. Ahn and S. C. Jun, "Multi-modal integration of EEG-fNIRS for brain-computer interfaces – Current limitations and future directions," *Front. Hum. Neurosci.*, vol. 11, no. October, pp. 1–6, 2017.

[5] J. Kwon, J. Shin, and C. H. Im, "Toward a compact hybrid brain-computer interface (BCI): Performance evaluation of multi-class hybrid EEG-fNIRS BCIs with limited number of channels," *PLoS One*, vol. 15, no. 3, pp. 1–14, 2020.

[6] A. R. Sereshkeh *et al.*, "Development of a ternary hybrid fNIRS-EEG brain – computer interface based on imagined speech," *Brain-Computer Interfaces*, vol. 00, no. 00, pp. 1–13, 2019.

[7] K.-S. Hong, M. J. Khan, and M. J. Hong, "Feature Extraction and Classification Methods for Hybrid fNIRS-EEG Brain-Computer Interfaces," *Front. Hum. Neurosci.*, vol. 12, no. June, pp. 1–25, 2018.

[8] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, "Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and fNIRS data," *Front. Hum. Neurosci.*, vol. 10, no. MAY2016, pp. 1–14, 2016.

[9] S. Ge *et al.*, "A Brain-Computer Interface Based on a Few-Channel EEG-fNIRS Bimodal System," *IEEE Access*, vol. 5, pp. 208–218, 2017.

[10] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring mental workload with EEG+fNIRS," *Front. Hum. Neurosci.*, vol. 11, no. July, pp. 1–20, 2017.

[11] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nat. Commun.*, vol. 10, no. 1, p. 3096, 2019.

[12] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.

[13] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[14] C. Cooney, R. Folli, and D. Coyle, "Optimizing Layers Improves CNN Generalization and Transfer Learning for Imagined Speech Decoding from EEG," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 1311–1316.

[15] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, Aug. 2020.

[16] P. Sirpal, A. Kassab, P. Pouliot, and D. K. Nguyen, "fNIRS improves seizure detection in multimodal EEG-fNIRS recordings," *J. Biomed. Opt.*, vol. 24, no. 05, p. 1, 2019.

[17] H. Ghonchi, M. Fateh, V. Abolghasemi, S. Ferdowsi, and M. Rezvani, "Deep recurrent-convolutional neural network for classification of simultaneous EEG-fNIRS signals," *IET Signal Process.*, vol. 14, no. 3, pp. 142–153, 2020.

[18] Y. Liu, H. Ayaz, and P. A. Shewokis, "Multisubject 'learning' for mental workload classification using concurrent EEG, fNIRS, and physiological measures," *Front. Hum. Neurosci.*, vol. 11, no. July, 2017.

[19] M. Teplan and others, "Fundamentals of EEG measurement," *Meas. Sci. Rev.*, vol. 2, no. 2, pp. 1–11, 2002.

[20] H. D. Nguyen, S. H. Yoo, M. R. Bhutta, and K. S. Hong, "Adaptive filtering of physiological noises in fNIRS data," *Biomed. Eng. Online*, vol. 17, no. 1, pp. 4–9, 2018.

[21] S. Ge *et al.*, "Neural Activity and Decoding of Action Observation Using Combined EEG and fNIRS Measurement," *Front. Hum. Neurosci.*, vol. 13, no. October, pp. 1–15, 2019.

[22] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018.

[23] M. Saadati, J. Nelson, and H. Ayaz, "Convolutional neural network for hybrid fNIRS-EEG mental workload classification," in *International Conference on Applied Human Factors and Ergonomics*, 2019, pp. 221–232.

[24] D. A. Moses, M. K. Leonard, and E. F. Chang, "Real-time classification of auditory sentences using evoked cortical activity in humans," *J. Neural Eng.*, vol. 15, no. 3, p. 036005, 2018.

[25] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nat. Neurosci.*, vol. 23, no. 4, pp. 575–582, 2020.

[26] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[27] C. H. Nguyen, G. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian Manifold features," *J. Neural Eng.*, vol. 15, no. 1, p. 016002, 2017.

[28] O. Iljina *et al.*, "Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 186–199, 2017.

[29] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface," *IScience*, vol. 8, pp. 103–125, 2018.

[30] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, p. 046006, 2010.

[31] U. Chaudhary, B. Xia, S. Silvoni, L. G. Cohen, and N. Birbaumer, "Brain–Computer Interface–Based Communication in the Completely Locked-In State," *PLoS Biol.*, vol. 15, no. 1, pp. 1–25, 2017.

[32] J. Derix, O. Iljina, J. Weiske, A. Schulze-Bonhage, A. Aertsen, and T. Ball, "From speech to thought: the neuronal basis of cognitive units in non-experimental, real-life communication investigated using ECoG," *Front. Hum. Neurosci.*, vol. 8, no. June, pp. 1–17, 2014.

[33] E. M. Mugler *et al.*, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *J. Neural Eng.*, vol. 11, no. 3, p. 035015, 2014.

[34] T. Kim, J. Lee, H. Choi, H. Lee, I. Y. Kim, and D. P. Jang,

"Meaning based covert speech classification for brain-computer interface based on electroencephalography," *Int. IEEE/EMBS Conf. Neural Eng. NER*, pp. 53–56, 2013.

[35] K. Brigham and B. V. K. V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy," *2010 4th Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2010*, pp. 1–4, 2010.

[36] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, "Synchronization between overt speech envelope and EEG oscillations during imagined speech," *Neurosci. Res.*, vol. 153, pp. 48–55, 2020.

[37] A. R. Sereshkeh, R. Yousefi, A. T. Wong, and T. Chau, "Online classification of imagined speech using functional near-infrared spectroscopy signals," *J. Neural Eng.*, vol. 16, no. 016005, 2018.

[38] C. Cooney, R. Folli, and D. Coyle, "Differential effects of stimulus modality on decoding overt and imagined speech from EEG," 2021.

[39] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," *IEEE Int. Conf. Intell. Robot. Syst.*, vol. 2015-Decem, pp. 681–687, 2015.

[40] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behav. Brain Sci.*, vol. 22, no. 01, pp. 1–75, 1999.

[41] P. Indefrey, "The spatial and temporal signatures of word production: a critical update," *Front. Psychol.*, vol. 2, no. October, 2011.

[42] E. Edwards *et al.*, "Spatiotemporal imaging of cortical activation during verb generation and picture naming," *Neuroimage*, vol. 50, no. 1, pp. 291–301, 2010.

[43] G. S. Dell, N. Martin, and M. F. Schwartz, "A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming," *J. Mem. Lang.*, vol. 56, no. 4, pp. 490–520, 2007.

[44] S. Martin, I. Iturrate, J. del R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis," *Front. Neurosci.*, vol. 12, no. JUN, pp. 1–10, 2018.

[45] F. Pulvermüller, C. Cook, and O. Hauk, "Inflection in action: Semantic motor system activation to noun- and verb-containing phrases is modulated by the presence of overt grammatical markers," *Neuroimage*, vol. 60, no. 2, pp. 1367–1379, 2012.

[46] L. Pylkkänen, D. K. Bemis, and E. B. Elorrieta, "Building phrases in language production: An MEG study of simple composition," *Cognition*, vol. 133, no. 2, pp. 371–384, 2014.

[47] M. Angrick *et al.*, "Speech Synthesis from ECoG using Densely Connected 3D Convolutional Neural Networks," *J. Neural Eng.*, vol. 16, no. 036019, 2019.

[48] S. Ikeda *et al.*, "Neural decoding of single vowels during covert articulation using electrocorticography," *Front. Hum. Neurosci.*, vol. 8, no. March, pp. 1–8, 2014.

[49] S. Martin *et al.*, "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, 2016.

[50] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 4320, no. c, pp. 1–1, 2020.

[51] W. B. Baker, A. B. Parthasarathy, D. R. Busch, R. C. Mesquita, J. H. Greenberg, and A. G. Yodh, "Modified Beer-Lambert law for blood flow," *Biomed. Opt. Express*, vol. 5, no. 11, p. 4053, 2014.

[52] M. Bhatt, K. R. Ayyalasomayajula, and P. K. Yalavarthy, "Generalized Beer–Lambert model for near-infrared light propagation in thick biological tissues," *J. Biomed. Opt.*, vol. 21, no. 7, p. 076012, 2016.

[53] F. Herold, P. Wiegel, F. Scholkmann, and N. Müller, "Applications of Functional Near-Infrared Spectroscopy (fNIRS) Neuroimaging in Exercise–Cognition Science: A Systematic, Methodology-Focused Review," *J. Clin. Med.*, vol. 7, no. 12, p. 466, 2018.

[54] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[55] T. W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources," *Neural Comput.*, vol. 11, no. 2, pp. 417–441, 1999.

[56] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Comput. Intell. Neurosci.*, vol. 2011, 2011.

[57] S. Ioffe and C. Szegedy, "Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv Prepr. arXiv1502.03167*, 2015.

[58] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," *arXiv Prepr. arXiv1409.1259*, pp. 103–111, 2015.

[59] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," no. NeurIPS, 2019.

[60] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 2, pp. 321–357, 2002.

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, 2010.

[62] K. He, X. Zhang, R. Shaoqing, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[63] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[64] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *arXiv Prepr. arXiv1511.07289*, pp. 1–14, 2015.

[65] J. W. Tukey, "Comparing Individual Means in the Analysis of Variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.

[66] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 2623–2631, 2015.

[67] J. Lüdtke, C. K. Friedrich, M. De Filippis, and B. Kaup, "Event-related potential correlates of negation in a sentence-picture verification paradigm," *J. Cogn. Neurosci.*, vol. 20, no. 8, pp. 1355–1370, 2008.

[68] H.-J. Hwang *et al.*, "Toward more intuitive brain–computer interfacing: classification of binary covert intentions using functional near-infrared spectroscopy," *J. Biomed. Opt.*, vol. 21, no. 9, p. 091303, 2016.

[69] N. Naseer and K. S. Hong, "Classification of functional near-infrared spectroscopy signals corresponding to the right- and left-wrist motor imagery for development of a brain-computer interface," *Neurosci. Lett.*, vol. 553, pp. 84–89, 2013.

[70] M. Jawad Khan, M. J. Hong, and K. S. Hong, "Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface," *Front. Hum. Neurosci.*, vol. 8, no. 1 APR, pp. 1–10, 2014.

[71] C. Herff, D. Heger, F. Putze, C. Guan, and T. Schultz, "Cross-subject classification of speaking modes using fNIRS," in *International Conference on Neural Information Processing*, 2012, pp. 417–424.

[72] H.-J. Hwang, J.-H. Lim, D.-W. Kim, and C.-H. Im, "Evaluation of various mental task combinations for near-infrared spectroscopy-based brain-computer interfaces," *J. Biomed. Opt.*, vol. 19, no. 7, p. 077005, 2014.

[73] A. Faress and T. Chau, "Towards a multimodal brain-computer interface: Combining fNIRS and fTCD measurements to enable higher classification accuracy," *Neuroimage*, vol. 77, pp. 186–194, 2013.