frontiers in Neuroscience

ORIGINAL RESEARCH
published: 09 October 2018
doi: 10.3389/fnins.2018.00695

Speech Recognition via fNIRS Based Brain Signals

Yichuan Liu [1,2] *and Hasan Ayaz* [1,2,3,4]*

[1] School of Biomedical Engineering, Drexel University, Science and Health Systems, Philadelphia, PA, United States, [2] Cognitive Neuroengineering and Quantitative Experimental Research (CONQUER) Collaborative, Drexel University, Philadelphia, PA, United States, [3] Department of Family and Community Health, University of Pennsylvania, Philadelphia, PA, United States, [4] The Division of General Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, United States

In this paper, we present the first evidence that perceived speech can be identified from the listeners' brain signals measured via functional-near infrared spectroscopy (fNIRS)—a non-invasive, portable, and wearable neuroimaging technique suitable for ecologically valid settings. In this study, participants listened audio clips containing English stories while prefrontal and parietal cortices were monitored with fNIRS. Machine learning was applied to train predictive models using fNIRS data from a subject pool to predict which part of a story was listened by a new subject not in the pool based on the brain's hemodynamic response as measured by fNIRS. fNIRS signals can vary considerably from subject to subject due to the different head size, head shape, and spatial locations of brain functional regions. To overcome this difficulty, a generalized canonical correlation analysis (GCCA) was adopted to extract latent variables that are shared among the listeners before applying principal component analysis (PCA) for dimension reduction and applying logistic regression for classification. A 74.7% average accuracy has been achieved for differentiating between two 50 s. long story segments and a 43.6% average accuracy has been achieved for differentiating four 25 s. long story segments. These results suggest the potential of an fNIRS based-approach for building a speech decoding brain-computer-interface for developing a new type of neural prosthetic system.

Keywords: BCI, fNIRS, prefrontal cortex (PFC), parietal lobe, speech perception, decoding

OPEN ACCESS

Edited by:
Mikhail Lebedev,
Duke University, United States

Reviewed by:
Kazutaka Takahashi,
University of Chicago, United States
Fernando Brunetti,
Universidad Católica Nuestra Señora de la Asunción, Paraguay
Antonio Oliviero,
Fundación del Hospital Nacional de Parapléjicos, Spain

***Correspondence:**
Hasan Ayaz
hasan.ayaz@drexel.edu

Specialty section:
This article was submitted to Neuroprosthetics, a section of the journal Frontiers in Neuroscience

Received: *10 May 2018*
Accepted: *18 September 2018*
Published: *09 October 2018*

Citation:
Liu Y and Ayaz H (2018) Speech Recognition via fNIRS Based Brain Signals. Front. Neurosci. 12:695. doi: 10.3389/fnins.2018.00695

INTRODUCTION

The decoding of speech from brain signals has attracted the attention of researchers in recent years (Chakrabarti et al., 2015; AlSaleh et al., 2016; Herff and Schultz, 2016). A device that can directly translate brain signals into texts that describe a person's thoughts may help people with disabilities and verbal communication deficits and enable a new communication channel with the outside world. Such brain-computer interfacing device may also help healthy people to directly interact with a machine without the need of using muscles and potentially expand the interaction bandwidth.

Most of the previous studies focused on phoneme-based decoding and adopted invasive or partially invasive technology that requires the implant of sensors during neurosurgery. For example, Brumberg et al. investigated the classification of intended phoneme production based on intracortical microelectrode recordings (Brumberg et al., 2011). Herff et al. decoded words from

continuously spoken speech from intracranial electrocorticographic (ECoG) signals recorded from epileptic patients based on the classification of phonemes (Herff et al., 2015). A 75% classification accuracy has been achieved for a dictionary of 10 words and a 40% accuracy for a dictionary of 100 words. Martin et al. investigated the classification of individual words from a pair of imagined word using ECoG and a 58% binary classification accuracy has been achieved (Martin et al., 2016). The authors also showed that the binary classification of listened and overt spoken words is much better, which achieved 89% and 86% accuracy, respectively. We refer to Chakrabarti et al. (2015) and Herff and Schultz (2016) for a more comprehensive review of the state of art in the field.

Despite of the promising results achieved, invasive technology requires the implantation of sensors which limits their applications, especially for the healthy populations. In the field of non-invasive brain-computer interface, studies mainly adopted electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) for the classification of speech related task conditions. For example, O'Sullivan et al. played two audio stories of 60 s long simultaneously to the subjects who were instructed to attend to one of the stories. They were able the identify which story the subjects were attended to with 89% accuracy (O'Sullivan et al., 2015). Yoshimura et al. investigated the 3-class classification of two imagined phonemes and no-imagery control and a 59% accuracy has been achieved adopting fMRI prior for current source estimation (Yoshimura et al., 2016). Vodrahalli et al. investigated the classification of movie scene from fMRI response (Vodrahalli et al., 2017). A shared response model (SRM) has been used for dimension reduction and a classification accuracy of 72% has been achieved over 4% chance level. A recent fMRI study investigated a similar task of identifying musical pieces (Hoefle et al., 2018). A multiple regression has been adopted to predict features of musical pieces for the task of differentiating between new pieces not used for training the predictive model.

In this study, we adopted fNIRS for the classification of listened stories. This approach is motivated from our recent fNIRS study (Liu et al., 2017c) and a number of other fMRI studies (Stephens et al., 2010; Lerner et al., 2011; Hasson et al., 2012) which show that listeners' brain mirror each other whenever they are listening to the same story and the listeners' brain mirror the speaker's brain with a delay. We also inspired from fNIRS BCI studies which revealed the potential of fNIRS in classifying mental conditions (Ayaz et al., 2009; Power et al., 2010, 2012; Fazli et al., 2012a,b; Khan et al., 2014; Liu et al., 2017b).

fNIRS is an optical based neuroimaging technique for measuring the cortical concentration changes in the oxygenated (oxy-Hb) and deoxygenated (deoxy-Hb) hemoglobin. It is portable, wearable (Piper et al., 2014; Mckendrick et al., 2015), non-invasive and can even be battery-operated and wireless, and hence particularly suitable for brain-computer interfacing in everyday settings (Ayaz et al., 2011, 2013; Liu et al., 2017a,b). Several studies in the literature adopted fNIRS to investigate the classification of neural signals during listening comprehension, speech production or related topics. Power et al. (2010) investigated the classification of music and mental arithmetic conditions and a 77% accuracy has been achieved (Power et al., 2010). The same group in 2012 investigated the three-class classification problem of differentiating mental arithmetic, mental singing and no-control state and a 56.2% accuracy has been achieved (Power et al., 2012). Telkemeyer et al. (2011) investigated the acoustic processing of non-linguistic sounds in infants combining EEG and fNIRS (Telkemeyer et al., 2011). Herff et al. (2012) adopted fNIRS for differentiating between speech and not speaking conditions (Herff et al., 2012). They achieved 71% and 61% for classifying overt speech/not speaking and silent speech/not speaking, respectively. Moghimi et al. (2012) investigated the classification of music excerpts with different emotional content using only fNIRS. They were able to differentiate excerpts with positive and negative emotions with 72% accuracy (Moghimi et al., 2012). Putze et al. (2014) combined EEG and fNIRS for differentiating visual and auditory perception processes from each other and achieved 98% accuracy (Putze et al., 2014).

To study fNIRS-based speech recognition, we used the data from our previous study which included fNIRS recordings while participants ($N = 19$) were listening to English stories (Liu et al., 2017c). We divided the fNIRS signals into several segments (each corresponding to a part of a story) and machine learning was applied for identifying which part of the stories a participant was listening to using only fNIRS signal.

A major obstacle in the classification of fNIRS signal is the individual variations caused by the different size and shape of the head/brain across the subjects. For some subjects, their head shape resulted in channels being rejected due to bad contact. Conventionally, this problem can be alleviated by acquiring additional information such as the 3D coordinates of the sensors and results from a structural magnetic resonance imaging scan. This information can be used to estimate the exact location of the brain a sensor was measuring from and transform all data to a standard brain space (Tsuzuki and Dan, 2014). However, the conventional approach can be time consuming, costly and it doesn't take into account the individual differences in brain activation regions which could also deteriorate the accuracy of fNIRS signal classification. As a solution to this problem, we applied spatial filters for extracting latent variables that have minimum between-subject variations. Spatial filters find linear combination of the optodes (i.e., fNIRS sensors) to linearly transform the raw data. We consider two spatial filtering techniques: GCCA (Shen et al., 2014) and SRM (Chen et al., 2015). Each of these techniques finds a set of subject-specific spatial filters to extract latent variables. GCCA extracts the 1st latent variable that maximizes the between-subject correlations in the signal time course. It then extracts the 2nd latent variable to maximize the between-subject correlations subject to the constrain that it is uncorrelated with the 1st latent variable. This procedure is repeated until no more latent variable could be found. For SRM, it finds spatial filters which minimize the sum of squared error between extracted latent variables and the estimated component time courses that are shared among all subjects.

METHODS

Participants

Nineteen participants from our previous study (Liu et al., 2017c) who have completed the listening comprehension of both story E1 and E2 were used for the analysis in this study.

Signal Acquisition

Two continuous wave optical brain imaging devices were used simultaneously on each participant to record brain activity from prefrontal cortex (PFC) and parietal cortex (PL) using 40 measurement locations (optodes) (**Figure 1**). Anterior prefrontal cortex was recorded in 2 Hz by a 16-optode continuous wave fNIRS system (fNIR Imager Model 1100; fNIR Devices, LLC) first described by Chance et al. (1998) and developed in our lab at Drexel University (Izzetoglu et al., 2005; Ayaz et al., 2011). Parietal cortex was recorded in 10 Hz using a 24-optode continuous wave Hitachi fNIRS system (ETG 4000; Hitachi Medical Systems). Please refer to Liu et al. (2017c) for a detailed description of the signal acquisition.

Signal Processing

Raw light intensities were converted into concentration changes in oxygenated-hemoglobin (HbO) and deoxygenated-hemoglobin (HbR) concentrations using the modified Beer-Lambert law (Cope and Delpy, 1988). The raw signal and hemoglobin concentration changes were inspected both visually and also using the automated SMAR algorithm (Ayaz et al., 2010), which uses a coefficient-of-variation based approach to assess signal quality, reject problematic channels with bad contact or saturated raw light intensity. Two optodes, 1 and 15, were over the hairline for most participants and hence were rejected from the study. One more optode (optode 37) was rejected from the study because it has been rejected in more than 40% of the subjects. The HbO and HbR signals were band-pass filtered from 0.01 to 0.1 Hz using a 4-th order zero-phase infinite impulse response (IIR) filter for reducing artifacts from physiological signals (Ayaz et al., 2011). The cut-off frequencies reflect common settings addressing global drifts (<0.01 Hz) and systemic interferences such as Mayer wave (∼0.1 Hz), respiration rate (>0.2 Hz) and cardiac cycles (>0.5 Hz). We then downsampled the signals to 1 Hz. We rejected the first 30 s of each story because they may be affected by transient global increases or decreases in response amplitude caused by the start of listening comprehension. For each story, the signal time courses were divided into segments of 100 s duration. A total of 9 segments were extracted from the two stories.

Feature Extraction

fNIRS signals vary considerably from subject to subject due to the different head size, head shape, and spatial locations of brain functional regions. For reducing the between-subject variations, two spatial filtering approaches were considered: the shared response model (SRM) and generalized canonical correlation analysis (GCCA).

Shared Response Model

The SRM was proposed by Chen et al. (2015) for identifying the shared brain responses among subjects by estimating subject-specific spatial filters. More specifically, the spatial filters $\boldsymbol{W_i}$ for subject i were estimated as below:

$$\begin{aligned} &\min_{wi,s} \sum_i ||\boldsymbol{W_i^T X_i} - \boldsymbol{S}||_F^2 \\ &s.t. \boldsymbol{W_i}^T \boldsymbol{W_i} = \boldsymbol{I_k} \end{aligned} \tag{1}$$

where $\boldsymbol{X_i} \in \mathbb{R}^{v_i \times t} (i = 1, \ldots, N)$ is the fNIRS signals of subject i with v_i channels and t time points. In this study $\max_i(v_i) = 37(optodes) \times 2\left(HbO/HbR\right) = 74$. For a subject, some optodes were rejected due to over the hairline, ambient light leakage or severe motion artifact contamination. $\boldsymbol{W_i} \in \mathbb{R}^{v_i \times k}$ is the spatial filters of subject i and the parameter k represents the number of spatial filters to be estimated. Parameter k needs to be selected by the experimenter. And $\boldsymbol{S} \in \mathbb{R}^{k \times t}$ is the corresponding time series of responses shared by all participants. For each subject, $\tilde{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{W_i^T X_i}$ is then used as feature for classification.

Generalized Canonical Correlation Analysis

GCCA estimates subject-specific spatial filters for extracting orthogonal components that are maximally correlated among the subjects. We denote $\boldsymbol{X_i} \in \mathbb{R}^{v_i \times t} (i = 1, \ldots, N)$ the fNIRS signals of subject i with v_i channels and t time points and $\boldsymbol{W_i} \in \mathbb{R}^{v_i \times k} (i = 1, \ldots, N)$ the spatial filters for subject i. GCCA maximizes the following:

$$\begin{aligned} \phi(\boldsymbol{W}) &= tr(\boldsymbol{W^T X X^T W}) \\ s.t. \boldsymbol{W^T D W} &= \boldsymbol{I_t} \end{aligned} \tag{2}$$

where $\boldsymbol{X} = \left[\boldsymbol{X_1^T}, \ldots, \boldsymbol{X_N^T}\right]^T \in \mathbb{R}^{v \times t}$ $(v = v_1 + \ldots + v_N)$, $\boldsymbol{W} = \left[\boldsymbol{W_1^T}, \ldots, \boldsymbol{W_N^T}\right]^T \in \mathbb{R}^{v \times k}$ and $\mathbf{D} = \begin{pmatrix} \boldsymbol{X_1 X_1^T} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{X_N X_N^T} \end{pmatrix}$ is a block diagonal matrix. This could be achieved by solving the following generalized eigenvalue problem:

$$\boldsymbol{X X^T w} = \lambda \boldsymbol{D w}$$

After the spatial filters were estimated, the latent variables $\tilde{\boldsymbol{X}}_i = \boldsymbol{W_i^T X_i}$ is used as features for classification.

Inter-subject Correlation

For gaining an intuitive understanding of the relative importance of the fNIRS channels in characterizing the story segments, we calculated the inter-subject correlation for each fNIRS channel j as follows:

$$r^j = \frac{1}{19} \sum_{i=1}^{19} \rho(\boldsymbol{x_i^j}, \bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{j}})$$

where $\rho(\cdot)$ denotes Pearson's correlation, $\boldsymbol{x_i^j}$ is the fNIRS time course of subject i, $\bar{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{j}}$ is the average time course of all other

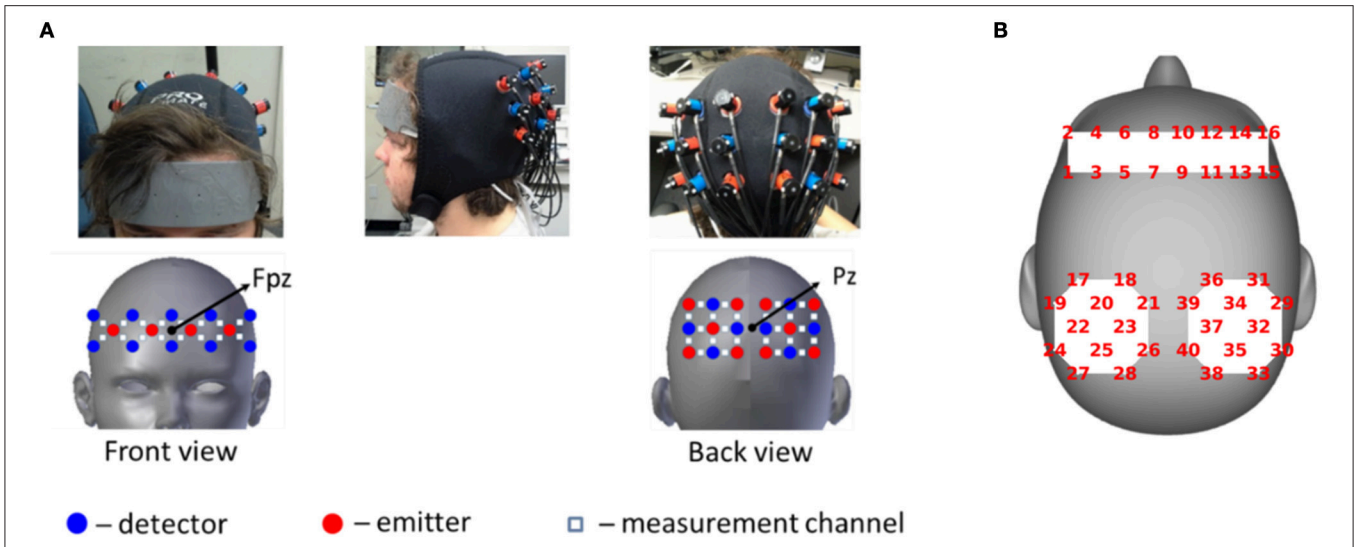

FIGURE 1 | (A) fNIRS acquisition setup. Red circles indicate emitters; blue circles indicate detectors; White squares indicate measurement channels between emitters and detectors. **(B)** Schematic representation of the same optode locations on head surface which is used to show results. Written informed consent was obtained from the participant and adapted from Liu et al. (2017c).

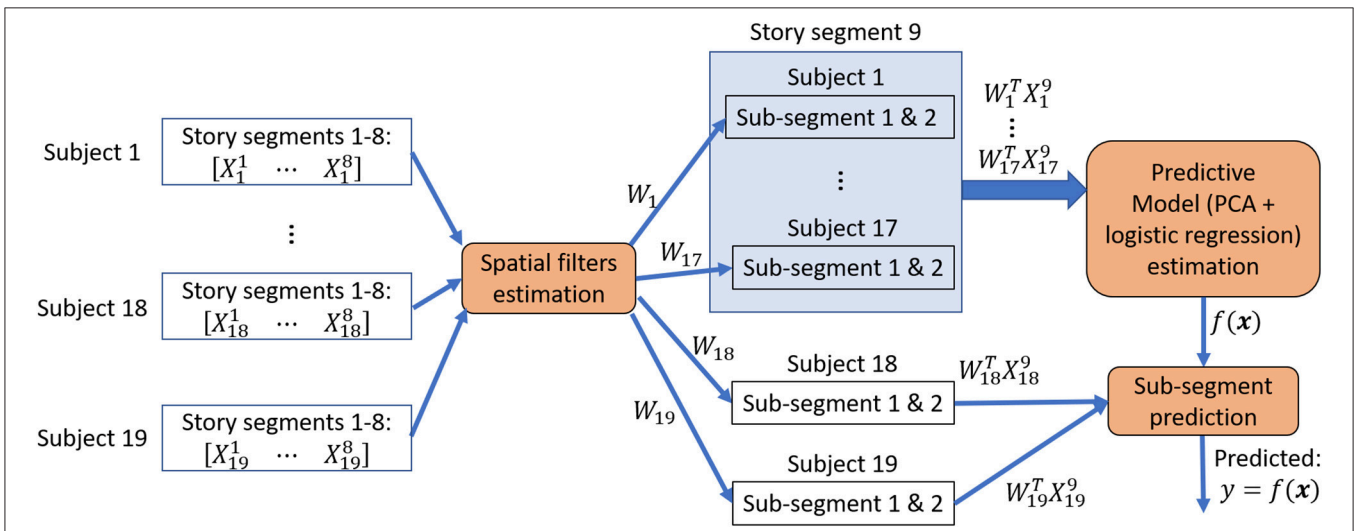

FIGURE 2 | Story sub-segments classification procedure. Spatial filters were estimated from 8 story segments to minimize between-subject variations and then applied to the 9th segment that has been left out. The 30 s. of data immediately before and after the 9th segment was rejected from training data. This procedure is repeated until all segment has been left out exactly once for a leave-one-segment-out cross-validation. During each iteration of the cross-validation, the spatially filtered segment that has been left out were further divided into 2 or 4 sub-segments, forming a 2-class or 4-class classification problem. The sub-segments from 18 or 19 subjects were used to estimate a predictive model and its classification performance is evaluated on the one or two subjects that has been left out for performing an inner 10-fold cross-validation.

subjects except subject *i*. The inter-subject correlation reflects the consistency of the signal cross different subjects. To test the significance of the inter-subject correlation, a phase-scrambling random permutation procedure as used in (Lerner et al., 2011) was adopted. More specifically, the Fourier transform was applied on a fNIRS time course, the phase of the frequency components was randomized and the inverse Fourier transform was applied to obtain a phase-scrambled time course. This procedure was repeated 1,000 times for estimating the null distributions of the inter-subject correlations. In our study, there are 34 (optodes) × 2 (HbO/HbR) = 74 fNIRS channels. To correct for multiple comparison, the maximum statistic approach was applied (Nichols and Holmes, 2002), i.e., in each of the 1,000 iterations, the maximum inter-subject correlation values among the 74 phase-scrambled fNIRS time courses were calculated ($r_{max} = \max_{j=1...74} r_j$) for estimating the null distribution of $r_{\max}$.

The null hypothesis of a fNIRS channel is rejected if its inter-subject correlation is higher than 95% of the samples in the null distribution of $r_{\max}$.

Classification and Performance Evaluation

For classification, the fNIRS time courses of the signal segments were used as feature, a principal components analysis was adopted for dimensional reduction and the logistic regression was adopted for classification. A leave-one-segment-out cross-validation was performed to apply spatial filtering for reducing between-subject variations and a 10-fold cross-validation was performed to evaluate story segments classification. We denote $\mathbf{X}_i^p (i = 1, \ldots, N; p = 1, \ldots, 9)$ the fNIRS time courses of subject i and story segment p. The performance evaluation procedure is as follows:

For story segment $p = 1, \ldots, 9$:

- Spatial filter training:
  All story segments except segment p $\{\mathbf{X}_i^j, j \neq p,\ i = 1, \ldots, N\}$ were used as training set for estimating K spatial filters for each subject ($\mathbf{W_p} = \left[\mathbf{W_{1p}^T}, \ldots, \mathbf{W_{Np}^T}\right]^T \in \mathbb{R}^{v \times K}$) adopting SRM or GCCA**.** Before applying spatial filtering, the fNIRS optodes were normalized to have a mean of zeros and standard deviation of ones. The 30 s. long data immediately before and after the testing segments were rejected from training data.
- Spatial filter testing:
  The spatial filters estimated in training were applied to the story segment p (the testing set) to extract spatial components $\tilde{X}_i^p$.
- Classification:
  The estimated $\tilde{\mathbf{X}}_i^p$ (or $\mathbf{X_i^p}$ if spatial filtering is not applied) were divided into four 25 s sub-segments (four-class classification problem) or two 50 s sub-segments (binary classification problem) for investigating story segments classification. For the four-class classification problem, there were 4 (classes) × 19 (subjects) = 76 samples and each sample included k (latent variables) × 25 (time points) = $25k$ features. For the two-class classification problem, there were 2 (classes) × 19 (subjects) = 38 samples and each sample included k (latent variables) × 50 (time points) = $50k$ features. The 10-fold cross-validation was then applied to the 76 (four-class problem) or 38 (two-class problem) samples for estimating the classification accuracy. More specifically, we randomly divided the subjects into 10 groups. Using the data from 9 groups as training subjects, we first applied PCA to reduce the dimensionality of the data. The largest principal components that explained 99.9% of the variance of the data were extracted. A logistic regression analysis was performed using the principal components as features. The classification accuracy of the group that has been left out was then estimated applying the PCA and classifier coefficients learned from training data. The classification accuracy using story segment p as testing segment and subject i as one of the testing subject is denoted as acc_i^p.

For each subject i, we compared the average classification accuracy $acc_i = \frac{1}{9}\sum_{p=1}^{9} acc_i^p$ achieved with SRM-estimated spatial filter, with GCCA-estimated spatial filter and without applying any spatial filter. **Figure 2** illustrates the procedure for estimating story segments classification accuracies.

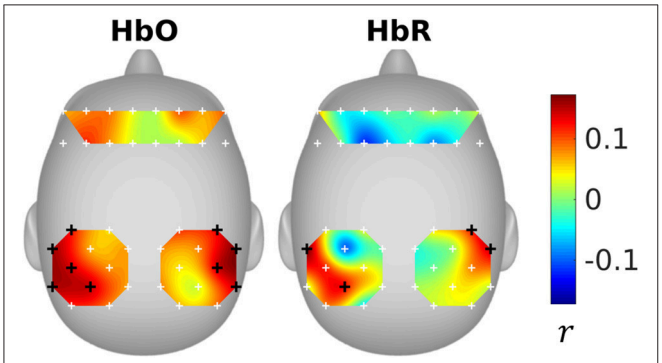

FIGURE 3 | Inter-subject correlation. Black "+" sign represents significant channels.

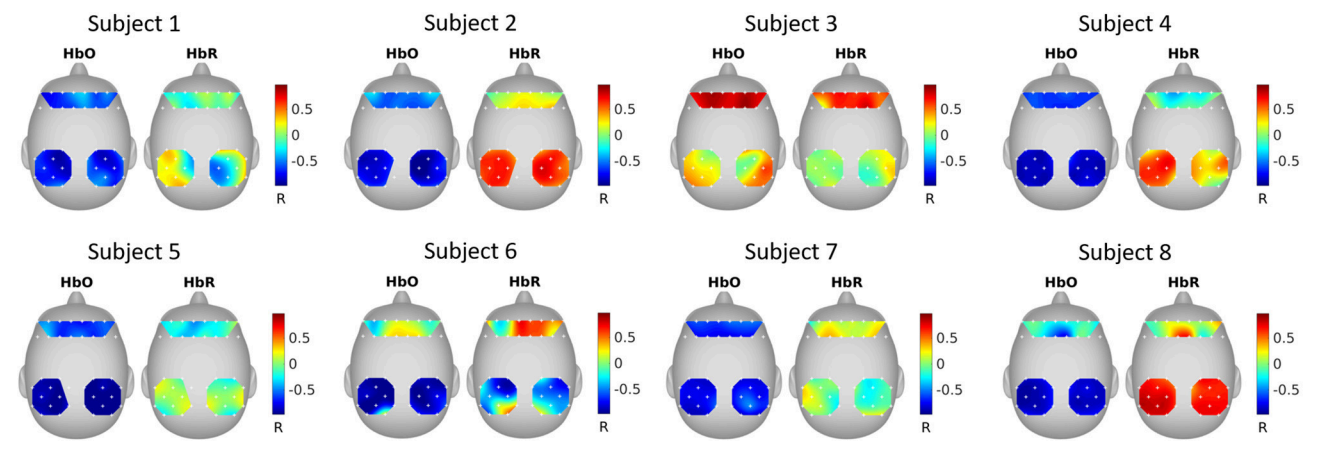

FIGURE 4 | SRM estimated spatial filters. The spatial filters for each subject were estimated with SRM $k = 1$ spatial filter. The topographical maps show the Pearson's correlation between the extracted latent variable and the fNIRS channels.

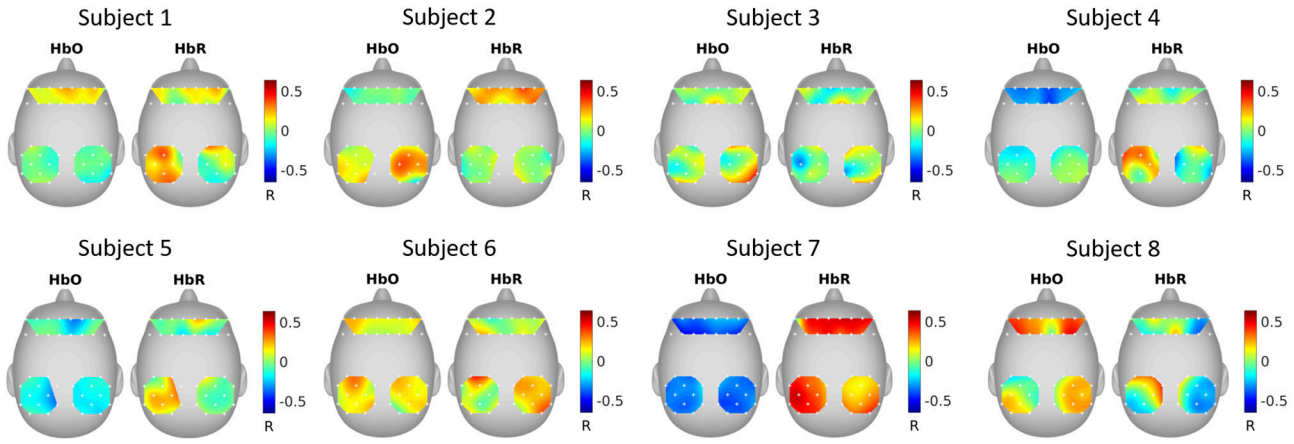

FIGURE 5 | GCCA estimated spatial filters. The topographical maps show the Pearson's correlation between the extracted latent variable that was most correlated among subjects and the fNIRS channels.

RESULTS

Inter-subject Correlation

The inter-subject correlations were shown in **Figure 3**. It can be seen that the subjects were significantly correlated in the parietal areas. The significant optodes cover parts of supramarginal gyrus, angular gyrus, superior parietal gyrus, and postcentral gyrus (please refer to Figure 8 and Table S1 of Liu et al., 2017c, for the fNIRS optode locations).

SRM Estimated Spatial Filters

Figure 4 shows the correlation between SRM extracted latent variable ($k = 1$ *spatial filter*) and fNIRS channels. It can be seen that the latent variable is mostly negatively correlated with HbO and positively correlated with HbR.

GCCA Estimated Spatial Filters

Figure 5 shows the correlation between GCCA extracted latent variable (the variable that is most correlated among subjects) and fNIRS channels. It can be seen that the correlation pattern varies for different subjects.

Classification

Figure 6 shows the classification results for the three approaches (SRM, GCCA or no spatial filtering) with different story segment duration. We first estimated using all 74 spatial components for classification [there are 37 (optodes) $\times$ 2 [HbO/HbR] $=$ 74 channels]. All three approaches achieved significantly better than chance level accuracy (FDR $q < 0.05$). The chance level accuracy is 50% for the binary classification and 25% for the 4-class classification problem. Without spatial filtering, 63.2 $\pm$ 11.8% and 38.0 $\pm$ 9.8% (mean $\pm$ standard deviation) accuracy have been achieved for the two-class (50 s. sub-segment) and four-class (25 s. sub-segment) problem, respectively. GCCA achieved 74.7 $\pm$ 8.5% and 43.6$\pm$13.2% accuracy for the two-class problem and four-class problem, respectively. For the two-class problem, GCCA significantly outperformed the accuracy achieved without spatial filtering (FDR $q < 0.05$). SRM achieved 72.0 $\pm$ 10.5% and 43.8 $\pm$ 12.1% accuracy for the two-class and four-class problem, respectively. No other significant differences between the three approaches has been found.

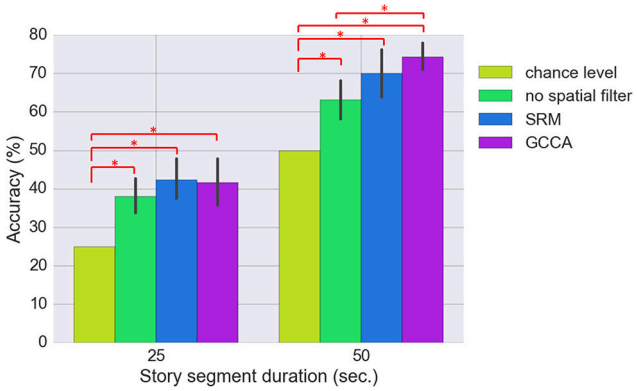

FIGURE 6 | Comparing the classification accuracy using SRM ($K = 74$ latent variables), GCCA ($K = 74$ latent variables), and without spatial filter. The results for two classification problems are shown: 4-class (25 s. sub-segment duration) and binary (50 s. sub-segment duration). The error bar stands for bootstrapped 95% confidence interval. Methods with statistically significant differences (FDR $q < 0.05$) are marked by red bars and asterisks.

The effect of different number of spatial components on GCCA and SRM accuracy is shown in **Figure 7**. It can be seen that the highest accuracy was achieved using all 74 spatial components.

DISCUSSION

In this study, we applied machine learning to identify among several story segments the one that was listened by a subject based on the brain's hemodynamic response measured with fNIRS. An inter-subject correlation analysis revealed that the time courses of fNIRS were significantly correlated at parietal areas, suggesting that signal was most consistent at parietal optodes,

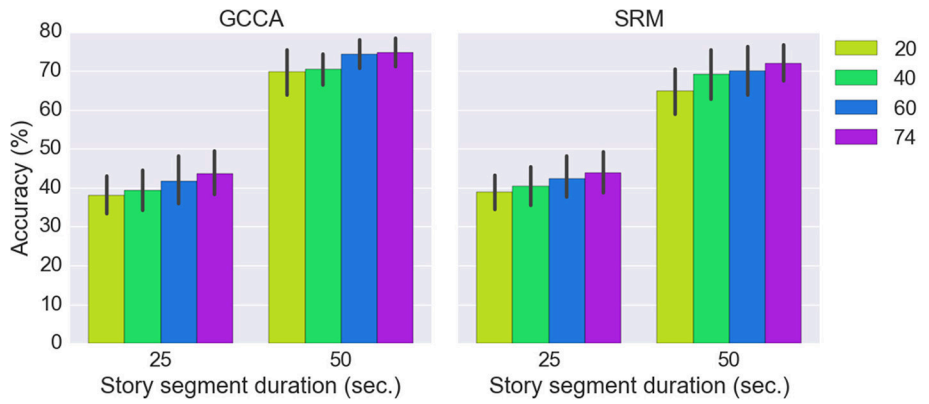

FIGURE 7 | Classification accuracies of GCCA and SRM with different number of spatial components.

and parietal optodes may have provided the most information for characterizing the story segments. To reduce the between-subject variations caused by inter-subject differences such as head size, head shape, and brain activation regions, spatial filters were applied to extract latent variables which are linear combinations of the fNIRS optodes. Without spatial filtering, a story segment classification accuracy of 63.2% and 38.0% have been achieved for the binary classification problem (50 s. story segment duration) and 4-class classification problem (25 s. story segment duration), respectively. After applying GCCA spatial filters, a classification accuracy of 74.7% and 43.6% have been achieved for the binary and 4-class classification problems, respectively. This is better than the results achieved without spatial filtering. Applying SMR spatial filters resulted in better classification accuracy compared to no spatial filter but it is not as effective as GCCA especially for the binary classification problem.

Although we performed the classification of fNIRS signals in response to the listening comprehension of stories, we speculate that it is also plausible to classify overt or even covert speech production based on the evidences that a listener's brain mirrors the speaker's brain with a delay (Stephens et al., 2010; Liu et al., 2017c). Further evidences can be found in an fNIRS-based speaking mode classification study (Herff et al., 2012). In the study, the binary classifications of overt speech/not speaking and silent speech/not speaking were investigated and an accuracy of 71% and 61% have been achieved, respectively.

It is worth pointing out that we only utilized fNIRS from 37 locations in the prefrontal and parietal lobe. With full head fNIRS coverage and increased optode density, we expect an even better story segment classification performance for allowing the identification of shorter speech from a larger pool of candidate speeches.

Despite the promising results, the current study is limited in the following aspects. First, the audio clips have been played to the subjects only once, i.e., the stories were novel to all subjects. How well the predictive models can perform for stories that has been repeatedly played to the subjects are still unknown. Second, we evaluated the performance of the spatial filters on story segments with a total duration of 900 s. The stories have been played to the participants on the same day within 2 h. How well the spatial filters can generalize to longer stories and for stories played on different days remain to be seen. We speculate that estimating the spatial filters using longer stories with more varieties in the story content and applying regularization techniques can help improve generalization. Finally, our accuracy still needs improvement for real-world setups and 25–50 s time course length is not suitable for most of the neuroprosthetic applications. Further development in fNIRS signal acquisition and processing is needed for improving decoding accuracy and decreasing the time course length. Incorporating information from other modalities such as EEG may also help.

In summary, we showed that it is possible to identify speech from fNIRS data with machine learning techniques. The application of spatial filters reduced the inter-subject variations in the data and help improved classification performance. The current study is a step toward a BCI that reconstructs speech contents from brain signals for helping locked-in syndrome patients or healthy individuals to augment human-computer interaction as a new type of neural prosthetic system. However, there is a long way before such BCI can be achieved. As the next step, studies could be conducted using high density fNIRS covering more areas of the brain and incorporating information from other modalities such as EEG for improving the classification accuracy of shorter speeches.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Institutional Review Board of Drexel University. The protocol was approved by the Institutional Review Board of Drexel University. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

YL performed the experiment, collected the fNIRS data, analyzed the data, and prepared/wrote the manuscript. HA initiated and supervised the study, designed the experiment, analyzed the data, discussed, and interpreted the results as well as prepared/revised the manuscript.

REFERENCES

AlSaleh, M. M., Arvaneh, M., Christensen, H., and Moore, R. K. (2016). "Brain-computer interface technology for speech recognition: a review," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (Jeju), 1–5. doi: 10.1109/APSIPA.2016.7820826

Ayaz, H., Izzetoglu, M., Shewokis, P. A., and Onaral, B. (2010). "Sliding-window motion artifact rejection for functional near-infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (Buenos Aires: IEEE), 6567–6570.

Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P. A., Mckendrick, R., and Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: Empirical examples and a technological development. *Front. Human Neurosci.* 7:871. doi: 10.3389/fnhum.2013.00871

Ayaz, H., Shewokis, A. P., Bunce, S., Schultheis, M., and Onaral, B. (2009). "Assessment of cognitive neural correlates for a functional near infrared-based brain computer interface system," in *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience,* eds D. Schmorrow, I. Estabrooke, and M. Grootjen. (San Diego, CA: Springer), 699–708.

Ayaz, H., Shewokis, P. A., Curtin, A., Izzetoglu, M., Izzetoglu, K., and Onaral, B. (2011). Using mazesuite and functional near infrared spectroscopy to study learning in spatial navigation. *J. Visual. Exp.* 56:e3443. doi: 10.3791/3443

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., and Kennedy, P. R. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Front. Neurosci.* 5:65. doi: 10.3389/fnins.2011.00065

Chakrabarti, S., Sandberg, H. M., Brumberg, J. S., and Krusienski, D. J. (2015). Progress in speech decoding from the electrocorticogram. *Biomed. Eng. Lett.* 5, 10–21. doi: 10.1007/s13534-015-0175-1

Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., et al. (1998). A novel method for fast imaging of brain function, non-invasively, with light. *Opt Express* 2, 411–423. doi: 10.1364/OE.2.000411

Chen, P.-H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J. V., and Ramadge, P. J. (2015). "A reduced-dimension fMRI shared response model," in *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Montreal: MIT Press).

Cope, M., and Delpy, D. T. (1988). System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Med. Biol. Eng. Comput.* 26, 289–294. doi: 10.1007/BF02447083

Fazli, S., Mehnert, J., Steinbrink, J., and Blankertz, B. (2012a). "Using NIRS as a predictor for EEG-based BCI performance," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (San Diego, CA), 4911–4914.

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., et al. (2012b). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage* 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., and Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends Cogn. Sci.* 16, 114–121. doi: 10.1016/j.tics.2011.12.007

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., et al. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* 9:217. doi: 10.3389/fnins.2015.00217

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012). Speaking mode recognition from functional near infrared spectroscopy. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2012:1715–8. doi: 10.1109/EMBC.2012.6346279

Herff, C., and Schultz, T. (2016). Automatic speech recognition from neural signals: a focused review. *Front. Neurosci.* 10:429. doi: 10.3389/fnins.2016.00429

Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., et al. (2018). Identifying musical pieces from fMRI data using encoding and decoding models. *Sci. Rep.* 8:2266. doi: 10.1038/s41598-018-20732-3

Izzetoglu, M., Izzetoglu, K., Bunce, S., Ayaz, H., Devaraj, A., Onaral, B., et al. (2005). Functional near-infrared neuroimaging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 13, 153–159. doi: 10.1109/TNSRE.2005.847377

Khan, M. J., Hong, M. J., and Hong, K.-S. (2014). Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface. *Front. Human Neurosci.* 8:244. doi: 10.3389/fnhum.2014.00244

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Liu, Y., Ayaz, H., and Shewokis, P. A. (2017a). Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy. *Brain Comput. Interf.* 4, 175–185. doi: 10.1080/2326263X.2017.1304020

Liu, Y., Ayaz, H., and Shewokis, P. A. (2017b). Multisubject "Learning" for mental workload classification using concurrent EEG, fNIRS, and physiological measures. *Front. Human Neurosci.* 11:389. doi: 10.3389/fnhum.2017.00389

Liu, Y., Piazza, E. A., Simony, E., Shewokis, P. A., Onaral, B., Hasson, U., et al. (2017c). Measuring speaker–listener neural coupling with functional near infrared spectroscopy. *Sci. Rep.* 7:43293. doi: 10.1038/srep43293

Martin, S., Brunner, P., Iturrate, I., Millán, J. D. R., Schalk, G., Knight, R. T., et al. (2016). Word pair classification during imagined speech using direct brain recordings. *Sci. Rep.* 6:25803. doi: 10.1038/srep25803

Mckendrick, R., Parasuraman, R., and Ayaz, H. (2015). Wearable functional Near Infrared Spectroscopy (fNIRS) and transcranial Direct Current Stimulation (tDCS): Expanding Vistas for Neurocognitive Augmentation. *Front. Syst. Neurosci.* 9:27. doi: 10.3389/fnsys.2015.00027

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. doi: 10.1088/1741-2560/9/2/026022

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Piper, S. K., Krueger, A., Koch, S. P., Mehnert, J., Habermehl, C., Steinbrink, J., et al. (2014). A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage* 85, 64–71. doi: 10.1016/j.neuroimage.2013.06.062

Power, S. D., Falk, T. H., and Chau, T. (2010). Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy. *J. Neural Eng.* 7:026002. doi: 10.1088/1741-2560/7/2/026002

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. doi: 10.1186/1756-0500-5-141

Putze, F., Hesslinger, S., Tse, C.-Y., Huang, Y., Herff, C., Guan, C., et al. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Front. Neurosci.* 8:373. doi: 10.3389/fnins.2014.00373

Shen, C., Sun, M., Tang, M., and Priebe, C. E. (2014). Generalized canonical correlation analysis for classification. *J. Multi. Analys.* 130, 310–322. doi: 10.1016/j.jmva.2014.05.011

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14425–14430. doi: 10.1073/pnas.1008662107

Telkemeyer, S., Rossi, S., Nierhaus, T., Steinbrink, J., Obrig, H., and Wartenburger, I. (2011). Acoustic processing of temporally modulated

sounds in infants: evidence from a combined near-infrared spectroscopy and EEG study. *Front. Psychol.* 2:62. doi: 10.3389/fpsyg.2011.00062

Tsuzuki, D., and Dan, I. (2014). Spatial registration for functional near-infrared spectroscopy: From channel position on the scalp to cortical location in individual and group analyses. *NeuroImage* 85, 92–103. doi: 10.1016/j.neuroimage.2013.07.025

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., et al. (2017). Mapping between fMRI responses to movies and their natural language annotations. *Neuroimage* 180(Pt A):223–231. doi: 10.1016/j.neuroimage.2017.06.042

Yoshimura, N., Nishimoto, A., Belkacem, A. N., Shin, D., Kambara, H., Hanakawa, T., et al. (2016). Decoding of covert vowel articulation using electroencephalography cortical currents. *Front. Neurosci.* 10:175. doi: 10.3389/fnins.2016.00175

Conflict of Interest Statement: fNIR Devices, LLC manufactures the optical brain imaging instrument and licensed IP and know-how from Drexel University. HA was involved in the technology development and thus offered a minor share in the new startup firm fNIR Devices, LLC.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.