Speaking Mode Recognition from Functional Near Infrared Spectroscopy

Christian Herff[1], Felix Putze[1], Dominic Heger[1], Cuntai Guan[2] and Tanja Schultz[1]

***Abstract*— Speech is our most natural form of communication and even though functional Near Infrared Spectroscopy (fNIRS) is an increasingly popular modality for Brain Computer Interfaces (BCIs), there are, to the best of our knowledge, no previous studies on speech related tasks in fNIRS-based BCI.**

We conducted experiments on 5 subjects producing audible, silently uttered and imagined speech or do not produce any speech. For each of these speaking modes, we recorded fNIRS signals from the subjects performing these tasks and distinguish segments containing speech from those not containing speech, solely based on the fNIRS signals. Accuracies between 69% and 88% were achieved using support vector machines and a *Mutual Information based Best Individual Feature* approach. We are also able to discriminate the three speaking modes with 61% classification accuracy. We thereby demonstrate that speech is a very promising paradigm for fNIRS based BCI, as classification accuracies compare very favorably to those achieved in motor imagery BCIs with fNIRS.

I. INTRODUCTION

A. Motivation

Speech has long been an established paradigm for human-computer interaction as it is intuitive and very efficient. However, speech has not been applied as a paradigm to functional Near Infrared Spectroscopy (fNIRS) based Brain Computer Interfaces (BCIs) to this point. We therefore investigate the feasibility of speech in different modes as a paradigm for BCIs, since it allows for intuitive passive and active BCI control.

fNIRS enables the robust measurement of brain activity and is less affected by movement artifacts than other modalities for BCIs such as electroencephalography (EEG). With the growing number of fNIRS research, advances towards even higher mobility can be expected. Furthermore, fNIRS and EEG are combinable to profit from the advantages of both modalities. In contrast to functional magnetic resonance imaging (fMRI), which relies on the same effects as fNIRS (see Section I-B), fNIRS systems are inexpensive and portable, which makes them particularly suitable for BCIs in real-life scenarios.

Recently, the feasibility of fNIRS for BCI using motor imagery has been shown by Coyle [1]. Ang et al. [2] successfully used mental arithmetics to demonstrate BCI capabilities of fNIRS, by distinguishing between levels of difficulty with high accuracies. Several fMRI studies have shown different activation patterns in speech related brain areas (e.g. [3]). Even though fNIRS has been used in a number of clinical studies investigating speech (e.g. [4]), there are only very limited studies using speech related tasks in combination with fNIRS for BCI control. Naito et al. [5] used a single-channel fNIRS system to detect imagined singing.

For speech to be used as modality for computer interaction and to study speech activation patterns, we investigated the discrimination of three different speaking modes in this paper: Normal audible speech (AUD_{Speech}), silently uttered speech, for which our subjects moved their articulatory muscles as if speaking, without producing actual sounds (SIL_{Speech}) and speech imagery, where the subjects conceived themselves speaking but only imagined the movement of their articulatory muscles (IMG_{Speech}). We expected to see different brain activation patterns between AUD_{Speech} and SIL_{Speech} since the latter lacks auditory feedback, and between SIL_{Speech} and IMG_{Speech}, since the latter involves no articulation execution but planning, memory and speech specific activations.

B. Functional Near Infrared Spectroscopy

fNIRS is a brain imaging technique based on the concentration changes of oxy-hemoglobin (*HbO*) and deoxy-hemoglobin (*HbR*) caused by neural activity in the brain's cortical areas. These hemodynamic responses can be recorded using light-sources and detector-optodes, which are placed on the subject's head. Sources emit at least two wavelengths of light in the near infrared range of the electromagnetic spectrum (620 nm - 1000 nm). The properties of the biological tissue allow the infrared light to disperse through the scalp, skull and cortical areas of the brain and exit again along the photon path [6]. At the end of this path, whose depth is determined by the source-detector distance, a detector measures the light intensities transmitted through the head. As *HbO* and *HbR* have different light absorption characteristics, the modified Beer-Lambert law [7] can be applied to transfer optical densities changes (ΔOD) into *HbO* and *HbR* differences, denoted as ΔHbO and ΔHbR, respectively. Given the source-detector distance l, path length b and the absorption coefficients for *HbO* and *HbR*, α_{HbO} and α_{HbR}, the concentration changes ΔHbO and ΔHbR can

Part of this work was performed during the invited visit of the first author at A*STAR, Singapore, for which we are very thankful. This project received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

[1]Christian Herff, Felix Putze, Dominic Heger and Tanja Schultz are with the Cognitive Systems Lab, Karlsruhe Institute of Technology, Adenauerring 4, 76131 Karlsruhe, Germany. christian.herff@kit.edu

[2]Cuntai Guan is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632.

be calculated from ΔOD using the following equations:

$$\Delta HbO = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbO}} \qquad \Delta HbR = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbR}} \quad (1)$$

Typically, a hemodynamic response to cortical activity rises on stimulus onset for *HbO* and decreases for *HbR*. Levels are expected to return to baseline after the end of the stimulus. Figure 1 shows a hemodynamic response, which was obtained by averaging over all SIL_{Speech} trials from subject 2 for a location on the lower motor cortex. It reflects well the expected typical hemodynamic response and value range.

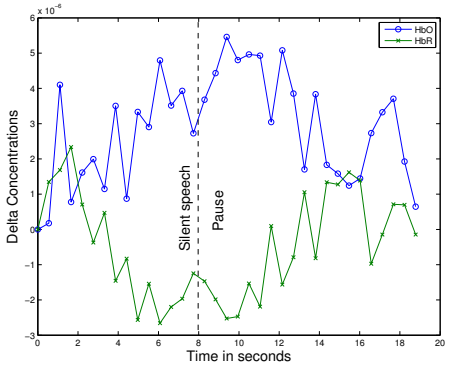


Fig. 1. Average hemodynamic response of subject 2 when speaking silently followed by pausing, for a location in the lower motor cortex.

C. Relevant Brain Areas

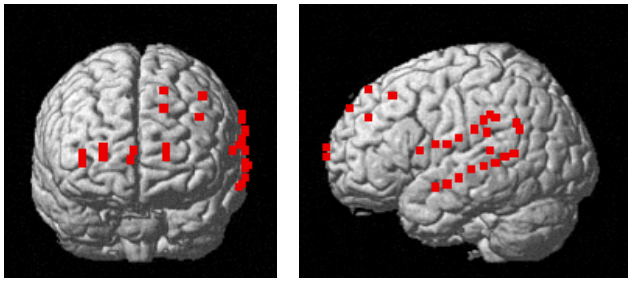
For approximately 90% of the population, the left hemisphere is dominant for speech and language processing. This lateralization is even larger for right-handed individuals (see [8]). To increase the probability of measuring relevant areas, we decided to focus on right-handed subjects in this pilot study. The prefrontal cortex is implicated in executive functions such as decision making, expectation management and the working memory, while the Broca's and Wernicke's areas are relevant for speech perception and production, and the lower motor cortex is identified with muscle control for the tongue and facial areas.

Thus, we recorded fNIRS signals from Broca's (4 optodes) and Wernicke's (10 optodes) areas, the prefontal (12 optodes) and lower motor cortex (6 optodes) to cover all relevant areas. We used an ANT Visor infrared camera system to register the positioning of the 32 optodes and plotted them onto the brain surface using the NIRS-SPM software [9]. See Figure 2 for the exact optode positions in our experiment.

II. EXPERIMENTS

A. Experimental Setup

We used a Dynot232 system designed by NIRX Medical Technologies with 32 optodes used both as sources and detectors, sampling at 1.81 Hz. The system outputs values for every source-detector pair of which we selected only pairs with distances between 2.5 and 4.5 cm. This way, we obtained 252 channels of raw optical densities. Wavelengths



(a) b)

Fig. 2. (a) Optode positions frontal view. (b) Optode positions left lateral view. Created with [9].

of 760 and 830 nm were used.

The subjects were placed 50 cm away from a computer screen with 48 cm screen size and had the NIRS-optodes fixed to their heads using a helmet to firmly keep the optodes at the desired positions. Five male students with a mean age of 27.6 years participated in our study. All of them were right-handed with a mean Edinburgh handedness score [10] of 86.

The experiment consisted of 10 sentences, with nearly equal lengths (roughly 66 characters) taken from the broadcast-news domain.

The subjects were asked to produce utterances in the three modes AUD, SIL, IMG, followed by pauses. The utterances were prompted from displaying sentences on the screen. Every utterance of a sentence is denoted as a trial. The pauses following the utterances are denoted as separate trials. The trials are named according to their respective mode names, i.e. AUD_{Speech}, SIL_{Speech}, IMG_{Speech} and AUD_{Pause}, SIL_{Pause}, IMG_{Pause}. Every sentence was repeated three times in each mode by every subject. Sentences were presented in blocks of 6, which had to be produced in the same mode. Mode order and sentence order were randomized to eliminate sequence effects. Each block had 4 steps. It started with (1) the instruction in which mode the following sentences had to be produced, i.e. either Audible, Silent or Imagine. (2) A beep indicated that a sentence was about to be displayed in 2 seconds. (3) The sentence was then displayed for a duration of 8 seconds in which the subject had to either read it out audibly, silently or imagine reading it out. The AUD_{Speech}, SIL_{Speech}, IMG_{Speech} trials were recorded in these periods. (4) Afterwards, a fixation cross was shown for 10 seconds. The respective $Pause$ trials were recorded in these intervals. These four steps were repeated 6 times to form a block. In between blocks, the subjects had 25 seconds to relax.

Table I summarizes the complete corpus characteristics.

B. Signal Preprocessing

The 252 channels of raw optical densities were sampled at 1.81 Hz, which is low enough to not require low-pass filtering. We used the HomER package to transfer raw optical densities to the ΔHbO and ΔHbR values.

Afterwards, we detrended the resulting 252 channels of

TABLE I
CORPUS CHARACTERISTICS

Subject-ID	1	2	3	4	5
AUD$_{Speech}$ trials	13	30	30	30	24
AUD$_{Pause}$ trials	13	30	30	30	24
SIL$_{Speech}$ trials	18	30	30	30	18
SIL$_{Pause}$ trials	18	30	30	30	18
IMG$_{Speech}$ trials	18	30	30	30	18
IMG$_{Pause}$ trials	18	30	30	30	18
Total recording time (minutes)	20.6	37.5	37.5	37.5	25.2

Δ*HbO* and Δ*HbR*. Trials were then extracted based on the experiment timing. A class label corresponding to the $Speech$ or $Pause$ mode was assigned to each trial.

C. Feature Extraction

Feature extraction assumes an idealized hemodynamic response, i.e. a rise in *HbO* and a decrease in *HbR* during speech activity trials and a return to baseline-levels for the subsequent $Pause$ trials (see Figure 1). Based on the idea for feature extraction by Leamy et al. [11], we take the mean μ of the first 7 samples of every trial (corresponding to roughly 4 seconds) and subtract the mean of samples 9 to 15 of the Δ*HbO* and Δ*HbR* signals in every channel i for each trial t.

$$f_{i,t}^{\Delta HbO} = \mu(\Delta HbO_{t,1:7}^{i}) - \mu(\Delta HbO_{t,9:15}^{i}) \quad (2)$$

$$f_{i,t}^{\Delta HbR} = \mu(\Delta HbR_{t,1:7}^{i}) - \mu(\Delta HbR_{t,9:15}^{i}) \quad (3)$$

In total, we extract 504 features per trial. After extraction, features of every channel were standardized to zero mean and unit standard deviation (z-normalization).

D. Feature Selection

We used a *Mutual Information based Best Individual Feature (MIBIF)* approach as presented by Ang et al. [12] to select the top $k = 30$ features out of the 504-dimensional feature space on the training data. The Mutual Information $I(X;Y)$ between two random variables X and Y, measures the amount of information the two variables share. Therefore, a high Mutual Information between features and the class labels should indicate features which contain highly relevant information. This would potentially lead to high classification accuracy assuming that the training data are representative of the test data. The Mutual Information $I(C;F)$ between class labels C and features F is defined as

$$I(C;F) = H(C) - H(C|F) \quad (4)$$

with $H(C)$ and $H(C|F)$ referring to the entropy and the conditional entropy, respectively. Using Bayes theorem and given the equal priors, the conditional probability $p(c|f)$ and the joint probability $p(c,f)$, which are needed to determine the entropies, can be calculated through $p(f|c)$. Ang et al. [12] describe a method to estimate the probability density function $p(f|c)$ from the training data. To estimate the conditional probability, kernel density estimation using Parzen windows is applied:

$$\hat{p}(f|c) = \frac{1}{n_c} \sum_{j \in I_c} \phi\left(f_j, h\right), \quad (5)$$

where n_c is the number of samples in class c, I_c is the set of sample indices in class c and ϕ being a smoothing kernel with parameter h. A univariate Gaussian kernel was employed for smoothing:

$$\phi(x,h) = \frac{1}{2\pi} e^{-\left(\frac{x^2}{2h^2}\right)} \quad (6)$$

MIBIF then selects the k features f_l with highest Mutual Information with the class labels $\arg\max_l(I(C,f_l))$. We set $k = 30$ after studying the distributions of Mutual Information of features with the class labels.

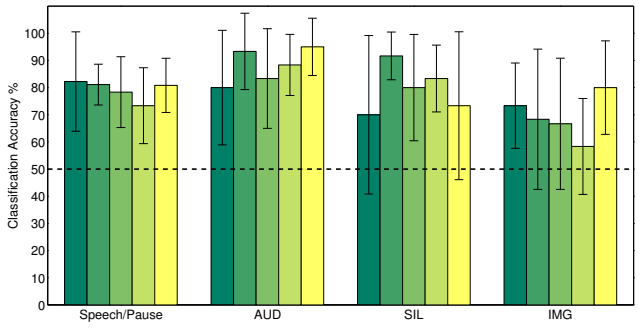
The *MIBIF* approach presents a fast feature selection technique that uses a high relevance criterion to reduce the dimensionality of the feature space. It is orders of magnitude faster than more complex *Mutual Information* based approaches as for example the *Mutual Information based features selection (MIFS)* by Battiti [13] and still yields comparable or even better results for BCI data (compare [12]).

E. Classification

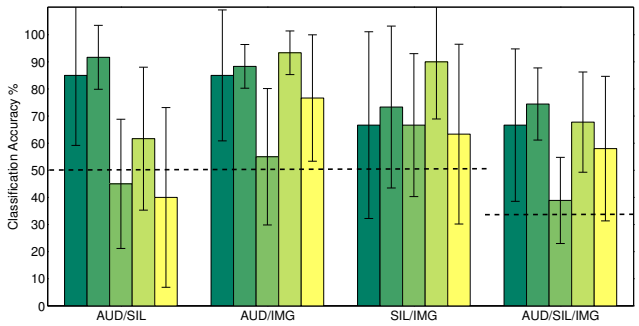
To evaluate our system, we applied a 10-fold person dependent cross-validation approach. For classification, we employ support vector machines with radial basis function kernels on the resulting 30-dimensional feature set S determined with *MIBIF*. SVM parameters c and γ are estimated via cross-validation on the training data using a grid search. We tested the three $Speech$ modes combined against the three $Pause$ modes combined to discriminate general speech activity from inactivity. Then, we classified every mode against its respective $Pause$ trials in a binary classification setup. Additionally, the three speaking modes AUD$_{Speech}$, SIL$_{Speech}$ and IMG$_{Speech}$ were discriminated from each other in binary and three-class experiments.

III. RESULTS

All classification results are presented in Figure 3. Part (a) of Figure 3 shows classification results of the modes against their respective $Pause$. Classifying combined $Speech$ (build from AUD$_{Speech}$, SIL$_{Speech}$ and IMG$_{Speech}$) from the combined $Pause$ worked very reliably for all subjects with an average accuracy of 79%. Next, we tested each of the three modes individually against their respective $Pause$. As expected, distinguishing between AUD$_{Speech}$ and AUD$_{Pause}$ worked best (88%) as most neuronal activity should be observed due to the acoustic feedback. Results for SIL$_{Speech}$ and SIL$_{Pause}$ are slightly lower, which might be explained by the fact that no acoustic signal has to be processed in the brain and thus the neural activity level of SIL$_{Speech}$ might be closer to the one in SIL$_{Pause}$. Yet, classification performance is still very high with 80% average accuracy. IMG$_{Speech}$ versus IMG$_{Pause}$ yielded lowest results (69%)



(a)



b)

Fig. 3. Classification results of all subjects for binary and three-class problems. (a) Binary classification experiments $Speech$ against $Pause$ in all modes. (b) Classification accuracies between $Speech$ of different speaking modes. Each color represents one subject. Whiskers indicate standard deviations. Dotted line stands for naive classification accuracies.

as execution of the actions is entirely missing in the brain activity.

Classification accuracies between the different speaking modes are illustrated in part (b) of Figure 3. IMG_{Speech} could be discriminated from AUD_{Speech} and SIL_{Speech} reliably with 80% and 72% on average. Differentiating between AUD_{Speech} and SIL_{Speech} yielded the lowest results with 65% accuracy on average and produced the only two results lower than naive classification. The three classes could be distinguished well with an average accuracy of 61% compared to a naive classification accuracy of 33%.

The fact that these high accuracies, which are at least as good as comparable experiments with motor imagery, were achieved with less than 9 minutes of training data in the binary experiments indicates the large potential of speech as a paradigm for fNIRS based BCI. The low inter-subject variances further support this fact.

Table II summarizes our findings by showing average results and standard deviations across all five subjects. All captured fNIRS signals strongly resemble expected hemodynamic responses (compare Figure 1). We obtained high accuracies for AUD_{Speech} versus SIL_{Speech} and IMG_{Speech} versus IMG_{Pause} and since our experimental design controls for artifacts, these results are indeed achieved based on brain activity patterns.

TABLE II

AVERAGE CLASSIFICATION RESULTS AND STANDARD DEVIATIONS ACROSS SUBJECTS IN %.

	$Speech/Pause$	AUD	SIL	IMG
Acc.	79	88	80	69
Std.	3.6	6.3	8.5	8.0

	AUD/SIL	AUD/IMG	SIL/IMG	AUD/SIL/IMG
Acc.	65	80	72	61
Std.	23.1	15.0	10.7	13.8

IV. SUMMARY

We have shown that the fNIRS signals captured while performing a speech related task has large potential to be used for BCI control with very high accuracies. This is a novel direction for NIRS-based BCIs which mainly relied on motor imagery to this point. Our results are highly significant and compare favorably to those achieved with motor imagery, while being natural, intuitive and do not require any prior learning. Moreover, our experimental setup allows for further investigations of brain activation patterns for speech related tasks.

REFERENCES

[1] SM Coyle, TE Ward, and CM Markham, "Brain-computer interface using a simplified functional near-infrared spectroscopy system," *Journal of Neural Engineering*, vol. 4, no. 3, pp. 219, 2007.

[2] KK Ang, C Guan, K Lee, JQ Lee, S Nioka, and B Chance, "A Brain-Computer Interface for Mental Arithmetic Task from Single-Trial Near-Infrared Spectroscopy Brain Signals," *Int. Conference on Pattern Recognition*, pp. 3764–3767, 2010.

[3] JR Binder, SJ Swanson, TA Hammeke, and DS Sabsevitz, "A comparison of five fMRI protocols for mapping speech comprehension systems," *Epilepsia*, vol. 49, pp. 1980–97, Dec. 2008.

[4] H Sato, T Takeuchi, and K L Sakai, "Temporal cortex activation during speech recognition: an optical topography study," *Cognition*, vol. 73, no. 3, pp. B55–66, Dec. 1999.

[5] M Naito, Y Michioka, K Ozawa, Y Ito, M Kiguchi, and T Kanazawa, "A communication means for totally locked-in als patients based on changes in cerebral blood volume measured with near-infrared light," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 7, pp. 1028–1037, July 2007.

[6] E Okada, M Firbank, M Schweiger, SR Arridge, M Cope, and DT Delpy, "Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head," *Appl. Opt.*, vol. 36, no. 1, pp. 21–31, Jan 1997.

[7] A Sassaroli and S Fantini, "Comment on the modified beerlambert law for scattering media," *Physics in Medicine and Biology*, vol. 49, no. 14, pp. N255, 2004.

[8] S Knecht, B Dräger, M Deppe, L Bobe, H Lohmann, A Flöel, E.-B. Ringelstein, and H Henningsen, "Handedness and hemispheric language dominance in healthy humans," *Brain*, vol. 123, no. 12, pp. 2512–2518, 2000.

[9] JC Ye, S Tak, KE Jang, J Jung, and J Jang, "Nirs-spm: Statistical parametric mapping for near-infrared spectroscopy," *NeuroImage*, vol. 44, pp. 428 – 447, 2009.

[10] RC Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia*, vol. 9, pp. 97–113, 1971.

[11] DJ Leamy, R Collins, and T Ward, "Combining fNIRS and EEG to Improve Motor Cortex Activity Classification during an Imagined Movement-Based Task," in *HCI (20)*, 2011, pp. 177–185.

[12] KK Ang, Z Yang Chin, H Zhang, and C Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *Neural Networks, 2008. IJCNN 2008.*, June 2008, pp. 2390 –2397.

[13] R Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 5, no. 4, pp. 537–50, Jan. 1994.