Speech Processes for Brain-Computer Interfaces

Christian Emanuel Herff

Kumulative Dissertation
zur Erlangung des Grades eines
Doktors der Ingenieurwissenschaften – Dr.-Ing. –

Vorgelegt im Fachbereich 3 (Mathematik und Informatik)
Universität Bremen

31. Oktober 2016

Datum des Promotionskolloquiums: 12. Dezember 2016

Gutachter

Prof. Dr. Tanja Schultz (Universität Bremen, Germany)
Prof. Dean Krusienski, Ph.D. (Old Dominion University, VA, USA)

Cover illustration courtesy of Britt Winkelmann, 2016

Abstract

Speech interfaces have become widely used and are integrated in many applications and devices. However, speech interfaces require the user to produce intelligible speech, which might be hindered by loud environments, concern to bother bystanders or the general inability to produce speech due to disabilities. Decoding a user's imagined speech instead of actual speech would solve this problem. Such a Brain-Computer Interface (BCI) based on imagined speech would enable fast and natural communication without the need to actually speak out loud. These interfaces could provide a voice to otherwise mute people.

This dissertation investigates BCIs based on speech processes using functional Near Infrared Spectroscopy (fNIRS) and Electrocorticography (ECoG), two brain activity imaging modalities on opposing ends of an invasiveness scale. Brain activity data have low signal-to-noise ratio and complex spatio-temporal and spectral coherence. To analyze these data, techniques from the areas of machine learning, neuroscience and Automatic Speech Recognition are combined in this dissertation to facilitate robust classification of detailed speech processes while simultaneously illustrating the underlying neural processes.

fNIRS is an imaging modality based on cerebral blood flow. It only requires affordable hardware and can be set up within minutes in a day-to-day environment. Therefore, it is ideally suited for convenient user interfaces. However, the hemodynamic processes measured by fNIRS are slow in nature and the technology therefore offers poor temporal resolution. We investigate speech in fNIRS and demonstrate classification of speech processes for BCIs based on fNIRS.

ECoG provides ideal signal properties by invasively measuring electrical potentials artifact-free directly on the brain surface. High spatial resolution and temporal resolution down to millisecond sampling provide localized information with accurate enough timing to capture the fast process underlying speech production. This dissertation presents the *Brain-to-Text* system, which harnesses automatic speech recognition technology to decode a textual representation of continuous speech from ECoG. This could allow to compose messages or to issue commands through a BCI. While the decoding of a textual representation is unparalleled for device control and typing, direct communication is even more natural if the full expressive power of speech - including emphasis and prosody - could be provided. For this purpose, a second system is presented, which directly synthesizes neural signals into audible speech, which could enable conversation with friends and family through a BCI. Up to now, both systems, the *Brain-to-Text* and synthesis system are operating on audibly produced speech. To bridge the gap to the final frontier of neural prostheses based on imagined speech processes, we investigate the differences between audibly produced and imagined speech and present first results towards BCI from imagined speech processes.

This dissertation demonstrates the usage of speech processes as a paradigm for BCI for the first time. Speech processes offer a fast and natural interaction paradigm which will help patients and healthy users alike to communicate with computers and with friends and family efficiently through BCIs.

Zusammenfassung

Sprachschnittstellen werden mittlerweile häufig benutzt und sind in zahlreiche Programme und Geräte integriert. Allerdings muss ein Nutzer einer Sprachschnittstelle verständliche Sprache produzieren, was durch Umgebungslärm, die Sorge, Umstehende zu stören oder durch ein generelles Unvermögen, Sprache zu produzieren, zum Beispiel durch eine Behinderung, unmöglich sein kann. Das Erkennen vorgestellter Sprache statt tatsächlicher Sprache eines Nutzers würde dieses Problem lösen. Ein solches Brain-Computer Interface (BCI), basierend auf vorgestellter Sprache, würde natürliche und schnelle Kommunikation erlauben ohne tatsächlich sprechen zu müssen. Diese Technologie könnte daher Menschen mit Spracheinschränkungen die Sprachfähigkeit zurückgeben.

Diese Dissertation realisiert BCIs basierend auf Sprachprozessen und untersucht dafür zwei Modalitäten zum Messen von Gehirnaktivität mit gegensätzlicher Benutzerfreundlichkeit. Um die Gehirnaktivitätsdaten mit sehr niedrigem Signal-Rausch-Abstand und komplexen zeitlichen, räumlichen und spektralen Zusammenhängen auswerten zu können, werden Techniken aus dem maschinellen Lernen, den Neurowissenschaften und automatischer Spracherkennung zusammengefügt. Durch diese Kombination lassen sich bestmögliche Ergebnisse erzielen und gleichzeitig neue Einsichten in die zugrunde liegenden neuronalen Prozesse erlangen.

Funktionale Nahinfrarotspektroskopie (fNIRS) misst die Durchblutung in zerebralen Gehirnarealen und ist schon mit einfachen Mitteln zu realisieren. Gute Aufnahmequalität kann schon in wenigen Minuten in alltäglichen Situationen erreicht werden. fNIRS bietet daher gute Eigenschaften für praktische Nutzerschnittstellen. Allerdings sind die hemodynamischen Prozesse, die fNIRS zugrunde liegen, langsam und die zeitliche Auflösung damit beschränkt. Diese Dissertation untersucht Sprache mit Hilfe von fNIRS und zeigt erfolgreiche Klassifikation von Sprachprozessen durch fNIRS-Daten mit denen BCIs entwickelt werden können.

Elektrokortikographie (ECoG) bietet ideale Signaleigenschaften, da elektrische Potenziale artefaktfrei direkt auf der Gehirnoberfläche aufgezeichnet werden. Die hohe räumliche und zeitliche Auflösung bietet die erforderliche Lokalisierung mit genauen zeitlichen Informationen, die erforderlich sind, um die schnellen Prozesse, die der Sprachproduktion zugrunde liegen, zu erforschen. Diese Dissertation präsentiert das *Brain-to-Text* System, das automatische Spracherkennungstechnologie verwendet um kontinuierlich gesprochene Sprache aus ECoG in eine textuelle Repräsentation zu dekodieren. Dies ermöglicht das Verfassen von Nachrichten oder die Eingabe von Befehlen durch das BCI. Während dieses Verfahren ideal für die Kontrolle von Geräten und die Eingabe von Texten ist, würde direkte Kommunikation für Menschen mit Locked-In-Syndrom noch natürlicher, wenn die gesamte Ausdrucksstärke von Sprache - inklusive Betonung und Prosodie - generiert werden könnte. Wir präsentieren ein System, das dies erreicht, indem es Sprache direkt aus neuronalen Signalen synthetisiert und damit Konversationen mit Freunden und Familie durch ein BCI ermöglicht. Bisher arbeiten sowohl das *Brain-to-Text*, als auch die Sprachsynthese aus neuronalen Daten, auf laut produzierter Sprache. Um die letzte Hürde für Neuroprothesen basierend auf vorgestellter Sprache zu meistern, untersucht diese Dissertation die Unterschiede zwischen laut artikulierter und vorgestellter Sprache und präsentiert erste Ergebnisse für BCIs basierend auf vorgestellter Sprache.

Diese Dissertation demonstriert, dass Sprachprozesse als Paradigma für BCIs genutzt werden können. Sprachprozesse bieten ein schnelles und natürliches Paradigma für BCIs, welches Patienten und gesunden Nutzern helfen wird, mit Computern, Freunden oder Familie effizient zu kommunizieren.

Contents

Chapter 1

Introduction

To prepare Brain-Computer Interfaces (BCIs) for the mass market, many fundamental problems still need to be solved. Among these, the definition of a comfortable and intuitive communication paradigm remains unresolved. Speech, as the primary means of communication, enables fast and natural information transfer. Speech is therefore an ideal candidate for an input paradigm to next generation BCIs. The research summarized in this cumulative dissertation offers a contribution towards BCIs based on speech processes and answers important questions about the representation of speech in the brain.

This chapter provides a short overview about the state of the art in BCI and discusses different brain measurement techniques in terms of their usefulness to investigate speech processes as input paradigm for BCI. Finally, the structure and contributions of this dissertation and the corresponding publications are described.

1.1 Motivation

BCIs are systems that send messages or commands to a computer or electronic device without using the normal output pathways of peripheral nerves and muscles by decoding brain activity instead (compare [Wolpaw and Wolpaw, 2012]). A BCI opens up an additional input channel to keyboard and mouse, which healthy users can use to interact with a computer without the need for muscle movement. Additionally, BCIs have access to parameters even the user might not be aware of such as stress or drowsiness. Therefore, BCIs can augment and complement traditional input for healthy users. Recently, the opposite direction of stimulating the brain to transmit information into the brain has gained interest [Flesher et al., 2016, Pais-Vieira et al., 2013, Rao et al., 2014]. Despite this innovative and promising idea, BCIs are currently only used by a small number of patients [Vaughan et al., 2006] suffering from neurodegenerative diseases, such as Amyotrophic Lateral Sclerosis (ALS), brainstem stroke, or spinal cord injuries which can result in a locked-in state without voluntary muscle control [Plum and Posner, 1982]. For such patients, a BCI might be the only opportunity to communicate with caregivers and family. The low number of BCI users can be explained by the very slow input rates, with 5.32 bits per second being the fastest known BCIs [Chen et al., 2015b], compared to about 29 bits per second for an average typist on a keyboard [Brown, 1998]. In a survey with BCI users [Huggins et al., 2011], users wished for at least three times faster input speeds. Additionally, BCIs so far suffer from very unnatural input paradigms for which users have to focus on single letters that blink

occasionally or flicker in certain frequencies. To make BCI a competitive input channel for non-impaired users, the same restrictions have to be overcome.

Automatic speech recognition software has been used intensively by medical doctors and lawyers for over a decade. With Apple's Siri [AS], Google Voice Search [GVS] and Amazon's Alexa Voice Search [AAV], to name only a few, speech-driven applications have recently entered the daily life of millions of users. Speech can be used for natural interfaces with electronic devices and is a fast means of text input, with input speeds between 40 and 60 bits per second [Reed and Durlach, 1998].

Combining the advantages of speech and BCI would yield next generation interfaces, which are very natural and fast while also providing an supplementary input channel for healthy users and means of communication for paralyzed users. As a BCI interprets brain activity a user would not necessarily need to speak audibly, but could also imagine themselves to speak. This could potentially offer all advantages of speech interfaces but without the need to actually produce sound. Speech interfaces without the need to audibly produce speech can be realized with other physiological measures than brain activity, including measurement of facial muscles using electromyography (EMG) [Schultz and Wand, 2010, Wand et al., 2013, Zahner et al., 2014, Diener et al., 2015, Wand and Schmidhuber, 2016], or measurement of tongue movement using ultrasound [Hueber et al., 2007, Bocquelet et al., 2015] or electromagnetic articulography (EMA) [Fagan et al., 2008, Gonzalez et al., 2016]. However, all these physiological measures require articulator movement. See the review about Silent speech interfaces for further reading [Denby et al., 2010]. Only silent speech interfaces based on brain activity can realize speech based interfacing without the need for peripheral muscle movement. Therefore, only brain activity measurements can empower the full potential of silent speech interfaces by creating an additional input channel not requiring any movements for healthy users, which can also be used by locked-in patients. This dissertation aims at developing Brain-Computer Interfaces based on speech processes.

1.2 A Short State of the Art in BCI

Current BCIs use a few major input paradigms for the interaction between humans and computers. To enter text, a user can employ one of several speller devices which work well for most people and have achieved reliable input speeds up to 5.32 bits per second [Chen et al., 2015b]. Spellers often use the P300 response, which occurs roughly 300 ms after a novel or infrequent event. The P300 response to infrequent events has been studied intensively and was used for the first BCIs [Farwell and Donchin, 1988, Sutter, 1992, Donchin et al., 2000, Krusienski et al., 2008]. To use the P300 response for a speller device, letters and numbers are arranged in a grid and the user is instructed to focus on the character they want to select. Figure 1.1 illustrates the grid layout of a P300 speller interface realized with OpenVibe [Renard et al., 2010]. Rows and columns flash consecutively and the selected character can be identified by estimating the column and row pair which elicited the characteristic P300 response in the user's brain activity. As at least one column and one row flash is needed to select the correct character, this process is rather slow and concentration on individually flashing characters for long periods of time very tiresome.

Approximately 73% of the general population can operate a P300 speller error free and up to 89% reach accuracies over 80% [Guger et al., 2009]. As these interfaces have reached such

Figure 1.1: Screenshot of a P300 Speller realized using OpenVibe [Renard et al., 2010].

a high reliability, the first commercial systems, like the Intendix device [Int], are available. The P300 response can also be elicited and used for BCI using tactile [Brouwer and van Erp, 2010] or auditory [Hill and Schölkopf, 2012] stimuli.

Alternatively, speller interfaces can be realized using Steady State Visually Evoked Potentials (SSVEP). These potentials are responses to observed visual stimuli. For example, when the retina is excited by a flickering light of a specific frequency between 3.5 and 75 Hz, the same frequency can be found in electric potentials measured from the brain. As evoked potentials occur very robustly, especially in the visual cortex, a speller can be realized by having different characters flickering in different frequencies [Müller-Putz et al., 2005, Bin et al., 2009, Bin et al., 2011]. A user focuses their attention on a flickering character. The system detects the prominent frequency in the measured electrical brain activity and outputs the character associated with that frequency. Auditory Steady-State Responses (ASSR) are similar responses elicited by auditory stimuli and have also been used for BCI [Higashi et al., 2011].

As an alternative approach to speller interfaces, BCIs often use cursor or directional control to enable users to steer a cursor. Event-related de-synchronization in the μ and β frequency bands is observed in the motor cortex during actual and imagined movement [Pfurtscheller and Aranibar, 1979, Pfurtscheller et al., 1993, McFarland et al., 2000]. The μ and β rhythms can be used to issue directional commands by associating body movements or imagined body movements with directions [Fabiani et al., 2004], e. g. left hand for left cursor movement, right hand for right cursor movement, tongue for down and both feet for upwards movements [Schlögl et al., 2005]. BCIs based on motor imagery, as imagined movement is often called, are not yet as reliable as P300 spellers. In a recent study [Ortner et al., 2015] only 55% of users reached accuracies over 80% with only 5% reaching perfect classification. Motor imagery based BCIs have profited greatly from advances in ma-

chine learning techniques and show large increases in classification accuracies [Blankertz et al., 2008, Ang et al., 2008].

Another common paradigm for the control of BCI are slow cortical potentials (SCP), which can be voluntarily controlled after a training phase through operand conditioning. SCP are several seconds in length and have been used successfully to control one dimensional cursors for ALS patients [Birbaumer et al., 1999]. The slow nature of SCP is drastically limiting information transfer rates.

Despite the great improvements of BCI technology over the last decades, interaction strategies are still unnatural and information transfer rates very low. Using speech as an input for BCI would be much more natural than focusing on individual letters and would simultaneously be much faster. This dissertation investigates BCIs based on speech processes to enable faster and more natural communication with and through computers and electronic devices.

1.3 Speech Processes

This dissertation investigates speech processes in neural recordings. Human speech can be produced in very different styles of speaking, as speech is much more than the production of isolated words. A news anchor reading out their teleprompter speaks continuously, very controlled and with very few hesitations or disfluencies. In spontaneous speech, hesitations and disfluencies are much more common. Additionally, sentences do not necessarily follow proper grammatical rules. Even more disfluencies can be found in conversational speech between two or more speakers, who interrupt each other. To minimize disfluencies and to make sure that proper sentences are being uttered, we focus on continuously produced, prompted speech in this dissertation.

Natural speech is a continuous stream of words and does therefore not provide a trivial segmentation into words or single sounds. To simplify the investigation of speech processes, researchers often focus on isolated aspects of speech production, such as single vowels [Ikeda et al., 2014, Yoshimura et al., 2016], syllables [Bouchard and Chang, 2014] or words [Kellis et al., 2010] instead of continuous speech. While these investigations have provided important insights into isolated aspects of speech production, only continuous speech production contains the full complexity of natural speech. To take all aspects of natural speech into account, this dissertation investigates continuously spoken speech as opposed to isolated aspects of speech.

We look at different modes of speech production besides normally articulated speech. In silently articulated speech, the articulators are moved as if producing speech but without producing sound. Silently articulated speech requires the same facial muscle movements as normally articulated speech but no perception of one's own voice takes places. This speaking mode therefore allows us to investigate motor aspects of speech production in detail without contamination by auditory processing. As the final goal of BCIs based on speech, we investigate imagined speech, the act of speaking out loud is only imagined, i. e. neither sound nor articulatory movement is present. The advantages of BCIs based on speech can only be harnessed if imagined speech is employed as a paradigm. Therefore, a thorough investigation of imagined speech is crucial. The processing of imagined speech gives rise to a large amount of problems, as experimenters have no control on how the task is performed.

This way, it is very difficult to obtain a ground truth of imagined speech. Additionally, speed of imagined speech might vary greatly from normally articulated speech.

An unsolved problem when investigating speech processes in neural data is the concept of fundamental building blocks of speech in the brain. The units of sound in human speech are called phones (compare [IPA, 1999]). If two phones lead to a difference in meaning in identical phonetic environments (i. e. surrounding sounds), they belong to different phonemes. Two such words which are only different in one phoneme are called a minimal pair. If two phones in the same environment never lead to differences in meaning, they are called allophones. There is agreement among scientists that utterances are made up of words and words are build by concatenating phonemes, which are then realizes as phones, but the fundamental building blocks of speech in the brain are largely unknown. The brain could process speech production in phonological categories such as phones or phonemes, or in larger sequences such as syllables. Articulatory gestures (also called phonetic features) define the set of vocal tract configurations necessary to produce phones, which is another likely candidate for the representation of speech in the human brain. We are shedding more light on the neural representations of speech within this dissertation using interpretable machine learning approaches.

Besides the spoken words, speech also conveys information through pitch, stress, accent, prosody and many more. A textual representation of the spoken words cannot completely capture these aspects. This dissertation investigates speech synthesis as an approach that can convey more information than just a textual representation to create neuroprostheses that allow to converse naturally.

1.4 Measuring Brain Activity

A large number of techniques for the measurement of brain activity exist. Broadly, brain activity measures can be divided into two categories: Measurement techniques based on blood flow related information are called metabolic signals and have very different characteristics than techniques based on the measurement of electrical potentials arising from neural activity, called electrophysiological signals. We will briefly describe both types of signals with the respective measurement techniques in the following section and give a rational for our selection of measurement techniques investigated in this dissertation.

1.4.1 Metabolic Signals

Techniques based on metabolic processes measure blood flow related parameters in the brain. With increased neural activity, the neurons' demand for energy increases, which has to be accommodated by oxygen, delivered through the blood supply. Thus, the amount of fresh oxygenated blood in a certain brain region can be used as an indirect marker of neural activity in that regions. The dynamics of the blood flow are referred to as hemodynamic processes and the change of blood flow associated with a cognitive process are called hemodynamic responses. As blood vessels form a very intricate network in the brain, they can regulate the blood flow to very localized regions. Brain measurement techniques based on metabolic processes can therefore achieve very high spatial resolutions. On the flip side, hemodynamic responses take several seconds to fully transpire, resulting in poor temporal

resolution. Metabolic processes can be measured in different ways: Functional Magnetic Resonance Imaging (fMRI) makes use of the different magnetic properties of oxygenated and deoxygenated hemoglobin. These different magnetic properties can be detected outside the head by strong magnetic fields produced in the large tube of the MRI. The measured change in magnetic properties produced by the neural activity is called Blood Oxygen Level Dependent (BOLD) effect [Ogawa et al., 1990]. The fMRI measures the BOLD signal in volumetric pixels, called voxels, of several cubic centimeters. The signal can be measured from the entire brain and can therefore be used to investigate deeper brain structures, as well as the cortex. The high spatial resolution over the entire brain facilitates detailed investigation of cognitive processes. These advantages render fMRI the de facto standard in neuroimaging. However, the slow nature of hemodynamic processes in combination with the extremely costly device and the very restrained measurement positions make it less than ideal for the investigation of the fast processes underlying continuous speech production. For general insights about the cortical regions involved in speech production, speech perception and reading, fMRI can be instrumental, even if exact timing cannot be resolved. See the reviews [Price, 2012, Talavage et al., 2014] for more on this topic.

Another way to measure metabolic activity makes use of properties of near infrared light. Light in the near infrared part of the electromagnetic spectrum disperses through most types of biological tissue, such as skin, skull and hair, but is absorbed by hemoglobin. In functional Near Infrared Spectroscopy (fNIRS), this property is used to indirectly measure neural activity, by shining near infrared light into the skull and measuring the absorption along designated photon paths. Due to scattering, fNIRS can only measure outer cortex layers and can therefore not be used to investigate deep brain structures like fMRI. As light emitters and photon detectors can be head-mounted and do not require electrode gel, fNIRS is very user friendly. The user-friendliness of fNIRS makes it a good candidate for BCI [Coyle et al., 2007, Sitaram et al., 2007a].

1.4.2 Electrophysiological Signals

Electric potentials can be measured non-invasively on the scalp using surface electrodes or invasively via penetrating electrodes. Single action potentials from individual neurons can be measured by needle electrodes which are either placed individually into the cortex or as part of a microarray. Microarrays provide very localized information with the fastest measurable temporal resolution. To place needle electrodes or microarray grids on the cortex, a surgical incision is necessary. As these electrodes are not needed for clinical purposes, they are only very rarely implanted in humans and are most often used in primate studies. Action potentials measured with single electrodes in the speech-motor cortex have been used to decode isolated phones and to synthesize vowels [Guenther et al., 2009, Brumberg et al., 2010] in a completely paralyzed subject.

Electrodes on the brain's cortex or scalp surface cannot measure individual action potentials of single neurons, but measure ensembles of neurons firing in synchrony. Electroencephalography (EEG) uses electrodes placed on the scalp to measure the synchronized activity of these neural ensembles, which consist of several million neurons with similar spatial orientation. The placement on the scalp can result in strong contamination by motion artifacts, especially from head movements. As speech requires intense movement of facial muscles and tongue movement, electromyographic and glossokinetic artifacts are superim-

posed on the EEG signal when investigating speech. Since electric potentials are conducted through the brain, cerebrospinal fluid, skull and scalp, large volume conduction effects result in superimposed signals from many sources at each electrode. These effects make localization of brain activity in EEG very difficult. Especially the strong influence of artifacts by facial movement prohibit the investigation of articulated speech in EEG. However, EEG has been used to study perceived speech [O'Sullivan et al., 2015, Di Liberto et al., 2015], classify rhythms in imagined syllables [Deng et al., 2010], investigate temporal effects in imagined speech [Porbadnigk et al., 2009] and to discriminate between two imagined vowels [Yoshimura et al., 2016]. Even though EEG is not ideally suited to investigate speech, it is the de facto standard for BCI due to its high signal quality and easy setup.

Albeit not using electrodes for measurement, Magnetoencephalography (MEG) measures similar activity as the EEG. Synchronized activity of large groups of neurons is measured through changes in the brain's magnetic field using magnetometers placed around the head. The strong magnetic fields induced by these magnetometers require intensive shielding around the device, resulting in large chambers that have to be built specifically for the MEG. The skull filters higher frequency bands in EEG signals, but MEG can reliably measure higher frequency bands, which are very useful to localize activity. However, MEG is as effected by movement artifacts as EEG, and speech production is therefore difficult to study with MEG. Speech perception on the other hand has been studied intensively using MEG. Differences between the processing of phonetic and musical sounds have been highlighted using MEG [al Tervaniemi et al., 1999]. MEG has also been used to discriminate between two aurally presented words and achieved good classification results [Guimaraes et al., 2007].

Electrocorticography (ECoG) measures cortical electric potentials directly on the brain surface. The necessary craniotomy and implantation of electrode grids are routinely performed clinically for epilepsy or tumor treatment. Signals measured using these electrodes are unfiltered by scalp and skull and measure activity only of the neural ensembles below the electrode and do not suffer from the effects of volume conduction as in scalp EEG. Electrodes are usually within 1 cm or less from each other, providing high-density spatial sampling of neural activity. As the electrodes are implanted, they are not affected by movements of facial muscles. ECoG is therefore ideally suited for the investigation of speech production.

1.4.3 Properties of Measurement Techniques

Brain measurement techniques can be characterized with regards to different properties. The spatial resolution describes how well nearby locations can be discriminated. The temporal resolution describes the time scale of neural activity that can be separated by the measurement technique. Besides temporal and spatial resolution, the brain area that can be covered by the measurement technique plays an integral role on which processes can be investigated. Measurement techniques covering the entire head can investigate wide spread processes that involve different areas of the brain, while techniques covering only few neurons can investigate localized processes such as motor control in detail. A last factor to keep in mind when choosing measurement techniques is the invasiveness of the technique. Techniques that are minimally invasive only require electrode gel in the participants' hair, tight contact to optical sensors or dry electrodes and are therefore already usable outside the lab. MEG and MRI require specific chambers housing the expensive and large devices, therefore these techniques

are only usable in laboratories. Highly invasive techniques require a clinical environment as surgical incisions are necessary to implant the electrodes.

Figure 1.2 visualizes spatial and temporal resolution of different brain measurement techniques. The location of the corresponding circle for each modality shows temporal (abscissa) and spatial (ordinate) resolution (data used from [Wolpaw and Wolpaw, 2012]) on a logarithmic scale, with the highest spatial and temporal resolution in the lower left corner of the plot. The color of the circle illustrates the size of the area that can be covered with the measurement technique. Circle size highlights invasiveness of the procedure. Smaller circles refer to techniques that are minimally invasive and can already be used outside of the lab. Medium sized circles refer to techniques that require large equipment in laboratory settings. Measurement techniques marked with a large circles require clinical environments and surgical incisions. In the bottom left corner of Figure 1.2 microarrays are plotted show-

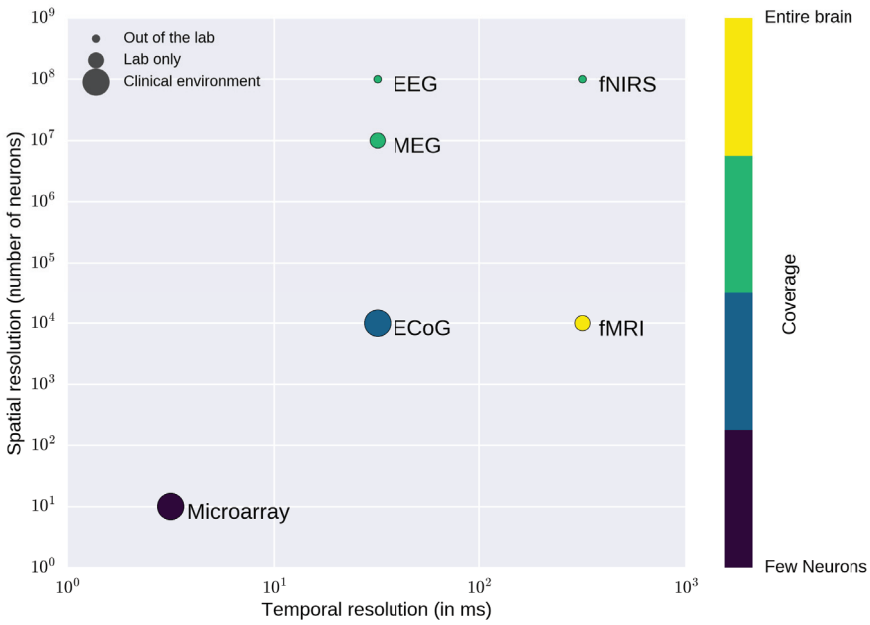


Figure 1.2: Spatial and temporal resolution of various brain imaging techniques (data from [Wolpaw and Wolpaw, 2012]). The circle color indicates area that can be covered. Circle size refers to invasiveness of measurement technique.

ing their extremely high spatial and temporal resolution. The implanted grids usually have only a few square millimeters of surface area and hence cover only a very small region of the cortex (shown by the purple shading of the circle). The clinical incision necessary to implant microarrays makes them only usable in clinical environment (indicated by the large circle). While ECoG does not penetrate the brain, the clinical incision necessary for implantation is still highly invasive. The flexible grids implanted for epilepsy and tumor treatment usually cover a much larger area of the cortex than microarrays. ECoG's temporal resolution is slightly more blurred than the single action potentials measured by microarrays. As ECoG measures the synchronized activity of thousands of neurons at the same time, the spatial resolution is significantly more coarse than that of microarrays. The spatial resolution of

fMRI can be as good as that of ECoG, when a strong enough magnetic field is created. The hemodynamic processes are inherently slow and take several seconds to complete, resulting in a much coarser temporal resolution. fMRI does not require a clinical incision and experiments can be carried out by trained experimenters in the laboratory without the need for a medical doctor. fMRI is therefore marked by a medium sized circle. fMRI can measure hemodynamic responses from the entire brain, including not only the cortex but also deeper brain structures such as thalamic and hippocampal regions (indicated by the yellow circle). MEG devices also require a large chamber and trained personal and experiments can only be conducted in these laboratory environments, as well. The electrical potentials measured by MEG are as fast as signals measured by ECoG, but have a coarser spatial resolution. MEG can measure signals from the entire cortex and is not restricted to a pre-selected area (indicated by the green circle). EEG measures the same electric potentials as ECoG and MEG and has the same temporal resolution but, due to filtering by skull and scalp and the corresponding superimposition of signals, a much coarser spatial resolution. Modern EEG devices are very easy to setup and can be used out of the lab easily [Debener et al., 2015]. The only cumbersome aspect of EEG montage is the electrode gel which is required to lower impedance between scalp and electrodes. The optical measurement technique fNIRS is as user-friendly as EEG and only requires firmly attached head-mounting for the optical sensors. With a spatial resolution comparable to EEG, it can yield interesting insights into many cortical regions, but cannot measure deeper brain structures. As fNIRS measures hemodynamic responses, temporal resolution is very coarse. Our review *"Automatic Speech Recognition from Neural Signals: A Focused Review"* [2] (See Appendix A.1) discusses the usability of different brain measurement techniques for automatic speech recognition from neural data.

In this dissertation, two measurement techniques on opposing sides of the invasiveness scale, namely fNIRS and ECoG, will be investigated. On the one hand, fNIRS can easily be used in practical real-world scenarios and can quickly be set up by the users themselves. ECoG on the other hand provides ideal signal characteristics but requires a surgical incision in clinical environments. Our motivation is to investigate the potential of a very user friendly technique compared to the technique with the best signal characteristics. Both techniques will be described in more detail in the respective chapters. Investigation and classification of speech processes using both fNIRS and ECoG will be presented and we will show how BCIs can be realized using both techniques.

1.5 Contributions of this Dissertation

This cumulative dissertation investigates speech processes using different brain measurement techniques. Each section summarizes the findings of one of the author's papers and embeds the results in the context of this dissertation. Contributions of the studies are highlighted and we explain to which extend the results advance towards the goal of speech in BCI.

Chapter 2 describes our contributions in the field of fNIRS, a measurement technique that is already usable out of the lab with very affordable hardware. We show how different speaking modes can be decoded using fNIRS and describe how the achieved results can be used for BCIs. Speech activity decoded with fNIRS could be used for a BCI that issues

binary commands like *"yes"* and *"no"* or as a wake command for a more powerful, but also more disruptive interface.

The limitations in temporal resolution suggest fNIRS to be used in non-time critical applications such as tutoring systems. We show results of several user states that can reliably be discriminated using fNIRS and describe how they can be utilized in interfaces. To improve classification accuracies, we investigate deep learning approaches for fNIRS data.

Chapter 3 highlights our work in modeling of continuous speech in invasively measured brain activity. We describe the first decoding system of continuous speech from neural signal, which we call *Brain-to-Text*. This system presents an important step towards BCIs based on continuous speech. Our approach could enable users to compose messages or control a computer using textual commands through a BCI as intuitive as natural speech. We also present an alternative approach to decoding a textual representation of speech, in which speech is directly synthesized into an audio waveform from ECoG data. This approach could enable locked-in patients to communicate with their friends and family by using the synthesized output of the BCI system, offering them the full expressive power of speech, including emphasis and prosody. Both approaches have advantages and have specific applications in which they excel.

The final frontier of speech based BCIs is to use imagined instead of normally articulated speech. In Chapter 4 we first summarize our work with fNIRS showing that imagined speech can have very different neural representations among different participants. We then discuss brain areas involved in the speech production process, which we identified in continuous speech using analysis from our *Brain-to-Text* approach. We argue which areas should show similar activation pattern in normally articulated and imagined speech and show that these areas can be used to decode speech from ECoG, to simulate decoding of imagined speech. This presents an important step towards BCI using imagined speech for communication and control.

The conclusion in Chapter 5 summarizes all contributions and provides an outlook on open issues for future research. Finally, all publications by the author relevant to the field of speech based BCI are cumulated in Appendix A. For each publication, the *List of Publications by the Author* (cf. pp. 59) contains a short description of the author's contribution. Publications cumulated in Appendix A are marked in bold.

Chapter 2

Speech in Non-Invasive BCI

Functional Near Infrared Spectroscopy is a relatively new brain imaging technique that has recently gained momentum in the BCI community. This chapter starts by briefly explaining the basic principles of functional Near Infrared Spectroscopy. It is then shown how information bearing features can be extracted from the measured signals to investigate and discriminate different types of speech processes. The classification of different speaking types can be used as a paradigm for BCI or to announce the general wish to operate a device. Shortcomings of this approach are highlighted and appropriate applications in the field of physiological computing are depicted. In an attempt to improve classification accuracies, we compare different classifiers - including state of the art deep learning models - on one of our datasets.

2.1 Basics of Functional Near Infrared Spectroscopy (fNIRS)

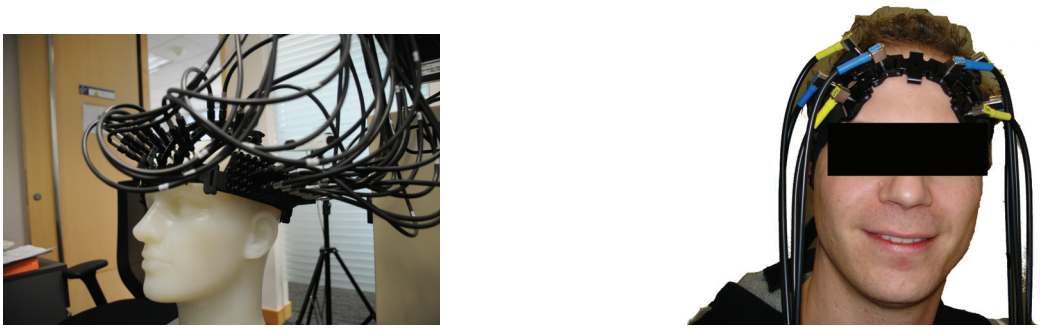
Functional Near Infrared Spectroscopy (fNIRS) is a relatively new brain imaging techniques pioneered by [Jobsis, 1977]. While light in the near infrared part of the light spectrum (620-700 nm) is not absorbed by most biological tissue, such as bones, skin and muscle, it is absorbed by hemoglobin, the oxygen carrying part of the blood. This property allows near-infrared light to disperse through the scalp, skull and brain tissue and reach outer layers of the cortex. Neural activity in specific brain areas yields an increased demand for energy which is supplied to the neurons through fresh oxygenated blood. The blood vessels in the brain form an intricate network which allows blood flow and thereby energy supply with hemoglobin to be controlled very localized to specific brain areas. The basic principle of near infrared spectroscopy is to shine light into the skull, and by measuring the amount of scattered light photons arriving at a detector, the absorption along the photon path between light emitter and detector can be estimated. The most commonly used technique for fNIRS utilizes continuous wave light emitter [Ferrari et al., 2004]. In this method, the light is shined at a constant amplitude and frequency, which is technically easiest to build. In fact, a continuous wave fNIRS device can be build for less than $500 following our description of the OpenNIRS system [3]. Continuous wave fNIRS devices require two different wavelength of near-infrared light, often one below and one above the isobestic point of hemoglobin at 808 nm. For example, the OpenNirs device uses wavelengths of 750 and 850 nm. While many commercial devices use laser diodes, simple LEDs have also been shown to work successfully [Ayaz et al., 2013, McKendrick et al., 2015], providing the

means for affordable or self-made fNIRS devices. Continuous wave NIRS can only measure the relative changes in light attenuation and can therefore not measure absolute hemoglobin levels. To estimate changes in hemoglobin levels, the modified Beer-Lambert law is applied [Sassaroli and Fantini, 2004]. The modified Beer-Lambert Law makes use of the different light absorption characteristics of oxygenated (HbO) and deoxygenated hemoglobin (HbR). Changes in hemoglobin levels in cortical areas of the brain can then be measured by placing light emitters and detectors on the scalp and measuring the amount of photons traveling along the photon path between light emitter and light detector. The relative changes ΔHbo and ΔHbR in oxygenated and deoxygenated hemoglobin, respectively, can be calculated from the changes in measured optical densities ΔOD at the light detectors. To achieve this, the absorption coefficients α_{HbO} and α_{HbR} of HbO and HbR are employed:

$$\Delta HbO = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbO}} \qquad \Delta HbR = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbR}} \tag{2.1}$$

where b is the length of the photon path between light emitter and detector and l is the distance between emitter and detector. For an emitter-detector pair with distance l, the measurement position is located in the middle between emitter and detector in a depth of approximately $l/2$ and is denoted as a channel.

Contrary to the practice in EEG of the well established 10-20 system for electrode montage, there is no standardized way to place light emitters and detectors on participants' heads. Optodes, as optical light emitters and detectors are often called in reference to electrodes, are usually placed in accordance to relevant areas measured for the proposed experiment, with distances and exact positioning designed by the experimenter. Figure 2.1 illustrates two different montages used in our experiments. Figure 2.1 (a) shows optode placement for our experiments investigating speech production which are further described in Section 2.3, while (b) shows a montage measuring signals from the prefrontal cortex which was used in the experiments described in Sections 2.4.1 and 2.4.2.



(a) (b)

Figure 2.1: Two examples of optode montages for fNIRS systems. (a) Shows the Dynot232 System with 32 optodes working as both emitters and detectors simultaneously. In this specific setup, 12 optodes cover the prefrontal cortex and 20 optodes are placed to cover inferior portions of the frontal cortex and superior temporal areas. (B) Shows the Artenis Oxymon Mk III with 4 emitters (yellow) and 4 detectors (blue), covering parts of the prefrontal cortex.

The changes of blood oxygenation in response to neural activity are called hemodynamic responses. These hemodynamic responses are slow in nature and take several seconds to fully develop. Figure 2.2 depicts an example for a typical hemodynamic response. After a stimulus, for example the onset of a mental task, oxygenated hemoglobin levels (solid purple line) slowly increase. This is due to the excessive supply of fresh oxygenated blood brought into the active regions. Similarly, deoxygenated hemoglobin values decrease (solid yellow line) in the same period, though the decrease is usually not as pronounced as the increase in oxygenated hemoglobin, as an excessive supply of oxygenated hemoglobin is provided and only the true demand for oxygen is represented in deoxygenated hemoglobin. After the end of the mental activity - indicated by the vertical line in the plot - both oxygenated and deoxygenated hemoglobin levels slowly return to baseline. This decay can take up to 15 seconds. In periods of no additional neural activity in the measured areas, hemoglobin levels stay very stable, which is shown by the dashed lines.

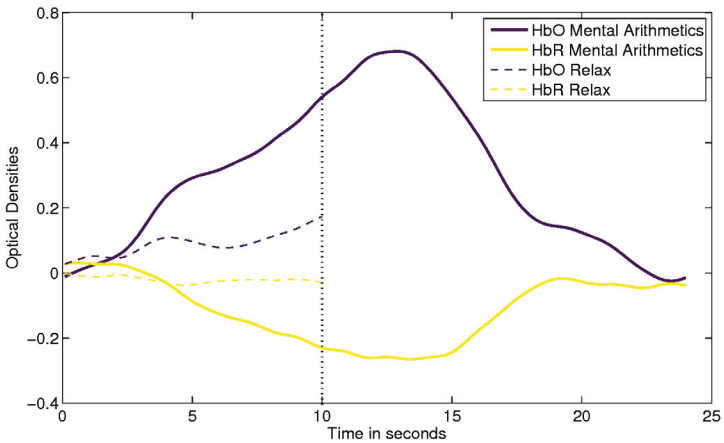


Figure 2.2: Average hemodynamic response of a participant performing mental arithmetics (solid lines) and relax tasks (dashed lines). Mental activity starts at second 0 and ends with the dotted line. Return to baseline is shown for the average mental arithmetics response after the dotted line from second 15 to 25. Figure modified from [18].

Hemodynamic responses occur very robustly and can be described by relatively few features. The increase or decrease of hemoglobin levels can be described by parameters such as the slope, the peak value or the latency. The robust nature of the hemodynamic responses and the possibility to easily target very specific measurement location make fNIRS a very good candidate for neuroscientific studies. However, the hemodynamic response is slow, which means that even with best possible feature extraction and classification algorithms, the delay between stimuli or mental activity and recognizable hemodynamic response will always be in the order of seconds.

2.2 Speech Processes Measured by fNIRS

Electrophysiological measures on the scalp severely suffer from motion artifacts produced by movements of facial muscle and the tongue, which makes it very difficult to investigate speech using EEG or MEG. The confined tube of the fMRI results in difficult recording

situations, especially for infants and elder populations. As fNIRS is less affected by motion artifacts and is very easily set up in non-laboratory environments, it is an ideal modality to investigate speech processes in non-clinical populations.

Speech has been studied intensively using fNIRS by the neuroscience community. Studies showed that fNIRS can be used to identify the dominant hemisphere for language function [Watanabe et al., 1998, Gallagher et al., 2007] and could become a non-invasive alternative to the WADA test [Wada and Rasmussen, 1960, Loring et al., 2012]. In the invasive WADA procedure, sodium amytal is injected intracarotid into a hemisphere, which induces temporary loss of function including speech aphasia if the dominant hemisphere was injected. A non-invasive alternative would clearly be of great benefit to the patients.

Many findings from behavioral studies and traditional neuroscience imaging techniques, such as fMRI, could be confirmed in fNIRS, such as robust activations in left temporal lobe for picture naming task [Hull et al., 2009]. [Bortfeld et al., 2009] showed that this cortical lateralization in the left temporal lobe could also be identified in children. A more fine-grained localization was presented in [Cannestra et al., 2003], localizing Broca's area using a covert object naming task and contrasting it with other tasks. [Wiggins et al., 2016] found that speech-evoked hemodynamic responses are very reliable. Children's cortical responses were investigated in [Bortfeld et al., 2007] using fNIRS. A study using speech segments played to infants showed that young children had different activation pattern for speech played forward or backward [Pena et al., 2003], illustrating that children can identify the typical sound of a language before being able to understand speech. Similarly, prosody processing was shown for infants [Homae et al., 2006] as different average fNIRS responses could be measured for normal and flattened speech. [Wartenburger et al., 2007] illustrated language comprehension by comparing hemodynamic responses to full sentences and suprasegmental information (prosody) by humming sentences.

fNIRS can also be instrumental to study different medical conditions. Dyslexia was investigated in children [Gan et al., 2003, Zhang et al., 2006] using fNIRS and different activation patterns were found between children with dyslexia and control subjects. Language aphasia after a stroke was investigated by comparing fNIRS activation pattern for poststroke-aphasic patients, poststroke-nonaphasic patients and healthy controls [Sakatani et al., 1998], finding the same pattern for poststroke-nonaphasic patients and controls and very different pattern for aphasia patients. Speech evoked hemodynamic responses in the auditory cortex were also studied in children newly equipped with cochlear implants using fNIRS [Sevy et al., 2010]. [Chen et al., 2015a] used fNIRS and EEG to study the functional reorganization between visual and auditory cortex in adult cochlear implant users. Auditory cortex and surrounding areas were investigated for patients with tinnitus in [Issa et al., 2016].

Optodes are far less affected by movement artifacts [Strangman et al., 2002] which makes fNIRS very well suited for the investigation of overt speech. [Fallgatter et al., 1998] investigated reading out loud using fNIRS. Robust activation in Broca's and Wernicke's areas during overt reading were found in [Sakatani et al., 1999, Horovitz and Gore, 2004, Lo et al., 2009]. Verbal fluency was evaluated using fNIRS and robust increases in frontal and temporal areas were found in various studies [Herrmann et al., 2003, Herrmann et al., 2005, Kakimoto et al., 2009]. For more reading on speech processes in fNIRS see the excellent reviews on speech [Dieler et al., 2012], speech investigation in infants and adults

[Quaresima et al., 2012] and the hemispheric differences during language development [Obrig et al., 2010].

Despite the large body of research on speech processes in fNIRS, speech processes had not previously been investigated in single trial classification. This is due to the mostly neuroscientific usage of fNIRS. An interface based on speech activity measured with fNIRS would need to react to each occurrence of speech activity, differences in average hemodynamic response are not sufficient to build interfaces. fNIRS is slowly gaining momentum in the BCI community. In the following section we will study whether the observed average hemodynamic responses to speech production can be robustly classified in single trial and will look into different types of speech production.

2.3 Decoding of Speech Processes in fNIRS

This section describes our study *"Speaking Mode Recognition from Functional Near Infrared Spectroscopy"* [22] (see Appendix A.2). The study will be shortly reviewed and the most important results recapped. We will highlight interesting aspects of the study and challenges overcome.

To investigate speech as a paradigm for BCIs using fNIRS, we recorded fNIRS activity during different types of speech production. This is meant as a first investigation to assess if neural responses to different types of speaking speaking can be classified in single trial. In our experiment, we asked participants to produce one of three speaking modes:

- **Normal audible speech (AUD)**: Participants were asked to read a displayed sentence out aloud.
- **Silently uttered speech (SIL)**: Participants were asked to silently mouth the sentences. This way, the articulatory muscles are moved as if producing speech, but without any sound production.
- **Imagined speech (IMG)**: Participants were asked to imagine to read and produce the sentences. Only imagined movement of articulatory muscles takes place.

These three speaking modes should yield different brain activity pattern, as they require the participation of different cortical regions. AUD requires movement planning, movement execution and processing of auditory stimuli (of the participants' own voice). In SIL no auditory processing is required, as no acoustic stimuli need to be processed, but motor planning and execution are present. IMG lacks both auditory processing and motor execution, but should require planning of articulator movement. All modes require a certain degree of memory processing as well as speech and language processing. The slow nature of the hemodynamic processes requires a resting period after each sentence, which we fixed at 10 seconds and denote as Pause. We asked our participants to read prompted sentences displayed on a screen for 8 seconds in one of the three modes (AUD, SIL, IMG) in randomized order. Our experiment consisted of 10 different sentences of almost equal length which had to be repeated three times for each of the modes.

To define an optimal optode layout for our speech experiment, we covered areas involved in speech perception and production, motor areas and areas associated with working memory. The left hemisphere is dominant for speech and language processing in over 90% of the

population and even more for right handed individuals [Knecht et al., 2000]. For a definite assessment of language lateralization, the invasive WADA test has to be employed. Even though this procedure is known to have very little side effects it can clearly not be employed for our experiment, as it would render the usability aspects of fNIRS void. To increase the likelihood of left hemisphere dominance for our participants, we only used strongly right-handed participants (mean handedness-score of 86) as assessed by the Edinburgh handedness inventory [Oldfield, 1971]. In total, we recorded data from 5 participants.

We recorded signals from the prefrontal cortex with 12 optodes. Language related areas in the inferior frontal gyrus (often called Broca's area) were covered by 4 optodes. Regions in the superior temporal gyrus, associated with Wernicke's area were measured using 10 optodes. Another 6 optodes covered lower parts of the left motor cortex, containing areas for muscle control of facial and tongue muscles. The resulting positioning of all 32 optodes covered relevant areas sufficiently well and was digitized using an ANT Visor infrared camera system.

The optodes of the utilized Dynot 232 system function as light emitters and detectors simultaneously. Each optode measures light intensities from all other emitters resulting in a total of $32 \times 31 = 992$ measured optical densities. We restricted this amount to those emitter-detector pairs with distances between 2.5 and 4.5 cm, as shorter distances result in light that has not traveled through the cortex and longer distances have to few photons arriving for a reliable measurement of cortical activity. This way, 252 channels of raw optical densities in two wavelength (760 and 830 nm) were obtained. Afterwards, the optical densities were transferred to changes in oxygenated and deoxygenated hemoglobin, ΔHbO and ΔHbR respectively using the HomER software package [Huppert et al., 2009] which automatically applies the modified Beer-Lambert Law. Each period of one of the speech modes or Pause is denoted as a trial. To extract meaningful information from the trials, we extracted a simple feature also used in previous work by [Leamy et al., 2011]. For this feature, the mean of the first half of the ΔHbO and ΔHbR values of a trial t is subtracted from the second half. This is performed for each channel i:

$$\begin{aligned} f_{i,t}^{\Delta HbO} &= \mu(\Delta HbO_{t,1:4}^{i}) - \mu(\Delta HbO_{t,4:8}^{i}) \\ f_{i,t}^{\Delta HbR} &= \mu(\Delta HbR_{t,1:4}^{i}) - \mu(\Delta HbR_{t,4:8}^{i}) \end{aligned} \tag{2.2}$$

A total of 504 features per trial is extracted this way. This high-dimensional feature vector would result in overfitting with most classifiers, due to the the small amount of trials. To prevent overfitting, we utilized the *Mutual Information based Best Individual Feature (MIBIF)* selection approach, presented by [Ang et al., 2008]. *MIBIF* relies on the Mutual Information $I(X;Y)$ between two random variables X and Y, which is a measure of the amount of information X and Y share. A feature sequence $F = [f_1, \ldots, f_n]$ that shares a lot of information with the corresponding class labels C should therefore be a good candidate for classification. The Mutual Information $I(C;F)$ for continuous features F and discrete class labels C can be calculated as:

$$I(C;F) = \int_f \sum_{c=1}^{N_c} p(c,f) \log\left(\frac{p(c,f)}{p(c)\,p(f)}\right) df \tag{2.3}$$

Since $p(c,f) = p(f|c) \cdot p(c)$, $I(C;F)$ can be estimated using the conditional probability $p(f|c)$, the discrete probability distributions of the classes $p(c)$ and the continuous distribution of

the features $p(f)$. While $p(c)$ and $p(f)$ can easily be calculated from the data, we apply kernel density estimation using Parzen windows to get $p(f|c)$:

$$\hat{p}(f|c) = \frac{1}{n_c} \sum_{j \in I_c} \phi\left(f_j, h\right), \tag{2.4}$$

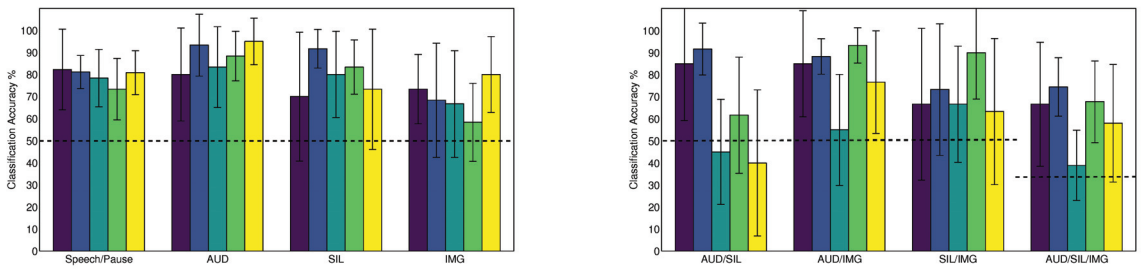
where n_c is the number of samples in class c, I_c is the set of sample indices in class c and ϕ being a smoothing kernel with parameter h. An univariate Gaussian kernel was employed for smoothing:

$$\phi(x, h) = \frac{1}{2\pi} e^{-\left(\frac{x^2}{2h^2}\right)} \tag{2.5}$$

Feature selection then works by choosing the k features f_l with the highest Mutual Information $\arg\max_l(I(C, f_l))$. The *MIBIF* approach is fast and yields comparable results to more computationally expensive procedures such as *Mutual Information based feature selection (MIFS)* [Battiti, 1994]. For this study, we selected the first $k = 30$ features, as the remaining features had only small Mutual Information with the class labels.

We evaluated the classification of speaking modes in fNIRS using a 10-fold cross-validation. Feature selection and classifier training were performed on 9/10 of the data and the remaining samples were used for testing. This procedure was repeated until all samples were used for testing once. A support vector machine (SVM) [Vapnik, 1998] with radial basis function kernel was used for classification. Parameters c and γ were optimized in a nested cross-validation. All classification models were trained participant-dependent.

To investigate whether speaking modes are viable candidates for Brain-Computer Interfacing, we first tested discriminating the different speaking modes from Pause. Classification results for the discrimination of different speaking modes from Pause can be found in Figure 2.3 (a). As a first experiment, we classified all three speaking modes combined (referred to



(a) (b)

Figure 2.3: Results for binary classification experiments of speaking modes against Pause (a) and among different speaking modes (b). Each color represents one subject. Dotted line indicates naive classification accuracy. Whiskers denote standard deviations. Figure modified from [22].

as Speech) against Pause, resulting in an average classification accuracy of 79% over all participants. This is a first indicator that speech processes can be discriminated reliably from segments without activity. Looking at the three speaking modes individually, the best

results were achieved when classifying AUD from PAUSE with an average of 88%. This is reasonable, as most cortical regions should be involved when speaking audibly. Classification results for this tasks were significantly better than chance level for all 5 participants (one sided t-tests, $p < 0.05$). Discrimination between silently voiced speech SIL and PAUSE yielded slightly lower results of 80% on average. We hypothesize that this is due to the lack of acoustic feedback and thus lower neural activity. Again, all 5 participants had results better than chance level. Finally, imagined speech (IMG) could be discriminated from PAUSE with an average accuracy of 69% and better than chance levels for 4 out of 5 participants. These results already highlight the fact that speech process can be classified from segments containing inactivity in single trial. Especially the results achieved with imagined speech make speech processes viable candidates for BCI.

As a next step we discriminated between the different speaking modes. The results of these investigations can be found in Figure 2.3 (b). Discrimination between normally produced speech (AUD) and silently voiced speech (SIL) worked better than chance level for only 2 out of 5 participants with an average accuracy of 65%. This could be explained by very similar neural activation pattern for the two speaking modes. Discrimination between AUD and IMG resulted in better than chance accuracies for all but one participant with an average of 80%. These two speaking modes have the least amount of shared cortical requirements, which might explain the high classification accuracies. Imagined speech (IMG) and silently uttered speech (SIL) were reliably discriminated for all participants with an average of 72%. The accuracy between the previous two results fits well with our hypothesis, as SIL and IMG differ in the amount of motor execution, but both lack auditory feedback. Classification between all three speaking modes resulted in a mean accuracy of 61%, which was significantly better than chance (33%) for all but one participant.

These results show that speech can be used for BCIs based on fNIRS by discriminating between different speaking modes. The robust classification of different types of speech classification could be used to issue directional commands by associating each speaking mode and *Pause* with one direction. Alternatively, fNIRS could be measured continuously and the detection of speaking activity could function as an idle switch to activate a more powerful, but also more disruptive system.

2.3.1 Continuous Decoding of Speaking Modes

This section summarizes our findings from *"Self-paced BCI with NIRS based on speech activity"* [27] (see Appendix A.3). Our findings are embedded in the topic of this dissertation and interesting aspects are highlighted.

For an intuitive BCI based on speech processes, users should decide themselves when they want to interact with the system. This requires a self-paced BCI which detects idle and voluntary control states [SCHERER et al., 2008]. While the identification of segments containing speech activity is a well studied field in speech technologies [LASKOWSKI and SCHULTZ, 2006], it has not been investigated in fNIRS before. Using the data recorded in our experiment described in 2.3, we implemented a system for continuous detection of speech activity. Rather than analyzing the data in a stimulus locked fashion, we dissected the data into overlapping 10 second long windows and extracted the features described in Equation 2.2. In a 10-fold cross-validation, we first selected the top 50 features using the *MIBIF* approach and trained SVMs to detect segments containing speech activation. We combined

all three speaking modes, to increase the amount of available training data, forming a general speech activity class. Conversely, PAUSE describes segments without speech activity. Figure 2.4 shows an example segmentation of our approach on a data excerpt.

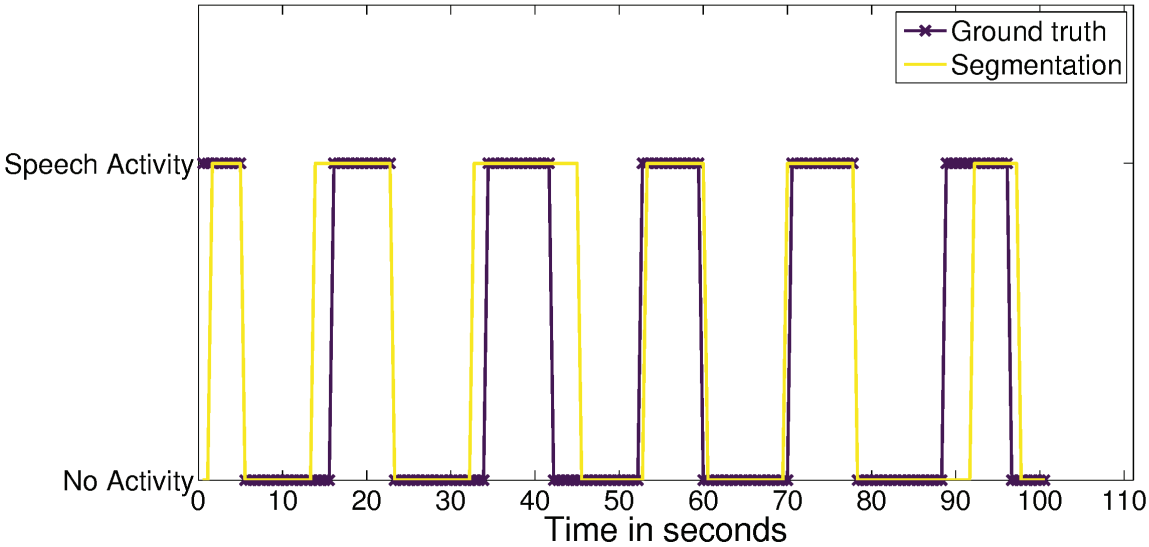


Figure 2.4: Segmentation of a participant's data into speech activity and non-speech. Figure modified from [27].

The short example in Figure 2.4 shows how accurately speech activity can be discriminated from segments without speech activity in a continuous fashion. A frame-based accuracy of 74% was reached when averaging over all participants, this compares very well to the 79% achieved in stimulus locked experiments. F-Scores are stable across all participants and significantly better than chance level (two-sided t-test, $p < 0.05$) with an average of 66%. This investigation highlights that speech processes can be discriminated from segments without speech activity without the need for stimulus locked evaluation. We regard this as an important step towards self-paced BCI using fNIRS.

Our results in discriminating speech activity from pauses and from each other show that speech activity is a viable candidate for BCI. Using imagined speech and pause could for example be used to issue two different directional commands or a simple 'Yes'/'No' interface. In the continuous decoding setup, speech activity could be used as switch to turn on a more complex BCI, probably based on EEG or combined EEG and fNIRS.

2.4 User State Estimation from fNIRS

The slow nature of hemodynamic responses make a more detailed classification of speech - down to the word or even phoneme level - impossible. As is, the transfer rates of one binary decision (speech or no speech activity) every 10 seconds - resulting in a information transfer rate of at most $\frac{1}{10} bit/s$ are far too slow for a realistic communication interface. Current EEG BCIs outperform these transfer rates by orders of magnitudes.

We thus conclude that fNIRS is better suited for user monitoring in situations when speed is not vital. Tutoring systems and user interfaces in general would benefit greatly if they could identify the cognitive or emotional condition of a user [Benyon and Murray, 1993]. Cognitive or emotional conditions are often called user states and can include a large variety of different conditions. We focus on the amount of workload a user is experiencing and the type of task a user is currently engaged in. We understand workload as the "perceived relationship between the amount of mental processing capability or resources and the amount required by the task" (compare [Hart and Staveland, 1988]). For an in-depth description of workload, see for example the cognitive load theory [van Merrienboer and Sweller, 2005, Kalyuga and Singh, 2015].

Interfaces, such as tutoring systems, could adapt their interaction strategy if too high mental workload was detected in the user [Putze, 2014]. This would result in more natural and efficient human-computer interaction. The idea to adapt user interfaces to the psychological state of a user through real time measurement of physiological signals is relatively new [Fairclough, 2009]. BCIs that do not target a classic control or communication strategy but use brain activity measurements to detect some type of user state are often referred to as passive BCIs [Cutrell and Tan, 2008, Zander and Kothe, 2011]. Brain activity measurements, such as EEG or fNIRS, were shown to identify these mental states more robustly than other physiological measurements such as electrodermal activity, heart-rate related parameters and eye-gaze measurements [Frey et al., 2014, Hogervorst et al., 2014]. For adaptive user interfaces and tutoring system, the reaction time is often not crucial. A feedback within 10 seconds whether a pupil is unable to cope with an exercise is still absolutely fast enough for many tutoring systems. One of the disadvantages of fNIRS does therefore not apply to passive BCI systems.

While EEG is by far the most common modality for passive BCIs and has been successfully used for discrimination of high and low workload [Berka et al., 2007, Kothe and Makeig, 2011, Brouwer et al., 2012] and discrimination of mental tasks [Friedrich et al., 2012], fNIRS is a promising modality for passive BCIs or physiological computing, as fNIRS headsets can be easily developed and are minimally intrusive.

The detection and classification of mental arithmetics for classroom application has been demonstrated using fNIRS [Ang et al., 2010a, Ang et al., 2010b]. Other studies investigated the intersession consistency of hemodynamic responses evoked by mental arithmetics in fNIRS [Power et al., 2012]. Two different types of mental task, namely mental singing and mental arithmetics were investigated in [Power et al., 2010]. Differences in averaged hemodynamic responses for different mental tasks were presented in [Hoshi and Tamura, 1997]. [Hirshfield et al., 2011] investigated different mental tasks to adapt user interfaces.

The possibility of workload estimation using physiological parameters was shown, among others, in a multi-modal experiment in a realistic driving scenario [Jarvis et al., 2011]. Distinct average hemodynamic pattern were found for different workload levels for the n-back tasks [Hoshi et al., 2003, Ayaz et al., 2007] and more realistic game scenarios [Izzetoglu et al., 2003]. [Ayaz et al., 2010, Ayaz et al., 2012] demonstrated that workload can be classified in a realistic air-traffic controller task, showing that fNIRS can be used for real-life adaptive interfaces. The cognitive workload of airline pilots [Çakir et al., 2016] and unmanned aerial vehicles [Izzetoglu et al., 2015] was effectively monitored using fNIRS to increase safety standards.

Another user state which is particularly interesting to detect for vehicle drivers is drowsiness [Wierwille et al., 1994] or mental fatigue [Kaplan, 2001, Kuo and Sullivan, 2001]. Drowsiness detection [Khan and Hong, 2015, Khan et al., 2016] and mental fatigue [Ahn et al., 2016] have been successfully detected using fNIRS. With better and more user friendly fNIRS systems available, fNIRS has also been used to evaluate user interfaces, for example in web layout [Lukanov et al., 2016] or to capture the emotional user experience [Pollmann et al., 2016]. Emotion is another form of well investigated user states and we have shown that they can be measured with fNIRS reliably [7][21]. However, emotions are not a focus of this dissertation.

For further readings, see the reviews on fNIRS for neuroergonomics [Derosière et al., 2013] and passive BCI [Strait and Scheutz, 2014]. In the following section we will present our results on passive BCIs to adapt interfaces to a user's mental state.

2.4.1 Identification of Activity Type

This section describes our study *"Classification of mental tasks in the prefrontal cortex using fNIRS"* [18] (see Appendix A.4). We describe how the study fits within the scope of this dissertation and touch on some interesting aspects of the study.

A system that can detect the type of activity a user is currently occupied with could potentially be useful in classroom or online tutoring settings. The type of detected activity could for example be a math problem, language exercises or spatial imagination. In these settings, the system could adapt to the user's specific demands and help with the type of problem, mathematical, language or orientation should the user require help. To evaluate the feasibility of such a system to be realized with fNIRS, we conducted a study with 10 participants.

The prefrontal cortex is generally linked to executive functions including planning of complex cognitive behavior, decision making and social behavior [Fuster, 1988, Miller and Cohen, 2001]. The brain activity in the prefrontal cortex is therefore highly important to detect the type of activity a user is currently undertaking or the amount of workload a user experiences. [Sato et al., 2013] illustrated that fNIRS measures the same activity pattern in the prefrontal cortex as fMRI by measuring strong correlations between fNIRS and BOLD signals.

Our experiment consisted of three different tasks requiring very different problem solving capacities:

- **Mental Arithmetics (MA):** Participants had to repeatedly subtract a presented minuend between 7 and 19 (10 excluded) from a starting number between 501 and 999. This required them, for example to start with 716 and repeatedly subtract 9, i. e. 716 - 707 - 698 - 689 and so forth. The calculation had to be performed mentally, no verbalization was allowed. This task puts high demands on mathematical capacities of the user.

- **Word Generation (WG):** Participants were asked to think of words starting with a presented letter. For example, given a "G", they could think of Giraffe - Gentleman - Geography and so forth. It is important to note that participants had to only imagine the words and not utter them audibly. This task requires language capacities [Klein et al., 1995].

- **Mental Rotation (MR):** Participants were shown a 3D object and had to visualize it rotate in the horizontal plane. Figure 2.5 shows an example object that had to be imagined rotating. This task requires spatial imagination [Bartoshuk et al., 1960].

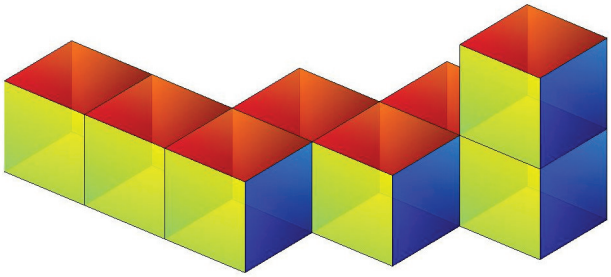


Figure 2.5: Example of a 3D object that participants had to imagine rotating.

Each task lasted for 10 seconds at a time, which we denote as a trial. After each trial, participants had to rest for 15 seconds to allow hemoglobin levels to return to baseline again. We recorded 30 repetitions of each task in a randomized order. To allow a comparison to inactivity, we extended the 15 seconds rest periods by another 10 seconds at 30 randomly chosen times in the experiment. These intervals will be used as Relax trials. This experiment setup is also useful to break the regular pattern of task and relax, in which slow oscillations might be misinterpreted as brain activity. We recorded fNIRS data while the participants were engaged in the tasks using a Artinis Oxymon Mark III system with 4 receiver and 4 transmitter optodes placed on the forehead. This results in a total of 8 channel of ΔHbO and ΔHbR. Data was sampled at 10 Hz. As none of the tasks require any user movement or interaction, we should not have any systematic artifacts in the recorded fNIRS data. A total of 10 participants were recorded for this study.

To preprocess the data, we first bandpass filtered the data between 0.001 and 0.6 Hz using and IIR filter with filter order 6. This filtering is applied to attenuate heartbeat artifacts and long period shifts. We tried to remove linear shift by applying linear detrending in blocks of 5 minutes. To reduce the influence of artifacts, we applied the wavelet artifact removal technique [Molavi and Dumont, 2010], which is suggested in the excellent review of artifact reduction techniques for fNIRS [Cooper et al., 2012]. For the wavelet denoising technique, ΔHbO and ΔHbR data $y(t)$ of every channel is transformed to the wavelet domain:

$$y(t) = \sum_k c_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_k d_{jk}\psi_{jk}(t) \tag{2.6}$$

where $c_{j_0 k}$ are the approximation and d_{jk} the detail coefficients of the wavelet transform. $\phi_{jk}(t)$ and $\psi_{jk}(t)$ correspond to scaling and wavelet functions. Dilation is represented by the parameter j, with j_0 being the coarsest scale in the decomposition. The translation parameter is denoted as k. Assuming the distribution of wavelet coefficients of fNIRS data to be Gaussian, one can simply estimate the probability for each encountered coefficient. Hemodynamic signals are expected to have a smooth distribution of wavelet coefficients and very low variance. Given this observations, artifacts can be removed by the removal of wavelet coefficients with occurrence probabilities below a threshold α. In this experiment,

a threshold of 10 times the interquartile distance was chosen. Once noisy coefficients are eliminated, the signal is transfered back into the time domain. After wavelet denoising, we extract baseline corrected trial data.

To extract meaningful information from the trial data, we extended the feature selection from [22]: Rising and falling trends can be observed in fNIRS data, which we capture by looking at the largest increase $f^{\uparrow}_{t,c}$ and decrease $f^{\downarrow}_{t,c}$ in the oxygenated hemoglobin data of a channel c in a trial t. As ΔHbO and ΔHbR are highly correlated [Cui et al., 2010], we limit our feature space to the ΔHbO information of each channel. A total of two features is therefore extracted for each channel c for every trial t:

$$\begin{aligned} f^{\uparrow}_{t,c} &= \max_{i \in [fs, len(t)-fs]} (\mu(\Delta HbO^{t}_{c,i:i+fs}) - \mu(\Delta HbO^{t}_{c,i-fs:i})) \\ f^{\downarrow}_{t,c} &= \max_{i \in [fs, len(t)-fs]} (\mu(\Delta HbO^{t}_{c,i-fs:i}) - \mu(\Delta HbO^{t}_{c,i:i+fs})), \end{aligned} \tag{2.7}$$

where fs is the length of the interval to be shifted. This feature extraction procedure results in a total of 16 features per trial.

We evaluated whether the different mental tasks could be discriminated from each other and from a relaxed state in a 10-fold cross-validation. A simple linear classifier suffices to discriminate in the relatively low dimensional feature space. We applied a Linear Discriminant Analysis (LDA) for classification. Figure 2.6 summarizes the classification results.

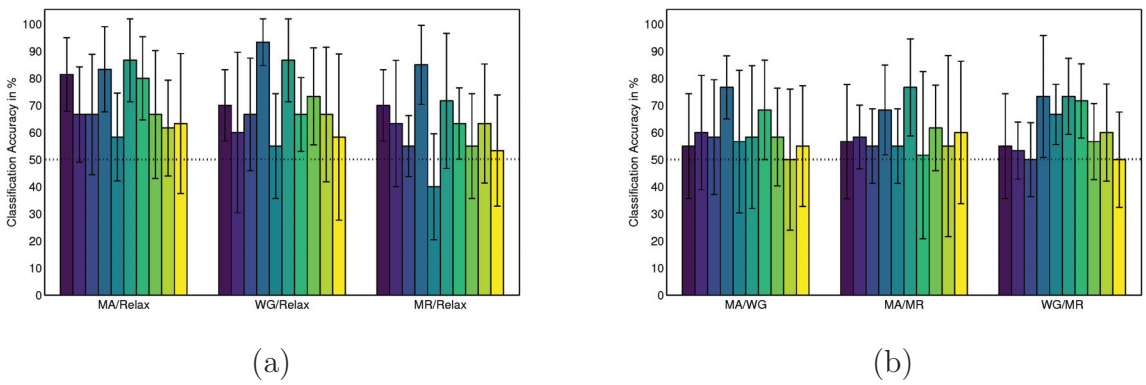


Figure 2.6: Classification results for discrimination of Mental Arithmetics (MA), Word Generation (WG) and Mental Rotation (MR) from Relax (a) and against each other (b). Each color represents one participant. Whiskers show standard deviations. Naive classification accuracy is indicated by the dotted line. Figure modified from [18].

All three mental tasks could be discriminated from Relax significantly better than chance level (one-sided t-test, $p < 0.01$). Activity during mental arithmetics (MA) achieved the best results with an average over all 10 participants of 71% classification accuracy. Word generation (WG) worked almost equally well with 70% accuracy on average. The mental rotation task (MR) yielded the lowest results with a mean accuracy of just 62% accuracy. Our participants filled out an questionnaire after participation in which they unanimously described mental rotation as the hardest task to engage in. This might be an explanation for the low classification results achieved in this task.

These results indicate that mental tasks can be discriminated from periods of inactivity. In a next step, we tried to identify which mental task a user was occupied with. Classification between MA and WG and between MA and MR worked with a mean accuracy of 60%. Discrimination between WG and MR yielded slightly better results with 61% accuracy. All mean classification accuracies are better than naive classification, as tested by a one-sided t-tests ($p < 0.01$). While these results are not suitable for a realistic Human-Computer Interface yet, they give a first indication of the feasibility of different mental tasks for BCI. While further investigations are clearly necessary, we think that tutoring systems might be augmented by the input of fNIRS by for example highlighting the task the user is currently occupied with. Before fNIRS augmented interfaces can be productively used, classification accuracies need to be improved further and more realistic tasks need to be tested.

2.4.2 Estimation of User Workload

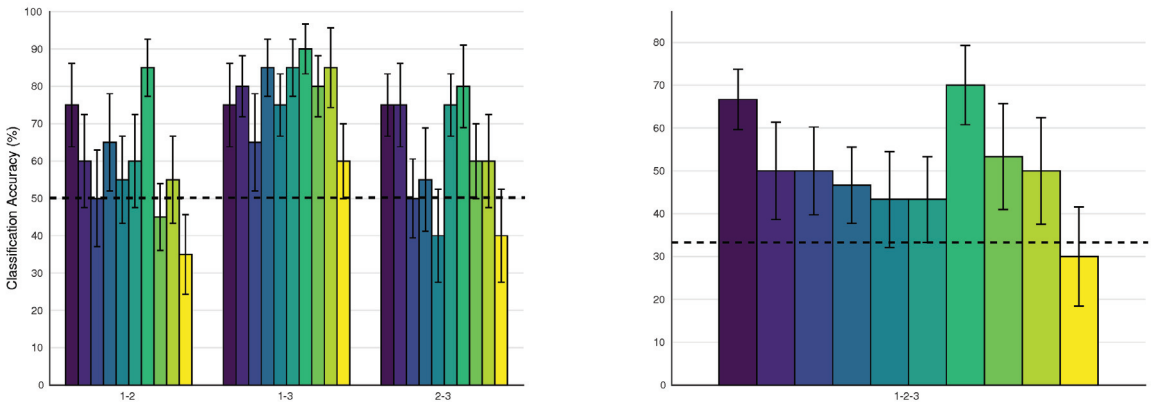
This section shortly describes our study *"Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS"* [6] (see Appendix A.5). Our findings are embedded within the topic of this dissertation and some interesting aspects are highlighted.

In this section, we extend the concept of detection of type of mental demand to quantifying the amount of mental workload a user is experiencing using fNIRS. For this purpose we conducted a study using the n-back task [Kirchner, 1958]. The n-back task has been used for classification of workload in EEG [Kothe and Makeig, 2011], but to the best of our knowledge, had not been used for classification of workload levels with fNIRS before. Albeit the n-back task has been investigated in averaged hemodynamic responses intensively [Sato et al., 2013], the extension to single trial classification is non-trivial, as grand averages have much better signals-to-noise ratios as noise is removed through averaging [Rugg and Coles, 1995]. In the n-back task, workload is induced by displaying a sequence of letters and asking the participants to push a button if the currently presented letter is the same as the n-th letter before the current letter. Letters, which are the same as the n-th previous letter are denoted as targets. This way, participants have to continuously remember the last n letters of the presented series and constantly update the remembered sequence by shifting it forward. Obviously, the task difficulty increases with increasing n, as more letters have to be remembered. User performance and subjective evaluations in our experiment clearly indicated that the task was demanding and that workload was higher with increasing n. In our experiment, we recorded fNIRS activity using the Artinis Oxymon Mk II device while participants underwent 10 trials each of 1-,2- and 3-back conditions. The 3-back task is already considered very hard without intensive training [Kane et al., 2007].

A total of 10 participants underwent 44 seconds of n-back task followed by 15-25 seconds of relaxation per trial. We investigated whether the different workload levels could be discriminated from each other and from Relax periods. We extracted the slope of a line fitted to the fNIRS data of ΔHbO and ΔHbR in a window as features. This is done using a least squares approach. All evaluations are performed using an LDA classifier in a 10-fold cross-validation.

We first discriminated the different workload levels from the Relax state. As expected, the classification of 3-back worked best, as the highest workload was present with an average accuracy of 81%. 2-back and 1-back yielded 80% and 72% accuracy, respectively. These results show that fNIRS can be used to discriminate workload from a Relax state using

brain activity measured with fNIRS. The initial goal of this study was to quantify the amount of workload, which we evaluate by testing whether the different workload conditions can be discriminated from each other. Figure 2.7 summarizes our results.



(a) (b)

Figure 2.7: Accuracies for each participant. Numbers in caption show n-back levels discriminated from each other. (a) Two class discrimination of workload level. (b) Discrimination between all three levels of workload. Each color represents one participant.

The classification between 1- and 3-back worked well for all participants, while only roughly half of the participants yield good results for the 1- versus 2-back and 2- versus 3-back scenarios. We hypothesize that this is due to the larger difference in mental workload for 1- versus 3-back compared to the 1- versus 2-back and 2- versus 3-back scenarios. These results indicate that the classification of high workload from low workload is feasible while a intermediate level of workload cannot be discriminated from either high or low workload reliably. Figure 2.7 (b) illustrates three class accuracies, which are better than chance for all but one participant.

Our results achieved in single trial classification of mental workload induced by the n-back task have since been replicated [Ang et al., 2014] with very different recording equipment. These results indicate that adaptive interfaces reacting to a user's mental workload can be realized using fNIRS.

Single trial analysis of fNIRS is a very young field and generally lacks standard algorithms and procedures. To speed-up the development of new methods and allow a good comparability of methods, we shared the recorded data from this study with the public and made it freely available for download[1]. The data has since been downloaded and used for improvements in fNIRS methodology [Zhongpeng and Wenxue, 2016] from fNIRS researchers around the world. We have shown improvements in fNIRS methodology by combining feature extraction and classification through regularized least-squares optimization on this data corpus [16].

[1] `http://www.csl.uni-bremen.de/cms/en/research/brain-activity-modeling`

2.4.3 Discrimination of More Workload Levels through Hybrid BCI

In this section, we summarize the results from our paper *"Hybrid fNIRS-EEG based discrimination of 5 levels of workload"* [11] (see Appendix A.6). The general approach is summed up and some interesting findings described.

Our previously presented study, as most of the literature in EEG [Brouwer et al., 2012] and combined EEG and fNIRS analysis [Coffey et al., 2012] investigate a maximum of 3 levels of workload. To extend the amount of classifiable levels of workload, we investigate the usage of a hybrid BCI [Pfurtscheller et al., 2010] that combines EEG and fNIRS measurements to classify a total of 5 levels of workload. The combination of EEG and fNIRS has been shown to enhance BCI performance [Fazli et al., 2012]. We have successfully used a hybrid BCI to classify auditory and visual perception processes [5].

To induce workload, we use the memory updating task, which was first proposed in [Salthouse et al., 1991] and updated by [Oberauer et al., 2000]. The memory updating task is known to reliably induce workload in more fine-grained levels than for example the n-back task [Lewandowsky et al., 2010]. In the memory updating task, participants are shown 1 to 5 boxes with an initial single digit that needs to be remembered. The number of shown boxes equals the number of digits that need to be remembered and constantly updated. More boxes induce higher workload and the number of boxes can thus be used as the difficulty level of the task. Over the course of the trial, digit operations are displayed in one of the boxes for 2.5 seconds and need to be applied to the initially remembered digits. Participants therefore need to update their memorized numbers with every displayed operation (hence the name). We recorded 10 trials for each of the 5 difficulty level (1 to 5 boxes) for each participants. A total of 10 participants were recorded.

Electrical brain activity was measured using three EEG electrodes on the midline at positions Fz, Cz and Pz according to the international 10-20 system [Jasper, 1958]. Four additional electrodes were placed around the eyes to measure electrooculography (EOG), which is used to remove EOG artifacts from the EEG data. All electrodes were referenced to the nose. Hemodynamic activity in the prefrontal cortex was measured using 15 detector and 28 emitter optodes on the forehead. Data was recorded using a frequency modulated oximeter (Imagent, ISS Inc.). Recorded differences in the optical densities of two wavelengths (690 nm and 830 nm) were transferred to ΔHbO and ΔHbR using the HomER software package [Huppert et al., 2009].

In the electrophysiological data, we extracted band-power features between 4 and 25 Hz in 1 Hz bins using Welch's method [Welch, 1967]. The data was cleaned previously using the EOG regression method proposed in [Schlögl et al., 2007] to reduce the impact of eye movement artifacts. In the hemoglobin data, we limited our analysis to the ΔHbO signals. Each detector measures light intensities from all 28 light emitters, resulting in a total of 420 channels with various emitter-detector distances. To limit our analysis to information bearing channels, we excluded all channels not having a clear pulse artifact, which should be present in clean fNIRS recordings. Channels with pulse artifacts were identified if a peak was detected in the log-power spectrum p in the frequency band between 0.8 and 1.7 Hz, which is the normal rate of heart beat for adults:

$$\max\{p(f)|f \in [0.8, 1.7]\} - \text{mean}\{p(f)|f \in [0.8, 1.7]\} > 0.5 \tag{2.8}$$

This procedure reduced the amount of channels to 13-47 channels depending on the participant. As a feature, we extracted the slope of a line fitted to the fNIRS data for each of the remaining channels.

To evaluate the feasibility of discriminating 5 levels of workload, we first compared the difficulty levels in binary experiments. All evaluations are performed for fNIRS, EEG and the feature-level fusion of both, which we call FUSION. For the FUSION approach, we simply combine the feature spaces of fNIRS and EEG. All evaluations were performed in a 10-fold cross-validation. A shrinkage regularized LDA was used for classification. Figure 2.8 (a) shows results for the binary classification experiments.

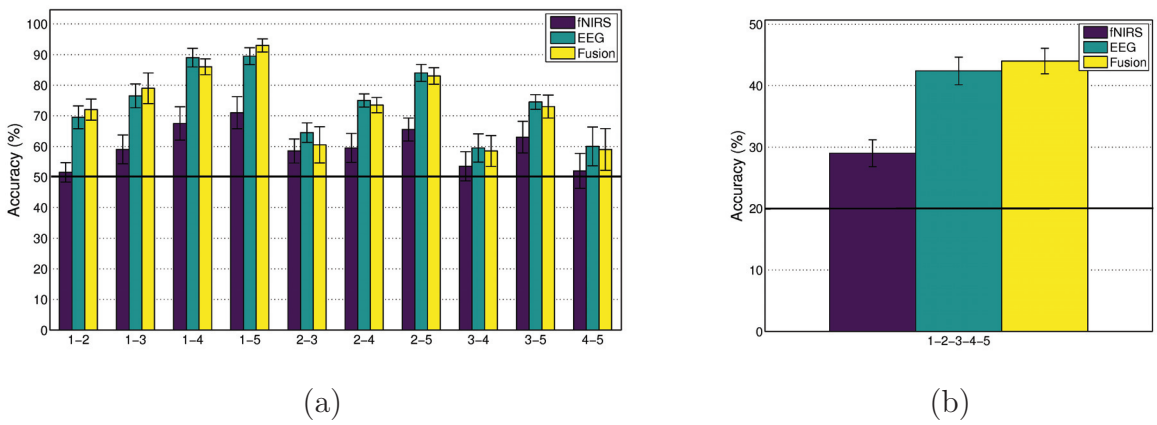


Figure 2.8: Mean classification accuracies over participants for discrimination between two levels of workload (a). 5-class discrimination accuracies. Numbers in captions indicate amount of boxes. Whiskers show standard error of the mean.

Our results show that all workload levels can be discriminated from each other with both EEG and fNIRS. Best results were achieved discriminating very high (5 boxes) from very low (1 box) workload with 71% accuracy for fNIRS, 90% for EEG and 93% for the FUSION approach. This is somewhat expected as the difference in workload is largest between those two conditions. Overall, achieved accuracies correlated strongly with the distance between workload levels (fNIRS $r = 0.38$, $p < 0.001$; EEG $r = 0.64$, $p < 0.001$; FUSION $r = 0.62$, $p < 0.001$). The further the workload levels are apart from each other, the better they could be discriminated. Therefore, lowest accuracies were achieved discriminating between levels 1 and 2, 2 and 3, 3 and 4 and between 4 and 5. The results of the five class classification problem are shown in Figure 2.8 (b). All approaches achieved better than naive classification (20%) with 29% for fNIRS, 42% for EEG and 44% for FUSION. These results show that more levels of workload than previously shown can be discriminated using EEG and fNIRS, which enables adaptive interfaces to react more fine-grained to the user's workload. Moreover, we have shown that the FUSION of EEG and fNIRS improves results in some cases.

2.5 Advanced Classification Approaches for fNIRS

In this section, we embed our study *"Investigating Deep Learning for fNIRS based BCI"* [12] (See Appendix A.7) in this dissertation. We summarize the general approach and repeat the most important findings.

For all signal classification approaches, the choice of the most appropriate classifier is a crucial decision. Different linear and non-linear classifiers have been investigated for EEG [Lotte et al., 2007] and fNIRS [Bauernfeind et al., 2014], with the conclusion that regularized classifiers seem a good choice for the noisy signals with usually low number of samples often found in BCI experiments.

With the recent success of Deep Neural Networks (DNN, [LeCun et al., 2015]) in diverse areas such as hand-written digit classification [Hinton and Salakhutdinov, 2006], image classification [Krizhevsky et al., 2012] and automatic speech recognition [Hinton et al., 2012], we investigated whether fNIRS BCIs could be improved using deep learning, as well. For this purpose, we conducted a study comparing deep learning to more classical BCI classifiers using the data from our mental tasks study [18] (see Section 2.4.1).

One of the many advantages of DNN is that features do not need to be hand-tailored but can be learned by the network efficiently from raw data [Martinez et al., 2013, Bengio, 2009]. This allows for good results without the need for domain knowledge. We therefore tested the results DNN could achieve on minimally preprocessed fNIRS data. Slow trends were removed by subtracting the mean of the surrounding 240 seconds from every sample in every channel. To attenuate higher frequency noise, especially the participants' heart beat, the data was low-pass filtered with a cutoff frequency of 0.5 Hz using an elliptic IIR filter with filter order 6. Data was downsampled to 1 Hz resulting in 10 measures of oxygenated and deoxygenated hemoglobin values for each channel in the 10 second trials of the mental tasks experiment. These values were stacked to form a feature vector for each trial. Our DNN consisted of an input layer with linear activation functions, between 1 and 3 hidden layers and a softmax output layer. Hidden layers used a logistic activation function. Network weights were pre-trained with layer wise, unsupervised restricted Boltzman machines which have been shown to speed up training and lead to better generalization [Hinton et al., 2012]. Training was then performed by minimizing the cross-entropy error using conjugate gradient, which is known for fast convergence and automatic estimation of learning rate [Martens, 2010, Hinton, 2012]. We analyzed the DNN performance with 1, 2 and 3 hidden layers. We did not investigate neural networks with more hidden layers, as no improvement in classification accuracy was found for more hidden layers.

For a comparison with standard BCI methods, we trained a LDA classifier with hand-tailored fNIRS features described in Section 2.4.1 and applied a shrinkage-LDA to the same high-dimensional feature vectors used for the DNN. In the shrinkage-LDA the covariance matrix is interpolated with the identity matrix to prevent overfitting. The optimal shrinkage parameter was estimated using the method by Ledoit and Wolf [Ledoit and Wolf, 2004]. All evaluations were done in a 10-fold cross-validation. Figure 2.9 compares classification results for the neural networks, the shrinkage LDA and the LDA with hand-tailored features.

It can be clearly seen that all classification approaches result in better than naive classification results for all experiments. Interestingly, the differences between the different approaches are only marginal. When averaging over all 7 experiments, DNN resulted in mean accuracies of 63.3%, 64.1% and 62.1% accuracy for 1, 2 and 3 hidden layers, respec-

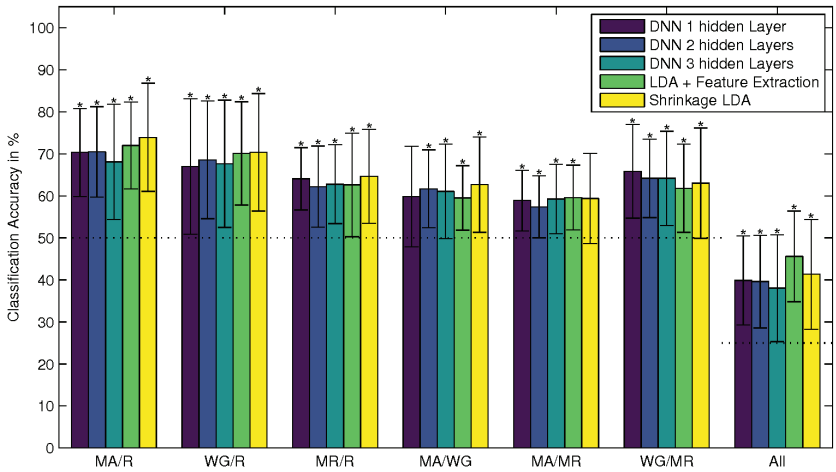


Figure 2.9: Classification accuracies of mental tasks averaged over participants. Whiskers denote standard deviations. Results marked with a "$*$" are significantly better ($p < 0.01$) than naive classification (dotted line).

tively. LDA with feature extraction on the other hand yielded an average of 64.3% and the regularized shrinkage-LDA achieved the highest accuracies with a mean of 65.7%. These results show that all classification approaches result in very similar classification results and the choice of classifier is therefore not crucial. When features are not hand-tailored, regularized models seem to be a good idea as they limit the amount of parameters that have to be learned which is a huge advantage when only very little training data is present. We hypothesize that DNN did not result in higher classification results for this problem domain, as only very little training data is available and the data is very noisy.

The successful application of DNN to fNIRS data show that BCIs based on fNIRS are possible without the need for hand-tailored features, greatly simplifying the process for new users of fNIRS.

2.6 Contributions of the Corresponding Publications

In our paper *"Speaking Mode Recognition from Functional Near Infrared Spectroscopy"* we showed for the first time that different speaking modes elicit consistent enough hemodynamic responses that they could be classified in single trial. The extension from grand average investigation of speech to single trial classification is a large step towards BCI using different speaking modes as a control paradigm. In our study *"Self-paced BCI with NIRS based on speech activity"*, we showed that speaking modes could be classified without the knowledge of stimulus timing, which is an important step towards self-paced BCI. Both these approaches render BCI based on speaking modes possible. These BCIs could use the different speaking modes as binary 'yes'/'no decisions or as an idle switch for another communication system.

The slow nature of hemodynamic responses are less of a disadvantage for so called passive BCIs. We present robust discrimination of type of mental activity in *"Classification of mental tasks in the prefrontal cortex using fNIRS"*, which could be used to detect the type of

task a user is currently occupied with to inform an adaptive interface. In *"Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS"*, we presented robust classification of 3 levels of workload induced by the n-back task for the first time and shared the data with the community. Through the combination with EEG, we showed that 5 levels of workload could be discriminated in *"Hybrid fNIRS-EEG based discrimination of 5 levels of workload"*. This is so far the largest number of memory levels discriminated in the literature. The robust quantification of workload using fNIRS could be used for interfaces that adapt to the cognitive workload of a user.

To improve upon the state of the art in fNIRS classification, we investigated deep learning approaches in *"Investigating Deep Learning for fNIRS based BCI"* and could show that deep learning can successfully be applied to fNIRS data, eliminating the need for hand-tailored features.

Chapter 3

Speech in Invasive BCI

Electrocorticographic measurements are possible through craniotomies, which are necessary in severe forms of epilepsy. The recordings obtained in these situations are of unparalleled signal quality and are ideally suited for the investigation of speech processes in the brain. This chapter describes electrocorticography and the distinctive features of the measured signals. Then, our approach for automatic speech recognition from neural signals is described. We show that in some situations speech synthesis is better suited for communication and demonstrate how intracranial activity can be used to resynthesize speech in real-time. To improve reconstruction quality, we investigate deep neural networks.

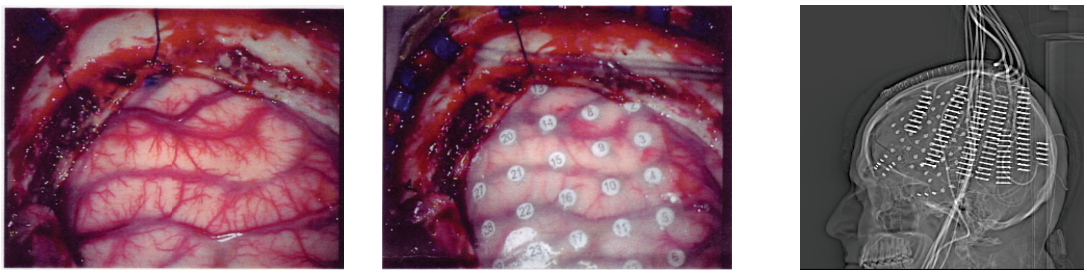
3.1 Basics of Electrocorticography (ECoG)

In severe cases of epilepsy, intracranial electrodes need to be implanted for surgical planning. Such patients usually suffer from frequent seizures and do not respond to antiepileptic medication. Implanted electrodes are needed to identify the seizure focus and simultaneously map the functional anatomy of the patients' eloquent cortex. The eloquent cortex are cortical regions, whose removal would result in loss of language, vision, motor function, sensory processing or memory. Given the seizure focus and mapping of eloquent cortex, the margin of surgical resection can be tailored to remove as much of the epileptogenic zone with the least amount of functional impairment. The process to localize a patient's seizure focus usually takes several days as a sufficient number of spontaneous seizures has to be recorded. Similarly, the mapping process typically takes several 2-hour sessions distributed over several days. Cortex mapping can be performed through electrocortical stimulation in which weak electrical currents are passed between pairs of electrodes. Alternatively, newer passive mapping approaches measure gamma activity [Brunner et al., 2009, Roland et al., 2010] during prompted tasks and create a mapping of the functional cortex without the need of stimulation. The clinical requirements for cortical mapping open the opportunity for scientific experiments, as the implanted electrode grids usually stay implanted for 5 to 10 days. During this period, patients are not permanently required to take part in clinical procedures and often consent to participate in scientific experiments. All studies have Institutional Review Board approval and do not cause any additional risk on the patients. Patients are free to stop participating at any time.

This clinical practice provides a unique opportunity to study electric brain activity directly on the brain surface. Intracranial electrodes provide measurements without the spatial blurring from dura matter, skull and scalp [Gevins et al., 1994, Cooper et al., 1965] observed

in EEG and with high spatial sampling of around one cm inter-electrode distance. Moreover, the skull acts as a low-pass filter in EEG [Pfurtscheller and Cooper, 1975], which further reduces the signal-to-noise ratio in relevant frequency bands for surface recordings. Additionally, artifacts, especially muscle movements, strongly occur in surface recordings [Akay and Daubenspeck, 1999] and are not present in invasive measurements further contributing to the higher quality recordings from ECoG.

Different types of electrodes are used for invasive recordings. Depth electrodes are stereotactically implanted through small burr holes. This procedure is especially useful to record from deep brain structures like hippocampus and amygdala and is safer due to the smaller surgical intervention. Activity from larger cortical areas such as motor cortex or perisylvian areas for speech and language processing can better be measured using electrodes placed on the brain surface. These subdural electrodes are placed direct on the brain surface and require a more invasive craniotomy. They are placed in strips of 4-8 electrodes or grids of 2-8 by 8 electrodes. Electrode grids consist of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) embedded in silicon with an inter-electrode distance of 0.6-1 cm. Full grids can be cut down to smaller grids to create an optimal coverage for the cortical region of interest. Figure 3.1 shows a craniotomy (a), the implanted grid (b) and a radiograph of the implanted electrodes (c) for a patient with a large grid on the left hemisphere.



(a) (b) (c)

Figure 3.1: Implantation of ECoG grids. (a) Brain after craniotomy. (b) Brain with electrode grid. (c) Radiograph of patient's head after implantation of ECoG grid.

To record the measured activity from the intracranial electrodes, biosignal amplifiers are necessary. Small voltage differences in the gamma band and fast dynamics require high sampling and adequate A/D converters. For all recordings used in this dissertation, several stacked 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria) were utilized.

3.1.1 High-Gamma in ECoG

The high signal quality of ECoG recordings enables the measurement of task related activity in the high-gamma band, a broad range above 70 Hz. [Crone et al., 1998] first observed activity in this frequency range in motor tasks in ECoG recordings. High-gamma activity shows highly task specific activity which is more localized and more specific to exact timing [Crone et al., 2006] than other frequency ranges. Gamma power is ubiquitous across functional and anatomical domains and likely reflects multi-unit firing rates [Ray and Maunsell, 2011].

Among other findings, high-gamma has been shown to reflect higher-order auditory processing [Crone et al., 2001, Edwards et al., 2005, Sinai et al., 2009] and speech perception [Edwards et al., 2009, Pasley et al., 2012, Martin et al., 2014]. [Leuthardt et al., 2011] could show that high-gamma is highly task relevant for word repetition.

Investigations of cortical gamma during music listening revealed strong correlations with sound intensity [Potes et al., 2012] and timbral and harmonic features [Sturm et al., 2014]. High-gamma activity was also found to be highly task-related in motor tasks [Miller et al., 2007]. While other frequency ranges also have great signal-to-noise ratios in ECoG recordings and have been studied intensively [Hermes et al., 2014, de Pesters et al., 2016, Coon et al., 2016], the high-gamma band can be uniquely investigated in invasive recordings. We focus on the high-gamma band in all our ECoG studies.

3.2 Speech Investigation in Invasive Recordings

Signals measured using EEG are far too contaminated by motion artifacts originating from facial muscles and do not provide sufficient signal-to-noise ratio in the high-gamma band to study speech production. The signal characteristics and unaffectedness by artifacts make ECoG an ideal candidate for the investigation of speech process. Speech processes have therefore been investigated intensively using intracranial recordings. See the excellent reviews for a literature overview [Chakrabarti et al., 2015, Martin et al., 2016b].

3.2.1 Speech Perception

Speech and audio perception have been investigated by a large number of studies using ECoG. The spatio-temporal dynamics of word processing were shown in [Canolty et al., 2007] by presenting a series of infrequent proper names in a stream of verbs and non-words. They could identify cortical areas involved in processing of words and the timing in the different locations. [Kubanek et al., 2013] tracked the envelope of perceived speech in gamma activity in the cortex. [Chang et al., 2010] had participants listen to synthesized speech and found phoneme and phonetic feature encoding [Mesgarani et al., 2014] in posterior superior temporal gyrus. Neural variability was shown to reduce upon onset of speech stimuli in [Dichter et al., 2016]. Extending the investigation of the neural basis of speech perception to a reconstruction approach, [Pasley et al., 2012] reconstructed spectral features from perceived speech. [Yang et al., 2015] reconstructed the speech spectrogram of perceived speech using Deep Neural Networks (DNN).

3.2.2 Speech Production

As ECoG is not influenced by motion artifacts of facial muscles, it is also ideally suited for the study of overt speech production. A plethora of studies investigates different aspects of speech production in ECoG. [Kellis et al., 2010] demonstrated that 10 spoken words could be discriminated from each other using ECoG signals. Since words are constructed from more fundamental units of continuous speech like phonemes or syllables, studies have investigated different units of speech production. One dimensional cursor control was demonstrated using the production of isolated phonemes in [Leuthardt et al., 2011]. Classification of

articulatory features in speech production was demonstrated in [Bouchard et al., 2013, Lotte et al., 2015]. [Bouchard et al., 2016] showed that articulator movement, captured using ultrasound imaging of the tongue and videos of facial movement, could be reconstructed using ECoG recordings. A set of four different phonemes was classified in [Blakely et al., 2008] and very fine spatial localization for the different phonemes was found in high-density ECoG recordings. [Bouchard and Chang, 2014] classified entire syllables successfully. The full set of all 39 American phonemes was decoded in [Mugler et al., 2014] which were manually segmented in isolated word production. Instead of classification of fundamental units of speech production, [Martin et al., 2014] reconstructed spectral features of the produced audio waveform using ECoG recordings.

The promising results achieved in these previous studies of isolated aspects of speech production in invasive recordings motivated us to advance to the recognition of continuous speech from neural signals, which we explain in detail in the next section.

3.3 Brain-to-Text: Decoding Continuous Speech into a Textual Representation

This section summarizes our study *"Brain-to-Text: Decoding spoken phrases from phone representations in the brain"* [4] (See Appendix A.8). The findings are embedded within this dissertation and interesting aspects are highlighted.

In our *Brain-to-Text* system, we showed that audibly articulated continuous speech can be decoded into a textual representation. This textual representation is for example suitable to process the content of spoken phrases. Including this approach into a BCI, users could compose messages or enter commands through the decoding of words from neural data. For this study, we utilized the ECoG data of seven patients producing continuous speech. Our participants read aloud different text excerpts that were shown to them on a screen. During this speech production, we recorded neural activity using ECoG grids in synchrony with acoustic waveforms measured by a dynamic microphone. Each reading of a text is denoted as a session. Each session is cut along pauses into 21-49 phrases. Due to the very small amount of data for each session of participant 2, we combined all three sessions. We excluded participant 4 from further analysis as no speech related activity was found. As the electrode montage is only determined by the clinical needs of the patients, each participant has individual number and positioning of electrodes. This way, all models have to be trained participant specific.

Phones are usually produced in 30 to 50 ms in continuously spoken speech rendering the manual segmentation of each individual phone in our recorded data very cost and time consuming. We therefore applied Automatic Speech Recognition (ASR) technology [15] to obtain a ground truth of phone timing in the acoustic data. Figure 3.2 depicts the experimental setup and the subsequent automatic phone labeling.

As ECoG data and audio waveform are recorded simultaneously, we can impose the phone timings on the ECoG recordings. This allows us to identify the corresponding neural signals to each produced phone. It is important to note that automatic labeling is not perfect and is therefore introducing some noise into our ground-truth. We use BCI2000, a software system

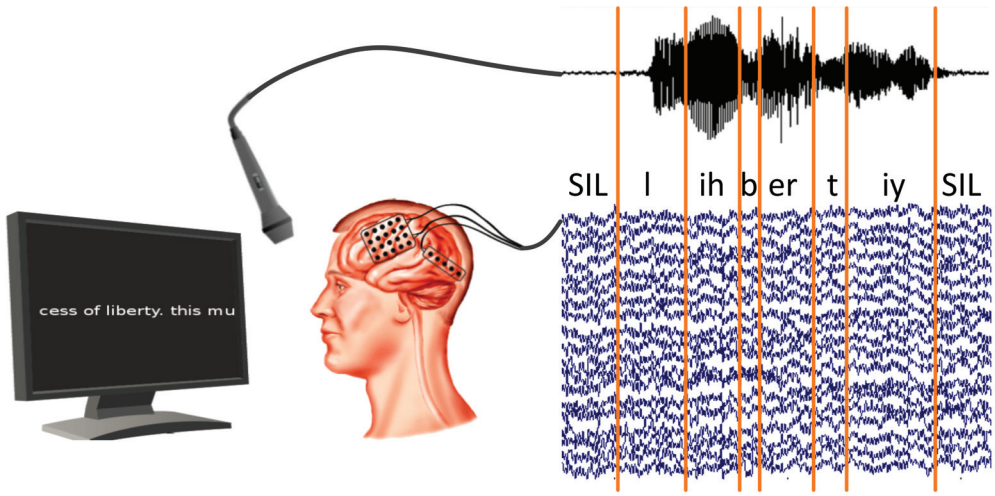


Figure 3.2: Experimental setup in our *Brain-to-Text* study. Participants read out texts displayed on a screen. ECoG signals and audio waveform are recorded simultaneously. Our in-house decoder BIOKIT [15] is used to extract the timing of spoken phones in the acoustic stream. These timings can then be imposed on the ECoG data. Figure used from [4].

for data recording, stimulus presentation and brain monitoring in BCI applications [Schalk et al., 2004], to record ECoG data and audio data in synchrony.

Once the phone timings are extracted, we segment the neural data into 50 ms long intervals with 50% overlap. This interval length enables us to capture the fast dynamics of speech production, while also allowing to extract broadband-gamma power reliably. For each of these intervals and each channel we extracted broadband-gamma power in the frequency range from 70 to 170 Hz. We applied the logarithm to the extracted activity to make the distribution more Gaussian. Various brain regions are participating in speech production with different timings [Sahin et al., 2009]. Speech planning and motor preparation are occurring prior to voice onset, while speech perception and auditory processing happen after speech production. This time course of neural activity makes capturing of context information absolutely crucial. We add context to our feature vectors by stacking the logarithmic broadband-gamma activity of all channels of neighboring windows up to 200 ms prior to and after the current interval. After feature calculation and stacking, we assign each feature vector the corresponding phone label from the acoustic stream.

To model phones in neural data, we describe the neural activity F_i associated with each phone j with a multivariate Gaussian distribution $p(F_i|\lambda_j) \sim \mathcal{N}(\mu_j, \Sigma_j)$. Mean μ_j and diagonal covariance Σ_j over all segments of a phone j describe the phone model λ_j. These generative models λ_j allow us to assign a likelihood that this activity was emitted by a phone model to a new segment of ECoG activity. Additionally, these models are easily interpretable. We chose to use Gaussian models, as they represent the underlying activity suitably well [Crone et al., 2001, Gasser et al., 1982] and can be estimated robustly from small amounts of data. Additionally, the computation of Gaussian models is very efficient and allows us to calculate the Kullback-Leibler-Divergence in a closed form.

The resulting phone models are sufficient to classify ECoG activity into a sequence of phones, by assigning each segment of ECoG activity to the phone with the highest likelihood. However, as neural data is very noisy, this would not yield great results. Additionally,

the sequence of decoded phones would somehow need to be translated to a sequence of words afterwards. This might be complicated, especially since phone combinations could be decoded that do not form proper words. ASR technology [Rabiner, 1989, Schultz and Kirchhoff, 2006] solves these issues by integrating the phone decoding models (called acoustic model in ASR) with a dictionary and a statistical language model into one decoding approach. These three knowledge sources are integrated into a combined decoding approach using Bayesian updating [Rabiner, 1989].

The pronunciation dictionary contains the mapping of phone sequences to words. The dictionary is used to guide the search for the correct words in ASR, as only words included in the dictionary can be recognized. As an example, Table 3.1 illustrates a dictionary covering all words of the phrase 'the slow undoing of those human rights' and some additional words.

Table 3.1: Example of a dictionary showing words and their pronunciation in a simplified notation for phones.

Word	Pronunciation
undoing	aa n t uw ih n
slow	s l ow
rights	r aa ih t s
of	aa v
human	hh ih uw m aa n
this	s ih s
the	s aa
those	s ow s
americans	aa m eh r aa k aa n s

A language model estimates how likely a word is, given its context. In N-gram language modeling [Jelinek, 1997, Stolcke, 2002], this is done by calculating probabilities of single words and probabilities for predicting words given the history of n - 1 previous words.

The decoding process of a phrase of ECoG feature vectors X is performed by finding the word sequence $\hat{W}$ that has the highest likelihood given the neural data. As this likelihood cannot be estimated directly, the decoding process uses Bayes rule to formulate:

$$\hat{W} = \arg\max_{W}\{P(W|X)\} = \arg\max_{W}\{\frac{p(X|W)P(W)}{P(X)}\} \tag{3.1}$$

with the probability for a word sequence $P(W)$ provided by the language model and the emittance probability for a sequence of feature vectors $p(X|W)$ provided by the ECoG phone models λ and the concatenation rules of the dictionary. The likelihood for the feature sequence $P(X)$ can be disregarded as it is the same for all word sequences and only the most likely word sequence is of interest, the absolute likelihood are not necessary.

Figure 3.3 illustrates the decoding process. The recorded *ECoG signals* are divided into 50 ms segments. After extracting *broadband gamma activity* in each segment, feature vectors are stacked to incorporate context information (*Signal processing*). Using the *ECoG phone models* emittance likelihoods for a phone given each feature vector can be estimated, resulting in *phone likelihoods over time*. The *Viterbi algorithm* calculates the *most likely word sequence* and corresponding *phone sequence* given these *ECoG phone models* together with the *language model* and *dictionary*. The decoding path can be visualized by highlighting the

most likely phone sequence in the *phone likelihoods over time* (red shaded areas). The complete systems produces a textual representation of the spoken phrase from recorded neural data.

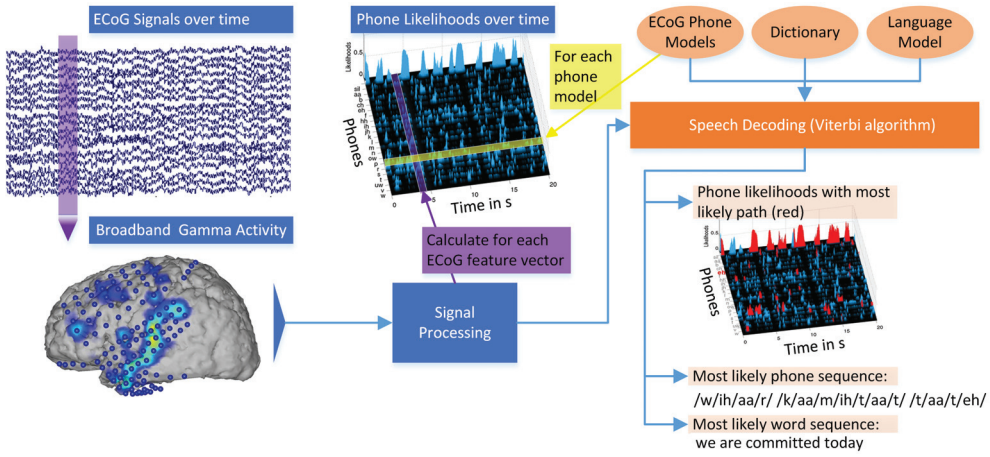


Figure 3.3: The *Brain-to-Text* decoding system. ECoG activity is recorded for every electrode. Broadband-gamma features are calculated in 50 ms intervals and stacked. Phone likelihoods over time are estimated. The combination of ECoG phone models, dictionary and language model is used in the Viterbi algorithm to decode phrases. Most likely word sequence and phone sequence are the decoding results. Red marked areas in the phone likelihoods correspond to the most likely phone path. Figure used from [4].

To evaluate our *Brain-to-Text* system we use a leave-one-phrase-out cross-validation. In this approach, we train our ECoG phone models λ_j on all but one phrase of ECoG data and decode the last remaining, unseen phrase. This approach is repeated until every phrase was used for testing once. We compare all our results against random models to evaluate statistical significance. For these random models, we shift the ECoG training data by half of its length. This way, the data still shows typical ECoG properties, but should not correspond to the labels any more. This approach also allows us to clearly identify the proportion, which language model and dictionary have in the success of our decoding approach. By applying the same processing pipeline to this shifted data we make sure that results are comparable and all analysis are statistically sound. The mean over all randomized results allows a robust estimation of chance level.

We judge decoding success using different metrics. For the decoding performance of single phones in the decoded sentence, we look at both, accuracy and average true positive over all phones.

As our approach extracts a textual representation of the spoken phrase, we also look at Word Error Rates for the complete phrases. The Word Error Rate (WER) between a decoded phrase and the corresponding reference phrase consists of the number of editing steps in terms of substitutions, deletions and insertions of words necessary to produce the reference from the predicted phrase, divided by the amount of words in the reference. A short example using the words from the dictionary in Table 3.1 illustrates the calculation of the WER:

Reference:	the	slow undoing of	those	human rights	
Predicted phrase:		slow undoing of	this	human rights	americans
Editing Step:	Deletion		Substitution		Insertion

To create the reference from the predicted phrase, one deletion ('the'), one substitution ('this' for 'those') and one insertion ('americans') are necessary. The word error rate is thus:

$$WER = \frac{|Insertions| + |Deletions| + |Substitutions|}{|Reference|} \cdot 100\% = \frac{1+1+1}{7} \cdot 100\% = 43\% \tag{3.2}$$

Figure 3.4 summarizes all decoding results of our experiment. As can be seen in Figure 3.4 (a) all sessions of all participants yielded higher single frame accuracies than chance level. Statistical significance is analyzed by comparing the decoding results of the ECoG models (purple) to those of the random models (yellow) using two-sided t-tests ($p < 0.05$). Clearly, participant 7 yielded the highest results with up to 52% accuracy for session 1. We hypothesize that participant 7's high results are explainable by the good electrode coverage of auditory regions in superior temporal gyrus with a high density grid and good coverage of interior frontal cortex (Broca's area) and speech motor cortex. Studies have shown that higher density grids yield better decoding performance, as neural information related to speech is very localized [Muller et al., 2016]. To further investigate the performance of participant 7, we first looked at a confusion matrix in the decoded phrases (Figure 3.4 (b)). The confusion matrix illustrates which phones in the decoded phrase correspond to which phones in the reference. It can clearly be seen that all phones are most often decoded as the correct phone, this means that our decoding approach reliably decodes all phones and does not rely on the correct classification of a small subset of phones. In Figure 3.4 (c), we look at WER (lines) depending on the size of the dictionary, i. e. the amount of words that can be recognized, and the average true positive rate over all phones depending on the size of the dictionary (bars). The average true positive rate remains relatively stable around 30% and only deteriorates slightly for larger dictionary sizes. Average true positive rates are far larger than the randomized controls for all dictionary sizes. Word recognition results for our system start at 25% WER for 10 words in the dictionary, meaning that in a phrase of 10 words, less than three words are recognized incorrectly or at the wrong position. Error rates increase steadily with dictionary size and reach a level of 60% for 100 words in the dictionary. These results outperform the random models for all dictionary sizes, but are not competitive to speech recognition systems using acoustic speech. This is due to far more noisy signals and far less data to train models from. While acoustic speech recognition systems are typically trained on thousands of hours of speech data, the *Brain-to-Text* systems only has few minutes available.

Our study presents the first system that applies ASR technology to neural data to decode a textual representation of spoken speech. Our system could be used in a BCI to compose messages or issue commands without the need to concentrate on individual flashing characters. We thereby demonstrate that continuous speech is a fast and natural paradigm for BCI. Recently, [Moses et al., 2016] were able to reproduce some of our results and successfully applied ASR technology to perceived speech.

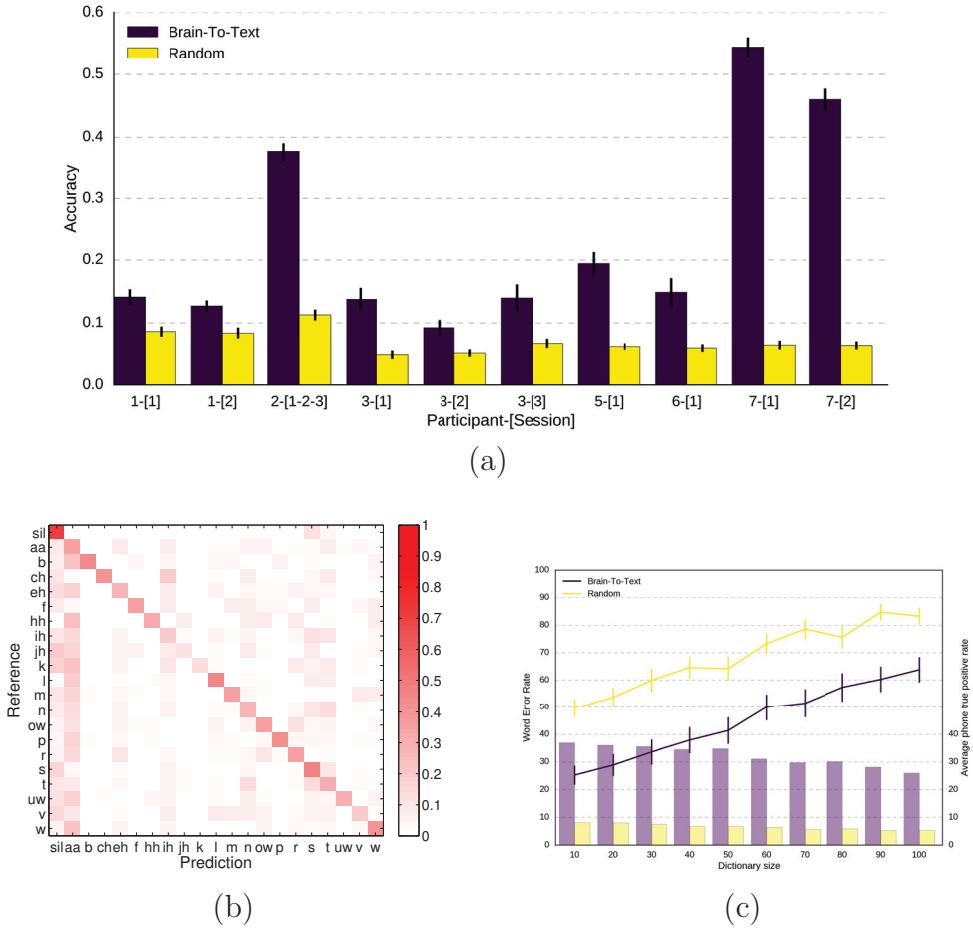


Figure 3.4: (a) Frame-wise accuracy for all sessions of six participants. Detailed results for participant 7 session 1: (b) Confusion matrix for classification of individual phones. The clearly visible diagonal illustrates that the *Brain-to-Text* system reliably recognized all phones. (c) shows Word Error Rates and average phone true positive rates over dictionary size.

3.4 Synthesis from Neural Signals

While a textual representation of decoded sentences is the ideal output to compose messages, e-mails or letters and to control a computer, it lacks some of the more subtle nuances of speech. Aspects such as emphasis, prosody and accent are lost when speech is written down. Capturing these aspects of speech would allow a user of a prosthetic device to harness the full expressive power of human speech. Especially to convey emotion, these additional information are crucial. For a locked-in patient that does not want to compose messages or control a computer, but converse with family or friends, a textual representation is also not the ideal output. In such a scenario, an audio waveform would enable a much more natural conversation. A simple approach would be to employ a text-to-speech system [Klatt, 1987] to create an audible representation from the decoded text. While this would be good to enable users to converse again, subtle nuances of speech would still be lost. A system that

transforms the measured neural signals directly into an audio waveform would allow the user to convey subtle nuances of speech and enable real conversation again. Such systems do exist converting EMG [Janke et al., 2012], EMA [Gonzalez et al., 2016] or ultrasound imaging of the tongue [Bocquelet et al., 2015] to audible signals and are called synthesis approaches. Through the direct synthesis from measured signal to audio waveform, subtle nuances of speech can be reconstructed.

In this dissertation we present the direct conversion from neural signals to audio waveform for the first time. We do so by first demonstrating the reconstruction of perceived rhythms from ECoG signals as a preliminary step towards speech synthesis from ECoG before we present a direct synthesis system reconstructing an audio waveform of speech from ECoG.

3.4.1 Music Envelope Reconstruction from ECoG

This section highlights some findings of our study *"Music rhythm reconstruction from ECoG"* [26] (See Appendix A.9) and embeds them within this dissertation.

As a first step towards synthesis of speech from neural signals we tried to reconstruct the sound envelope of music participants listened to from ECoG recordings. The reconstruction of perceived music is a first step towards the reconstruction of speech as both are auditory phenomena. However, music perception has far fewer cortical regions involved than speech production and perception, as no motor planning and execution for articulator movement are required and no language information and meaning has to be extracted from the perceived auditory stream. Music perception is ubiquitous in humans and should therefore elicit strong neural responses. We chose simple drum rhythms for this reconstruction analysis, as they only contain rhythmic information and lack the complex harmonic and melodic information of complete songs. The chosen stimuli thus enable an in depth analysis of this isolated aspect of music. In a study with 7 participants, ECoG data was recorded while participants listened to drum rhythms, which would occasionally stop for a few meters. Participants were instructed to follow the drum rhythm closely and imagine it to continue during the breaks. Audio data was time aligned to the recorded ECoG signals. To capture the audio amplitude, we extracted the Hilbert envelope in 50 ms intervals. For the ECoG signals, we extracted broadband-gamma activity in synchronized 50 ms intervals and added context information through feature stacking up to 200 ms prior to and after the current interval. We build linear regression models with an L_1-norm of the parameter vector (commonly referred to as the Least Absolute Shrinkage and Selection Operator (LASSO) regression) to reconstruct the audio envelope purely from ECoG signals. All evaluations are performed in a 10-fold cross-validation. Figure 3.5 shows an example of reconstructed envelope from neural data (purple line) and the original envelope (yellow line) for comparison.

This short excerpt clearly shows that basic rhythm perception can be captured in ECoG recordings and that rhythmic stimuli can be accurately reconstructed from neural data. To quantify reconstruction quality, we calculated correlation coefficients (Spearman's ρ) between actual sound intensity and reconstructed envelope. Correlations coefficients up to 0.45 could be achieved.

With these promising results for the reconstruction of perceived drum rhythms, we tried to tackle the more difficult task of reconstructing the entire speech spectrogram from neural recordings in the next section.

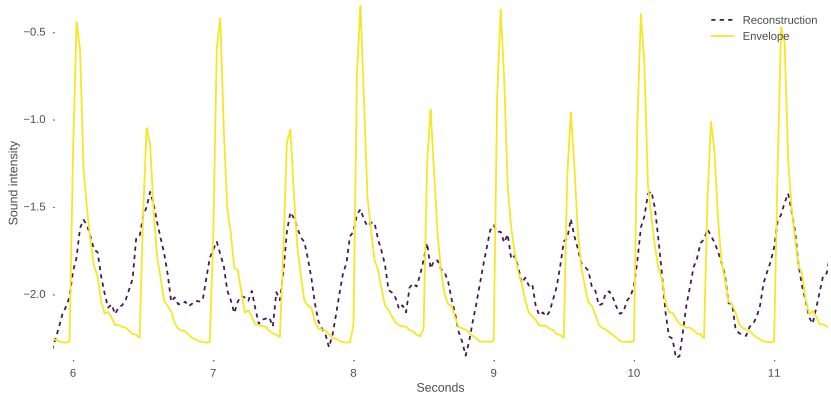


Figure 3.5: Example of original (yellow line) drum envelope and reconstruction (purple line) based on ECoG signals. Figure modified from [26].

3.4.2 Direct Speech Synthesis from ECoG

This section summarizes our findings from *"Towards direct speech synthesis from ECoG: A pilot study"* [9] (See Appendix A.10). The basic approach is explained and some results highlighted.

To enable a locked-in patient to converse with family and friends again in a natural way, we present a direct synthesis approach that transforms recorded ECoG data into an audio waveform. Contrary to the *Brain-to-Text* system, the neural data is not decoded into a sequence of words but directly mapped to audible output. This type of neuroprosthesis would allow the user to produce speech via ECoG with the possibility to stress and emphasize certain parts of the words or even create completely new words that are not predefined in a dictionary.

In our pilot study, we analyzed the data of one ECoG patient with 16 electrodes covering inferior regions of the frontal cortex and some coverage of the temporal lobe. Despite the low number of electrodes and suboptimal montage - no motor areas and only very few electrodes on auditory regions - we could show that our method is feasible.

Similar to the *Brain-to-Text* experiment, we recorded audio data and ECoG recordings simultaneously. In this experiment however, the participants had to repeat a sentence that was presented to them both visually (on a computer screen) and audibly.

Figure 3.6 illustrates the synthesis approach. We record ECoG data and audio signals synchronously. ECoG gamma activity is extracted and feature vectors are stacked to capture context information and temporal offsets in neural processing. Each individual ECoG feature vectors is transferred to an audio spectrum using Lasso regression models. Here, each spectral coefficient is created from a specific multivariate regression model from the neural data. These regression models for each individual spectral bin are trained previously on the distinct training data. Lasso regularization coefficients are determined in a nested 10-fold cross-validation. We use Lasso regression models instead of non-regularized regression models to prevent overfitting in these high-dimensional datasets. The reconstructed spectra from each feature vector can then be combined to form a reconstructed spectrogram. The reconstructed spectrogram created this way does not contain phase information, as this cannot be

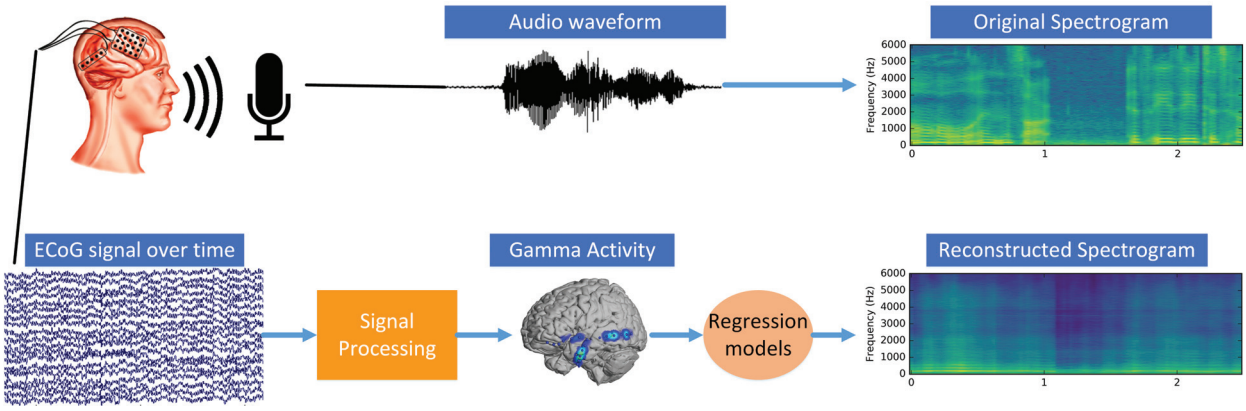


Figure 3.6: Direct speech synthesis from ECoG using regression models. The reconstructed magnitude spectrogram can be resynthesized to an audio waveform.

reconstructed from neural signals. Griffin and Lim [Griffin and Lim, 1984] proposed Algorithm 1 to reconstruct the waveform from the spectrogram without phase information. Given the spectrogram f reconstructed from the measured ECoG activity, one can reconstruct an audio waveform by iteratively modifying the spectral coefficients of a signal initialized with noise.

Algorithm 1: Waveform reconstruction

Data: Spectrogram f
Result: Waveform w
$w \leftarrow$ noise;
for $i \leftarrow 1$ **to** l **do**
  $X \leftarrow \mathrm{STFT}(w)$;
  $Z \leftarrow f \exp(i \angle X)$;
  $w \leftarrow \mathrm{ISTFT}(Z)$;

Here l refers to the number of iterations the procedure is repeated for. STFT and ISTFT are the Short-Term Fourier Transform and the Inverse Short-Term Fourier Transform. For our pilot study, we chose a value of $l = 8$, as no improvement could be achieved with more iterations and computation is still very fast for 8 iterations. Algorithm 1 can be used both on a complete pre-computed spectrogram in offline reconstruction or on isolated frames for online synthesis. Even though our analysis for this study was performed offline, all processing steps are fast enough to be used in real-time. In fact, we successfully used the method for online synthesis from EMG [10].

The reconstruction of spectrograms or a waveform is not a standard classification approach and can thus not be evaluated using standard accuracy metrics. To evaluate our reconstruction, we look at the Spearman correlation ρ of the spectral power over time for each spectral coefficient between the original and reconstructed spectrograms. Figure 3.7 shows Spearman correlations between original and reconstructed spectrogram for each spectral coefficient (purple). Rank correlations above 200 Hz are significantly better than chance (evaluated using randomization tests) and reach a level of 0.4 around 300 Hz. As the first

formant of vowels can usually be found in this frequency range, good reconstruction is especially important in this range. Correlations remain high up to 5 kHz. Above 5 kHz only little speech information can be found. The mean overall correlation over all frequency bins is $\rho = 0.36$, this is comparable to correlations achieved in [Martin et al., 2014].

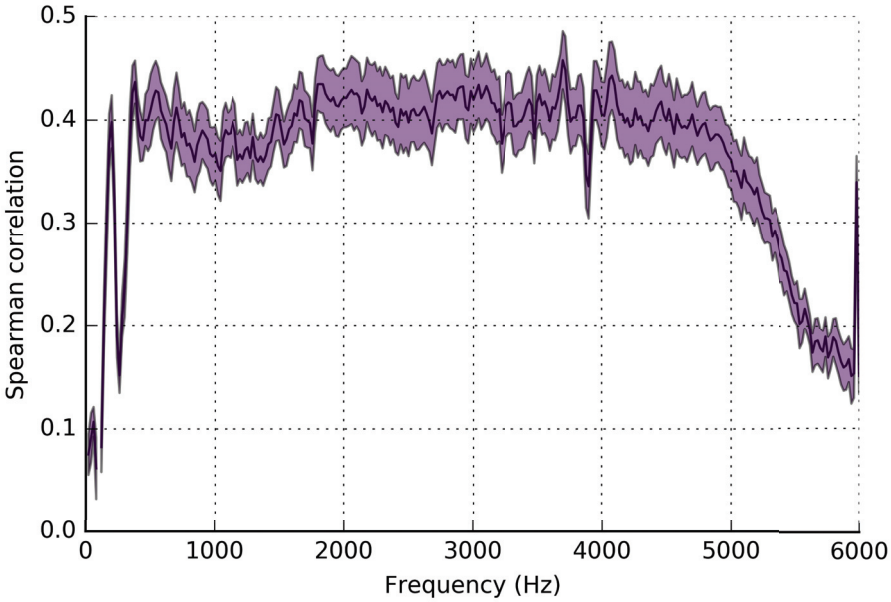


Figure 3.7: Spearman correlations between original and reconstructed spectrograms for each spectral coefficient. Shaded region shows standard error of the mean over the 10-fold cross-validation. Figure used from [9].

Our results extend the previously demonstrated reconstruction of spectrograms [Martin et al., 2014] as we show how to resynthesize the spectrogram to an audio waveform. To evaluate this last step in the processing pipeline, we calculate rank correlations between the mean absolute Hilbert envelope of the original and reconstructed waveform. This yielded a Spearman correlation of $\rho = 0.41$, which is significantly better than chance (randomization tests). It is important to note that a large proportion of this correlation is due to the correct discrimination of speech from silence with high spectral energy during speech and low spectral energy during silence.

Of course, the ultimate evaluation is to judge the produced audio through listening tests. For our pilot study participant, the reconstructed audio is not intelligible yet. This might be due to suboptimal electrode montage and the very low number of electrodes. Despite the fact that further improvements are necessary, our study has presented the first approach creating an audio waveform directly from neural signals, which could enable patients to converse with family and friends again.

3.4.3 Advanced Regression Models

Since the recording of the pilot participant described in the previous section, another three participants were recorded. In this section, we validate our approach on the newly recorded

data and compare the Lasso regression approach to DNN, which have shown promising results for spectral reconstruction in ECoG [Yang et al., 2015] and EMG based speech synthesis [Diener et al., 2015].

Instead of using the magnitude spectrogram used in the pilot study, we use logarithmic mel scaled spectrograms which should better represent acoustic information perceived by the listeners. Logarithmic mel spectrograms are extracted by taking the magnitude spectrogram and mapping it onto the mel scale using triangular overlapping filter banks. The mel (melody) scale was first introduced by [Stevens et al., 1937] and represents pitches in perceptually equal distances as rated by listeners instead of frequencies. It should therefore better represent the neural activity in the auditory cortex. The resulting mel coefficients are then logarithmized to make their distribution more Gaussian. We use 23 triangular filters in this study to represent most of the information in the 300 frequency bins in the previous study. This decreases the model complexity necessary to map ECoG features to spectral information by a factor of 13. Mel-based features like Mel-frequency cepstral coefficients (MFCCs) [Davis and Mermelstein, 1980] are widely used in ASR and speech synthesis as they drastically reduce the feature space while still containing most relevant information.

We compare the reconstruction performance of the previously used Lasso regression to two different neural networks. The first evaluated DNN is a Feed-Forward DNN (which we denote as FF). We use three fully connected hidden layers with rectified linear units [Nair and Hinton, 2010]. Rectified Linear Units (often called ReLU) lead to less computational complexity in the backpropagation step and usually result in sparse activations for the neurons, as many inputs can be zero. A dropout with probability of 0.5 was applied [Srivastava et al., 2014] to help generalization. Models were initialized using a Gaussian distribution to break symmetry [Glorot and Bengio, 2010]. A linear decreasing learning rate between 0.01 and 0.0001 over 200 epochs was applied. See the very informative guide to gradient-based deep learning [Bengio, 2012] for practical hints on DNN usage.

We also evaluated the usage of Long Short-Term Memory (LSTM) neural networks [Hochreiter and Schmidhuber, 1997, Greff et al., 2015a], which should be well suited for our system, as they remember the last states of the decoding process and thereby incorporate context information which is very important in speech, as samples are not independent of each other. LSTMs have been successfully applied to a variety of time-series including speech recognition [Graves et al., 2013], lip-reading [Wand et al., 2016] and other applications such as language modeling [Sundermeyer et al., 2012] and text-to-speech systems [Fan et al., 2014, van den Oord et al., 2016].

For our LSTM model, we use a fully connected feed-forward hidden layer followed by an LSTM layer with *tanh* activation function. Again, a dropout with probability 0.5 was included. Our LSTM model was trained for 300 epochs with a linear decreasing learning rate between 0.05 and 0.001 over the first 200 epochs. Weights were initialized following a uniform distribution between -0.05 and 0.05. Both neural networks were build using the Brainstorm framework [Greff et al., 2015b].

All results are evaluated in a 5-fold cross-validation, further dividing the training data into training and development sets. For this comparison, the Lasso approach was also used to reconstruct mel scaled spectrograms. Figure 3.8 shows results of our reconstruction experiments in terms of Spearman correlation coefficients.

The first thing to note in the results are the chance level results for participant 2 and the first two sessions of participant 4. Participant 2 only had bilateral temporal depth electrodes

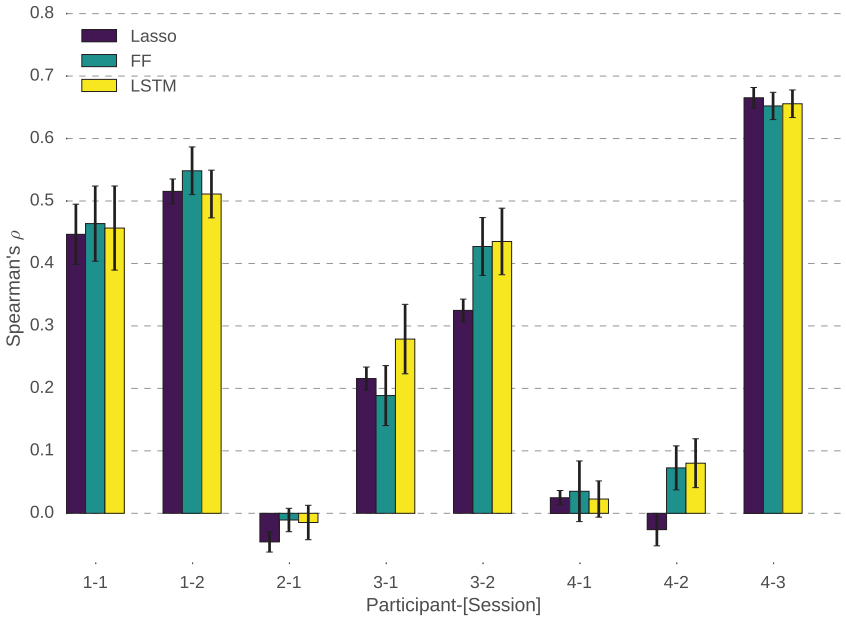


Figure 3.8: Mean Spearman correlation ρ for all sessions of all participants between original and reconstructed mel scaled spectrograms. We compare Lasso regression (Lasso), Feed-forward (FF) and Long Short-Term Memory (LSTM) neural networks .

which seem to have not measured signals from any areas involved in speech production. In the first two sessions of participant 4, one amplifier did not work properly and 16 electrodes covering perisylvian areas did not record any data, resulting in no usable information for these sessions. These results also highlight some of the disadvantages of ECoG data obtained during epilepsy surgery for speech investigations. Electrode montages are purely determined by clinical needs and might thus be not ideal for the targeted investigation. Especially with the current trend towards depth electrodes, the coverage of speech relevant areas gets more rare. While depth electrodes allow for very interesting investigations of deeper brain structures such as hippocampus and insula [Krusienski and Shih, 2011, Shih and Krusienski, 2012], they yield little information for the investigation and decoding of speech processes. Another disadvantage is the sparse sampling by ECoG electrodes. This on the one hand provides very localized information, but on the other hand a lot of crucial information are omitted. A suboptimal electrode coverage thus immediately has no task relevant information. Volume conduction, as appearing in EEG and MEG thus has advantages too, as relevant information can be measured when not ideally located, even if it is blurred by signals from other sources.

Results for the other sessions show reasonable reconstruction performance with Spearman correlations above 0.2. For the higher-performing sessions of participant 1 and session 3 of participant 4 no significant difference between the performance of the three regression approaches can be observed (Wilcoxon signed-rank test, $p > 0.05$). For session 1 of participant 3, LSTM yields better results than both FF and Lasso and for session 2 of participant 3 both FF and LSTM outperform the Lasso approach. Improvements of up to 0.1 can

be observed. Overall, no consistent improvement can be observed for the DNN. At the same time, the far more time-consuming training process and necessary parameter tuning are large disadvantages of these approaches. Another large disadvantage is the lack of interpretable models in deep learning. While some advances have been made to interpret deep learning models for digit recognition [Erhan et al., 2009], EEG-based BCI [Sturm et al., 2016] and EMG-based speech recognition [Wand and Schultz, 2014], it is currently not possible to interpret what the model has learned. Interpretability is especially useful when working with neural signals to verify the physiological basis of the results or even increase our understanding of neural processes. Figure 3.9 visualizes forward models [Haufe et al., 2014] corresponding to the Lasso backward ward models learned in our pilot study (see Section 3.4.2). It can be seen that most reconstruction power originates from perisylvian areas associated with auditory processing, validating that indeed neural activity is decoded and reconstruction does not rely on artifacts.

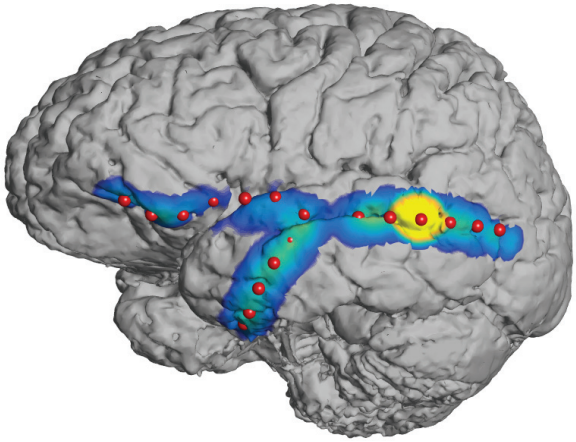


Figure 3.9: Average activation pattern of Lasso regression models for spectrogram reconstruction. Hot colors (red) mark areas with highly speech related activity. Figure used from [9].

We hypothesize that DNN cannot show their full potentials for this task as training data is very limited and due to the very noisy signals. Alternative regression models that are promising for this type of problem are gradient boosted regression trees [Friedman, 2001, Friedman, 2002], which were recently found to be very successful in a variety of machine learning problems while having the advantage of interpretable models. Another candidate for good spectrum reconstruction are Kalman Filter [Kalman, 1960] based approaches that are the standard for movement prediction from invasive recording on motor cortex [Wu et al., 2006].

Our comparison of interpretable linear models to DNN show that with the current data, no robust improvements can be obtained using the more complex DNN, which at the same time result in the loss of interpretability.

3.5 Contributions of the Corresponding Publications

The high spatial and temporal resolution combined with unaffectedness by motion artifacts make ECoG an ideal candidate for the investigation of speech production. In our publication

"Brain-to-Text: Decoding spoken phrases from phone representations in the brain", we have shown for the first time that phones can be modeled in neural data accurately enough to be decoded in continuous speech. These results provide an important step towards BCIs based on speech processes. The textual representation of a spoken phrase which we decode from ECoG could be used to issue commands to a computer or to compose messages or texts. Our approach could therefore be used for very natural and fast BCIs based on speech production.

Instead of producing a textual representation of speech, a device targeting to enable patients to converse again would offer greater expressive power by directly producing audible speech. As a first step towards this goal, we reconstruct simple repetitive audio stimuli. In our abstract *"Music rhythm reconstruction from ECoG"* we lay the foundation of envelope reconstruction from perceived rhythms from ECoG signals. This is a first step towards the more complex reconstruction of speech from neural signals. Our pilot study *"Towards direct speech synthesis from ECoG: A pilot study"* demonstrated for the first time that an audio waveform can be reconstructed from measured ECoG signals during speech. This is a alternative approach to speech recognition, which is potentially more natural for communication but not as adequate for control of computers and devices. Using the synthesis approach, BCIs could be developed that enable patients to interact with friends and family in a natural manner again.

Chapter 4

Towards Imagined Speech Processes in BCI

Despite the promising results achieved with audibly produced speech, the transfer to imagined speech processes is not self-evident. While a plethora of studies show that overtly performed and imagined movement yield similar activations in motor areas of the brain [Decety, 1996, Jeannerod, 1994, Sitaram et al., 2007b, McFarland et al., 2000, Schnitzler et al., 1997, Pfurtscheller and Neuper, 1997, Pfurtscheller and Neuper, 2001, Pfurtscheller et al., 2006, Roth et al., 1996, Decety et al., 1994], the same is not necessarily true for overt and imagined speech. One could hypothesize that motor activations are similar between overt and covert speech, but activity in the auditory cortex is distinctly different whether a person perceives there own voice or not [Pei et al., 2011b]. The differences in brain activity pattern between audibly produced and imagined speech need to be accounted for to be able to use the same techniques used for audible speech for imagined speech. Another problem is the lack of control over the participants' performance in the imagined condition. This condition does not provide the experimenter with any control neither on the speed of imagined speech production nor on whether the participant is performing the task at all. In this chapter, we will discuss the specifics of imagined speech and investigate a common representation of imagined speech between participants. A hypothesis which cortex areas might yield similar activation pattern for imagined and audible speech is developed by looking at discriminability between phones in different cortical areas and temporal offsets. The identified subregions are then tested for their results when used to decode continuous speech from neural signals to simulate the decoding of imagined speech.

4.1 Imagined Speech Processes

The classification and detection of imagined speech processes has been studied intensively using invasive recordings. In ECoG recordings, one-dimensional cursor control using imagined phone production versus a resting state was shown in [Leuthardt et al., 2011]. Extending these results, [Pei et al., 2011a] decoded limited pairings of vowels and consonants in imagined and audibly spoken words. Covert articulation of single vowels was decoded in [Ikeda et al., 2014]. Above chance-level discrimination between pairing of imagined word production was presented in [Martin et al., 2016a]. Instead of asking able-bodied participants to imagine speech production, [Brumberg et al., 2011] classified attempted phoneme production in a completely paralyzed participant with a Neurotrophic device [Bartels et al., 2008] implanted in left precentral gyrus yielding two channels of recorded activity. Transferring some of these results to the non-invasive EEG, [Deng et al., 2010] decoded three

different rhythms of imagined syllables. [Yoshimura et al., 2016] discriminated between two imagined vowels using EEG.

These initial results give hope that continuous imagined speech production might be decoded from ECoG. The following sections will investigate the concept of imagined speech further and present initial result towards solving this challenging problem.

4.2 Common Representation of Imagined Speech

This section looks at selected results from our study *"Cross-subject classification of speaking modes using fNIRS"* [23] (See Appendix A.11) and embeds them within the scope of this dissertation.

In this section we interpret results from one of our fNIRS studies to see whether the instruction to imagine speech production yields consistent hemodynamic responses across participants on a sentence level. The consistency across hemodynamic responses can give interesting insights into the representation of imagined speech processes. To explore whether a common representation of imagined speech between different persons exists, we conducted another study on the different speaking modes from our fNIRS experiments [22, 27]. To identify the common representation, we tried to classify the speaking modes across participants. All data processing was performed as in the previously described papers, but instead of training recognition models for each participants, we tried to build a recognition model on 4 of the 5 participants and classify the data of the last remaining persons. This process was repeated until every person was used for testing once, resulting in a leave-one-out cross-validation. Results of this approach can be found in Figure 4.1. The results indicate that audibly spoken speech (AUD) and silently mouthed speech (SIL) have a common representation across participants resulting in better than chance accuracies when discriminating from Pause. Imagined speech seems to have no such common representation among the 5 recorded participants in this study, resulting in chance level classification of IMG from Pause for all 5 participants.

This small analysis is an important indicator on the degree of difficulty of the measurement of imagined speech. Despite the fact that all participants were instructed in the same way to imagine themselves reading out the displayed sentences, no common neural representation could be found in the fNIRS data, while such common pattern seem to be present for audibly produced and silently uttered speech. This is not altogether clear in the first place either, as speaking rate and mother tongue varied greatly between participants. It seems that speaking audibly has clear neural markers which cannot easily be detected in imagined speech. It is therefore very important to create instructions for the participants that help them to easily imagine speech production.

4.3 Brain Areas Involved in the Speech Process

The results from this chapter are taken from our *"Brain-to-Text: Decoding spoken phrases from phone representations in the brain"* publication [4] (See Appendix A.8).

To design models that might be transferable from audibly produced to imagined speech, it is important to understand all cortical areas involved in speech production. The feature vec-

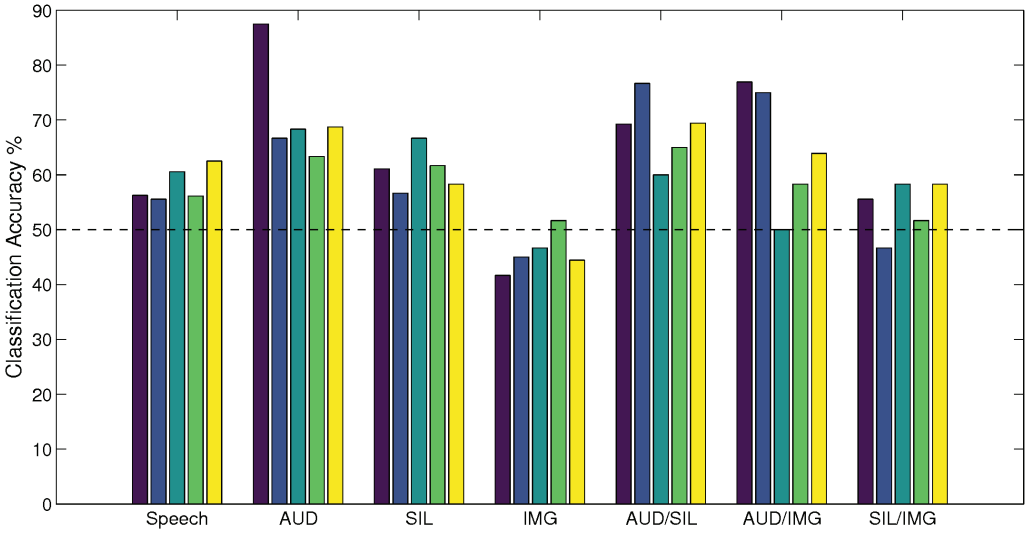


Figure 4.1: Classification accuracies for cross-participant classification of speaking modes against PAUSE and against each other. Speech denotes the three speaking modes combined. Each color represents accuracies achieved on one participant with models trained on the remaining four participants. Dotted line indicates naive classification accuracy. Figure modified from [23].

tors extracted for our *Brain-to-Text* system [4] (See Section 3.3) enable an in-depth analysis of the cortical areas and their temporal dynamics during continuous speech production. We identified regions and timings of high relevance by calculating the mean symmetrized Kullback-Leibler divergence [KULLBACK and LEIBLER, 1951] between all phone models for each temporal offset and at every electrode position. The Kullback-Leibler divergence measures the difference between two distributions P and Q and can be interpreted as the amount of extra bits needed to code samples from P when using Q as a code. We interpret the Kullback-Leibler divergence between phone models at an electrode location with a temporal offset as the amount of discriminability between phones in bits. The assumption that the logarithmic broadband gamma power features are approximately Gaussian allows us to employ a closed form calculation of the Kullback-Leibler divergence. The Kullback-Leibler divergence between normal distributed P and Q can easily be calculated as:

$$D_{KL}(N_P||N_Q) = \frac{1}{2}(tr(\Sigma_Q^{-1}\Sigma_P) + (\mu_Q - \mu_P)^T\Sigma_Q^{-1}(\mu_Q - \mu_P) - d - log_2(\frac{det(\Sigma_P)}{det(\Sigma_Q)}) \quad (4.1)$$

with d being the dimensionality of the distributions. Here, μ_P and μ_Q are the means and Σ_P and Σ_Q the covariances of distributions P and Q, respectively. For an electrode position and time interval we can now estimate the discriminability of the corresponding feature through the mean Kullback-Leibler divergence between all pairs of phone models for this particular feature. As the Kullback-Leibler divergence is non-symmetric, the divergence between two phones has to be calculated in both directions. The mean between all these divergences then symmetrizes the divergences and results in one average discriminability in bits between the phone models.

Figure 4.2 illustrates the spatial and temporal dynamics of the mean discriminabilities of phone models on an averaged brain (Talairach model [Talairach and Tournoux, 1988]). A total of 9 temporal offsets - 4 prior to and 4 after the currently produced phone - are visualized for all co-registered electrodes of all participants. Visualized discriminabilities exceed 99% of randomized mean Kullback-Leibler divergences. These randomized Kullback-Leibler divergences were obtained by shifting the data by half its length while leaving the phone labels in place.

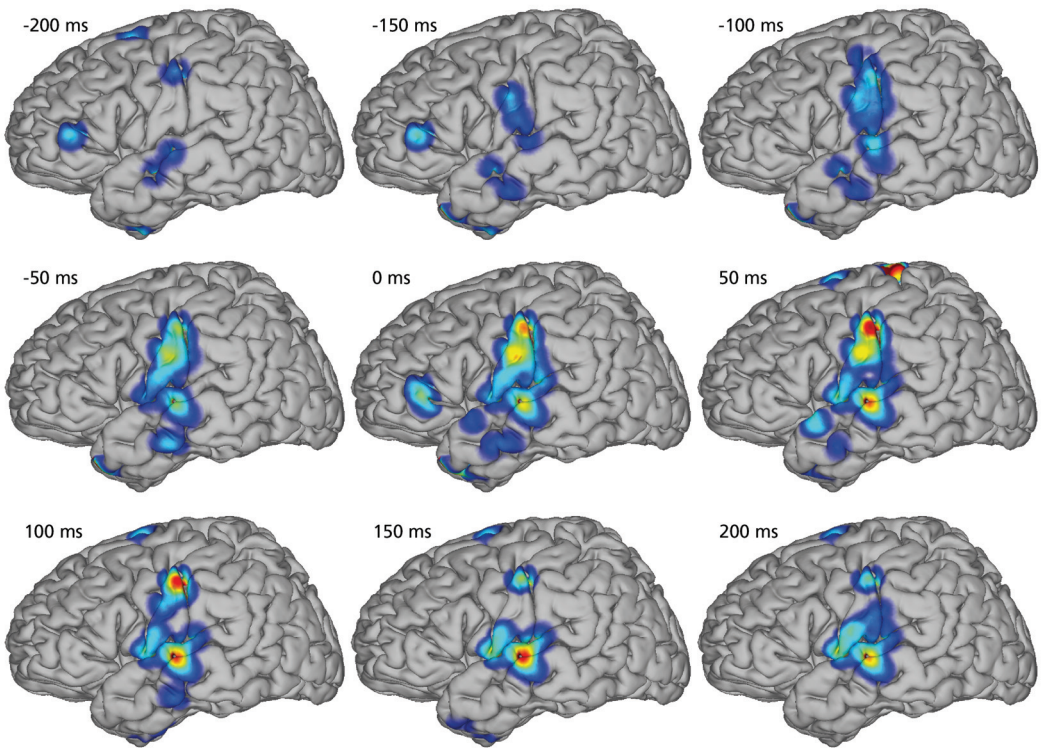


Figure 4.2: Mean Kullback-Leibler divergence for each temporal offset and electrode position of all participants. Heat maps on rendered average brain show regions of high discriminability (red). Only discriminabilities exceeding chance level are plotted. Figure used from [4] and created using [Kubanek and Schalk, 2014].

Earliest discriminability can be observed 200 ms prior to phone production in various areas including inferior frontal gyrus, which contains Broca's area. Broca's area is generally associated with planning of speech production [Broca, 1861, Sahin et al., 2009]. While discriminability in Broca's area remains high 150 ms prior to phone production, additional areas in the primary somatosensory cortex and primary motor cortex and regions in the temporal gyrus show increasing discriminability. Drawing closer to current phone production, discriminability in the inferior frontal gyrus vanished while discriminability in motor areas increases until peaking concurrent with actual phone production. Auditory areas in superior

temporal gyrus show increasing discriminability after phone production peaking 150 ms afterwards. These findings confirm the temporal dynamics that had been previously identified in traditional neuroscience studies [Penfield and Roberts, 2014, Binder et al., 1997].

The identified cortical regions illustrate the wide-spread network underlying speech production. Following the ideas from imagined movement, some of the identified areas might show similar activations for audibly produced and imagined speech. Other areas however will yield distinctly different activations whether participants hear their own voice or not. Especially the regions in the superior temporal gyrus, associated with auditory processing and speech understanding, will have different activation pattern when no auditory stimulus is present. The next section will investigate how well decoding can be performed with only parts of the cortical areas involved in the speech process to simulate the absence of auditory excitation in imagined speech.

4.4 Decoding Speech from Productive Brain Areas

This section presents and extends findings from our book chapter *"Towards continuous speech recognition for BCI"* [1] (See Appendix A.12).

To simulate imagined speech, we used the data from our *Brain-to-Text* study and limited the analysis to certain temporal offsets. This way, only certain types of information can be used in the decoding process and the absence of perceptive or productive information can be simulated. We used the same preprocessing and decoding approach as in the original study, but instead of using all temporal offsets, we looked at two different temporal subsets. As the same feature selection based on Kullback-Leibler divergences is applied, only features showing high discriminability will be selected (see regions of high discriminability in Figure 4.2). First, we used only temporal offsets prior to and concurrent to actual phone production (temporal offsets -200, -150, -100, -50 and 0 ms) to simulate imagined speech production, as auditory regions do not show discriminability in these temporal offsets, they will not be selected for other temporal offsets either. Therefore, no information from auditory perception of the participants' own voice is present in this analysis, as the auditory perception starts after the phone has been produced. We call this analysis *Production only*. As a comparison and extending [1], we also analyzed how well decoding could be performed when only the remaining temporal offsets (50 ms to 150 ms) were used. This analysis is termed *Perception only* and shows how well continuous speech can be decoded from the perceptive information of hearing one's own voice. Figure 4.3 shows phone accuracies for the complete *Brain-to-Text* system, *Production only* and *Perception only* and compares them to randomized results as a baseline. Purple bars show results for the complete *Brain-to-Text* system which combine productive and perceptive temporal offsets. They are better than chance level (randomized models, yellow bars) for all participants and sessions. The simulated imagined speech condition (*Production only*) outperforms chance level for all sessions of all participants, but is considerably worse than the full system. Especially for the outstanding performances achieved for participant 7, the *Production only* results are very promising with very high accuracies. The dark blue bars represent results for the perceptive temporal offsets (*Perception only*). *Perception only* outperforms the chance levels and *Production only* for almost all participants and yields results almost reaching the level of the combination of all temporal offsets.

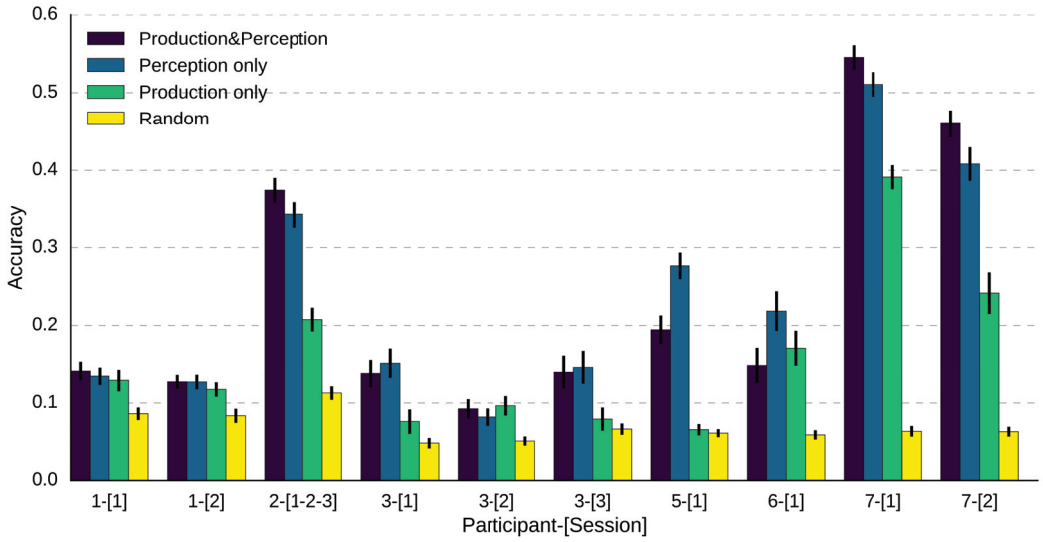


Figure 4.3: Average phone classification accuracies for all sessions and participants. Error bars show standard error of the mean.

These findings illustrate that while speech decoding from productive areas alone is possible, the knowledge from the perceptive regions improves the decoding process a great deal. The surprising results for a few sessions that perception only outperforms the combination of all temporal offsets can be explained by suboptimal automatic feature selection, which selects features until no sufficient improvement in Kullback-Leibler divergence is found anymore.

To investigate the results in more detail, we looked at Word Error Rates and Average True Positive Rates for phones. Figure 4.4 shows Word Error Rates and Average Phone True Positive Rates over dictionary size for session 1 of participant 7. As in the accuracy results, the combination of productive and perceptive temporal offsets yields the best results with lowest WER and highest average true positive rates. *Perception only* yields slightly lower results but outperforms *Production only* for all dictionary sizes. All results are better than randomized results. While *Production only* yields higher WER than *Perception only*, it still results in convincing WER of just under 40% for a dictionary of 10 words. Despite the fact that improvements are still necessary for practical applications, these results indicate that *Production only*, which might have similar activation pattern for audible and imagined speech, contains enough information to decode continuous speech from ECoG recordings.

These results support the hypothesis that speech imagery is a viable candidate for BCI. Due to the problematic acquisition of a ground truth, we have not investigated imagined speech, yet. A textual representation of the spoken words could be decoded from the neural data even when neural activity associated with speech perception of the participants' own voice was omitted, which simulates imagined speech recognition from neural activity. This is an important step towards realistic BCIs based on speech processes, as only imagined speech combines the advantages of BCIs and speech.

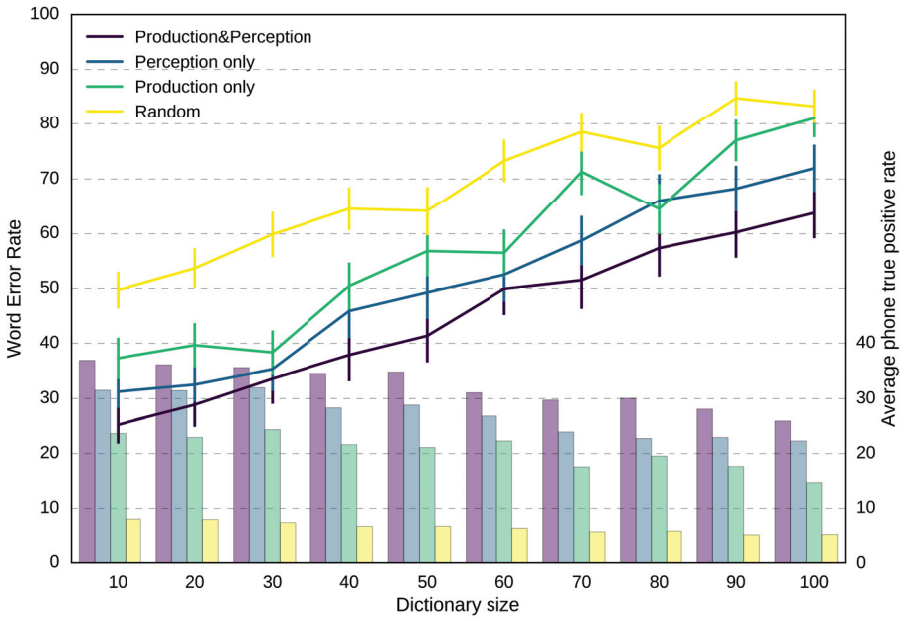


Figure 4.4: Detailed results for participant 7, session 1. Lines show Word Error Rates over dictionary size. Bars show average true positive rates for phone recognition in the extracted decoding path.

4.5 Contributions of the Corresponding Publications

In our fNIRS study *"Cross-subject classification of speaking modes using fNIRS"* we investigated the consistency of hemodynamic responses across participants. We could show that audible speech and silently mouthed speech had a common representation across participants and could be classified without the need for participant specific training data. However, no such common representation for imagined speech was found, highlighting difficulties in building BCI based on imagined speech.

The illustration of discriminative temporal offsets and cortical regions from our *"Brain-to-Text: Decoding spoken phrases from phone representations in the brain"* paper identifies regions which are likely involved in speech planning and motor movement and regions and offsets responsible for speech perception of the participant's own voice. These illustration allow us to investigate the neural basis of speech production and hypothesize which areas might show similar activation during imagined speech.

To simulate BCI based on imagined speech, our book chapter *"Towards continuous speech recognition for BCI"* excludes perceptive temporal offsets of brain activity and shows that word decoding can successfully be performed on productive temporal offsets only. This is an important prerequisite for BCI based on imagined speech.

Chapter 5

Conclusion

This dissertation presents substantial contributions to BCIs based on speech processes. The focus is on two very different brain activity measurement techniques. With fNIRS, we developed speech based BCIs that are usable in realistic scenarios and that are affordable. With ECoG, we measured the best possible signals for neuroprostheses based on speech.

In fNIRS, we have shown that different types of speech production, namely normally articulated speech, silently uttered speech and imagined speech can reliably be discriminated from periods without speech activity and from each other. This extends previous findings in averaged brain activity which could show increases in average hemodynamic activity associated with speech. Using these different styles of speech production, a BCI could be designed that uses the speaking styles as directional commands or which enables a more sophisticated, but disruptive system, which for example uses the combination of EEG and fNIRS. The slow nature of the hemodynamic response will always limit the information transfer rates of fNIRS-based BCI for communication and control. We therefore looked at applications in which the slow nature of fNIRS was unproblematic, such as tutoring systems and presented several approaches to detect and classify user states from fNIRS, which can be used to adapt interfaces to a user's psychological condition.

With the ideal signal properties of ECoG, we could present the first system extracting a textual representation of continuous speech from neural signals. The successful application of automatic speech recognition technology to ECoG data is a big leap towards BCIs using speech.Using our technology, brain activity could directly be transformed into written text for control of a device or to compose messages. To further improve the naturalness of potential prosthetic devices, we present a direct speech synthesis approach which generates an audio waveform from ECoG data. This approach would be ideal for patients suffering from speech pathologies or being in a locked-in state, as they could communicate with family and friends through the neuroprosthetic device.

Obviously, decoding audibly articulated speech from neural signals is only a first step towards BCIs based on speech activity, as a BCI would need to be based on imagined speech processes. We investigate the neural substrate of speech production to identify regions that might show similar activations during imagined speech production. As continuous imagined speech is currently too difficult to be investigated due to the lack of ground truth, we simulated imagined speech and present preliminary results. Results achieved in this simulation underline that imagined speech might possibly be decoded from continuous speech in the future. The design of better experiments to investigate imagined speech processes further and tackle the decoding of imagined speech are important steps for future research.

Besides the important insights these investigations give for BCIs and speech neuroprostheses it is also important to note that our classification approach is very helpful for neuroscientific research, as well. Single trial analysis as utilized in BCI and *Brain-to-Text* yield resilient results without the need to aggregate large cohorts. Generative models are easily interpretable while simultaneously granting important insights into the underlying brain functions, without typical statistical problems associated with large numbers of variables [Eklund et al., 2016].

Even with the promising results achieved in this dissertation in decoding speech process from neural data, it is very important to note that thoughts remain a person's property and the demonstrated techniques are not targeting thought decoding. A clear distinction has to be drawn between speech processes and inner voice or thought. While speech processes, including imagined speech, include the actual or imagined movement of tongue, larynx and lips, inner voice does not involve these processes, but contains the representation of the meaning of the words. The semantic representation of the meaning of different words is far more wide spread in the cortex [Huth et al., 2016] than the process of producing speech utilized in this dissertation.

Overall, this dissertation has presented an effective alternative approach to current BCI paradigms which is more natural, while potentially providing much higher information transfer rates. The work presented in this dissertations lays the foundation for the development of future invasive and non-invasive speech neuroprosthetics.

List of Publications by the Author

Reviewed Publications

Publications printed in boldface fall within the scope of this dissertation and are attached in the Appendix A.

Book chapters

[1] Christian Herff, Adriana de Pesters, Dominic Heger, Peter Brunner, Gerwin Schalk, and Tanja Schultz. **Towards continuous speech recognition for BCI**. In Christoph Guger, Brendan Z. Allison, and Junichi Ushiba, editors, *Brain-computer interface research: a state-of-the-art summary 5*. Springer, 2017.

My share: I conducted all data analysis and wrote the manuscript.

Scientific journals

[2] Christian Herff and Tanja Schultz. **Automatic Speech Recognition from Neural Signals: A Focused Review**. *Frontiers in Neuroscience*, 10(429), 2016.

My share: I wrote this review article.

[3] Alexander von Lühmann, Christian Herff, Dominic Heger, and Tanja Schultz. Towards a wireless open source instrument:functional Near-Infrared Spectroscopy in mobile neuroergonomics and BCI applications. *Frontiers in Human Neuroscience*, 9(617), 2015.

My share: This study resulted from a master thesis I supervised. I helped to write the manuscript and conducted the BCI evaluation.

[4] Christian Herff, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. **Brain-to-text: Decoding spoken phrases from phone representations in the brain**. *Frontiers in Neuroscience*, 9(217), 2015.

My share: This paper is joint first-authorship with Dominic Heger. We conducted all data analysis together and wrote the manuscript.

[5] Felix Putze, Sebastian Hesslinger, Chun-Yu Tse, YunYing Huang, Christian Herff, Cuntai Guan, and Tanja Schultz. Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Frontiers in Neuroscience*, 8(373), 2014.

My share: I only had a very minor share in this publication by assisting in fNIRS data analysis and commenting on the manuscript.

[6] Christian Herff, Dominic Heger, Ole Fortmann, Johannes Hennrich, Felix Putze, and Tanja Schultz. **Mental workload during n-back task - quantified in the prefrontal cortex using fNIRS**. *Frontiers in Human Neuroscience*, 7(935), 2014.

My share: I designed the experiment, analyzed the data and wrote the manuscript for this study.

[7] Dominic Heger, Christian Herff, Felix Putze, Reinhard Mutter, and Tanja Schultz. Continuous affective states recognition using functional near infrared spectroscopy. *Brain-Computer Interfaces*, 1(2):113–125, 2014.

My share: I helped with fNIRS data analysis and commented on the manuscript.

Conference proceedings

[8] Jochen Weiner, Christian Herff, and Tanja Schultz. Speech-Based Detection of Alzheimer's Disease in Conversational German. In *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.

My share: I only had a very minor share in this publication by helping with data classification.

[9] Christian Herff, Garett Johnson, Lorenz Diener, Jerry Shih, Dean Krusienski, and Tanja Schultz. **Towards direct speech synthesis from ECoG: A pilot study**. In *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

My share: I performed all data analysis and wrote the manuscript.

[10] Lorenz Diener, Christian Herff, Matthias Janke, and Tanja Schultz. An Initial Investigation into the Real-Time Conversion of Facial Surface EMG Signals to Audible Speech. In *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

My share: I wrote the synthesis code and helping with the manuscript.

[11] Christian Herff, Ole Fortmann, Chun-Yu Tse, Xiaoqin Cheng, Felix Putze, Dominic Heger, and T. Schultz. **Hybrid fNIRS-EEG based discrimination of 5 levels of memory load**. In *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*, pages 5–8, April 2015.

My share: This study resulted from a bachelor thesis I supervised. I wrote the manuscript and extended the data analysis.

[12] Johannes Hennrich, Christian Herff, Dominic Heger, and Tanja Schultz. **Investigating Deep Learning for fNIRS Based BCI**. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, Aug 2015.

My share: This study was created from a bachelor thesis I supervised. Additionally, I helped to write the manuscript.

[13] Dominic Heger, Christian Herff, Felix Putze, and Tanja Schultz. Joint optimization for discriminative, compact and robust Brain-Computer Interfacing. In *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*, pages 82–85, April 2015.

My share: I only had a very minor share in this publication and commented on the manuscript.

[14] Dominic Heger, Christian Herff, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Continuous Speech Recognition from ECoG. In *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015.

My share: I provided the baseline system for this study and helped to write the manuscript.

[15] Dominic Telaar, Michael Wand, Dirk Gehrig, Felix Putze, Christoph Amma, Dominic Heger, Ngoc Thang Vu, Mark Erhardt, Tim Schlippe, Matthias Janke, Christian Herff, and Tanja Schultz. BioKIT - Real-time Decoder For Biosignal Processing. In *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014.

My share: I only had a very minor share in this publication by helping with the manuscript.

[16] Dominic Heger, Christian Herff, and Tanja Schultz. Combining feature extraction and classification for fNIRS BCIs by regularized least squares optimization. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2012–2015, Aug 2014.

My share: I provided the baseline results and helped to write the manuscript.

[17] Felix Putze, Jutta Hild, Rainer Kärgel, Christian Herff, Alexander Redmann, Jürgen Beyerer, and Tanja Schultz. Locating User Attention Using Eye Tracking and EEG for Spatio-temporal Event Selection. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 129–136, New York, NY, USA, 2013. ACM.

My share: I only had a very minor share in this publication by helping with the manuscript and conducting some experiments on EOG regression for cursor positions.

[18] Christian Herff, Dominic Heger, Felix Putze, Johannes Hennrich, Ole Fortmann, and Tanja Schultz. **Classification of mental tasks in the prefrontal cortex using fNIRS**. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 2160–2163, July 2013.

My share: I designed the experiment, analyzed the data and wrote the manuscript for this study.

[19] Christian Herff, Dominic Heger, Felix Putze, Cuntai Guan, and Tanja Schultz. **Self-paced BCI with NIRS based on speech activity**. In *International BCI Meeting 2013, Asilomar, USA*, 2013.

My share: I conducted all data analysis and wrote the manuscript for this study.

[20] Dominic Heger, Felix Putze, Christian Herff, and Tanja Schultz. Subject-to-subject transfer for CSP based BCIs: Feature space transformation and decision-level fusion. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5614–5617, July 2013.

My share: I only had a very minor share in this publication by helping with the manuscript.

[21] Dominic Heger, Reinhard Mutter, Christian Herff, Felix Putze, and Tanja Schultz. Continuous Recognition of Affective States by Functional Near Infrared Spectroscopy Signals. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 832–837, Sept 2013.

My share: I helped with fNIRS data analysis for this study.

[22] Christian Herff, Felix Putze, Dominic Heger, Cuntai Guan, and Tanja Schultz. **Speaking mode recognition from functional Near Infrared Spectroscopy**. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 1715–1718, Aug 2012.

My share: I designed the experiment and collected the data. I performed all data analysis and wrote the manuscript for this study.

[23] Christian Herff, Dominic Heger, Felix Putze, Cuntai Guan, and Tanja Schultz. **Cross-Subject Classification of Speaking Modes Using fNIRS**. In Tingwen Huang, Zhigang Zeng, Chuandong Li, and ChiSing Leung, editors, *Neural Information Processing*, volume 7664 of *Lecture Notes in Computer Science*, pages 417–424. Springer Berlin Heidelberg, 2012.

My share: I performed all data analysis and wrote the manuscript.

[24] Christian Herff, Matthias Janke, Michael Wand, and Tanja Schultz. Impact of Different Feedback Mechanisms in EMG-based Speech Recognition. In *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association*, 2011.

My share: I recorded the data for this study, performed a few of the experiments and wrote parts of the manuscript.

[25] Jan Niehues, Mohammed Mediani, Teresa Herrmann, Michael Heck, Christian Herff, and Alex Waibel. The KIT Translation System for IWSLT 2010. In *International Workshop on Spoken Language Translation (IWSLT) 2010*, 2010.

My share: I developed German-English translation system and wrote the corresponding section in the manuscript.

Peer-reviewed abstracts

[26] Christian Herff, Garett Johnson, Jerry Shih, Tanja Schultz, and Dean Krusienski. **Music rhythm reconstruction from ECoG**. In *International BCI Meeting 2016, Asilomar, USA*, 2016.

My share: I conducted the data analysis and wrote the manuscript.

[27] Christian Herff, Dominic Heger, Felix Putze, Cuntai Guan, and Tanja Schultz. **Self-paced BCI with NIRS based on speech activity**. In *International BCI Meeting 2013, Asilomar, USA*, 2013.

My share: I conducted all data analysis and wrote the manuscript for this study.

[28] Dominic Heger, Christian Herff, Felix Putze, and Tanja Schultz. Towards Biometric Person Identification using fNIRS. In *International BCI Meeting 2013, Asilomar, USA*, 2013.

My share: I helped with fNIRS data analysis.

References

[AAV] *Amazon Alexa Voice Search*. https://developer.amazon.com/alexa. Accessed: 2016-10-13.

[AS] *Apple Siri*. http://www.apple.com/ios/siri/. Accessed: 2016-10-04.

[GVS] *Google Voice Search*. https://support.google.com/websearch/answer/2940021?co=GENIE.Platform%3DAndroid&hl=en. Accessed: 2016-10-04.

[Int] *Intendix by g.tec*. http://www.intendix.com/. Accessed: 2016-10-04.

[Ahn et al., 2016] Ahn, S., T. Nguyen, H. Jang, J. Kim and S. Jun (2016). *Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and fNIRS data*. Frontiers in human neuroscience, 10.

[Akay and Daubenspeck, 1999] Akay, M. and J. Daubenspeck (1999). *Investigating the contamination of electroencephalograms by facial muscle electromyographic activity using matching pursuit*. Brain and language, 66(1):184–200.

[Ang et al., 2008] Ang, K.K., Z. Chin, H. Zhang and C. Guan (2008). *Filter bank common spatial pattern (FBCSP) in brain-computer interface*. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pp. 2390–2397. IEEE.

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

[Ang et al., 2010b] Ang, K.K., C. Guan, K. Lee, J. Lee, S. Nioka and B. Chance (2010b). *Application of rough set-based neuro-fuzzy system in NIRS-based BCI for assessing numerical cognition in classroom*. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–7. IEEE.

[Ang et al., 2014] Ang, K.K., J. Yu and C. Guan (2014). *Single-trial classification of NIRS data from prefrontal cortex during working memory tasks*. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2008–2011. IEEE.

[Ayaz et al., 2007] Ayaz, H., M. Izzetoglu, S. Bunce, T. Heiman-Patterson and B. Onaral (2007). *Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy*. In *Neural Engineering, 2007. CNE '07. 3rd International IEEE/EMBS Conference on*, pp. 342–345.

[Ayaz et al., 2013] Ayaz, H., B. Onaral, K. Izzetoglu, P. Shewokis, R. McKendrick and R. Parasuraman (2013). *Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development*. Frontiers in Human Neuroscience, 7:871.

[Ayaz et al., 2012] Ayaz, H., P. Shewokis, S. Bunce, K. Izzetoglu, B. Willems and B. Onaral (2012). *Optical brain monitoring for operator training and mental workload assessment*. NeuroImage, 59(1):36 – 47.

[Ayaz et al., 2010] Ayaz, H., B. Willems, B. Bunce, P. Shewokis, K. Izzetoglu, S. Hah, A. Deshmukh and B. Onaral (2010). *Cognitive workload assessment of air traffic controllers using optical brain imaging sensors*. Advances in understanding human performance: neuroergonomics, human factors design, and special populations, pp. 21–31.

[Bartels et al., 2008] Bartels, J., D. Andreasen, P. Ehirim, H. Mao, S. Seibert, E. Wright and P. Kennedy (2008). *Neurotrophic electrode: method of assembly and implantation into human motor speech cortex*. Journal of neuroscience methods, 174(2):168–176.

[Bartoshuk et al., 1960] Bartoshuk, L.M., M. Harned and L. Parks (1960). *Mental rotation of three-dimensional objects*. Anim. Behav, 8:54.

[Battiti, 1994] Battiti, R. (1994). *Using mutual information for selecting features in supervised neural net learning*. IEEE Transactions on neural networks, 5(4):537–550.

[Bauernfeind et al., 2014] Bauernfeind, G., D. Steyrl, C. Brunner and G. Mueller-Putz (2014). *Single trial classification of fNIRS-based brain-computer interface mental arithmetic data: A comparison between different classifiers*. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 2004–2007. IEEE.

[Bengio, 2009] Bengio, Y. (2009). *Learning deep architectures for AI*. Foundations and trends® in Machine Learning, 2(1):1–127.

[Bengio, 2012] Bengio, Y. (2012). *Practical recommendations for gradient-based training of deep architectures*. In *Neural Networks: Tricks of the Trade*, pp. 437–478. Springer.

[Benyon and Murray, 1993] Benyon, D. and D. Murray (1993). *Adaptive systems: from intelligent tutoring to autonomous agents*. Knowledge-Based Systems, 6(4):197 – 219.

[Berka et al., 2007] Berka, C., D. Levendowski, M. Lumicao, A. Yau, G. Davis, V. Zivkovic, R. Olmstead, P. Tremoulet and P. Craven (2007). *EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks*. Aviation, Space, and Environmental Medicine, 78(5):B231–B244.

[Bin et al., 2011] Bin, G., X. Gao, Y. Wang, Y. Li, B. Hong and S. Gao (2011). *A high-speed BCI based on code modulation VEP*. Journal of neural engineering, 8(2):025015.

[Bin et al., 2009] Bin, G., X. Gao, Z. Yan, B. Hong and S. Gao (2009). *An online multi-channel SSVEP-based brain–computer interface using a canonical correlation analysis method*. Journal of neural engineering, 6(4):046002.

[Binder et al., 1997] Binder, J.R., J. Frost, T. Hammeke, R. Cox, S. Rao and T. Prieto (1997). *Human brain language areas identified by functional magnetic resonance imaging*. The Journal of Neuroscience, 17(1):353–362.

[Birbaumer et al., 1999] Birbaumer, N., N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub and H. Flor (1999). *A spelling device for the paralysed*. Nature, 398(6725):297–298.

[Blakely et al., 2008] Blakely, T., K. Miller, R. Rao, M. Holmes and J. Ojemann (2008). *Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids*. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 4964–4967. IEEE.

[Blankertz et al., 2008] Blankertz, B., R. Tomioka, S. Lemm, M. Kawanabe and K.-R. Müller (2008). *Optimizing spatial filters for robust EEG single-trial analysis*. IEEE Signal Processing Magazine, 25(1):41–56.

[Bocquelet et al., 2015] Bocquelet, F., T. Hueber, L. Girin, C. Savariaux, B. Yvert et al. (2015). *Real-time Control of a DNN-based Articulatory Synthesizer for Silent Speech Conversion: a pilot study*. In *Interspeech 2015 (16th Annual Conference of the International Speech Communication Association)*, pp. 2405–2409.

[Bortfeld et al., 2009] Bortfeld, H., E. Fava and D. Boas (2009). *Identifying cortical lateralization of speech processing in infants using near-infrared spectroscopy*. Developmental neuropsychology, 34(1):52–65.

[Bortfeld et al., 2007] Bortfeld, H., E. Wruck and D. Boas (2007). *Assessing infants' cortical response to speech using near-infrared spectroscopy*. Neuroimage, 34(1):407–415.

[Bouchard and Chang, 2014] Bouchard, K.E. and E. Chang (2014). *Neural Decoding of Spoken Vowels from Human Sensory-Motor Cortex with High-Density Electrocorticography*. In *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE.

[Bouchard et al., 2016] Bouchard, K.E., D. Conant, G. Anumanchipalli, B. Dichter, K. Chaisanguanthum, K. Johnson and E. Chang (2016). *High-Resolution, Non-Invasive Imaging of Upper Vocal Tract Articulators Compatible with Human Brain Recordings*. PloS one, 11(3):e0151327.

[Bouchard et al., 2013] Bouchard, K.E., N. Mesgarani, K. Johnson and E. Chang (2013). *Functional organization of human sensorimotor cortex for speech articulation*. Nature, 495(7441):327–332.

[Broca, 1861] Broca, P. (1861). *Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau*. Bull Soc Anthropol, 2(1):235–238.

[Brouwer and van Erp, 2010] Brouwer, A.-M. and J. van Erp (2010). *A tactile P300 brain-computer interface*. Frontiers in neuroscience, 4:19.

[Brouwer et al., 2012] Brouwer, A.-M., M. Hogervorst, J. van Erp, T. Heffelaar, P. Zimmerman and R. Oostenveld (2012). *Estimating workload using EEG spectral power and ERPs in the n-back task*. Journal of Neural Engineering, 9(4):045008.

[Brown, 1998] Brown, C.M. (1998). *Human-computer interface design guidelines*. Intellect Books.

[Brumberg et al., 2010] Brumberg, J.S., A. Nieto-Castanon, P. Kennedy and F. Guenther (2010). *Brain–computer interfaces for speech communication*. Speech communication, 52(4):367–379.

[Brumberg et al., 2011] Brumberg, J.S., E. Wright, D. Andreasen, F. Guenther and P. Kennedy (2011). *Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex*. Frontiers in neuroscience, 5.

[Brunner et al., 2009] Brunner, P., A. Ritaccio, T. Lynch, J. Emrich, J. Wilson, J. Williams, E. Aarnoutse, N. Ramsey, E. Leuthardt, H. Bischof et al. (2009). *A practical procedure for real-time functional mapping of eloquent cortex using electrocorticographic signals in humans*. Epilepsy & Behavior, 15(3):278–286.

[Çakir et al., 2016] Çakir, M.P., M. Vural, S. Koç and A. Toktaş (2016). *Real-Time Monitoring of Cognitive Workload of Airline Pilots in a Flight Simulator with fNIR Optical Brain Imaging Technology*. In *International Conference on Augmented Cognition*, pp. 147–158. Springer.

[Cannestra et al., 2003] Cannestra, A.F., I. Wartenburger, H. Obrig, A. Villringer and A. Toga (2003). *Functional assessment of Broca's area using near infrared spectroscopy in humans*. Neuroreport, 14(15):1961–1965.

[Canolty et al., 2007] Canolty, R.T., M. Soltani, S. Dalal, E. Edwards, N. Dronkers, S. Nagarajan, H. Kirsch, N. Barbaro and R. Knight (2007). *Spatiotemporal dynamics of word processing in the human brain*. Frontiers in neuroscience, 1:14.

[Chakrabarti et al., 2015] Chakrabarti, S., H. Sandberg, J. Brumberg and D. Krusienski (2015). *Progress in speech decoding from the electrocorticogram*. Biomedical Engineering Letters, 5(1):10–21.

[Chang et al., 2010] Chang, E.F., J. Rieger, K. Johnson, M. Berger, N. Barbaro and R. Knight (2010). *Categorical speech representation in human superior temporal gyrus*. Nature neuroscience, 13(11):1428–1432.

[Chen et al., 2015a] Chen, L.-C., P. Sandmann, J. Thorne, M. Bleichner and S. Debener (2015a). *Cross-modal functional reorganization of visual and auditory cortex in adult cochlear implant users identified with fNIRS*. Neural plasticity, 2016.

[Chen et al., 2015b] Chen, X., Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung and S. Gao (2015b). *High-speed spelling with a noninvasive brain–computer interface*. Proceedings of the National Academy of Sciences, 112(44):E6058–E6067.

[Coffey et al., 2012] Coffey, E.B.J., A.-M. Brouwer and J. van Erp (2012). *Measuring workload using a combination of electroencephalography and near infrared spectroscopy*. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, pp. 1822–1826. SAGE Publications.

[Coon et al., 2016] Coon, W.G., A. Gunduz, P. Brunner, A. Ritaccio, B. Pesaran and G. Schalk (2016). *Oscillatory phase modulates the timing of neuronal activations and resulting behavior*. NeuroImage, 133:294 – 301.

[Cooper et al., 2012] Cooper, R., J. Selb, L. Gagnon, D. Phillip, H. Schytz, H. Iversen, M. Ashina and D. Boas (2012). *A Systematic Comparison of Motion Artifact Correction Techniques for Functional Near-Infrared Spectroscopy*. Frontiers in Neuroscience, 6(147).

[Cooper et al., 1965] Cooper, R., A. Winter, H. Crow and W. Walter (1965). *Comparison of subcortical, cortical and scalp activity using chronically indwelling electrodes in man*. Electroencephalography and clinical neurophysiology, 18(3):217–228.

[Coyle et al., 2007] Coyle, S.M., T. Ward and C. Markham (2007). *Brain–computer interface using a simplified functional near-infrared spectroscopy system*. Journal of neural engineering, 4(3):219.

[Crone et al., 2001] Crone, N.E., D. Boatman, B. Gordon and L. Hao (2001). *Induced electrocorticographic gamma activity during auditory perception*. Clinical Neurophysiology, 112(4):565–582.

[Crone et al., 1998] Crone, N.E., D. Miglioretti, B. Gordon, J. Sieracki, M. Wilson, S. Uematsu and R. Lesser (1998). *Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization.*. Brain, 121(12):2271–2299.

[Crone et al., 2006] Crone, N.E., A. Sinai and A. Korzeniewska (2006). *High-frequency gamma oscillations and human brain mapping with electrocorticography*. Progress in brain research, 159:275–295.

[Cui et al., 2010] Cui, X., S. Bray and A. Reiss (2010). *Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics*. NeuroImage, 49(4):3039–46.

[Cutrell and Tan, 2008] Cutrell, E. and D. Tan (2008). *BCI for passive input in HCI*. In *Proceedings of CHI*, vol. 8, pp. 1–3. Citeseer.

[Davis and Mermelstein, 1980] Davis, S. and P. Mermelstein (1980). *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. IEEE transactions on acoustics, speech, and signal processing, 28(4):357–366.

[Debener et al., 2015] Debener, S., R. Emkes, M. De Vos and M. Bleichner (2015). *Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear*. Scientific reports, 5.

[Decety, 1996] Decety, J. (1996). *Neural representations for action*. Reviews in the Neurosciences, 7(4):285–297.

[Decety et al., 1994] Decety, J., D. Peranif, M. Jeannerod, V. Bettinardif, B. Tadary and R. Woods (1994). *Mapping motor representations with positron emission*. Nature, 371:13.

[Denby et al., 2010] Denby, B., T. Schultz, K. Honda, T. Hueber, J. Gilbert and J. Brumberg (2010). *Silent speech interfaces*. Speech Communication, 52(4):270–287.

[Deng et al., 2010] Deng, S., R. Srinivasan, T. Lappas and M. D'Zmura (2010). *EEG classification of imagined syllable rhythm using Hilbert spectrum methods*. Journal of neural engineering, 7(4):046006.

[Derosière et al., 2013] Derosière, G., K. Mandrick, G. Dray, T. Ward and S. Perrey (2013). *NIRS-measured prefrontal cortex activity in neuroergonomics: strengths and weaknesses*. Frontiers in human neuroscience, 7:583.

[Di Liberto et al., 2015] Di Liberto, G.M., J. O'Sullivan and E. Lalor (2015). *Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing*. Current Biology, 25(19):2457–2465.

[Dichter et al., 2016] Dichter, B.K., K. Bouchard and E. Chang (2016). *Dynamic Structure of Neural Variability in the Cortical Representation of Speech Sounds*. The Journal of Neuroscience, 36(28):7453–7463.

[Dieler et al., 2012] Dieler, A.C., S. Tupak and A. Fallgatter (2012). *Functional near-infrared spectroscopy for the assessment of speech related tasks*. Brain and language, 121(2):90–109.

[Diener et al., 2015] Diener, L., M. Janke and T. Schultz (2015). *Direct conversion from facial myoelectric signals to speech using Deep Neural Networks*. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE.

[Donchin et al., 2000] Donchin, E., K. Spencer and R. Wijesinghe (2000). *The mental prosthesis: assessing the speed of a P300-based brain-computer interface*. Rehabilitation Engineering, IEEE Transactions on, 8(2):174–179.

[Edwards et al., 2005] Edwards, E., M. Soltani, L. Deouell, M. Berger and R. Knight (2005). *High gamma activity in response to deviant auditory stimuli recorded directly from human cortex*. Journal of Neurophysiology, 94(6):4269–4280.

[Edwards et al., 2009] Edwards, E., M. Soltani, W. Kim, S. Dalal, S. Nagarajan, M. Berger and R. Knight (2009). *Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex*. Journal of neurophysiology, 102(1):377–386.

[Eklund et al., 2016] Eklund, A., T. Nichols and H. Knutsson (2016). *Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates*. Proceedings of the National Academy of Sciences, 113(28):7900–7905.

[Erhan et al., 2009] Erhan, D., Y. Bengio, A. Courville and P. Vincent (2009). *Visualizing higher-layer features of a deep network*. University of Montreal, 1341.

[Fabiani et al., 2004] Fabiani, G.E., D. McFarland, J. Wolpaw and G. Pfurtscheller (2004). *Conversion of EEG activity into cursor movement by a brain-computer interface (BCI)*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 12(3):331–338.

[Fagan et al., 2008] Fagan, M.J., S. Ell, J. Gilbert, E. Sarrazin and P. Chapman (2008). *Development of a (silent) speech recognition system for patients following laryngectomy*. Medical engineering & physics, 30(4):419–425.

[Fairclough, 2009] Fairclough, S.H. (2009). *Fundamentals of physiological computing*. Interacting with computers, 21(1):133–145.

[Fallgatter et al., 1998] Fallgatter, A.J., T. Müller and W. Strik (1998). *Prefrontal hypooxygenation during language processing assessed with near-infrared spectroscopy*. Neuropsychobiology, 37(4):215–218.

[Fan et al., 2014] Fan, Y., Y. Qian, F.-L. Xie and F. Soong (2014). *TTS synthesis with bidirectional LSTM based recurrent neural networks.*. In *Interspeech*, pp. 1964–1968.

[Farwell and Donchin, 1988] Farwell, L.A. and E. Donchin (1988). *Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials*. Electroencephalography and clinical Neurophysiology, 70(6):510–523.

[Fazli et al., 2012] Fazli, S., J. Mehnert, J. Steinbrink, G. Curio, A. Villringer, K.-R. Müller and B. Blankertz (2012). *Enhanced performance by a hybrid NIRS–EEG brain computer interface*. Neuroimage, 59(1):519–529.

[Ferrari et al., 2004] Ferrari, M., L. Mottola and V. Quaresima (2004). *Principles, techniques, and limitations of near infrared spectroscopy*. Canadian journal of applied physiology, 29(4):463–487.

[Flesher et al., 2016] Flesher, S.N., J. Collinger, S. Foldes, J. Weiss, J. Downey, E. Tyler-Kabara, S. Bensmaia, A. Schwartz, M. Boninger and R. Gaunt (2016). *Intracortical microstimulation of human somatosensory cortex*. Science Translational Medicine.

[Frey et al., 2014] Frey, J., C. Mühl, F. Lotte, M. Hachet et al. (2014). *Review of the Use of Electroencephalography as an Evaluation Method for Human-Computer Interaction*. In *PhyCS 2014-International Conference on Physiological Computing Systems*.

[Friedman, 2001] Friedman, J.H. (2001). *Greedy function approximation: a gradient boosting machine*. Annals of statistics, pp. 1189–1232.

[Friedman, 2002] Friedman, J.H. (2002). *Stochastic gradient boosting*. Computational Statistics & Data Analysis, 38(4):367–378.

[Friedrich et al., 2012] Friedrich, E.V.C., R. Scherer and C. Neuper (2012). *The effect of distinct mental strategies on classification performance for brain-computer interfaces*. International Journal of Psychophysiology, 84(1):86 – 94.

[Fuster, 1988] Fuster, J.M. (1988). *Prefrontal cortex*. In *Comparative Neuroscience and Neurobiology*, pp. 107–109. Springer.

[Gallagher et al., 2007] Gallagher, A., M. Thériault, E. Maclin, K. Low, G. Gratton, M. Fabiani, L. Gagnon, K. Valois, I. Rouleau, H. Sauerwein et al. (2007). *Near-infrared spectroscopy as an alternative to the Wada test for language mapping in children, adults and special populations*. Epileptic Disorders, 9(3):241–255.

[Gan et al., 2003] Gan, Z., C. Li, H. Gong, Q. Luo, B. Yao, R. Song and H. Wu (2003). *On children's dyslexia with NIRS*. In *Third International Conference on Photonics and Imaging in Biology and Medicine*, pp. 521–525. International Society for Optics and Photonics.

[Gasser et al., 1982] Gasser, T., P. Bächer and J. Möcks (1982). *Transformations towards the normal distribution of broad band spectral parameters of the EEG*. Electroencephalography and clinical neurophysiology, 53(1):119–124.

[Gevins et al., 1994] Gevins, A., B. Cutillo, J. Desmond, M. Ward, S. Bressler, N. Barbero and K. Laxer (1994). *Subdural grid recordings of distributed neocortical networks involved with somatosensory discrimination*. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 92(4):282–290.

[Glorot and Bengio, 2010] Glorot, X. and Y. Bengio (2010). *Understanding the difficulty of training deep feedforward neural networks*. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256.

[Gonzalez et al., 2016] Gonzalez, J.A., L. Cheah, J. Gilbert, J. Bai, S. Ell, P. Green and R. Moore (2016). *A silent speech system based on permanent magnet articulography and direct synthesis*. Computer Speech & Language, 39:67–87.

[Graves et al., 2013] Graves, A., N. Jaitly and A.-R. Mohamed (2013). *Hybrid speech recognition with deep bidirectional LSTM*. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 273–278. IEEE.

[Greff et al., 2015a] Greff, K., R. Srivastava, J. Koutník, B. Steunebrink and J. Schmidhuber (2015a). *LSTM: A search space odyssey*. arXiv preprint arXiv:1503.04069.

[Greff et al., 2015b] Greff, K., R. Srivastava and J. Schmidhuber (2015b). *Brainstorm: Fast, Flexible and Fun Neural Networks, Version 0.5*.

[Griffin and Lim, 1984] Griffin, D.W. and J. Lim (1984). *Signal estimation from modified short-time Fourier transform*. Acoustics, Speech and Signal Processing, IEEE Transactions on, 32(2):236–243.

[Guenther et al., 2009] Guenther, F.H., J. Brumberg, E. Wright, A. Nieto-Castanon, J. Tourville, M. Panko, R. Law, S. Siebert, J. Bartels, D. Andreasen et al. (2009). *A wireless brain-machine interface for real-time speech synthesis*. PloS one, 4(12):e8218.

[Guger et al., 2009] Guger, C., S. Daban, E. Sellers, C. Holzner, G. Krausz, R. Carabalona, F. Gramatica and G. Edlinger (2009). *How many people are able to control a P300-based brain-computer interface (BCI)?*. Neuroscience Letters, 462(1):94 – 98.

[Guimaraes et al., 2007] Guimaraes, M.P., D. Wong, E. Uy, L. Grosenick and P. Suppes (2007). *Single-trial classification of MEG recordings*. IEEE Transactions on Biomedical Engineering, 54(3):436–443.

[Hart and Staveland, 1988] Hart, S.G. and L. Staveland (1988). *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*. Advances in psychology, 52:139–183.

[Haufe et al., 2014] Haufe, S., F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz and F. Biessmann (2014). *On the interpretation of weight vectors of linear models in multivariate neuroimaging*. Neuroimage, 87:96–110.

[Hermes et al., 2014] Hermes, D., K. Miller, M. Vansteensel, E. Edwards, C. Ferrier, M. Bleichner, P. van Rijen, E. Aarnoutse and N. Ramsey (2014). *Cortical theta wanes for language*. Neuroimage, 85:738–748.

[Herrmann et al., 2003] Herrmann, M.J., A.-C. Ehlis and A. Fallgatter (2003). *Frontal activation during a verbal-fluency task as measured by near-infrared spectroscopy*. Brain Research Bulletin, 61(1):51–56.

[Herrmann et al., 2005] Herrmann, M.J., A.-C. Ehlis, P. Scheuerpflug and A. Fallgatter (2005). *Optical topography with near-infrared spectroscopy during a verbal-fluency task*. Journal of Psychophysiology, 19(2):100–105.

[Higashi et al., 2011] Higashi, H., T. Rutkowski, Y. Washizawa, A. Cichocki and T. Tanaka (2011). *EEG auditory steady state responses classification for the novel BCI*. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4576–4579. IEEE.

[Hill and Schölkopf, 2012] Hill, N.J. and B. Schölkopf (2012). *An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli*. Journal of neural engineering, 9(2):026011.

[Hinton, 2012] Hinton, G.E. (2012). *A practical guide to training restricted boltzmann machines*. In *Neural Networks: Tricks of the Trade*, pp. 599–619. Springer.

[Hinton et al., 2012] Hinton, G.E., L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath et al. (2012). *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*. IEEE Signal Processing Magazine, 29(6):82–97.

[Hinton and Salakhutdinov, 2006] Hinton, G.E. and R. Salakhutdinov (2006). *Reducing the dimensionality of data with neural networks*. Science, 313(5786):504–507.

[Hirshfield et al., 2011] Hirshfield, L.M., R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, R. Ward, T. Williams and R. Jacob (2011). *This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 373–382. ACM.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and J. Schmidhuber (1997). *Long short-term memory*. Neural computation, 9(8):1735–1780.

[Hogervorst et al., 2014] Hogervorst, M.A., A.-M. Brouwer and J. van Erp (2014). *Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload*. Frontiers in neuroscience, 8.

[Homae et al., 2006] Homae, F., H. Watanabe, T. Nakano, K. Asakawa and G. Taga (2006). *The right hemisphere of sleeping infant perceives sentential prosody*. Neuroscience research, 54(4):276–280.

[Horovitz and Gore, 2004] Horovitz, S.G. and J. Gore (2004). *Simultaneous event-related potential and near-infrared spectroscopic studies of semantic processing*. Human brain mapping, 22(2):110–115.

[Hoshi and Tamura, 1997] Hoshi, Y. and M. Tamura (1997). *Near-infrared optical detection of sequential brain activation in the prefrontal cortex during mental tasks*. NeuroImage, 5(4):292–297.

[Hoshi et al., 2003] Hoshi, Y., B. Tsou, V. Billock, M. Tanosaki, Y. Iguchi, M. Shimada, T. Shinba, Y. Yamada and I. Oda (2003). *Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks*. NeuroImage, 20(3):1493 – 1504.

[Hueber et al., 2007] Hueber, T., G. Aversano, G. Cholle, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel and M. Stone (2007). *Eigentongue feature extraction for an ultrasound-based silent speech interface*. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1, pp. I–1245. IEEE.

[Huggins et al., 2011] Huggins, J.E., P. Wren and K. Gruis (2011). *What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis*. Amyotrophic lateral sclerosis, 12(5):318–324.

[Hull et al., 2009] Hull, R., H. Bortfeld and S. Koons (2009). *Near-infrared spectroscopy and cortical responses to speech production*. The Open Neuroimaging Journal, 3(1).

[Huppert et al., 2009] Huppert, T.J., S. Diamond, M. Franceschini and D. Boas (2009). *HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain*. Applied optics, 48(10):D280–D298.

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

[Ikeda et al., 2014] Ikeda, S., T. Shibata, N. Nakano, R. Okada, N. Tsuyuguchi, K. Ikeda and A. Kato (2014). *Neural decoding of single vowels during covert articulation using electrocorticography*. Frontiers in human neuroscience, 8.

[IPA, 1999] IPA, International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

[Issa et al., 2016] Issa, M., S. Bisconti, I. Kovelman, P. Kileny and G. Basura (2016). *Human Auditory and Adjacent Nonauditory Cerebral Cortices Are Hypermetabolic in Tinnitus as Measured by Functional Near-Infrared Spectroscopy (fNIRS)*. Neural plasticity, 2016.

[Izzetoglu et al., 2015] Izzetoglu, K., H. Ayaz, J. Hing, P. Shewokis, S. Bunce, P. Oh and B. Onaral (2015). *UAV operators workload assessment by optical brain imaging technology (fNIR)*. In *Handbook of Unmanned Aerial Vehicles*, pp. 2475–2500. Springer.

[Izzetoglu et al., 2003] Izzetoglu, K., S. Bunce, M. Izzetoglu, B. Onaral and K. Pourrezaei (2003). *fNIR spectroscopy as a measure of cognitive task load*. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 4, pp. 3431–3434 Vol.4.

[Janke et al., 2012] Janke, M., M. Wand, K. Nakamura and T. Schultz (2012). *Further investigations on EMG-to-speech conversion*. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 365–368. IEEE.

[Jarvis et al., 2011] Jarvis, J., F. Putze, D. Heger and T. Schultz (2011). *Multimodal person independent recognition of workload related biosignal patterns*. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pp. 205–208, New York, NY, USA. ACM.

[Jasper, 1958] Jasper, H.H. (1958). *The ten twenty electrode system of the international federation*. Electroencephalography and clinical neurophysiology, 10:371–375.

[Jeannerod, 1994] Jeannerod, M. (1994). *The representing brain: Neural correlates of motor intention and imagery*. Behavioral and Brain sciences, 17(02):187–202.

[Jelinek, 1997] Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

[Jobsis, 1977] Jobsis, F.F. (1977). *Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters*. Science, 198(4323):1264–1267.

[Kakimoto et al., 2009] Kakimoto, Y., Y. Nishimura, N. Hara, M. Okada, H. Tanii and Y. Okazaki (2009). *Intrasubject reproducibility of prefrontal cortex activities during a verbal fluency task over two repeated sessions using multi-channel near-infrared spectroscopy*. Psychiatry and clinical neurosciences, 63(4):491–499.

[Kalman, 1960] Kalman, R.E. (1960). *A new approach to linear filtering and prediction problems*. Journal of basic Engineering, 82(1):35–45.

[Kalyuga and Singh, 2015] Kalyuga, S. and A.-M. Singh (2015). *Rethinking the boundaries of cognitive load theory in complex learning*. Educational Psychology Review, pp. 1–22.

[Kane et al., 2007] Kane, M.J., A. Conway, T. Miura and G. Colflesh (2007). *Working memory, attention control, and the N-back task: a question of construct validity.*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3):615.

[Kaplan, 2001] Kaplan, S. (2001). *Meditation, restoration, and the management of mental fatigue*. Environment and Behavior, 33(4):480–506.

[Kellis et al., 2010] Kellis, S., K. Miller, K. Thomson, R. Brown, P. House and B. Greger (2010). *Decoding spoken words using local field potentials recorded from the cortical surface*. Journal of neural engineering, 7(5):056007.

[Khan and Hong, 2015] Khan, M.J. and K.-S. Hong (2015). *Passive BCI based on drowsiness detection: an fNIRS study*. Biomedical optics express, 6(10):4063–4078.

[Khan et al., 2016] Khan, M.J., X. Liu, M. Bhutta and K.-S. Hong (2016). *Drowsiness detection using fNIRS in different time windows for a passive BCI*. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pp. 227–231. IEEE.

[Kirchner, 1958] Kirchner, W.K. (1958). *Age differences in short-term retention of rapidly changing information.*. Journal of experimental psychology, 55(4):352.

[Klatt, 1987] Klatt, D.H. (1987). *Review of text-to-speech conversion for English*. The Journal of the Acoustical Society of America, 82(3):737–793.

[Klein et al., 1995] Klein, D., B. Milner, R. Zatorre, E. Meyer and A. Evans (1995). *The neural substrates underlying word generation: a bilingual functional-imaging study.*. Proceedings of the National Academy of Sciences, 92(7):2899–2903.

[Knecht et al., 2000] Knecht, S., B. Dräger, M. Deppe, L. Bobe, H. Lohmann, A. Flöel, E.-B. Ringelstein and H. Henningsen (2000). *Handedness and hemispheric language dominance in healthy humans*. Brain, 123(12):2512–2518.

[Kothe and Makeig, 2011] Kothe, C.A. and S. Makeig (2011). *Estimation of task workload from EEG data: New and current tools and perspectives*. In *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, pp. 6547–6551.

[Krizhevsky et al., 2012] Krizhevsky, A., I. Sutskever and G. Hinton (2012). *Imagenet classification with deep convolutional neural networks*. In *Advances in neural information processing systems*, pp. 1097–1105.

[Krusienski et al., 2008] Krusienski, D.J., E. Sellers, D. McFarland, T. Vaughan and J. Wolpaw (2008). *Toward enhanced P300 speller performance*. Journal of neuroscience methods, 167(1):15–21.

[Krusienski and Shih, 2011] Krusienski, D.J. and J. Shih (2011). *Control of a brain–computer interface using stereotactic depth electrodes in and adjacent to the hippocampus*. Journal of neural engineering, 8(2):025006.

[Kubanek et al., 2013] Kubanek, J., P. Brunner, A. Gunduz, D. Poeppel and G. Schalk (2013). *The tracking of speech envelope in the human cortex*. PloS one, 8(1):e53398.

[Kubanek and Schalk, 2014] Kubanek, J. and G. Schalk (2014). *NeuralAct: A Tool to Visualize Electrocortical (ECoG) Activity on a Three-Dimensional Model of the Cortex*. Neuroinformatics, pp. 1–8.

[Kullback and Leibler, 1951] Kullback, S. and R. Leibler (1951). *On information and sufficiency*. The annals of mathematical statistics, 22(1):79–86.

[Kuo and Sullivan, 2001] Kuo, F.E. and W. Sullivan (2001). *Aggression and violence in the inner city effects of environment via mental fatigue*. Environment and behavior, 33(4):543–571.

[Laskowski and Schultz, 2006] Laskowski, K. and T. Schultz (2006). *Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings*. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I. IEEE.

[Leamy et al., 2011] Leamy, D.J., R. Collins and T. Ward (2011). *Combining fNIRS and EEG to improve motor cortex activity classification during an imagined movement-based task*. In *International Conference on Foundations of Augmented Cognition*, pp. 177–185. Springer.

[LeCun et al., 2015] LeCun, Y., Y. Bengio and G. Hinton (2015). *Deep learning*. Nature, 521(7553):436–444.

[Ledoit and Wolf, 2004] Ledoit, O. and M. Wolf (2004). *A well-conditioned estimator for large-dimensional covariance matrices*. Journal of multivariate analysis, 88(2):365–411.

[Leuthardt et al., 2011] Leuthardt, E.C., C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears and G. Schalk (2011). *Using the electrocorticographic speech network to control a brain–computer interface in humans*. Journal of neural engineering, 8(3):036004.

[Lewandowsky et al., 2010] Lewandowsky, S., K. Oberauer, L.-X. Yang and U. Ecker (2010). *A working memory test battery for MATLAB*. Behavior Research Methods, 42(2):571–585.

[Lo et al., 2009] Lo, Y.L., H. Zhang, C. Wang, Z. Chin, S. Fook-Chong, C. Gabriel and C. Guan (2009). *Correlation of near-infrared spectroscopy and transcranial magnetic stimulation of the motor cortex in overt reading and musical tasks*. Motor control, 13(1):84–99.

[Loring et al., 2012] Loring, D.W., K. Meador, G. Lee and D. King (2012). *Amobarbital effects and lateralized brain function: the Wada test*. Springer Science & Business Media.

[Lotte et al., 2015] Lotte, F., J. Brumberg, P. Brunner, A. Gunduz, A. Ritaccio, C. Guan and G. Schalk (2015). *Electrocorticographic Representations of Segmental Features in Continuous Speech*. Frontiers in Human Neuroscience, 9(97).

[Lotte et al., 2007] Lotte, F., M. Congedo, A. Lécuyer, F. Lamarche and B. Arnaldi (2007). *A review of classification algorithms for EEG-based brain–computer interfaces*. Journal of neural engineering, 4(2):R1.

[Lukanov et al., 2016] Lukanov, K., H. Maior and M. Wilson (2016). *Using fNIRS in usability testing: understanding the effect of web form layout on mental workload*. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4011–4016. ACM.

[Martens, 2010] Martens, J. (2010). *Deep learning via Hessian-free optimization*. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 735–742.

[Martin et al., 2014] Martin, S., P. Brunner, C. Holdgraf, H.-J. Heinze, N. Crone, J. Rieger, G. Schalk, R. Knight and B. Pasley (2014). *Decoding spectrotemporal features of overt and covert speech from the human cortex*. Frontiers in Neuroengineering, 7(14).

[Martin et al., 2016a] Martin, S., P. Brunner, I. Iturrate, J. Millán, G. Schalk, R. Knight and B. Pasley (2016a). *Word pair classification during imagined speech using direct brain recordings*. Scientific reports, 6.

[Martin et al., 2016b] Martin, S., J. Millán, R. Knight and B. Pasley (2016b). *The use of intracranial recordings to decode human language: Challenges and opportunities*. Brain and Language.

[Martinez et al., 2013] Martinez, H.P., Y. Bengio and G. Yannakakis (2013). *Learning deep physiological models of affect*. IEEE Computational Intelligence Magazine, 8(2):20–33.

[McFarland et al., 2000] McFarland, D.J., L. Miner, T. Vaughan and J. Wolpaw (2000). *Mu and beta rhythm topographies during motor imagery and actual movements*. Brain topography, 12(3):177–186.

[McKendrick et al., 2015] McKendrick, R., R. Parasuraman and H. Ayaz (2015). *Wearable functional near infrared spectroscopy (fNIRS) and transcranial direct current stimulation (tDCS): expanding vistas for neurocognitive augmentation*. Frontiers in Systems Neuroscience, 9:27.

[van Merrienboer and Sweller, 2005] Merrienboer, J.J.G. van and J. Sweller (2005). *Cognitive load theory and complex learning: Recent developments and future directions*. Educational psychology review, 17(2):147–177.

[Mesgarani et al., 2014] Mesgarani, N., C. Cheung, K. Johnson and E. Chang (2014). *Phonetic Feature Encoding in Human Superior Temporal Gyrus*. Science, p. 1245994.

[Miller and Cohen, 2001] Miller, E.K. and J. Cohen (2001). *An integrative theory of prefrontal cortex function*. Annual review of neuroscience, 24(1):167–202.

[Miller et al., 2007] Miller, K.J., E. Leuthardt, G. Schalk, R. Rao, N. Anderson, D. Moran, J. Miller and J. Ojemann (2007). *Spectral changes in cortical surface potentials during motor movement*. The Journal of neuroscience, 27(9):2424–2432.

[Molavi and Dumont, 2010] Molavi, B. and G. Dumont (2010). *Wavelet based motion artifact removal for Functional Near Infrared Spectroscopy*. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 5 –8.

[Moses et al., 2016] Moses, D.A., N. Mesgarani, M. Leonard and E. Chang (2016). *Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity*. Journal of Neural Engineering, 13(5):056004.

[Mugler et al., 2014] Mugler, E.M., J. Patton, R. Flint, Z. Wright, S. Schuele, J. Rosenow, J. Shih, D. Krusienski and M. Slutzky (2014). *Direct classification of all American English phonemes using signals from functional speech motor cortex*. Journal of Neural Engineering, 11(3):035015.

[Muller et al., 2016] Muller, L., S. Felix, K. Shah, K. Lee, S. Pannu and E. Chang (2016). *Thin-Film, High-Density Micro-Electrocorticographic Decoding of a Human Cortical Gyrus*. In *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*.

[Müller-Putz et al., 2005] Müller-Putz, G.R., R. Scherer, C. Brauneis and G. Pfurtscheller (2005). *Steady-state visual evoked potential (SSVEP)-based communication: impact of harmonic frequency components.*. Journal of neural engineering, 2(4):123–130.

[Nair and Hinton, 2010] Nair, V. and G. Hinton (2010). *Rectified linear units improve restricted boltzmann machines*. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814.

[Oberauer et al., 2000] Oberauer, K., H.-M. Süss, R. Schulze, O. Wilhelm and W. Wittmann (2000). *Working memory capacity - facets of a cognitive ability construct*. Personality and Individual Differences, 29(6):1017–1045.

[Obrig et al., 2010] Obrig, H., S. Rossi, S. Telkemeyer and I. Wartenburger (2010). *From acoustic segmentation to language processing: evidence from optical imaging*. Frontiers in Neuroenergetics, 2:13.

[Ogawa et al., 1990] Ogawa, S., T.-M. Lee, A. Kay and D. Tank (1990). *Brain magnetic resonance imaging with contrast dependent on blood oxygenation*. Proceedings of the National Academy of Sciences, 87(24):9868–9872.

[Oldfield, 1971] Oldfield, R.C. (1971). *The assessment and analysis of handedness: the Edinburgh inventory*. Neuropsychologia, 9(1):97–113.

[van den Oord et al., 2016] Oord, A. van den, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu (2016). *WaveNet: A generative model for raw audio*. arXiv preprint arXiv:1609.03499.

[Ortner et al., 2015] Ortner, R., J. Scharinger, A. Lechner and C. Guger (2015). *How many people can control a motor imagery based BCI using common spatial patterns?*. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 202–205.

[O'Sullivan et al., 2015] O'Sullivan, J.A., A. Power, N. Mesgarani, S. Rajaram, J. Foxe, B. Shinn-Cunningham, M. Slaney, S. Shamma and E. Lalor (2015). *Attentional selection in a cocktail party environment can be decoded from single-trial EEG*. Cerebral Cortex, 25(7):1697–1706.

[Pais-Vieira et al., 2013] Pais-Vieira, M., M. Lebedev, C. Kunicki, J. Wang and M. Nicolelis (2013). *A brain-to-brain interface for real-time sharing of sensorimotor information*. Scientific reports, 3.

[Pasley et al., 2012] Pasley, B.N., S. David, N. Mesgarani, A. Flinker, S. Shamma, N. Crone, R. Knight and E. Chang (2012). *Reconstructing speech from human auditory cortex*. PLoS biology, 10(1):e1001251.

[Pei et al., 2011a] Pei, X., D. Barbour, E. Leuthardt and G. Schalk (2011a). *Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans*. Journal of neural engineering, 8(4):046028.

[Pei et al., 2011b] Pei, X., E. Leuthardt, C. Gaona, P. Brunner, J. Wolpaw and G. Schalk (2011b). *Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition*. Neuroimage, 54(4):2960–2972.

[Pena et al., 2003] Pena, M., A. Maki, D. Kovaćić, G. Dehaene-Lambertz, H. Koizumi, F. Bouquet and J. Mehler (2003). *Sounds and silence: an optical topography study of language recognition at birth*. Proceedings of the National Academy of Sciences, 100(20):11702–11705.

[Penfield and Roberts, 2014] Penfield, W. and L. Roberts (2014). *Speech and brain mechanisms*. Princeton University Press.

[de Pesters et al., 2016] Pesters, A. de, W. Coon, P. Brunner, A. Gunduz, A. Ritaccio, N. Brunet, P. de Weerd, M. Roberts, R. Oostenveld, P. Fries and G. Schalk (2016). *Alpha power indexes task-related networks on large and small scales: A multimodal {ECoG} study in humans and a non-human primate*. NeuroImage, 134:122 – 131.

[Pfurtscheller et al., 2010] Pfurtscheller, G., B. Allison, C. Brunner, G. Bauernfeind, T. Solis-Escalante, R. Scherer, T. Zander, G. Mueller-Putz, C. Neuper and N. Birbaumer (2010). *The hybrid BCI*. Frontiers in neuroscience, 4.

[Pfurtscheller and Aranibar, 1979] Pfurtscheller, G. and A. Aranibar (1979). *Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement*. Electroencephalography and clinical neurophysiology, 46(2):138–146.

[Pfurtscheller et al., 2006] Pfurtscheller, G., C. Brunner, A. Schlögl and F. Da Silva (2006). *Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks*. Neuroimage, 31(1):153–159.

[Pfurtscheller and Cooper, 1975] Pfurtscheller, G. and R. Cooper (1975). *Frequency dependence of the transmission of the EEG from cortex to scalp*. Electroencephalography and clinical neurophysiology, 38(1):93–96.

[Pfurtscheller et al., 1993] Pfurtscheller, G., D. Flotzinger and J. Kalcher (1993). *Brain-computer interface-a new communication device for handicapped persons*. Journal of Microcomputer Applications, 16(3):293–299.

[Pfurtscheller and Neuper, 1997] Pfurtscheller, G. and C. Neuper (1997). *Motor imagery activates primary sensorimotor area in humans*. Neuroscience letters, 239(2):65–68.

[Pfurtscheller and Neuper, 2001] Pfurtscheller, G. and C. Neuper (2001). *Motor imagery and direct brain-computer communication*. Proceedings of the IEEE, 89(7):1123–1134.

[Plum and Posner, 1982] Plum, F. and J. Posner (1982). *The diagnosis of stupor and coma*, vol. 19. Oxford University Press, USA.

[Pollmann et al., 2016] Pollmann, K., M. Vukelić, N. Birbaumer, M. Peissner, W. Bauer and S. Kim (2016). *fNIRS as a Method to Capture the Emotional User Experience: A Feasibility Study*. In *International Conference on Human-Computer Interaction*, pp. 37–47. Springer.

[Porbadnigk et al., 2009] Porbadnigk, A., M. Wester, J. Callies and T. Schultz (2009). *EEG-based Speech Recognition - Impact of Temporal Effects*. In *2nd International Conference on Bio-inspired Systems and Signal Processing, Porto, Portugal*. Biosignals 2009.

[Potes et al., 2012] Potes, C., A. Gunduz, P. Brunner and G. Schalk (2012). *Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening*. NeuroImage, 61(4):841 – 848.

[Power et al., 2010] Power, S.D., T. Falk and T. Chau (2010). *Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy*. Journal of neural engineering, 7(2):026002.

[Power et al., 2012] Power, S.D., A. Kushki and T. Chau (2012). *Intersession Consistency of Single-Trial Classification of the Prefrontal Response to Mental Arithmetic and the No-Control State by NIRS*. PLoS ONE, 7(7):e37791.

[Price, 2012] Price, C.J. (2012). *A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading*. Neuroimage, 62(2):816–847.

[Putze, 2014] Putze, F. (2014). *Adaptive Cognitive Interaction Systems*. PhD thesis, Karlsruhe Institute of Technology.

[Quaresima et al., 2012] Quaresima, V., S. Bisconti and M. Ferrari (2012). *A brief review on the use of functional near-infrared spectroscopy (fNIRS) for language imaging studies in human newborns and adults*. Brain and Language, 121(2):79–89.

[Rabiner, 1989] Rabiner, L. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 77(2):257–286.

[Rao et al., 2014] Rao, R.P.N., A. Stocco, M. Bryan, D. Sarma, T. Youngquist, J. Wu and C. Prat (2014). *A direct brain-to-brain interface in humans*. PloS one, 9(11):e111332.

[Ray and Maunsell, 2011] Ray, S. and J. Maunsell (2011). *Different origins of gamma rhythm and high-gamma activity in macaque visual cortex*. PLoS Biol, 9(4):e1000610.

[Reed and Durlach, 1998] Reed, C.M. and N. Durlach (1998). *Note on information transfer rates in human communication*. Presence: Teleoperators and Virtual Environments, 7(5):509–518.

[Renard et al., 2010] Renard, Y., F. Lotte, G. Gibert, M. Congedo, E. Maby, V. Delannoy, O. Bertrand and A. Lécuyer (2010). *Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments*. Presence, 19(1):35–53.

[Roland et al., 2010] Roland, J., P. Brunner, J. Johnston, G. Schalk and E. Leuthardt (2010). *Passive real-time identification of speech and motor cortex during an awake craniotomy*. Epilepsy & Behavior, 18(1):123–128.

[Roth et al., 1996] Roth, M., J. Decety, M. Raybaudi, R. Massarelli, C. Delon-Martin, C. Segebarth, S. Morand, A. Gemignani, M. Décorps and M. Jeannerod (1996). *Possible involvement of primary motor cortex in mentally simulated movement: a functional magnetic resonance imaging study.*. Neuroreport, 7(7):1280–1284.

[Rugg and Coles, 1995] Rugg, M.D. and M. Coles (1995). *Electrophysiology of mind: Event-related brain potentials and cognition.*. Oxford University Press.

[Sahin et al., 2009] Sahin, N.T., S. Pinker, S. Cash, D. Schomer and E. Halgren (2009). *Sequential processing of lexical, grammatical, and phonological information within Broca's area*. Science, 326(5951):445–449.

[Sakatani et al., 1999] Sakatani, K., W. Lichty, Y. Xie, S. Li and H. Zuo (1999). *Effects of aging on language-activated cerebral blood oxygenation changes of the left prefrontal cortex: near infrared spectroscopy study*. Journal of Stroke and Cerebrovascular Diseases, 8(6):398–403.

[Sakatani et al., 1998] Sakatani, K., Y. Xie, W. Lichty, S. Li and H. Zuo (1998). *Language-activated cerebral blood oxygenation and hemodynamic changes of the left prefrontal cortex in poststroke aphasic patients a near-infrared spectroscopy study*. Stroke, 29(7):1299–1304.

[Salthouse et al., 1991] Salthouse, T.A., R. Babcock and R. Shaw (1991). *Effects of adult age on structural and operational capacities in working memory.*. Psychology and aging, 6(1):118.

[Sassaroli and Fantini, 2004] Sassaroli, A. and S. Fantini (2004). *Comment on the modified Beer-Lambert law for scattering media*. Physics in Medicine and Biology, 49(14):N255.

[Sato et al., 2013] Sato, H., N. Yahata, T. Funane, R. Takizawa, T. Katura, H. Atsumori, Y. Nishimura, A. Kinoshita, M. Kiguchi, H. Koizumi et al. (2013). *A NIRS–fMRI investigation of prefrontal cortex activity during a working memory task*. Neuroimage, 83:158–173.

[Schalk et al., 2004] Schalk, G., D. McFarland, T. Hinterberger, N. Birbaumer and J. Wolpaw (2004). *BCI2000: a general-purpose brain-computer interface (BCI) system*. Biomedical Engineering, IEEE Transactions on, 51(6):1034–1043.

[Scherer et al., 2008] Scherer, R., F. Lee, A. Schlögl, R. Leeb, H. Bischof and G. Pfurtscheller (2008). *Toward self-paced brain–computer communication: navigation through virtual worlds*. IEEE Transactions on Biomedical Engineering, 55(2):675–682.

[Schlögl et al., 2007] Schlögl, A., C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb and G. Pfurtscheller (2007). *A fully automated correction method of EOG artifacts in EEG recordings*. Clinical neurophysiology, 118(1):98–104.

[Schlögl et al., 2005] Schlögl, A., F. Lee, H. Bischof and G. Pfurtscheller (2005). *Characterization of four-class motor imagery EEG data for the BCI-competition 2005*. Journal of Neural Engineering, 2(4):L14.

[Schnitzler et al., 1997] Schnitzler, A., S. Salenius, R. Salmelin, V. Jousmäki and R. Hari (1997). *Involvement of primary motor cortex in motor imagery: a neuromagnetic study*. Neuroimage, 6(3):201–208.

[Schultz and Kirchhoff, 2006] Schultz, T. and K. Kirchhoff (2006). *Multilingual Speech Processing*. Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5.

[Schultz and Wand, 2010] Schultz, T. and M. Wand (2010). *Modeling coarticulation in EMG-based continuous speech recognition*. Speech Communication, 52(4):341–353.

[Sevy et al., 2010] Sevy, A.B.G., H. Bortfeld, T. Huppert, M. Beauchamp, R. Tonini and J. Oghalai (2010). *Neuroimaging with near-infrared spectroscopy demonstrates speech-evoked activity in the auditory cortex of deaf children following cochlear implantation*. Hearing research, 270(1):39–47.

[Shih and Krusienski, 2012] Shih, J.J. and D. Krusienski (2012). *Signals from intraventricular depth electrodes can control a brain–computer interface*. Journal of neuroscience methods, 203(2):311–314.

[Sinai et al., 2009] Sinai, A., N. Crone, H. Wied, P. Franaszczuk, D. Miglioretti and D. Boatman-Reich (2009). *Intracranial mapping of auditory perception: event-related responses and electrocortical stimulation*. Clinical Neurophysiology, 120(1):140–149.

[Sitaram et al., 2007a] Sitaram, R., H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu and N. Birbaumer (2007a). *Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface*. NeuroImage, 34(4):1416–1427.

[Sitaram et al., 2007b] Sitaram, R., H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu and N. Birbaumer (2007b). *Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain-computer interface*. NeuroImage, 34(4):1416 – 1427.

[Srivastava et al., 2014] Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. Journal of Machine Learning Research, 15:1929–1958.

[Stevens et al., 1937] Stevens, S.S., J. Volkmann and E. Newman (1937). *A Scale for the Measurement of the Psychological Magnitude Pitch*. The Journal of the Acoustical Society of America, 8(3):185–190.

[Stolcke, 2002] Stolcke, A. (2002). *SRILM - An extensible language modeling toolkit*. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*.

[Strait and Scheutz, 2014] Strait, M. and M. Scheutz (2014). *What we can and cannot (yet) do with functional near infrared spectroscopy*. Frontiers in Neuroscience, 8(117).

[Strangman et al., 2002] Strangman, G., D. Boas and J. Sutton (2002). *Non-invasive neuroimaging using near-infrared light*. Biological psychiatry, 52(7):679–693.

[Sturm et al., 2016] Sturm, I., S. Bach, W. Samek and K.-R. Müller (2016). *Interpretable Deep Neural Networks for Single-Trial EEG Classification*. arXiv preprint arXiv:1604.08201.

[Sturm et al., 2014] Sturm, I., B. Blankertz, C. Potes, G. Schalk and G. Curio (2014). *ECoG high gamma activity reveals distinct cortical representations of lyrics passages, harmonic and timbre-related changes in a rock song*. Frontiers in human neuroscience, 8:798.

[Sundermeyer et al., 2012] Sundermeyer, M., R. Schlüter and H. Ney (2012). *LSTM Neural Networks for Language Modeling.*. In *Interspeech*, pp. 194–197.

[Sutter, 1992] Sutter, E.E. (1992). *The brain response interface: communication through visually-induced electrical brain responses*. Journal of Microcomputer Applications, 15(1):31–45.

[Talairach and Tournoux, 1988] Talairach, J. and P. Tournoux (1988). *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. Thieme.

[Talavage et al., 2014] Talavage, T.M., J. Gonzalez-Castillo and S. Scott (2014). *Auditory neuroimaging with fMRI and PET*. Hearing research, 307:4–15.

[al Tervaniemi et al., 1999] Tervaniemi, M. al, A. Kujala, K. Alho, J. Virtanen, R. Ilmoniemi and R. Näätänen (1999). *Functional specialization of the human auditory cortex in processing phonetic and musical sounds: A magnetoencephalographic (MEG) study*. Neuroimage, 9(3):330–336.

[Vapnik, 1998] Vapnik, V.N. (1998). *Statistical learning theory*. Wiley, 1 ed.

[Vaughan et al., 2006] Vaughan, T.M., D. McFarland, G. Schalk, W. Sarnacki, D. Krusienski, E. Sellers and J. Wolpaw (2006). *The Wadsworth BCI research and development program: at home with BCI*. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 14(2):229–233.

[Wada and Rasmussen, 1960] Wada, J. and T. Rasmussen (1960). *Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance: experimental and clinical observations*. Journal of Neurosurgery, 17(2):266–282.

[Wand et al., 2016] Wand, M., J. KoutnÃk and J. Schmidhuber (2016). *Lipreading with Long Short-Term Memory*. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*.

[Wand and Schmidhuber, 2016] Wand, M. and J. Schmidhuber (2016). *Deep Neural Network Frontend for Continuous EMG-based Speech Recognition*. In *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*.

[Wand et al., 2013] Wand, M., C. Schulte, M. Janke and T. Schultz (2013). *Array-based Electromyographic Silent Speech Interface.*. In *BIOSIGNALS*, pp. 89–96.

[Wand and Schultz, 2014] Wand, M. and T. Schultz (2014). *Pattern learning with deep neural networks in EMG-based speech recognition*. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4200–4203. IEEE.

[Wartenburger et al., 2007] Wartenburger, I., J. Steinbrink, S. Telkemeyer, M. Friedrich, A. Friederici and H. Obrig (2007). *The processing of prosody: evidence of interhemispheric specialization at the age of four*. Neuroimage, 34(1):416–425.

[Watanabe et al., 1998] Watanabe, E., A. Maki, F. Kawaguchi, K. Takashiro, Y. Yamashita, H. Koizumi and Y. Mayanagi (1998). *Non-invasive assessment of language dominance with near-infrared spectroscopic mapping*. Neuroscience Letters, 256(1):49 – 52.

[Welch, 1967] Welch, P.D. (1967). *The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms*. IEEE Transactions on audio and electroacoustics, 15(2):70–73.

[Wierwille et al., 1994] Wierwille, W.W., S. Wreggit, C. Kirn, L. Ellsworth and R. Fairbanks (1994). *Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. final report*. Technical Report

[Wiggins et al., 2016] Wiggins, I.M., C. Anderson, P. Kitterick and D. Hartley (2016). *Speech-evoked activation in adult temporal cortex measured using functional near-infrared spectroscopy (fNIRS): Are the measurements reliable?*. Hearing Research, 339:142 – 154.

[Wolpaw and Wolpaw, 2012] Wolpaw, J. and E. Wolpaw (2012). *Brain-computer interfaces: principles and practice*. OUP USA.

[Wu et al., 2006] Wu, W., Y. Gao, E. Bienenstock, J. Donoghue and M. Black (2006). *Bayesian population decoding of motor cortical activity using a Kalman filter*. Neural computation, 18(1):80–118.

[Yang et al., 2015] Yang, M., S. Sheth, C. Schevon, G. II and N. Mesgarani (2015). *Speech reconstruction from human auditory cortex with deep neural networks*. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[Yoshimura et al., 2016] Yoshimura, N., A. Nishimoto, A. Belkacem, D. Shin, H. Kambara, T. Hanakawa and Y. Koike (2016). *Decoding of Covert Vowel Articulation Using Electroencephalography Cortical Currents*. Frontiers in Neuroscience, 10:175.

[Zahner et al., 2014] Zahner, M., M. Janke, M. Wand and T. Schultz (2014). *Conversion from facial myoelectric signals to speech: a unit selection approach.*. In *INTERSPEECH*, pp. 1184–1188.

[Zander and Kothe, 2011] Zander, T.O. and C. Kothe (2011). *Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general*. Journal of Neural Engineering, 8(2):025005.

[Zhang et al., 2006] Zhang, Z., T. Li, Y. Zheng, Q. Luo, R. Song and H. Gong (2006). *Study the left prefrontal cortex activity of Chinese children with dyslexia in phonological processing by NIRS*. In *Biomedical Optics 2006*, pp. 607833–607833. International Society for Optics and Photonics.

[Zhongpeng and Wenxue, 2016] Zhongpeng, Z. and H. Wenxue (2016). *A new signal analysis method for functional near-infrared spectroscopy*. In *2016 8th IEEE International Conference on Communication Software and Networks (ICCSN)*, pp. 100–106.

Appendix A

Accumulated Publications

frontiers in Neuroscience

FOCUSED REVIEW
published: 27 September 2016
doi: 10.3389/fnins.2016.00429

Automatic Speech Recognition from Neural Signals: A Focused Review

Christian Herff and Tanja Schultz*

Cognitive Systems Lab, Department for Mathematics and Computer Science, University of Bremen, Bremen, Germany

Speech interfaces have become widely accepted and are nowadays integrated in various real-life applications and devices. They have become a part of our daily life. However, speech interfaces presume the ability to produce intelligible speech, which might be impossible due to either loud environments, bothering bystanders or incapabilities to produce speech (i.e., patients suffering from locked-in syndrome). For these reasons it would be highly desirable to not speak but to simply envision oneself to say words or sentences. Interfaces based on imagined speech would enable fast and natural communication without the need for audible speech and would give a voice to otherwise mute people. This focused review analyzes the potential of different brain imaging techniques to recognize speech from neural signals by applying Automatic Speech Recognition technology. We argue that modalities based on metabolic processes, such as functional Near Infrared Spectroscopy and functional Magnetic Resonance Imaging, are less suited for Automatic Speech Recognition from neural signals due to low temporal resolution but are very useful for the investigation of the underlying neural mechanisms involved in speech processes. In contrast, electrophysiologic activity is fast enough to capture speech processes and is therefor better suited for ASR. Our experimental results indicate the potential of these signals for speech recognition from neural data with a focus on invasively measured brain activity (electrocorticography). As a first example of Automatic Speech Recognition techniques used from neural signals, we discuss the *Brain-to-text* system.

Keywords: ASR, automatic speech recognition, ECoG, fNIRS, EEG, speech, BCI, brain-computer interface

OPEN ACCESS

Edited by:
Giovanni Mirabella, Sapienza University of Rome, Italy

Reviewed by:
Andrea Brovelli, Centre National de la Recherche Scientifique (CNRS), France
Elizaveta Okorokova, National Research University Higher School of Economics, Russia

****Correspondence:***

Christian Herff
is currently a research assistant/Ph.D. student in the Cognitive Systems Lab at University of Bremen supervised by Tanja Schultz. Previously, he obtained his Diploma in Computer Science from the Karlsruhe Institute of Technology. His research interest lays in using machine learning techniques to analyze speech processes in neural data.
christian.herff@uni-bremen.de

Received: *15 April 2016*
Accepted: *05 September 2016*
Published: *27 September 2016*

Citation:
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

1. INTRODUCTION

With services like Siri and Google Voice Search, speech-driven applications arrived in our daily life and are used by millions of users every day. These speech interfaces allow for natural interaction with electronic devices and enable fast input of texts. **Brain-computer interfaces (BCIs)** (Wolpaw et al., 2002) on the other hand are currently only used by a small number of patients (Vaughan et al., 2006). This is in part due to the unnatural paradigms which have to be employed to enter commands or texts via the BCI. Motor imagery based BCIs (McFarland et al., 2000) use imagined movement of hands, arms or feet to issue directional commands. To spell out texts, users often

KEY CONCEPT 1 | Brain-computer interfaces (BCIs)
A Brain-Computer Interface is a system which sends messages or commands to a computer without using the brain's normal output pathways of peripheral nerves and muscles.

have to focus on a single letter at a time which is then selected (Farwell and Donchin, 1988; Sutter, 1992; Donchin et al., 2000; Müller-Putz et al., 2005). Even though these are the fasted currently known BCIs, they are still rather slow and very unnatural. Using speech as a paradigm for BCIs would solve these problems and enable very natural communication. A BCI based on speech would enable communication without the need for acoustic voice production, while maintaining the same advantages as ordinary speech interfaces. Brain activity is not the only approach possible for silent speech interfaces, see the review (Denby et al., 2010) for a description of other approaches to silent speech interfaces. However, only silent speech interfaces based on brain activity would enable severely disabled persons (i.e., locked-in syndrome) to communicate with the outside world.

The intention of this focused review is to investigate the potential of neural signals—captured by different brain imaging techniques—as input for **Automatic Speech Recognition (ASR)**. Brain imaging techniques can be broadly divided into two categories. Imaging methods based on metabolic processes measure the amount of oxygenated and/or deoxygenated blood in certain areas of the brain. We will discuss functional Magnetic Resonance Imaging (fMRI) and functional Near Infrared Spectroscopy (fNIRS) from this category of imaging techniques, as they are the most commonly used in neuroimaging.

KEY CONCEPT 2 | Automatic Speech Recognition (ASR)
Automatic Speech Recognition is a technology that enables the recognition of spoken language into a textual representation by computers. These technologies often rely on statistical models like Hidden-Markov-Models and can now be found in a large variety of consumer electronics from cars to mobile phones.

Measurement of electric potentials is possible both on the scalp and invasively. We will be discussing electroencephalography (EEG) and magnetoencephalography (MEG) as non-invasive and electrocorticography (ECoG) and microarrays as invasively measured examples of electrophysiological signals.

1.1. Metabolic Signals

Brain imaging techniques based on metabolic processes measure the amount of oxygen-carrying blood in certain areas of the brain. Active neurons have a higher demand for energy in the form of oxygen, resulting in increased blood flow to these active regions to satisfy the increased demand. Thus, the amount of fresh oxygenated blood can be used as an indirect marker of neural activity in very small regions, called voxels. Blood vessels form a very intricate network in the brain and can thus regulate the supply to very specific regions in the brain. Brain imaging techniques based on metabolic processes can therefor measure activity with a very high spatial resolution. On the flip side, these metabolic processes are slow in nature and take several seconds to complete. Continuous speech processes, like the production of single vowels or consonants, happen as fast as 50 ms, which makes them impossible to be measured with metabolic-based imaging techniques.

1.1.1. Functional Magnetic Resonance Imaging

Hemoglobin, the oxygen carrying part of the blood, has different magnetic properties when oxygenated or deoxygenated. These different properties can be detected by the strong magnetic fields produced in the large tube of the MRI. Observing the changes in these relative hemoglobin concentrations allows for the estimation of neural activity in a voxel. fMRI is instrumental in a large variety of neuroimaging studies. The high spatial resolution over the entire brain enables detailed investigations of neural processes during all sorts of cognitive processes.

The inherently slow natures of metabolic processes rule out fMRI to be used for continuous speech recognition, as phones change much too quick for the slow hemodynamic responses. However, fMRI can be used in neuroscientific studies to learn more about speech perception, speech production and reading. See the excellent reviews (Price, 2012; Talavage et al., 2014) for more on this topic. Besides neuroscientific breakthroughs, it has been shown that fMRI recordings can be used to classify isolated **phones** or attended speaker (Formisano et al., 2008).

KEY CONCEPT 3 | Phone
A phone is a distinct speech sound that can be perceptually differentiated from other speech sounds.

Moreover, the sheer size and cost of the apparatus and the fact that subjects have to remain motionless in it for extended periods of time make it ill-suited for real-life interfaces. Nevertheless, fMRI studies are indispensable for neuroscience, due to their unparalleled spatial resolution.

1.1.2. fNIRS

Light in the near infrared part of the light spectrum (~700–900 nm) disperses through skin, bones and tissue, but is absorbed by hemoglobin. It can be used to indirectly estimate brain activity by shining it through the skull and measuring how much of the re-emerging light is attenuated. The more light is absorbed, the more oxygenated hemoglobin and thus the more active the specific brain region. fNIRS measures similar physiological signals as fMRI with much cheaper devices, which can be head-mounted and do not require the subject to lay motionless. It provides signals on the same temporal scale as fMRI measurements, but with a far coarser spatial resolution. Additionally, fNIRS is only able to measure the hemodynamic response in outer areas of the cortex and is not able to provide signals from the entire brain.

While fNIRS can be used for BCIs both for direct control (Coyle et al., 2007; Sitaram et al., 2007) and passive monitoring of user states (Heger et al., 2013; Herff et al., 2013b, 2015a; Heger et al., 2014; Hennrich et al., 2015), it is not well suited for ASR, as recorded processes are far too slow to capture the fast dynamics of speech.

To investigate speech processes with fNIRS, some studies (Herff et al., 2012a,b, 2013a) discriminated the type of speech production that a user currently undertook, such as audible speech, silently-mouthed speech and speech imagery. These studies show that fNIRS can be used to study speech processes in

the brain, but is not suitable for continuous speech recognition from neural signals.

1.2. Electrophysiological Signals

Measurement of electrophysiological signals from the brain can be carried out both invasively or non-invasively. Electrodes can either measure ensembles of neurons firing in synchrony, which is done by MEG, EEG, and ECoG, or needle electrodes can be used to measure single action potentials (spikes) from individual neurons. Obviously the spatial and temporal resolution of single neuron measurements using microarrays is unparalleled, but it comes at the disadvantage of only covering small areas and thus not measuring all areas involved in speech production. MEG, EEG, and ECoG can cover larger areas or even the entire brain, but with coarser spatial resolution.

1.2.1. Microarrays

Microarrays provide high resolution information of very small brain areas with a size of few square milimeters. The spatial and temporal resolution down to single action potentials is unparalleled. Microarrays in the speech-motor cortex have successfully been used to decode intended phone production (Brumberg et al., 2011) for a number of isolated phones or to synthesize vowels (Guenther et al., 2009; Brumberg et al., 2010). As microarrays cover only very small areas of the cortex, they might miss crucial information from other parts of the brain involved in the speech production process and might thus not be well suited in the combination with ASR technology.

1.2.2. Electroencephalography (EEG)

Electroencephalography measures electric potentials of large ensembles of neurons firing at the same time by placing electrodes on the scalp. With these scalp electrodes, experiments are easy to setup and do not require a clinical environment. EEG is the de-facto standard for BCIs as the technique is non-invasive and easy to setup, while still providing high-quality signals with good temporal resolution.

However, the placement on the scalp makes EEG very prune to motion artifacts, especially from head movements. Muscle movements in the face as appearing from spoken speech yield large electromyographic and glossokinetic artifacts in the EEG that are not produced by brain activity. In fact, EMG activity in facial muscles alone can be used to accurately decode speech by itself (Schultz and Wand, 2010; Herff et al., 2011). Additionally, due to volume conduction effects, each EEG electrode measures signals from a variety of superimposed sources, making localization of brain activity very difficult.

While EEG is the de-facto standard for current BCIs, it can currently not be used for ASR from neural signals, as the first step for speech interfaces, namely speech decoding from audible speech is not possible due to artifact contamination. However, studies have used EEG successfully to investigate perceived speech (Di Liberto et al., 2015; O'Sullivan et al., 2015) or to classify limited numbers of imagined isolated phones (Yoshimura et al., 2016).

1.2.3. Magnetencephalography (MEG)

Magnetencephalography measures synchronized activity of large groups of neurons using magnetometers placed around the head, requiring extensive magnetic shielding around the device. MEG provides high temporal and acceptable spatial resolution and is less distorted by the scalp than EEG. However, movement, especially of the facial muscles yield large artifacts in the MEG signals, it is thus difficult to investigate overt speech production with MEG.

The high spatial and temporal resolution of MEG allow for thorough investigation of speech process, including the comparison between speech production and perception (Houde et al., 2002) and the comparison of processing of phonetic and musical sounds (Tervaniemi et al., 1999). Heinks-Maldonado et al. (2006) presented evidence for a forward model in speech production. MEG has been used for classification of speech processes, Guimaraes et al. (2007) showed single trial classification between two aurally presented words, but is difficult to be used with overt speech production, as would be needed for ASR.

Due to the large chambers needed for MEG devices, they are not ideally suited for future prosthetic devices.

1.2.4. Electrocorticography (ECoG)

Electrocorticography measures electrical potentials directly on the brain surface. ECoG grids are normally used in the process of epilepsy surgery and are not originally intended for neuroscientific studies or BCIs. ECoG provides high spatial and high temporal resolution while not being affected by motion or glossokinetic artifacts. It provides signals unfiltered by scalp and skin. Electrode positions are usually within 1 cm or less from each other and thus provide high-density neural recordings from large areas of the cortex. These characteristics make ECoG ideally suited for the investigation of speech, as artifacts of natural speech production do not affect the neural recordings. ECoG has been used to investigate the differences between speech production and perception (Cheung et al., 2016). Neural representations of phonetic features during speech production are documented in Chang et al. (2010) and Mesgarani et al. (2014).

Isolated aspects of speech have successfully been decoded. Lotte et al. (2015) demonstrated that phonetic features can be decoded from ECoG data. Syllables (Bouchard and Chang, 2014) and isolated words (Kellis et al., 2010) were shown to be distinguishable from neural data. Extending upon these ideas, Mugler et al. (2014) showed that a complete set of manually labeled phones can be classified from ECoG recordings.

An alternative approach to ASR from neural signals is the reconstruction of the acoustic waveform from neural signals. This would allow users to produce normal acoustic speech from imagined speech, which would be the most natural way to restore communication for locked-in patients. For other applications, such as human-computer interaction, recognition of a textual representation is better suited as a waveform would disturb bystanders and would have to be recognized by the computer. Pasley et al. (2012) have shown that perceived speech could be reconstructed from ECoG recordings. Martin et al. (2014) showed that the spectrogram of spoken speech can be

reconstructed from ECoG. See Chakrabarti et al. (2015) for a review on speech decoding and synthesis from ECoG.

The combination of the ideal characteristics of ECoG for ASR—such as high temporal and spatial resolution, robustness toward artifacts and being unfiltered by skull and scalp—together with the rich literature on speech processes investigated using ECoG make ECoG and ideal candidate to be used for ASR from neural signals. In our *Brain-to-text* study (Heger et al., 2015; Herff et al., 2015b) we could show that ECoG could indeed be used to decode continuously spoken speech from neural signals.

2. MATERIALS AND METHODS

In our *Brain-to-text* study (Herff et al., 2015b), we obtained data from seven patients undergoing surgery for epilepsy treatment. The treatment required the patients to have electrode grids implanted on the brain surface. Each patient had very different placement of the grids depending on his or her clinical needs. The electrode grids stay implanted for periods between a few days and a couple of weeks and patients agreed to take part in our experiment during this time.

In our experiment, patients were asked to read out texts that were shown on a computer screen in front of them. Texts included political speeches, fan-fiction and children rhymes. While the participants read the text, ECoG data and acoustic data were recorded simultaneously using BCI2000 (Schalk et al., 2004). All patients gave informed consent to participate in the study, which was approved by the Institutional Review Board of Albany Medical College and the Human Research Protections Office of the US Army Medical Research and Materiel Command. Once the data was recorded, we used ASR software (Telaar et al., 2014) to mark the beginning and ending of every spoken phone. See **Figure 1** for a visualization of the experiment setup.

To extract meaningful information from the ECoG data, we calculated logarithmic broadband gamma power between 70 and 170 Hz. Gamma power has been shown to contain highly localized task specific information (Miller et al., 2007; Leuthardt et al., 2011; Pei et al., 2011; Potes et al., 2012). As ECoG data and acoustic data are recorded simultaneously, we can use the timings of the phones in the neural data, as well. This enables us to calculate an ECoG phone model for the prototypical neural activity related to each individual phone. This prototypical activity is characterized by the mean and covariance of gamma power for each selected electrode and temporal offset. The best temporal offsets and electrodes are selected on the training data using the discriminability between phones as a criterion. **Figure 1** illustrates the training process for **ECoG phone models**.

KEY CONCEPT 4 | ECoG Phone Models
ECoG phone models can be used to estimate the likelihood that an internal of ECoG activity is a certain phone. This generative models might for example return that newly recorded data have a probability of 0.6 of being a /l/, but only a probability of 0.1 of being a /b/.

These models for each phone can be used to estimate the likelihood of a certain phone given a piece of ECoG

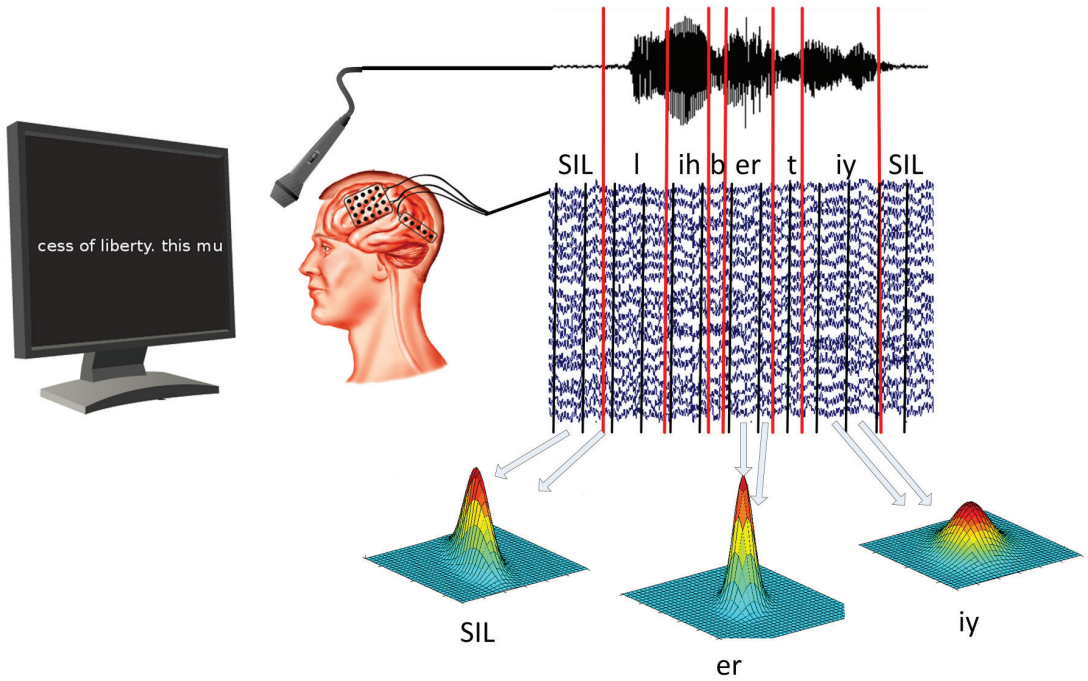


FIGURE 1 | ECoG and audio data are recorded at the same time. Speech decoding software is then used to determine timing of vowels and consonants in acoustic data. ECoG models are then trained for each phone individually by calculating the mean and covariance of all segments associated with that particular phone.

data. Additionally, the calculated generative models for each phone can be used to gain insights into the neural basis of speech production for different phones. Even though these ECoG phone models alone could be used to pick the most likely phone for each interval of ECoG activity, ASR software works by adding crucial information through a statistical **language model** (Jelinek, 1997; Stolcke, 2002) and a pronunciation **dictionary**. The combination of these three ingredients yields the great results known from speech interfaces. The ASR software extracts the search result by identifying the sequence of words from the dictionary that has the best score combination from language model and the ECoG phone models. Using these ideas from ASR, our *Brain-to-text* system is able to create a textual representation of spoken words from neural data. See **Figure 2** for a graphical explanation of the decoding process.

KEY CONCEPT 5 | Language Model
A language model estimates how likely a word is given the preceding words. In N-gram language modeling, this is done by calculating probabilities of single words and probabilities for predicting words given the history of $n-1$ previous words. The language model would thus contain that "I am" is very likely, while "I is" is rather unlikely.

KEY CONCEPT 6 | Dictionary
A pronunciation dictionary contains the mapping of phone sequences to words, for example, describing that the word liberty comprises of the phone sequence "/l/ /ih/ /b/ /er/ /t/ /iy/." The dictionary is used to guide the search for the correct words in ASR, as only words included in the dictionary can be recognized.

3. RESULTS

We evaluated our *Brain-to-text* system by training the phone models on all but one spoken phrase of a participant and then decoding the last remaining, unknown phrase. This procedure is repeated so that each phrase is decoded once. As electrode montages and brain physiologies are very different between participants, the ECoG phone models are trained for each participant individually. Acoustic speech recognition systems are trained on thousands of hours of data, while only a few minutes have to suffice for our system. To correct for this very limited amount of data, we evaluate our system with only between 10 and 100 words that can be recognized (i.e., that are in the dictionary).

For 10 words in the dictionary, we achieved up to 75% of correct words, meaning that in a phrase of 10 words, only 3 words were wrong or at the wrong position. When the system could choose between 100 words, still 40% of words were placed correctly at the appropriate position in a sentence. We used randomization tests to check whether this results were better than guessing and could show that all results were better than chance. Breaking down the decoded phrases further, we could show that on average, up to 54% of the ECoG intervals were assigned the correct phone. When looking at true positive rates for each phone, it was shown that each phone yielded better than chance true positive rates. This means that all phones worked reliably and that decoding was not based on the detection of a small subset of phones.

This results show that applying ASR to neural data is possible when the participant is speaking loudly. This is a first step toward ASR from imagined speech processes, but there are still a lot of challenges until imagined continuous speech can be decoded into a textual representation. While speech production and imagined speech production might yield similar neural responses in brain motor areas and speech planning areas, the observed neural activity in the brain's auditory cortex is distinctly different, as participants do not hear their own voice when only imagining to speak.

4. CONCLUSION AND DISCUSSION

In this focused review, we argue why only few brain imaging techniques can be used for ASR to produce textual

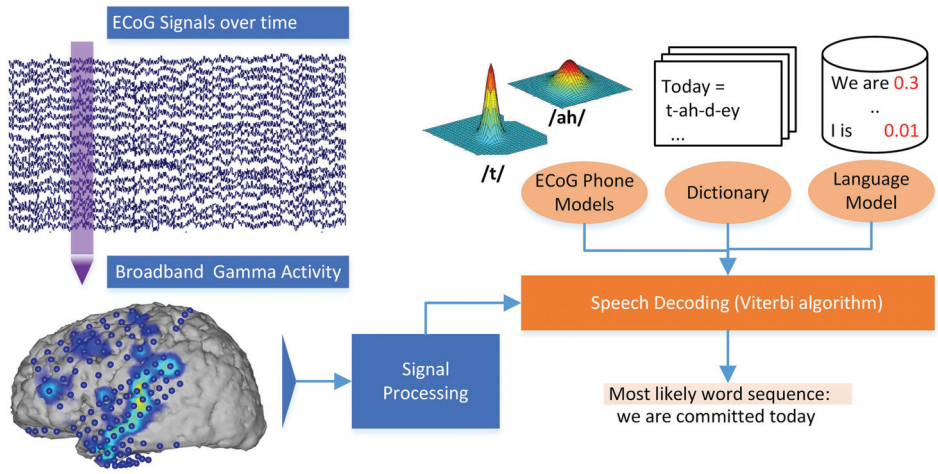


FIGURE 2 | Decoding process in the *Brain-to-text* system. Broadband gamma power is extracted for a phrase of ECoG data. The most likely word sequence is then decoded by combining the knowledge of ECoG phone models, dictionary and language model.

representations from imagined words. While no reconstruction of continuously imagined speech to a textual representation has been shown yet, we argue that measurement techniques based on electrophysiological signals are generally better suited than those based on metabolic processes. We show that ECoG is the most promising technique and demonstrate how audibly spoken speech can be recognized from ECoG data using ASR technology in our *Brain-to-text* system. Despite these first promising results, there still are a lot of open research questions to be addressed before neuroprostheses based on imagined speech processes become a reality. While having a lot of similar characteristics, imagined speech production is also distinctly different form overt speech yielding challenges for future decoding approaches. Also, initial alignment for model training is very difficult, when no audible waveform for alignment is present. These challenges need to be solved before ASR can be applied to neural signals for real life applications.

Besides the direct implications for neural prothesis based on speech processes, the successful results of the *Brain-to-text* system show promises for other areas, as well. The *Brain-to-text* systems demonstrates that leveraging advanced technology from non-adjacent areas can drastically increase decoding performance and enable new paradigms. Without the refined decoding approaches and knowledge sources from the Automatic Speech Recognition community, the results in our study could not present the entire decoding pipeline from neural signals to textual representation of words.

For neuroscience, the single trial analysis approach utilized in BCI and *Brain-to-text* yield resilient results without the need to aggregate large cohorts. Especially usage of generative models yields easily interpretable models that can grant important insights into complex brain functions without typical statistical problems associated with large numbers of variables (Eklund et al., 2016).

A fear often associated with BCI in general and the speech decoding in *Brain-to-text* in particular is that private thoughts could be read and thereby freedom of thought not be guaranteed any longer. In *Brain-to-text* activations associated with the production of speech are decoded, from planning to articulate speech prior to voice onset, to control of facial muscles, to processing of heared sounds. Thought processes or internal voice, while being formulated in words as well, do not make use of areas associated with the movement of articulatory muscles. So even if neural prothesis based on imagined speech processes become a reality, there is still a large distinction between thought processes and the process of imagining oneself to speak.

AUTHOR CONTRIBUTIONS

CH wrote the manuscript. TS supervised the research and revised the manuscript.

ACKNOWLEDGMENTS

We are very grateful for the fruitful collaboration with Adriana de Pesters, Peter Brunner, and Gerwin Schalk from the Schalk Lab which made the *Brain-to-text* system possible.

REFERENCES

Bouchard, K., and Chang, E. (2014). "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE). doi: 10.1109/embc.2014.6945185

Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., and Guenther, F. H. (2010). Brain–computer interfaces for speech communication. *Speech Commun.* 52, 367–379. doi: 10.1016/j.specom.2010.01.001

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

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Cheung, C., Hamiton, L. S., Johnson, K., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife* 5:e12577. doi: 10.7554/elife.12577

Coyle, S. M., Ward, T. E., and Markham, C. M. (2007). Brain–computer interface using a simplified functional near-infrared spectroscopy system. *J. Neural Eng.* 4:219. doi: 10.1088/1741-2560/4/3/007

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. (2010). Silent speech interfaces. *Speech Commun.* 52, 270–287. doi: 10.1016/j.specom.2009.08.002

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Donchin, E., Spencer, K. M., and Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a p300-based brain-computer interface. *IEEE Trans. Rehabil. Eng.* 8, 174–179. doi: 10.1109/86.847808

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7900–7905. doi: 10.1073/pnas.1602413113

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "who" is saying "what"? brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., et al. (2009). A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE* 4:e8218. doi: 10.1371/journal.pone.0008218

Guimaraes, M. P., Wong, D. K., Uy, E. T., Grosenick, L., and Suppes, P. (2007). Single-trial classification of meg recordings. *IEEE Trans. Biomed. Eng.* 54, 436–443. doi: 10.1109/TBME.2006.888824

Heger, D., Herff, C., Pesters, A. D., Telaar, D., Brunner, P., Schalk, G., et al. (2015). "Continuous speech recognition from ECOG," in *Sixteenth Annual Conference of the International Speech Communication Association* (Dresden).

Heger, D., Herff, C., Putze, F., Mutter, R., and Schultz, T. (2014). Continuous affective states recognition using functional near infrared spectroscopy. *Brain Comput. Interf.* 1, 113–125. doi: 10.1080/2326263X.2014.912884

Heger, D., Mutter, R., Herff, C., Putze, F., and Schultz, T. (2013). "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (Geneva), 832–837. doi: 10.1109/ACII.2013.156

Heinks-Maldonado, T. H., Nagarajan, S. S., and Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport* 17:1375. doi: 10.1097/01.wnr.0000233102.43526.e9

Hennrich, J., Herff, C., Heger, D., and Schultz, T. (2015). "Investigating deep learning for fnirs based BCI," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (Milan). doi: 10.1109/embc.2015.7318984

Herff, C., Fortmann, O., Tse, C.-Y., Cheng, X., Putze, F., Heger, D., et al. (2015a). "Hybrid fnirs-EEG based discrimination of 5 levels of memory load," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on* (Montpellier), 5–8. doi: 10.1109/ner.2015.7146546

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

Herff, C., Heger, D., Putze, F., Guan, C., and Schultz, T. (2012a). "Cross-subject classification of speaking modes using fnirs," in *Neural Information Processing, volume 7664 of* Lecture Notes in Computer Science, eds T. Huang, Z. Zeng, C. Li, and C. Leung (Berlin; Heidelberg: Springer), 417–424.

Herff, C., Heger, D., Putze, F., Guan, C., and Schultz, T. (2013a). "Self-paced bci with nirs based on speech activity," in *International BCI Meeting 2013, Asilomar* (Pacific Grove, CA).

Herff, C., Heger, D., Putze, F., Hennrich, J., Fortmann, O., and Schultz, T. (2013b). "Classification of mental tasks in the prefrontal cortex using fnirs," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka), 2160–2163. doi: 10.1109/embc.2013.6609962

Herff, C., Janke, M., Wand, M., and Schultz, T. (2011). "Impact of different feedback mechanisms in emg-based speech recognition," in *12th Annual Conference of the International Speech Communication Association* (Florence). Interspeech 2011.

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012b). "Speaking mode recognition from functional near infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (San Diego, CA), 1715–1718. doi: 10.1109/embc.2012.6346279

Houde, J. F., Nagarajan, S. S., Sekihara, K., and Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: an MEG study. *J. Cogn. Neurosci.* 14, 1125–1138. doi: 10.1162/089892902760807140

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., and Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* 7:056007. doi: 10.1088/1741-2560/7/5/056007

Leuthardt, E. C., Pei, X.-M., Breshears, J., Gaona, C., Sharma, M., Freudenberg, Z., et al. (2011). Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. *Front. Hum. Neurosci.* 6:99. doi: 10.3389/fnhum.2012.00099

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., and Schalk, G. (2015). Electrocorticographic representations of segmental features in continuous speech. *Front. Hum. Neurosci.* 9:97. doi: 10.3389/fnhum.2015.00097

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., et al. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* 7:14. doi: 10.3389/fneng.2014.00014

McFarland, D. J., Miner, L. A., Vaughan, T. M., and Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topogr.* 12, 177–186. doi: 10.1023/A:1023437823106

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Miller, K. J., Leuthardt, E. C., Schalk, G., Rao, R. P., Anderson, N. R., Moran, D. W., et al. (2007). Spectral changes in cortical surface potentials during motor movement. *J. Neurosci.* 27, 2424–2432. doi: 10.1523/JNEUROSCI.3886-06.2007

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., et al. (2014). Direct classification of all american english phonemes using signals from functional speech motor cortex. *J. Neural Eng.* 11:035015. doi: 10.1088/1741-2560/11/3/035015

Müller-Putz, G. R., Scherer, R., Brauneis, C., and Pfurtscheller, G. (2005). Steady-state visual evoked potential (ssvep)-based communication: impact of harmonic frequency components. *J. Neural Eng.* 2, 123–130. doi: 10.1088/1741-2560/2/4/008

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. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., and Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage* 54, 2960–2972. doi: 10.1016/j.neuroimage.2010.10.029

Potes, C., Gunduz, A., Brunner, P., and Schalk, G. (2012). Dynamics of electrocorticographic (ecog) activity in human temporal and frontal cortical areas during music listening. *NeuroImage* 61, 841–848. doi: 10.1016/j.neuroimage.2012.04.022

Price, C. J. (2012). A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading. *Neuroimage* 62, 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Trans. Biomed. Eng.* 51, 1034–1043. doi: 10.1109/TBME.2004.827072

Schultz, T., and Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Commun.* 52, 341–353. doi: 10.1016/j.specom.2009.12.002

Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., et al. (2007). Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface. *NeuroImage* 34, 1416–1427. doi: 10.1016/j.neuroimage.2006.11.005

Stolcke, A. (2002). "SriLM – an extensible extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)* (Denver, CO).

Sutter, E. E. (1992). The brain response interface: communication through visually-induced electrical brain responses. *J. Microcomput. Appl.* 15, 31–45. doi: 10.1016/0745-7138(92)90045-7

Talavage, T. M., Gonzalez-Castillo, J., and Scott, S. K. (2014). Auditory neuroimaging with fMRI and pet. *Hear. Res.* 307, 4–15. doi: 10.1016/j.heares.2013.09.009

Telaar, D., Wand, M., Gehrig, D., Putze, F., Amma, C., Heger, D., et al. (2014). "BioKIT - real-time decoder for biosignal processing," in *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)* (Singapore).

Tervaniemi, M. A., Kujala, A., Alho, K., Virtanen, J., Ilmoniemi, R., and Näätänen, R. (1999). Functional specialization of the human auditory cortex in processing phonetic and musical sounds: a magnetoencephalographic (MEG) study. *Neuroimage* 9, 330–336. doi: 10.1006/nimg.1999.0405

Vaughan, T. M., McFarland, D. J., Schalk, G., Sarnacki, W. A., Krusienski, D. J., Sellers, E. W., et al. (2006). The wadsworth BCI research and development program: at home with BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* 14, 229–233. doi: 10.1109/TNSRE.2006.875577

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

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

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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.

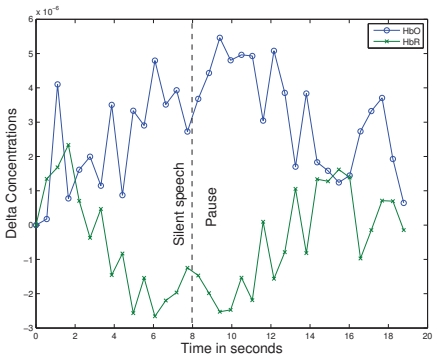


Fig. 1. Average hemodynamic response of subject 2 when speaking silently followed by pausing, for a location in the lower motor cortex.

C. Relevant Brain Areas

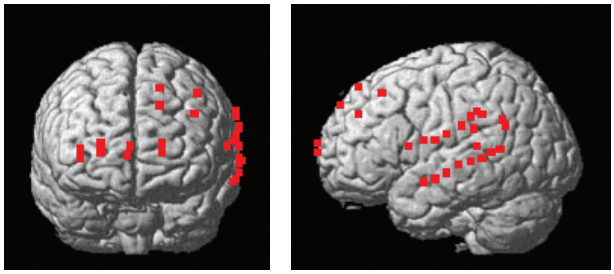
For approximately 90% of the population, the left hemisphere is dominant for speech and language processing. This lateralization is even larger for right-handed individuals (see [8]). To increase the probability of measuring relevant areas, we decided to focus on right-handed subjects in this pilot study. 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:

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

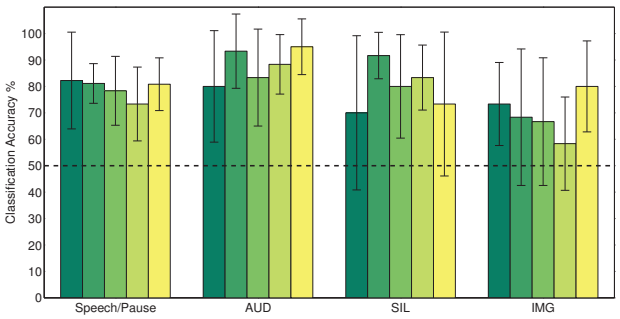
MIBIF then selects the k features f_l with highest Mutual Information with the class labels $\arg\max_l(I(C, f_l))$. We set $k = 30$ after studying the distributions of Mutual Information of features with the class labels.
The *MIBIF* approach presents a fast feature selection technique that uses a high 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

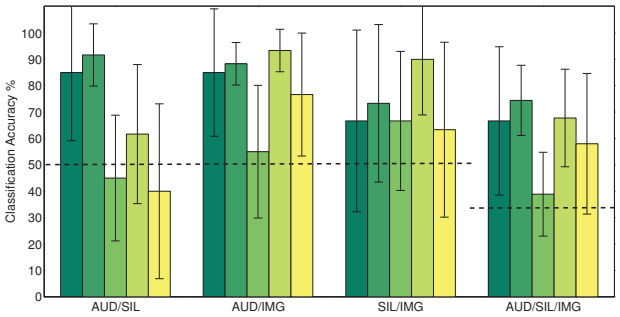
To evaluate our system, we applied a 10-fold person dependent cross-validation approach. For classification, we employ support vector machines with radial basis function kernels on the resulting 30-dimensional feature set S determined with *MIBIF*. SVM parameters c and γ are estimated via cross-validation 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.

Self-paced BCI with NIRS based on speech activity

C. Herff[1], D. Heger[1], F. Putze[1], C. Guan[2], T. Schultz[1]

[1]*CSL, Karlsruhe Institute of Technology, Germany;* [2]*Institute for Infocomm Research, Singapore*

Correspondence: C. Herff, Karlsruhe Institute of Technology, Institute for Anthropomatics, Cognitive Systems Lab (CSL), Adenauerring 4 (50.21), 76131 Karlsruhe, Germany. E-mail: christian.herff@kit.edu

Abstract. To enable Brain Computer Interfaces (BCIs) to be used intuitively, they should use an input paradigm which is as natural as possible, while the usage of the device is as convenient as possible. In this study, we show that functional near infrared spectroscopy (fNIRS) signals can be used to automatically detect the user's intent to use the system by detecting asynchronous speech activity, which is a very natural form of communication. Thereby, we take a first step towards self-paced BCIs with fNIRS based on speech activity.

Keywords: fNIRS, near infrared spectroscopy, asynchronous, self-paced, speaking modes

1 Introduction

Functional near infrared spectroscopy (fNIRS) is rapidly gaining attention as an imaging modality for Brain Computer Interfaces (BCIs) [Matthews et al., 2008]. First commercial systems are already available [Naito et al., 2007]. The majority of studies on fNIRS based BCIs rely on motor imagery and are operated stimulus locked. This means that the user can only interact with the system in predefined intervals. For an intuitive BCI, the users should be able to decide on their own when they want to operate the BCI. This is leading to a self-paced BCI, detecting idle and voluntary control states. Identifying segments containing activity is a known field of research in speech technologies [Laskowski and Schultz, 2006]. In this study, we show that the fNIRS signals measured during different speaking tasks can be automatically segmented into segments containing speech activity and segments without speech activity.

2 Material and Methods

We recorded 5 male subjects using a Dynot232 system with 32 transmitters and 32 receivers, sampling at 1.81 Hz. On the left hemisphere, four optodes were placed on Broca's area, 10 on Wernicke's area and six on 6 on the lower motor cortex. Additionally, we covered the prefrontal cortex with 12 optodes. Limiting our analysis to channels with an inter-optode distance between 2.5 and 4.5cm, we considered 252 channels of oxygenated and deoxygenated hemoglobin values. All subjects were recorded over an interval of 37.5 minutes in which they conducted three types of speech activity as prompted by messages displayed on a screen. These were: Normal audible speech; silent speech, which consisted of moving the articulatory muscles as if speaking, but without sound production and speech imagery, for which the subjects had to imagine themselves reading out the displayed sentence. Users were asked to relax when they were not prompted to conduct speech activity. See [Herff et al., 2012b] for more details on the experiment design.
The continuous hemoglobin values were then dissected into almost completely overlapping 10 second long windows, allowing for continuous decoding. A simple feature, measuring the difference between the mean of the first 4.5 seconds and the second 4.5 seconds in every window, was extracted for oxygenated and deoxygenated hemoglobin of every channel resulting in a total of 504 features. Using the *Mutual Information based Best Individual Feature (MIBIF)* algorithm [Ang et al., 2008], we selected a subset of 50 features which contained the most relevant information.
The data was labeled as containing speech activity (of any of the three modes) or not containing speech activity based on the experiment timings. This approach yielded more reliable results than identifying the modes individually, as more training data for the classes is available. Nevertheless, previous results [Herff et al., 2012b] show that all three modes can be discriminated from inactivity. We trained Support Vector Machines on these selected features to create an automated segmentation method for speech activity based on the fNIRS data.
A 10 fold cross-validation approach was used to evaluate our segmentation method. Both training and test set features were z-normalized with the mean and standard deviation of the training set. No data from the test set was used for feature selection, normalization or training of the classifier.

3 Results

Figure 1 shows an example segmentation of our method for one of the folds of subject 3. The cross-marked line line shows speech activity as labeled using the experiment prompts and the red line shows speech segments predicted

by the proposed algorithm. Note that the ground truth might be flawed as well, since we could not control whether our subjects were actually performing the speech tasks when prompted. This typical example clearly shows that our

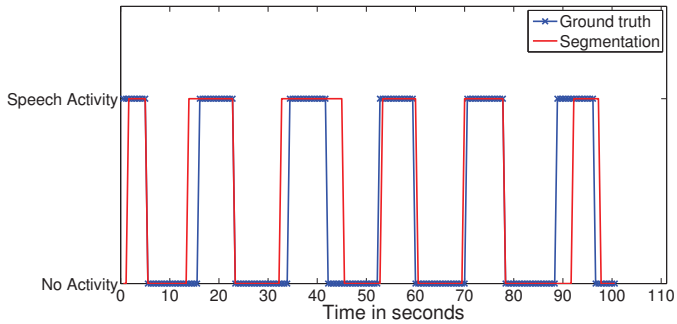


Figure 1: *Segmentation of subject 3's data into speech activity and non-activity. The cross-marked line is ground truth while the red line shows segmentation by our method.*

method extracts most of the segments containing speech activity reliably. The performance is promising on all 5 subjects and is significantly better than chance ($p < 0.05$). Precision and recall are very stable over all subjects and false positive rates are low across all subjects, as well. Frame based accuracies are high with an average of 74% and are only slightly lower than the 79% average accuracy achieved in a stimulus locked experiment on the same dataset [Herff et al., 2012b]. Table 1 lists all results for all subjects.

Table 1: *Results for speech activity detection with fNIRS across 5 subjects.*

Subject	Accuracy	True Positive Rate	False Positive Rate	True Negative Rate	Precision	Recall
Subject 1	0.72	0.61	0.20	0.80	0.69	0.61
Subject 2	0.74	0.62	0.17	0.83	0.71	0.62
Subject 3	0.79	0.63	0.11	0.89	0.79	0.63
Subject 4	0.73	0.57	0.16	0.84	0.72	0.57
Subject 5	0.74	0.63	0.17	0.83	0.72	0.63
Average	**0.74**	**0.61**	**0.16**	**0.84**	**0.73**	**0.61**

4 Discussion

We have shown that segments containing speech activity can be continuously decoded from those not containing speech activity based on fNIRS signals alone. Thereby, we make an important step towards self-paced BCI with fNIRS. All methods used in this analysis can be easily transferred to an online scenario. The results in this first study can possibly be extended to make use of the cross-subject capabilities of fNIRS [Herff et al., 2012a].

Acknowledgments

This project received financial support by the Concept for the Future of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

References

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE International Joint Conference on Neural Networks. IJCNN*, pages 2390–2397. IEEE.

Herff, C., Heger, D., Putze, F., Guan, C., and Schultz, T. (2012a). Cross-subject classification of speaking modes using fnirs. In Huang, T., Zeng, Z., Li, C., and Leung, C., editors, *Neural Information Processing*, volume 7664 of *Lecture Notes in Computer Science*, pages 417–424. Springer Berlin Heidelberg.

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012b). Speaking mode recognition from functional near infrared spectroscopy. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 1715 –1718.

Laskowski, K. and Schultz, T. (2006). Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE.

Matthews, F., Pearlmutter, B., and Ward, T. (2008). Hemodynamics for brain-computer interfaces. *Signal Processing*, pages 87–94.

Naito, M., Michioka, Y., Ozawa, K., Ito, Y., Kiguchi, M., and Kanazawa, T. (2007). 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.*, pages 1028–1037.

Classification of mental tasks in the prefrontal cortex using fNIRS

Christian Herff, Dominic Heger, Felix Putze, Johannes Hennrich, Ole Fortmann and Tanja Schultz

Abstract— **Functional near infrared spectroscopy (fNIRS) is rapidly gaining interest in both the Neuroscience, as well as the Brain-Computer-Interface (BCI) community. Despite these efforts, most single-trial analysis of fNIRS data is focused on motor-imagery, or mental arithmetics. In this study, we investigate the suitability of different mental tasks, namely mental arithmetics, word generation and mental rotation for fNIRS based BCIs. We provide the first systematic comparison of classification accuracies achieved in a sample study. Data was collected from 10 subjects performing these three tasks.**
An optode template with 8 channels was chosen which covers the prefrontal cortex and only requires less than 3 minutes for setup. Two-class accuracies of up to 71% average across all subjects for mental arithmetics, 70% for word generation and 62% for mental rotation were achieved discriminating these tasks from a relax state.
We thus lay the foundation for fNIRS based BCI using additional mental strategies than motor imagery and mental arithmetics. The tasks were chosen in a way that they might be used for user state monitoring, as well.

I. INTRODUCTION

A. Motivation

Functional Near Infrared Spectroscopy (fNIRS) is a state-of-the-art non-invasive brain imaging technology based on hemodynamic responses to cortical activities. The effects that can be measured using fNIRS (see Section I-B) are the same ones observed with fMRI, the de facto standard in neuroimaging. Compared to fMRI, fNIRS is portable, cheap and does not confine the subjects. Measuring the very reliable hemodynamic responses and offering a very good spatial resolution, fNIRS has advantages over EEG, the standard in Brain-Computer-Interface (BCI) research, as well.
The paradigm used for BCI control can affect classification and recognition accuracies in EEG significantly [1]. Even though this has been studied in detail in EEG, there is, to the best of our knowledge, no systematic comparison of paradigms and the resulting accuracies for classification in fNIRS. To investigate the suitability of different mental tasks for BCI control and the discriminability of the tasks from relax and from each other, we conducted experiments with three mental tasks, namely mental arithmetics, word generation and mental rotation. An optode layout on the forehead, measuring hemodynamic responses in the prefrontal cortex, was used to allow for fast setup times.
Besides being useful in BCIs, the robust classification of these tasks might also enable user state monitoring, as we could classify which type of task is currently occupying the user. fNIRS could be used to classify user states, which are currently non observable. This might be useful in classroom settings, where it could help evaluate what type of problem, mathematical, language or orientation, the student is currently struggling with.
So far, motor imagery is the most popular paradigm for BCI research in EEG and has been published first for fNIRS based BCI, as well [2]. Accuracies achieved in these BCIs are usually good and motor imagery suits most users. However, setup of EEG caps is usually very time consuming. In fNIRS experiments, the identification of relevant areas for optode placement on the motor cortex is complex and cumbersome. To measure fNIRS signals, optodes require skin contact, a constraint often hard to meet on the motor cortex, where hair has to be moved aside under the optodes in a lengthy procedure. Our optode template, on the other hands, requires little anatomical knowledge and can be setup in less than three minutes.
Mental arithmetic [3], word generation [4] and mental rotation [5] have been shown to create hemodynamic responses in the prefrontal cortex, our area of interest. Mental arithmetic has been used successfully in single trial analysis of fNIRS data [6]. Ogata et al. have conducted first single trial experiments with different mental tasks in the prefrontal cortex [7]. However, in their study with only 10 trials per subject and task, they neither discriminate the tasks from one another nor compare classification accuracies.

B. Functional Near Infrared Spectroscopy

Light in the near infrared range of the electromagnetic spectrum (620 nm - 1000 nm) disperses through most biological tissues like bones and skin. Hemoglobin, the oxygen-carrying part of blood, on the other hand absorbs near infrared light. As changes in blood oxygenation in cortical areas are triggered by neural activity, hemoglobin levels change with neural activation. fNIRS makes use of this effect to measure cortical activity by shining near infrared light into the subject's head with light-sources and measuring the light intensities transmitted through the head with detector-optodes. For a source-detector pair with distance l, the measurement position is located in the middle between the two in a depth of approximately $l/2$ and is denoted as a channel. Oxygenated hemoglobin (*HbO*) and deoxygenated hemoglobin (*HbR*) have different light absorption characteristics (with absorption coefficients α_{HbO} and α_{HbR}) and thus their concentration changes (denoted as Δ*HbO* Δ*HbR*) can be calculated from the changes in light intensities (ΔOD) using the modified Beer-Lambert law [8]:

$$\Delta HbO = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbO}}, \qquad \Delta HbR = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbR}},$$

All authors are with the Cognitive Systems Lab, Karlsruhe Institute of Technology, Adenauerring 4, 76131 Karlsruhe, Germany. christian.herff@kit.edu

where b is the length of the photon path between sources and detectors, along which the light travels.

A typical hemodynamic response to cortical activity rises

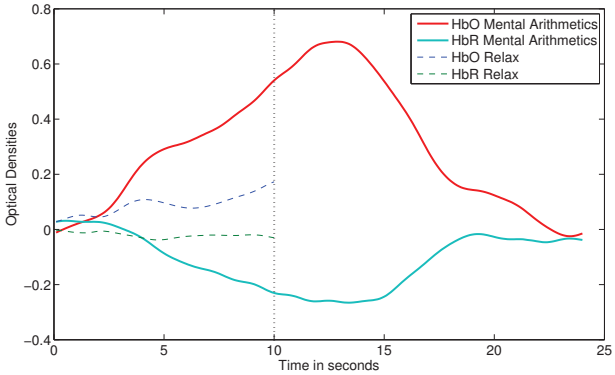


Fig. 1. Average hemodynamic response of subject 4 in channel 7 performing mental arithmetics (solid lines) and relax tasks (dashed lines). The dotted line indicates end of mental task.

during activity for *HbO* and returns to baseline after the end of the activation. *HbR* levels should respond inverted, i.e. decrease upon activity and rise back to baseline in rest periods. Figure 1 shows an average hemodynamic response of subject 4 to mental arithmetics and the return to baseline when resting. The Figure also illustrates that *HbO* and *HbR* do not change significantly during the averaged relax trial.

II. MATERIAL AND METHODS

A. Experimental Setup

To measure the hemodynamic responses in the prefrontal cortex, we used an Oxymon Mark III by Artinis Medical Systems. Four transmitter optodes, transmitting at two wavelength of 765 nm and 856 nm each, and 4 receiver optodes were placed on the subjects' foreheads. Transmitter and receiver optodes were placed 3.5 cm apart. In this setup, every receiver optode was measuring light intensities from two transmitter optodes resulting in a total of 8 channels of ΔHbO and ΔHbR data. Figure 2 illustrates our optode setup. An experienced assistant needs less than three minutes to fix the optode holder to the subject's forehead while assuring high data quality. Data was sampled at 10 Hz.

The experiment consisted of three different tasks the subjects

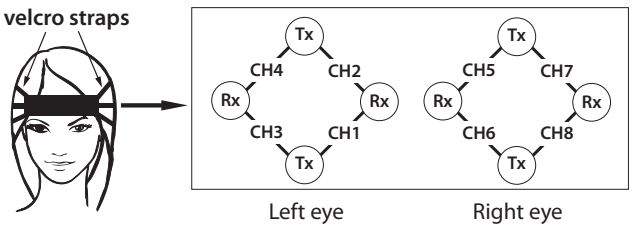


Fig. 2. Optode placement on subject's head. Transmitter optodes are labelled Tx. Receiver optodes are marked Rx.

had to process during trials. These were:

- **Mental Arithmetics (MA):** The subjects were asked to repeatedly subtract a given minuend between 7 and 19 (10 excluded) starting with a given number between 501 and 999.
- **Word Generation (WG):** The subjects were asked to imagine words starting with a given letter.
- **Mental Rotation (MR):** The subjects were asked to visualize rotating the shown 3D object around the x-axis.

Trials were presented to the subjects on a screen for 10 seconds in a random order. After every trial of one of the three tasks, the subjects had to rest for 15 seconds to ensure that hemoglobin levels could return to baseline levels. None of the tasks required any input by the user assuring that there are no systematic motion artifacts in our data. We recorded 30 trials of each task for every subject. A total of 30 relax trials were inserted randomly after resting periods to gather data during a mental relax state without prior activation. The subjects continued to rest motionlessly in these intervals and did not receive specific instructions. In total, we collected 120 trials per subject. Longer pauses of 5 minutes were included after each 15 minute block in which the subjects could drink and talk to the experiment supervisor. This resulted in a total recording time of 52.5 minutes per subject. In total, we recorded 10 right handed subjects (3 female) with a mean age of 23 and a mean Edinburgh handedness score [9] of 83. All subjects were informed prior to the experiment and gave written consent.

B. Signal Preprocessing

To remove heartbeat artifacts and long period shifts from the 8 channels of ΔHbO and ΔHbR data, we bandpass filtered the signals from 0.01 Hz to 0.6 Hz using elliptic IIR filter with filter order 6. Subsequently, linear trends were removed in 5 minute blocks using linear detrending.

In an excellent comparison of movement artifact reduction techniques for fNIRS, Cooper et. al. [10] suggest the wavelet denoising technique as the most reliable to remove movement artifacts from fNIRS data. We applied the wavelet artifact removal technique suggested in [11] to our signals. For this procedure, the ΔHbO and ΔHbR data $y(t)$ of every channel is transformed using the general wavelet transformation:

$$y(t) = \sum_k c_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_k d_{jk}\psi_{jk}(t)$$

with $c_{j_0 k}$ and d_{jk} being the approximation and detail coefficients and $\phi_{jk}(t)$ and $\psi_{jk}(t)$ being scaling and wavelet functions. The parameter j represents the dilation, with j_0 being the coarsest scale in the decomposition, k is the translation parameter. Assuming a normal distribution of wavelet coefficients, we can easily estimate the probability of coefficients higher than a given coefficient. Hemodynamic signals should have a smooth probability distribution and very low variance. Based on these observations, one can remove artifacts by removing wavelet coefficients with probabilities smaller than a cutoff threshold α. We used a

threshold of 10 times the interquartile distance. As none of our tasks contain any systematic movement and since our subjects were sitting relatively still, we applied a very low threshold to filter only the most unlikely wavelet coefficients. After preprocessing, trials were extracted based on the experiment timing. Each 10 second trial was baseline normalized by subtracting the mean of the 5 seconds prior to the trial. A label corresponding to one of the 3 tasks or relax was assigned to each of the trials. We did not include the resting periods directly after each trial, in which hemoglobin levels returned to baseline, nor the long pauses of 5 minutes in our analysis.

C. Feature Extraction

Feature extraction for single trial fNIRS analysis usually uses simple features based on a typical hemodynamic response. Rising and falling trends in the trials are often extracted by subtracting the mean μ of the first half of the trials from the mean of the second half of the trial [12], [13]. We extend on this idea, by looking for the largest increase and decrease between the mean of two adjacent frames of size *fs*. As beginning and end of the hemodynamic response vary between subjects and even trials, we extract both decrease and increase. These features are denoted as $f^{\uparrow}_{t,c}$ and $f^{\downarrow}_{t,c}$, respectively.
In typical hemodynamic responses, ΔHbO and ΔHbR are strongly negatively correlated [14] with changes more pronounced in the ΔHbO data. To reduce the size of the feature space, we only extracted features for ΔHbO and did not include the mostly redundant ΔHbR data. In total, we extract two features for every trial t in every channel c in the following manner:

$$f^{\uparrow}_{t,c} = \max_{i \in [fs, len(t)-fs]} (\mu(\Delta HbO^{t}_{c,i:i+fs}) - \mu(\Delta HbO^{t}_{c,i-fs:i}))$$

$$f^{\downarrow}_{t,c} = \max_{i \in [fs, len(t)-fs]} (\mu(\Delta HbO^{t}_{c,i-fs:i}) - \mu(\Delta HbO^{t}_{c,i:i+fs}))$$

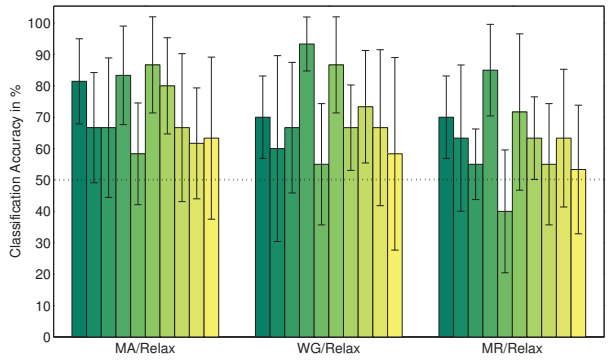
We chose a framesize of 3.5 seconds in this study.
In total, we thus extracted 16 features for each of the 120 trials.

D. Evaluation

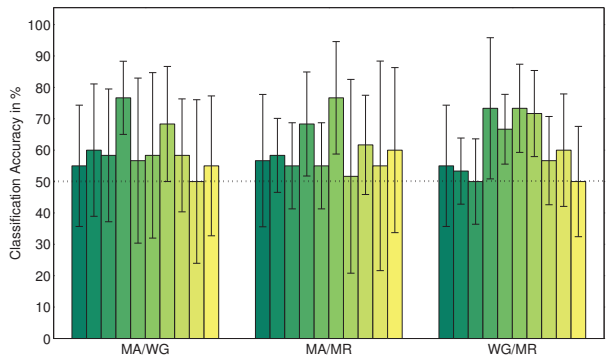
To judge the suitability of the different mental tasks for fNIRS based BCI or user state monitoring, we evaluated our system using a 10-fold cross-validation approach. We divided the data into 10 equally sized folds and trained a Linear Discriminant Analysis (LDA) classifier on the features of 9 of these folds and tested on the features of the remaining fold. This was repeated 10 times in a round-robin manner.
We evaluated classification accuracies of all mental tasks (MA, WG, MR) against relax and of the mental tasks against each other.

III. RESULTS

Figure 3 illustrates the differences in average hemodynamic responses which serve as basis for our classification. Decreases and increases in ΔHbO occur in different channels and at different points in time for the different mental tasks, leading to earlier returns to baseline and different amplitude of hemodynamic responses. Our extracted features $f^{\uparrow}_{t,c}$ and $f^{\downarrow}_{t,c}$ make use of this fact and allow us to distinguish reliably between the tasks and the relax state.
A complete overview of all classification results is presented in Figure 4. Part (a) depicts classification results of the three



(a)



b)

Fig. 4. Classification results of all 10 subjects for experiments against relax (a) and against each other (b). Each bar represents one subject in one experiment. Whiskers indicate standard deviations. Dotted line denotes naive classification rate.

tasks (MA, WG, MR) against relax for all 10 participants. Classifying mental arithmetics from relax worked with an average of 71% accuracy. This result is comparable to that achieved in [15]. Differentiating between word generation and relax yielded 70% average accuracy. Accuracies for mental rotation were lower with an average of 62%. There was no significant difference in the classification performance of mental arithmetics and word generation, as tested by a Wilcoxon rank-sum test ($p = 0.4280$), but both were significantly better than mental rotation ($p < 0.01$). All three tasks could be discriminated from the relax state significantly better than naïve classification ($p < 0.01$). These results show that all three mental tasks are effective paradigms for fNIRS based BCI or user state monitoring, but word generation and mental arithmetics work more reliably than mental rotation.
Classification results among the three different tasks are shown in part (b) of Figure 4. Mental arithmetics was

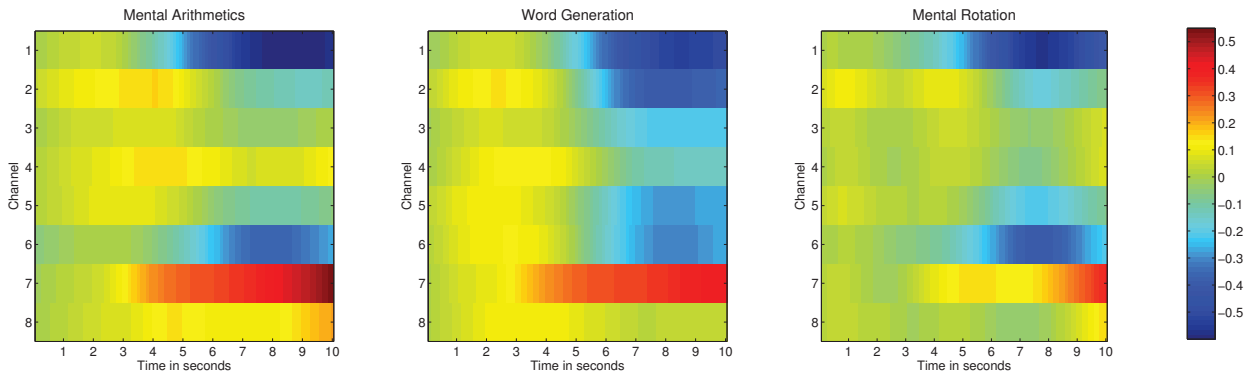


Fig. 3. Average hemodynamic responses in *HbO* of subject 4 during all tasks.

discriminated from word generation with an average performance of 60%. The LDA classifier achieved 60% between mental arithmetics and mental rotation. Word generation and mental rotation yielded an average result of 61%. Results for discrimination between the three different tasks were significantly better than naïve classification ($p < 0.01$), as well. The fact that all these experiments yield significant results shows that fNIRS based BCIs with multiple tasks is feasible and might reach results comparable to 4-class systems in EEG [16]. Table I summarizes our findings with average results and standard deviations across all 10 folds of all 10 subjects.

TABLE I

AVERAGE CLASSIFICATION RESULTS AND STANDARD DEVIATIONS ACROSS SUBJECTS IN %. RESULTS MARKED WITH * ARE SIGNIFICANTLY BETTER THAN NAIVE CLASSIFICATION.

	MA	WG	MR	MA/WG	MA/MR	WG/MR
Acc.	71*	70*	62*	60*	60*	61*
Std.	10.3	12.1	12.2	7.6	7.5	9.5

IV. CONCLUSION

In a study with 10 subjects, we have shown that fNIRS signals in response to three different mental tasks can be reliably discriminated both from a relax state and from each other. The optode template we used supplied us with good measurements of hemodynamic responses in relevant parts of the prefrontal cortex and can be set up quickly. We thus present the first systematic comparison of classification accuracies of the mental tasks mental arithmetics, word generation and mental rotation and prove that all tasks are suitable for fNIRS based BCI. All of these tasks have been of prior interest to the neuroscientific community, but have never been evaluated in single-trial analysis with fNIRS. Classification accuracies and hemodynamic patterns lay the foundation for fNIRS based BCI using different mental tasks than the established motor imagery paradigm and might be used for state monitoring.

REFERENCES

[1] E VC Friedrich, R Scherer, and C Neuper, "The effect of distinct mental strategies on classification performance for braincomputer interfaces," *International Journal of Psychophysiology*, vol. 84, no. 1, pp. 86 – 94, 2012.

[2] 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.

[3] M Tanida, K Sakatani, R Takano, and K Tagai, "Relation between asymmetry of prefrontal cortex activities and the autonomic nervous system during a mental arithmetic task: near infrared spectroscopy study," *Neuroscience Letters*, vol. 369, no. 1, pp. 69 – 74, 2004.

[4] E Watanabe, A Maki, F Kawaguchi, K Takashiro, Y Yamashita, H Koizumi, and Y Mayanagi, "Non-invasive assessment of language dominance with near-infrared spectroscopic mapping," *Neuroscience Letters*, vol. 256, no. 1, pp. 49 – 52, 1998.

[5] N Shimoda, K Takeda, I Imai, J Kaneko, and H Kato, "Cerebral laterality differences in handedness: A mental rotation study with nirs," *Neuroscience Letters*, vol. 430, no. 1, pp. 43 – 47, 2008.

[6] 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.

[7] H Ogata, T Mukai, and T Yagi, "A study on the frontal cortex in cognitive tasks using near-infrared spectroscopy," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, aug. 2007, pp. 4731 –4734.

[8] 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.

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

[10] R Cooper, J Selb, L Gagnon, D Phillip, H W Schytz, H K Iversen, M Ashina, and D A Boas, "A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy," *Frontiers in Neuroscience*, vol. 6, no. 147, 2012.

[11] B Molavi and GA Dumont, "Wavelet based motion artifact removal for functional near infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 31 2010-sept. 4 2010, pp. 5 –8.

[12] C Herff, F Putze, D Heger, C Guan, and T Schultz, "Speaking mode recognition from functional near infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 1715 –1718.

[13] C Herff, D Heger, F Putze, C Guan, and T Schultz, "Cross-subject classification of speaking modes using fnirs," in *Neural Information Processing*, T Huang, Z Zeng, C Li, and C Leung, Eds., vol. 7664 of *Lecture Notes in Computer Science*, pp. 417–424. Springer Berlin Heidelberg, 2012.

[14] X Cui, S Bray, and AL Reiss, "Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics," *NeuroImage*, vol. 49, no. 4, pp. 3039–46, Feb. 2010.

[15] S D Power, A Kushki, and T Chau, "Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by nirs," *PLoS ONE*, vol. 7, no. 7, pp. e37791, 07 2012.

[16] E VC Friedrich, R Scherer, and C Neuper, "Long-term evaluation of a 4-class imagery-based braincomputer interface," *Clinical Neurophysiology*, 2013.

frontiers in
HUMAN NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 16 January 2014
doi: 10.3389/fnhum.2013.00935

Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS

Christian Herff*, Dominic Heger, Ole Fortmann, Johannes Hennrich, Felix Putze and Tanja Schultz

Cognitive Systems Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

Edited by:
Leonid Perlovsky, Harvard University and Air Force Research Laboratory, USA

Reviewed by:
Hasan Ayaz, Drexel University, USA
Megan Strait, Tufts University, USA

***Correspondence:**
Christian Herff, Cognitive Systems Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology, Adenauerring 4, 76131 Karlsruhe, Germany
e-mail: christian.herff@kit.edu

When interacting with technical systems, users experience mental workload. Particularly in multitasking scenarios (e.g., interacting with the car navigation system while driving) it is desired to not distract the users from their primary task. For such purposes, human-machine interfaces (HCIs) are desirable which continuously monitor the users' workload and dynamically adapt the behavior of the interface to the measured workload. While memory tasks have been shown to elicit hemodynamic responses in the brain when averaging over multiple trials, a robust single trial classification is a crucial prerequisite for the purpose of dynamically adapting HCIs to the workload of its user. The prefrontal cortex (PFC) plays an important role in the processing of memory and the associated workload. In this study of 10 subjects, we used functional Near-Infrared Spectroscopy (fNIRS), a non-invasive imaging modality, to sample workload activity in the PFC. The results show up to 78% accuracy for single-trial discrimination of three levels of workload from each other. We use an *n*-back task ($n \in \{1, 2, 3\}$) to induce different levels of workload, forcing subjects to continuously remember the last one, two, or three of rapidly changing items. Our experimental results show that measuring hemodynamic responses in the PFC with fNIRS, can be used to robustly quantify and classify mental workload. Single trial analysis is still a young field that suffers from a general lack of standards. To increase comparability of fNIRS methods and results, the data corpus for this study is made available online.

Keywords: fNIRS, near-infrared spectroscopy, prefrontal cortex, workload, mental states, user state monitoring, n-back, passive BCI

1. INTRODUCTION

Functional Near-Infrared Spectroscopy (fNIRS) is an imaging modality measuring hemodynamic processes in the brain. It provides insights into the same activation patterns as functional Magnetic Resonance Imaging (fMRI), the de facto standard in neuroscience research, while not confining the subject in a small space. Thereby, it allows for measurements of large subject populations outside of clinical environments. Besides montages covering the whole head, fNIRS sources and detector optodes can also be placed on the subjects head to measure exactly the parts of the cortex that contain relevant activations for the investigated task. When the region of interest is known beforehand, this can be used to design optode holders that can be fixed in place in less than 1 min. Potentially, fNIRS could thus be used in real world scenarios, as well.

Most fNIRS studies investigate differences in average activation patterns for different conditions. Only very recently has fNIRS been used to classify single-trial activations for Brain-Computer Interfacing (Coyle et al., 2007). A Brain-Computer Interface is a communication channel between the brain and a computer through interpretation of neural activation pattern (Wolpaw et al., 2002). Nearly all existing single-trial studies differentiate fNIRS patterns of subjects performing a cognitive task from the rest state or no-control state. The most frequently used paradigm is motor-imagery (Sitaram et al., 2007).

Recently, neural signals have been used to adapt and complement traditional input sources, such as keyboard and mouse, by adapting the interface to the users' state instead of directly controlling the interface. These so called passive Brain-Computer Interfaces (Cutrell and Tan, 2008; Zander and Kothe, 2011) mostly use the Electroencephalogram (EEG). Passive Brain-Computer Interfaces (BCIs) often measure a user's state and adapt a user interface accordingly. In fNIRS, multiple studies investigate mental arithmetics (Ang et al., 2010a) to monitor users' engagement in arithmetic tasks. Power et al. (2012) investigate the consistency of mental arithmetic classification across different sessions. Instead of recognizing mental arithmetics, Power et al. (2010) show that mental arithmetic and music imagery lead to distinct activation patterns that can be classified in single trial analysis. Following up on this idea, Herff et al. (2013) differentiate three different mental tasks, namely mental arithmetics, mental rotation and word generation. Girouard et al. (2009) distinguish between two difficulty levels in the popular game Pac-Man, instead of discriminating from a rest state. Ang et al. (2010b) show robust classification for three difficulty levels in mental arithmetics using fNIRS to evaluate numerical cognition class-room settings. While Ang et al. focus on the differentiation of difficulty levels, our focus is on the classification of mental workload induced by a memory task. Recently, Hirshfield et al. (2011) evaluated the type of cognitive demand placed on a user by different types of tasks. The focus of their study is on the type of workload, while we are aiming at the quantification of workload in this study.

In a multi-modal study using blood volume pressure, respiration measures, electrodermal activity and EEG, Jarvis et al. (2011) measured workload in a driving simulator to adapt a driving assistant. Workload has been of interest in the fNIRS community, as well. Cognitive workload has been assessed for air-traffic controllers in several studies Ayaz et al. (2010, 2012). Izzetoglu et al. (2003) show that task load in the Warship Commander tasks yield distinct hemodynamic responses on average. Aiming at a usage for BCI, Ayaz et al. (2007) analyze workload induced by the *n*-back tasks, but limit their results to grand averages, as well. However, these studies look at average hemodynamic responses and do not attempt single trial analysis. To use these findings to adapt interfaces to the user's current workload, the hemodynamic responses have to be analyzed in single trial. Proving that a cognitive task yiels hemodynamic responses on average does not automatically mean that the activations can be robustly recognized in single trial, which is necessary if interfaces should be adapted. In this work, we provide evidence that different levels of workload yield hemodynamic responses that can be robustly classified without averaging.

Findings in EEG Brouwer et al. (2012); Berka et al. (2007) show that workload induced by the *n*-back task can be classified in single trial. Baldwin and Penaranda (2012) demonstrate how the models trained on one workload condition can be transferred to others in EEG. In this study, we show that the workload induced by different *n*-back conditions results in hemodynamic responses that are consistent enough to be classified on a single trial basis. We use an *n*-back task to induce different levels of workload, forcing subjects to continuously remember the last one, two, or three of rapidly changing items. To enable realistic passive BCIs, we not only evaluate whether a user is engaged in a task, but quantify the level of mental workload the user experiences during the *n*-back task ($n \in \{1, 2, 3\}$). Thereby, we quantify workload using fNIRS.

In functional imaging studies, the prefrontal cortex (PFC) has been identified to be among the relevant areas for memory related tasks (Smith and Jonides, 1997). The PFC has been found to be relevant both in PET (Smith and Jonides, 1997) and fMRI studies (Cohen et al., 1997). An in depth meta-analysis of *n*-back studies using fMRI (Owen et al., 2005) confirms the importance of the PFC for *n*-back. Hoshi et al. (2003) show spatio temporal changes for working memory tasks in the PFC using fNIRS. Their analysis is based on averages and does not include single trial analysis, but confirms that fNIRS is ideally suited for measurements of the PFC. An fNIRS headset can be quickly fixed to the forehead and enables measurements of the PFC within minutes, while guaranteeing high data quality. In an investigation using finger tapping and fNIRS, Cui et al. (2010b) show that the delay in fNIRS-based BCIs can be reduced to further improve the usability of fNIRS in real-life scenarios. Workload induced by a memory task and fNIRS-based measurement of the PFC are thus an ideal combination for a realistic passive BCI to monitor workload levels.

2. MATERIALS AND METHODS

2.1. *n*-BACK

In the *n*-back task, users have to continuously remember the last *n* of a series of rapidly flashing letters. The *n*-back task requires subjects to react when a stimulus is the same as the *n*-th letter before the stimulus letter. We denote a (letter) stimulus, which is the same as the one *n* previously as a target. Subjects had to press the space key on a keyboard when they encountered a target. With increasing *n* the task difficulty increases, as the subjects have to remember more letters and continuously shift the remembered sequence. Performance in this task can be evaluated by measuring the amount of missed targets, when the subjects do not press the key for a target and through the amount of wrong reactions, when the subjects incorrectly identify a stimulus letter as a target.

2.2. NIRS DATA RECORDING

Like fMRI, fNIRS measures changes in blood oxygenation in brain areas triggered by neural activity. Using light in the near-infrared range of the electromagnetic spectrum (620–1000 nm), which disperses through most biological tissue but is absorbed by hemoglobin, the level of oxygenated and deoxygenated hemoglobin (*HbO* and *HbR*) can be estimated using the modified Beer-Lambert law (Sassaroli and Fantini, 2004).

We used an Oxymon Mark III by Artinis Medical Systems to measure fNIRS signals. The system uses two wavelength of 765 and 856 nm and outputs concentration changes of *HbO* and *HbR*. To measure hemodynamic activity in the PFC, we attached four transmitter and four receiver optodes to the forehead. Each detector measures time-multiplexed from two sources, located at a distance of 3.5 cm, resulting in a total of 8 channels of *HbO* and *HbR*. Our signals were sampled at 25 Hz.

Figure 1 shows the placement of our optodes on the subjects' forehead. The recording setup on the forehead is very simple and needs less than 3 min to be fixed in place and to assess data quality.

2.3. EXPERIMENT DESIGN

In our experiment, we investigated 10 trials each of 1-,2-, and 3-back tasks. Each trial contained 3 ± 1 targets. The experiment was presented to the subjects on a screen, which was placed in front of them in 50 cm distance.

A trial consisted of 5 s of instruction, informing the subject which task (1-,2- or 3-back) was about to start. The trial then presented a new letter every 2 s. Every letter was displayed for 500 ms. The screen was left blank for the remaining 1.5 s. A total of 22 letters was presented during every trial resulting in a trial length of 44 s. Subsequently, a cross was displayed for 15 s during which the subjects were asked to relax to ensure that hemoglobin levels returned to baseline. We excluded these periods from our analysis, as they are strongly influenced by the previous hemodynamic responses. After half of the trials, an additional 10 s of the resting cross were displayed to have data periods with no activity to be used as RELAX trials. We intentionally use periods with true relax signals for our analysis instead of periods in which *HbO* and *HbR* returned to baseline. **Figure 2** shows the experiment protocol. The order of the different *n*-back conditions was pseudo-randomized. A 150 s break during which the subjects could drink or chat was included after 15 trials. The entire experiment had a recording time of 37 min (30 trials of 64 s, 15 relax trials of 10 s and 150 s in the middle).

The fNIRS data was recorded continuously during the entire session. The trials were segmented afterwards based on the

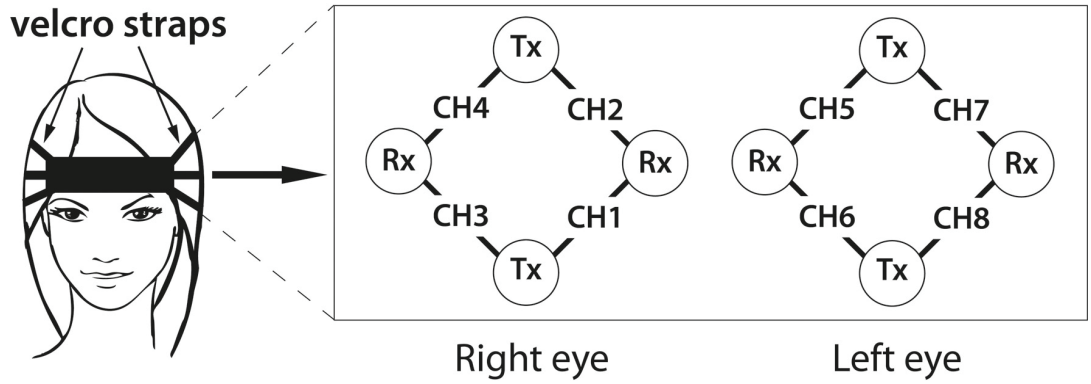


FIGURE 1 | Optode placement in our experiment. Transmitter optodes are marked as Tx, while Rx indicates receiver optode positions.

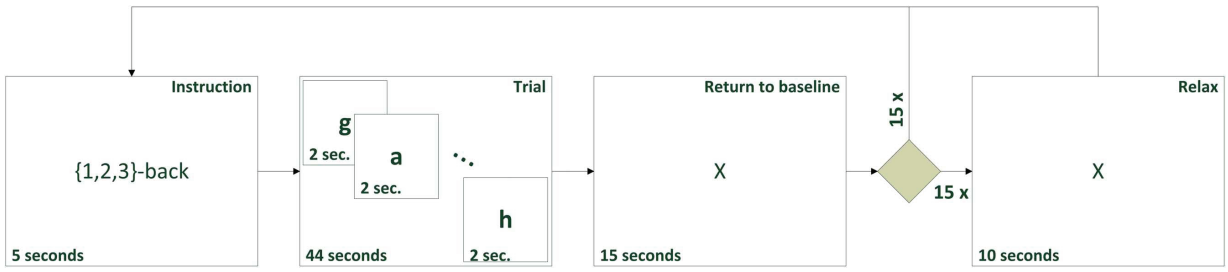


FIGURE 2 | Experimental design for *n*-back task.

time sequence induced by the described experimental setup. In addition to the recorded fNIRS data, subjects filled out a questionnaire regarding their age, occupation, handedness and a series of questions about the experiment on a 6-point Likert scale. The scale ranged from "no agreement" (1) to "complete agreement" (6) for a given statement. We asked our subjects how much they agreed with the statements "The *n*-back task was demanding," to evaluate subjective workload. Subjects were asked to judge their level of concentration during the first and second half of the experiment by indicating their agreement with the statement "I was very concentrated." Additionally, subjects indicated their agreement with the phrase "The system is comfortable to wear." Lastly, we evaluated whether our participants thought that the duration of the experiment was appropriate. Section 3.1 contains results of the questionnaire evaluation.

2.4. PARTICIPANTS

In this study, we recorded 10 subjects (4 females) with a mean age of 22 years. Using the Edinburgh handedness inventory Oldfield (1971), we evaluated the handedness of our subjects. In total, we had 8 right-handed and 2 left-handed participants. All subjects had normal or corrected to normal vision. The participants were informed prior to the experiment and gave written consent. None of the subjects had ever taken part in an *n*-back study before to ensure that no training effects are present.

To increase comparability between fNIRS methods and results, the complete data collected in this study will be shared with the community (see Section 4.1).

2.5. SIGNAL PROCESSING AND ARTIFACT REMOVAL

The signals measured by fNIRS are subject to biological and technical artifacts. Cardiovascular effects like heart-beat, respiration and slow waves (e.g., Mayer Waves) influence the recorded data. Movement artifacts which alter the position of the optodes and lift them off the scalp, causing spikes in the recordings, are present in most fNIRS datasets, as well. A general overview of fNIRS artifacts and artifact removal techniques can be found in Cooper et al. (2012).

To attenuate trends and Mayer Wave like effects, we used a moving average filter, which subtracted the mean of the 120 s before and after every sample from every *HbO* and *HbR* datapoint. Moving average filters have been used successfully before to remove slow trends in experiments with long trials (Heger et al., 2013). Heart-beat and faster frequency signals are attenuate using an elliptical IIR low-pass filter with cutoff frequency of 0.5 Hz and filter order of 6, which robustly reduces heart-beat influences in the data. Finally, we used a wavelet artifact removal method (Molavi and Dumont, 2010) to reduce the effect of movement artifacts.

The trials were then extracted based on the experiment timings and associated with a label according to the *n*-back condition

or RELAX. Each trial of any of the *n*-back conditions is 44 s long, while the relax trials are 10 s long.

2.6. FEATURE EXTRACTION AND SELECTION

Typical hemodynamic responses increase for *HbO* with neural activity in a specific region and return to baseline afterward. In *HbR*, signals typically behave opposite and decrease upon stimulus onset and increase back to baseline after the end of the stimulus. This typical behavior is often used in the feature extraction. The mean value of the signal (Heger et al., 2013) in a specific window or the increase in mean value between different windows (Herff et al., 2012) is often used as a simple, but effective feature. In this study, we use the slope of a straight line fitted to the data in a window as the feature. The line was fitted using linear regression with a least-square approach. Window sizes were varied in the experiments. Even though *HbO* and *HbR* signals of every channel are strongly negatively correlated (Cui et al., 2010a), we extract the slope feature for *HbO* and *HbR* of every channel. Including both *HbO* and *HbR* signals often yields more robust classification results. This results in 16 features per window, as we extract one feature for *HbO* and one for *HbR* for each of the 8 channels.

To reduce the feature set size, we only include features with a high relevance for classification in the feature set. We calculate the Mutual Information between each continuous feature and the discrete labels on the training data using non-parametric probability density functions. These were estimated using kernel methods (Parzen windows). See Ang et al. (2008) for a more detailed description of feature selection methods using Mutual Information. In this study, we limit our feature set to the 8 features containing the highest Mutual Information with the labels, as the remaining half of the features only contained little to no relevance.

2.7. EVALUATION

To classify the data, we used a Linear Discriminant Analysis (LDA) classifier. For the multi-class experiments, we used a one-vs-one multi-class classifying approach (Duda et al., 2012). To evaluate classification accuracy in our experiment, we used a 10-fold cross-validation. For this, the data of one subject is divided into 10 equally sized parts and in a round-robin manner, 9 parts are used for feature selection and training, while the last part is used for evaluation. Presented accuracies are then averaged over all 10 folds. We only evaluate subject dependend systems in this paper. As we use a 10-fold approach and have 10 trials per class, we never use any data shortly before or after the testing data, which could be problematic given the high auto-correlation of fNIRS signals. To evaluate our data set, we first classified the three *n*-back classes from RELAX. The RELAX trials are only 10 s long, while the *n*-back trials last 44 s. We only extracted 10 s long windows from *n*-back classes for this task, as well. Therefore, we evaluated the effect on classification accuracy resulting from different offsets from the start of a trial.

To really quantify mental workload we evaluate classification between the three *n*-back classes. We evaluate classification accuracy depending on window length in which we extract the slope feature.

3. RESULTS

3.1. USER PERFORMANCE AND SUBJECTIVE RATING

To confirm that our subjects perceived the different *n*-back conditions as different, we analyzed the user performance. **Figure 3** shows user performance and subjective evaluation of the experiment.

We evaluated the amount of missed targets, when a subject failed to press the key when a target stimulus was presented. A One-Way ANOVA shows significant differences between the three *n*-back levels in the amount of missed targets ($F = 16.3151$; $p < 0.001$). The percentage of targets missed by the subjects increased from 5.7% on average for the 1-back condition to 16.7% for 2-back to 33.7% for the 3-back task. This clearly shows that the three tasks have significantly different difficulty levels (tested by one-sided t-tests, $p < 0.01$ after Bonferroni correction all three comparisons). Additionally, this clarifies that even in the 3-back tasks our subjects identified two thirds of the targets. Next, we evaluated the amount of wrong reactions, when subjects incorrectly identified a letter as a target and pressed the space key. The amount of wrong reactions is significantly influenced by the *n*-back level (tested by ANOVA, $F = 9.613$; $p < 0.001$). Again, the number of wrong reactions increases from 1.4 on average to 1.9 to 4.5. The differences in wrong reactions between 1 and 3-back and 2 and 3-back are significant (tested by one-sided t-test $p < 0.01$ after Bonferrroni correction), while the difference between 1 and 2-back is not statistically significant. The subjective evaluation of the subjects agreeing with the phrase "The *n*-back task was demanding," clearly shows the different mental workload levels of the three conditions (statistically significant as tested by One-Way ANOVA, $F = 25.8540$; $p < 0.001$). While the average agreement was 1.6 (1 meaning no agreement) for 1-back, subjects answered 3.1 for 2-back and 5.1 on average for 3-back (6 being total agreement). All differences between the three classes are significant (tested by one-sided t-tests $p < 0.01$ after Bonferroni correction). This clearly shows the different levels of workload induced by the three *n*-back conditions.

Subjects stated that they were highly concentrated during the first half of the experiment, answering that they agreed with 4.9 with the phrase "I was concentrated during this half of the experiment." This decreased slightly to 4.0 for the second half. The fNIRS system was judged as being comfortable to wear (3.9 in agreement to a comfortable system) in the first half, which decreased to a medium 2.7 for the second half. Our subjects evaluated the duration of the experiment as appropriate (agreement of 4.7).

3.2. HEMODYNAMIC RESPONSES

To see whether the Hemodynamic responses for the three *n*-back conditions yield any differences, we first analyze the grand averages of all subjects. For this analysis, we baseline every trial by subtracting the mean of the 10 s prior to the trial for *HbO* and *HbR* of every channel. The trials are not baseline normalized for the remaining classification analyses. **Figure 4** shows grand averages for all channels and all *n*-back conditions.

Gray lines show grand averages for individual channels, while the black line shows the mean over all channels. In the *HbO* channels, there is little activity for 1- and 2-back, but a clear increase

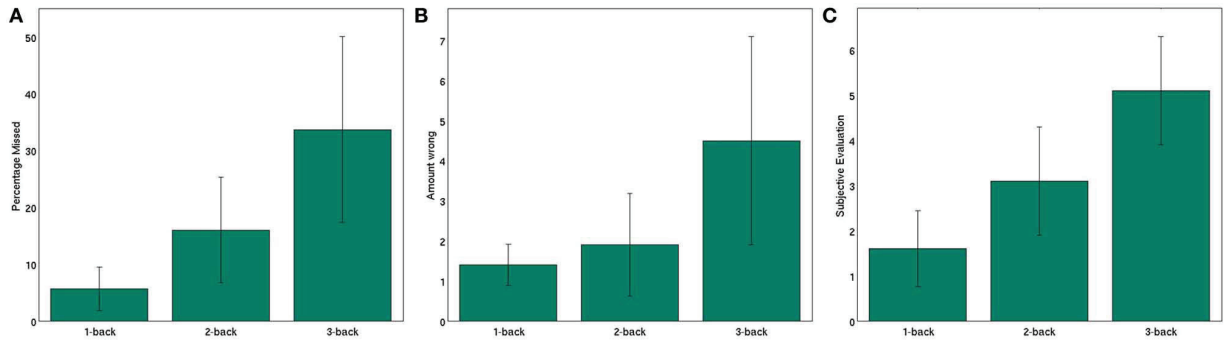


FIGURE 3 | User performance and subjective evaluation in the *n*-back task (A) average number of missed targets (B) average number of wrong reactions (C) average subjective evaluation of task difficulty. Whiskers show standard deviations between subjects. All differences between the conditions are significant (tested by one-sided *t*-tests, $p < 0.05$ after Bonferroni correction), except for the difference between 1 and 2-back in **(B)**.

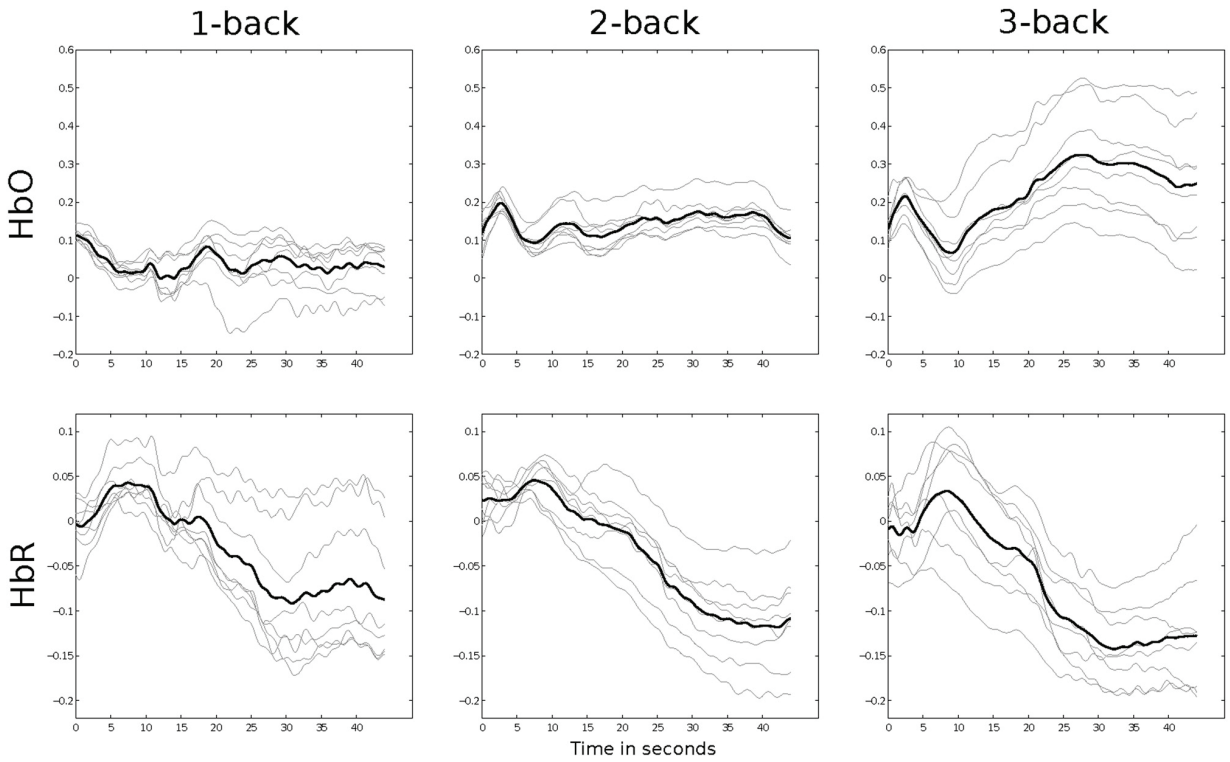


FIGURE 4 | Grand averages of all 10 subjects in the three *n*-back conditions. Gray lines indicate single channels. The black line presents the mean of all channels.

for most channels in the 3-back conditions. It is obvious that a feature derived from the slope of those grand averages could discriminate the 3-back trials from the others. In *HbR* the typical decrease can be seen for all three conditions. While the slope is negative for all three tasks, it is clearly steeper in the 2-back grand average than in the 1-back and steepest for the 3-back averages. These grand averages show that we have different activation patterns for the three conditions and visualize the basis of our classification.

3.3. *n*-BACK vs. RELAX

To evaluate the data set we first classified our *n*-back trials from the RELAX trials collected after the signals returned to baseline. Since our relax trials are only 10 s long, while our *n*-back trials are 44 s in length, we evaluated the effect the offset from the beginning of the trial has on classification accuracies. **Figure 5** shows the classification accuracies depending on the offset from the beginning of the trial when extracting the 10 s long windows.

Extracting the 10 s long window directly after the beginning of the trial yields the worst results for all conditions. This can be explained by the fact that subjects are only beginning to memorize the stimuli and are not experiencing workload yet. After an offset of 10 s the results remain relatively stable. All results are significantly better than chance level (tested by Wilcoxon rank-sum). Even in the four-class classification task we could achieve accuracies up to 45% (chance 25%). As expected, classifying 3-back against RELAX yielded the best results of up to 81% accuracy. For 2-back, we could achieve 80% accuracy for classification against RELAX and 72% for 1-back, respectively. These results show that the single trial data can be robustly discriminated from a relax state.

Table 1 summarizes classification accuracies of each of the conditions against relax and for the four class experiment with an offset of 10 s. These results can be used to compare with previous studies which focus on discriminating from the RELAX state.

3.4. QUANTIFYING MENTAL WORKLOAD

To quantify workload it is necessary to discriminate different levels of workload from each other and not only from a RELAX state. We investigate the three *n*-back conditions against each other in two class and three class scenarios. To evaluate the window length necessary for robust classification of mental workload, we show classification accuracies depending on window length in **Figure 6**.

Part (**A**) of **Figure 6**, shows accuracies for the two class discrimination between two levels of workload, while part (**B**) shows the three class accuracies of all three workload levels. Note that with increasing window size, the amount of instances reduces. While we can extract 80 instances for a window length of 5 s, this amount reduces to 10 for window lengths larger than 25 s. The little amount of training and testing data sets explains the unstable results for window lengths longer than 25 s.

Results increase for increasing window lengths and peak for the length of 25 s. The discrimination between 1- and 3-back works best, which can easily be explained as the degree of difficulty is most different in those two conditions. Classification between 1- and 2-back and 3- and 2-back yield comparable results as the difference in difficulty level across these conditions is similar. For longer window lengths, these results are significantly better than chance level. The three class experiment is above chance for all window lengths and peaks at 50% accuracy for 25 s window length. The detailed results for every subject for window length of 25 s can be found in **Figure 7**. It can be seen that all subjects yield good results for the discrimination between 1-3 back, while only roughly half of the subjects work well for the other two scenarios. The results across subjects are significantly better than chance level for all classification scenarios (tested by Wilcoxon rank-sum tests).

Table 2 summarizes the mean results across all subjects for window lengths of 25 s and 15 s. We present the results for window length of 15 s as well, as this length has been used for workload evaluation with EEG before (Kothe and Makeig, 2011). The results for 25 s long windows clearly show that fNIRS signals can be used to robustly quantify different levels of workload. This is a large step toward passive BCIs using fNIRS for workload monitoring.

4. DISCUSSION

In this study of 10 subjects, we show that fNIRS signals measured from the PFC with an easy to setup montage can be used to robustly quantify users' workload. The analysis of user performance show significant differences in the amount of missed targets and wrong reactions depending of the *n*-back level. Additionally, the subjective evaluation of the users show big differences in perceived difficulty level between the *n*-back levels, as well.

Using 8 channels on the forehead, we were able to classify the different levels of workload induced by *n*-back tasks from a relax state with accuracies up to 81%. As expected, 3-back could be discriminated best from the relax state (81% accuracy), as the mental workload induced by this condition is the largest.

Table 1 | Classification accuracies of the conditions against a relax state.

	1-back	2-back	3-back	1-2-3-relax
Mean	71.5%	80.3%	80.5%	44.5%
Standard deviation	17.7	10.5	13.8	10.0
Chance level	50%	50%	50%	25%

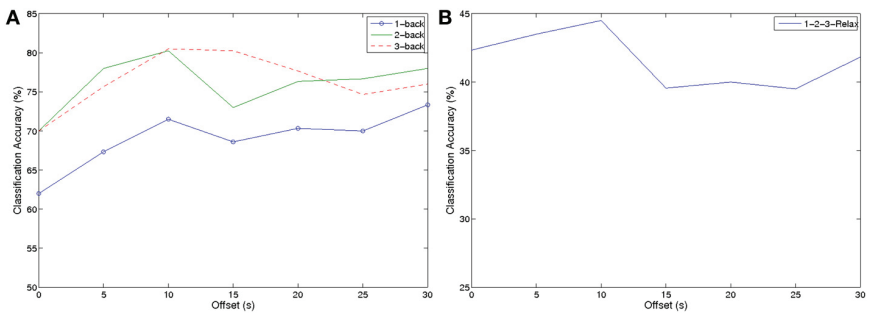


FIGURE 5 | Classification accuracies for *n*-back tasks depending on the offset from trial start (A) two class problems of classification accuracy of 1-,2-,3-back against Relax (B) four class classification between all three *n*-back and RELAX.

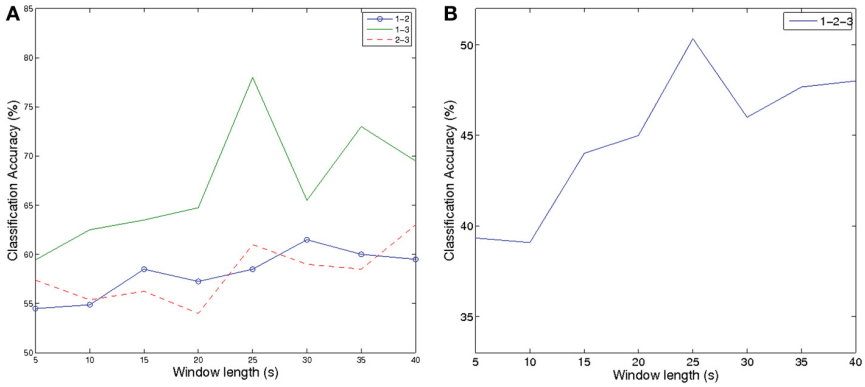


FIGURE 6 | Classification accuracies depending on window length (A) two class problems between different workload levels (B) three class classification of all three workload levels.

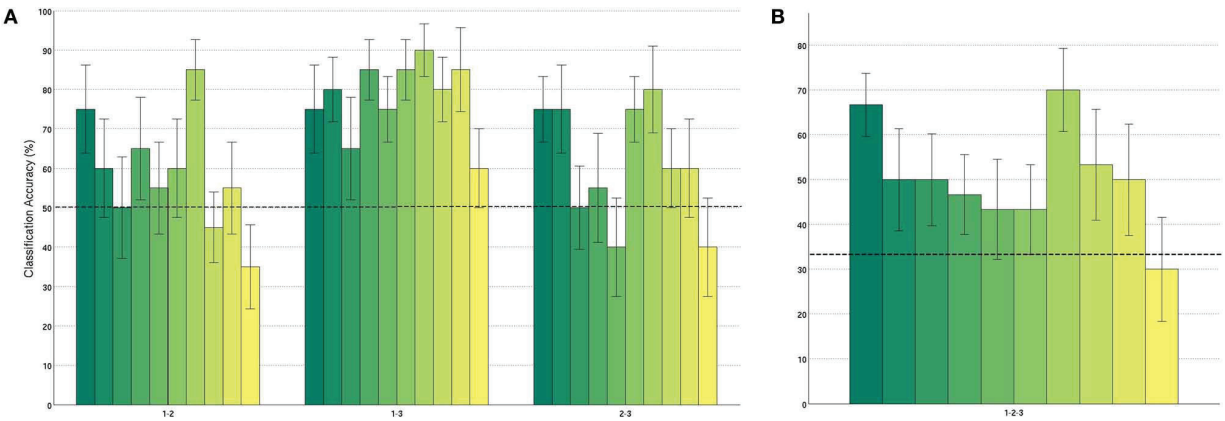


FIGURE 7 | Classification accuracies for each subject with window length of 25 s (A) two class problems (B) three class classification. Each bar represents classification accuracies of one subject. The dotted line denotes naive classification accuracy. Whiskers show standard error in the cross-validation.

Table 2 | Classification accuracies of the conditions against each other.

Window length	1-2	1-3	2-3	1-2-3
15 s	58.5%	63.5%	56.3 %	44.0%
25 s	58.5%	78.0%	61.0%	50.3%
Chance level	50%	50%	50%	33.3%

However, classification of 2-back and 1-back against relax still yielded mean accuracies of 80 and 72%, respectively. These results show that even the workload induced by relatively simple tasks can be robustly discriminated from a resting state.

More importantly, the hemodynamic responses measured in the PFC are consistent enough to be used to discriminate between three levels of workload. While the classification of high vs. low workload (1 vs. 3-back) worked well for all 10 subjects and yielded an average of 78% accuracy, the discrimination between 1 and 2-back only resulted in usable results for half of the subjects (average of 58.5%). Classification between the workload induced by 2 and 3-back tasks resulted in an average of 61% accuracy. These results mirror the subjective and user performance evaluation, as the difference between 1 and 3-back is largest and the difference in workload induced by 1 and 2-back seems to be smallest (no significant difference in the amount of errors between those two conditions).

We thereby show the potential of fNIRS as a modality for passive BCI and user state monitoring, despite the fact that further investigation is necessary to differentiate between more levels of workload with higher accuracies. The simple optode montage and the robust results encourage fNIRS to be used in real-life scenarios like car navigation and class-room settings. In this study, the data was analyzed in an offline manner and especially the moving average filter needs to be adapted for usage in an online system. Instead of only classifying whether a subject was engaged in a task or not, we were able to reliably show the degree of workload a subject was experiencing. The presented results thus show the feasibility of using fNIRS to quantify workload in single trial.

4.1. DATA SHARING

Single-trial analysis of fNIRS data is still a very young field and to the best of our knowledge, there are only very few publicly available data sets of single trial fNIRS experiments. To increase comparability of single trial fNIRS methods and allow for benchmarking, the data corpus used in this study will be publicly available on the authors' website [1]. The fNIRS time courses for all 10 subjects and for all *n*-back conditions and RELAX can be downloaded in both MATLAB™and Comma-Separated-Value (CSV) file formats. The questionnaire and behavior results will be included, as well. Thereby, we hope to provide a common data set for evaluation and testing of fNIRS methods and algorithms.

ACKNOWLEDGMENTS

The fNIRS equipment used in this study is part of the DFG funded Karlsruhe Design and Decision Laboratory, a collaboration laboratory of Economic Sciences, Psychology and Computer Science to investigate decision processes in groups. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

REFERENCES

Ang, K., Guan, C., Lee, K., Lee, J., Nioka, S., and Chance, B. (2010a). "A brain-computer interface for mental arithmetic task from single-trial near-infrared spectroscopy brain signals," in *International Conference on Pattern Recognition* (Istanbul), 3764–3767.

Ang, K. K., Guan, C., Lee, K., Lee, J. Q., Nioka, S., and Chance, B. (2010b). "Application of rough set-based neuro-fuzzy system in nirs-based bci for assessing numerical cognition in classroom," in *The 2010 International Joint Conference on Neural Networks (IJCNN)* (Barcelone), 1–7.

Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence)* (Hong Kong), 2390–2397.

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., and Onaral, B. (2007). "Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy," in *3rd International IEEE/EMBS Conference on Neural Engineering, 2007. CNE '07* (Kohala Coast, HI), 342–345.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Ayaz, H., Willems, B., Bunce, B., Shewokis, P. A., Izzetoglu, K., Hah, S., et al. (2010). "Cognitive workload assessment of air traffic controllers using optical brain imaging sensors," in *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*, eds T. Marek, W. Karwowski, and V. Rice (CRC Press Taylor & Francis Group), 21–31.

Baldwin, C. L., and Penaranda, B. (2012). Adaptive training using an artificial neural network and eeg metrics for within-and cross-task workload classification. *Neuroimage* 59, 48–56. doi: 10.1016/j.neuroimage.2011.07.047

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78, B231–B244. Available online at: http://www.ingentaconnect.com/content/asma/asem/2007/00000078/A00105s1/art00032

Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using eeg spectral power and erps in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Cohen, J., Perlstein, W., Braver, T., Nystrom, L., Noll, D., Jonides, J., et al. (1997). Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604. doi: 10.1038/386604a0

Cooper, R., Selb, J., Gagnon, L., Phillip, D., Schytz, H. W., Iversen, H. K., et al. (2012). A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. *Front. Neurosci.* 6:147. doi: 10.3389/fnins.2012.00147

Coyle, S. M., Ward, T. E., and Markham, C. M. (2007). Brain–computer interface using a simplified functional near-infrared spectroscopy system. *J. Neural Eng.* 4, 219. doi: 10.1088/1741-2560/4/3/007

Cui, X., Bray, S., and Reiss, A. (2010a). Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage* 49, 3039–3046. doi: 10.1016/j.neuroimage.2009.11.050

Cui, X., Bray, S., and Reiss, A. L. (2010b). Speeded near infrared spectroscopy (nirs) response detection. *PLoS ONE* 5:e15474. doi: 10.1371/journal.pone.0015474

Cutrell, E., and Tan, D. (2008). "Bci for passive input in hci," in *Proceedings of CHI*, Vol. 8 (Citeseer), 1–3.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. New York, NY: John Wiley & Sons.

Girouard, A., Solovey, E., Hirshfield, L., Chauncey, K., Sassaroli, A., Fantini, S., et al. (2009). "Distinguishing difficulty levels with non-invasive brain activity measurements," in *Human-Computer Interaction INTERACT 2009, Volume 5726 of Lecture Notes in Computer Science*, eds T. Gross, J. Gulliksen, P. Kotz, L. Oestreicher, P. Palanque, R. Prates, and M. Winckler (Berlin; Heidelberg: Springer), 440–452.

Heger, D., Mutter, R., Herff, C., Putze, F., and Schultz, T. (2013). "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (Geneva: IEEE), 832–837.

Herff, C., Heger, D., Putze, F., Hennrich, J., Fortmann, O., and Schultz, T. (2013). "Classification of mental tasks in the prefrontal cortex using fnirs," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka), 2160–2163.

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012). "Speaking mode recognition from functional near infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (San Diego), 1715–1718.

Hirshfield, L. M., Gulotta, R., Hirshfield, S., Hincks, S., Russell, M., Ward, R., et al. (2011). "This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM) (Vancouver, BC), 373–382.

Hoshi, Y., Tsou, B. H., Billock, V. A., Tanosaki, M., Iguchi, Y., Shimada, M., et al. (2003). Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks. *Neuroimage* 20, 1493–1504. doi: 10.1016/S1053-8119(03)00412-9

Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., and Pourrezaei, K. (2003). "fNIR spectroscopy as a measure of cognitive task load," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, Vol. 4, (Cancun), 3431–3434.

Jarvis, J., Putze, F., Heger, D., and Schultz, T. (2011). "Multimodal person independent recognition of workload related biosignal patterns," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11 (New York, NY: ACM), 205–208.

Kothe, C., and Makeig, S. (2011). "Estimation of task workload from eeg data: New and current tools and perspectives," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (Boston), 6547–6551.

Molavi, B., and Dumont, G. (2010). "Wavelet based motion artifact removal for functional near infrared spectroscopy," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE* (Buenos Aires), 5–8.

Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59. doi: 10.1002/hbm.20131

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

[1] http://csl.anthropomatik.kit.edu/english/2506.php

Power, S. D., Kushki, A., and Chau, T. (2012). Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by nirs. *PLoS ONE* 7:e37791. doi: 10.1371/journal.pone.0037791

Sassaroli, A., and Fantini, S. (2004). Comment on the modified beerlambert law for scattering media. *Phys. Med. Biol.* 49:N255. doi: 10.1088/0031-9155/49/14/N07

Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., et al. (2007). Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a braincomputer interface. *Neuroimage* 34, 1416–1427. doi: 10.1016/j.neuroimage.2006.11.005

Smith, E. E., and Jonides, J. (1997). Working memory: a view from neuroimaging. *Cogn. Psychol.* 33, 5–42. doi: 10.1006/cogp.1997.0658

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Zander, T. O., and Kothe, C. (2011). Towards passive braincomputer interfaces: applying braincomputer interface technology to humanmachine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2013; accepted: 25 December 2013; published online: 16 January 2014.

Citation: Herff C, Heger D, Fortmann O, Hennrich J, Putze F and Schultz T (2014) Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Front. Hum. Neurosci. ***7****:935. doi: 10.3389/fnhum.2013.00935*

This article was submitted to the journal Frontiers in Human Neuroscience.

Hybrid fNIRS-EEG based discrimination of 5 levels of memory load

Christian Herff[1], Ole Fortmann[1], Chun-Yu Tse[2], Xiaoqin Cheng[3],
Felix Putze[1], Dominic Heger[1] and Tanja Schultz[1]

Abstract— **In this study, we show that both electroencephalograhy (EEG) and functional Near-Infrared Spectroscopy (fNIRS) can be used to discriminate between 5 levels of memory load. We induce memory load with the memory updating task, which is known to robustly generate memory load and allows us to define 5 different levels of load. Typical experiments only discriminate between low and high workload or up to a maximum of three classes. To the best of our knowledge, the memory updating task has not been used in combination with brain activity measurements before.**
Here, accuracies of up to 93% are achieved for the binary classification between very high and very low workload. On average, two levels of workload could be discriminated with 74% accuracy. Classification between the full five classes yielded 44% accuracy on average. Despite the fact that EEG results consistently outperformed the results obtained with fNIRS, we could show that the feature-level fusion of both modalities increased robustness of classification results. A reliable discrimination between different levels of memory load could be used to adapt user interfaces or present the right amount of information to a learner.

I. INTRODUCTION

The adaptation of user interfaces using physiological measurements has recently gained increased attention [1]. If inherently internal information like emotion, or, as in this study, the memory load, could be robustly determined, user interfaces could react accordingly and thus greatly increase efficiency and naturalness of human-computer interaction.
Brain activity measurements have been shown to allow the identification of these mental states more robustly than other physiological measurements like electrodermal activity, heart-rate related parameters and eye-gaze measurements [2], [3]. The most common modality for such, so called, passive Brain-Computer Interfaces (BCIs) [4] is the electroencephalogram (EEG), which has been used to robustly discriminate between high and low workload [5]. An alternative method for the measuring of brain activity is functional Near-Infrared Spectroscopy (fNIRS), which indirectly measures concentration changes of oxygenated and deoxygenated hemoglobin through the absorption of near-infrared light. While EEG has high temporal resolution, signals at each measurement location are a summation of various brain and non-brain sources, yielding a low spatial resolution. Hemodynamic changes measured by fNIRS take several seconds to be measurable and thus offer a poor temporal resolution, but can be localized to a very small area, giving high spatial resolution. These two modalities thus complement each other in spatial and temporal resolution. Besides classic BCI paradigms, fNIRS has been successfully used, among others, to classify emotion [6], the type of task a user is engaged in [7] and has been applied to more realistic settings such as workload monitoring of air traffic controllers [8]. See [9] for a review of fNIRS for brain imaging in general and [10] for passive BCIs specifically.
The combination of several brain measurement modalities is called hybrid BCI [11]. Hybrid BCIs have been used to identify the type of attention (visual or auditory) participants are engaged in [12] and have been shown to enhance performance in a motor imagery task over single modality interfaces [13].
Different levels of memory load have been evaluated with fNIRS [14], EEG [15] and the combination of both [16] employing the n-back paradigm, with a maximum of 3 levels of memory load. In this study, a larger number of levels was targeted. For this purpose, we chose a variation of the memory updating task. The memory updating task was first presented by Salthouse et al. [17] and adapted by Oberauer et al. [18]. In this task, memory load is induced with a number of digits that have to be remembered and constantly updated with arithmetic operations. By varying the number of digits, the memory demand can be adapted and thus, different levels of memory load induced. This procedure allows for a larger number of distinct load levels [19] than the regularly used n-back paradigm. To the best of our knowledge, the memory updating task has not been used in combination with brain activity measurements. Additionally to the novelty of the task, the number of memory load levels is usually far lower than the five classes investigated here.

II. MATERIAL AND METHODS

A. Experiment

A variation of the memory updating task [17], [18], [19] was used to induce 5 different levels of memory load. During a trial, a participant is shown a row of boxes on the screen. The number of boxes depends on the difficulty level and ranges between one and five in our study. At the beginning of each trial, the boxes are shown empty for one second. The initial digit is then shown for 2.5 seconds in each box. Afterward simple additions and subtractions (between -7 and $+7$) are displayed for 2.5 seconds in a randomly chosen box. Participants have to apply the displayed operation to the currently remembered digit and hence update their

[1]Christian Herff, Ole Fortmann, Felix Putze, Dominic Heger and Tanja Sschultz are with the Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany christian.herff@kit.edu
[2]Chun-Yu Tse is with the Department of Psychology & Center for Cognition and Brain Studies, The Chinese University of Hong Kong, China
[3]Xiaoqin Cheng is with the Department of Psychology and LSI Neurobiology/Aging Programme, National University of Singapore, Singapore

remembered number. The number of operations is fixed for each difficulty level to keep the trials at a fixed length of 31s. After the trial, the recall phase starts, in which participants have 3 seconds per box to recall and type in the final result. This recall phase is ignored in EEG and fNIRS analysis. The recall phase is followed by either 15 or 25 seconds of pause, during which a fixation cross was displayed and the participants were asked to relax. We focus on the differences between memory updating levels and ignore the pause trials for this analysis.

Participants were asked to avoid unnecessary motion during trials. We recorded 10 trials per difficulty level and participant, resulting in a total of 50 trials per participant. Figure 1 illustrates the experiment procedure.

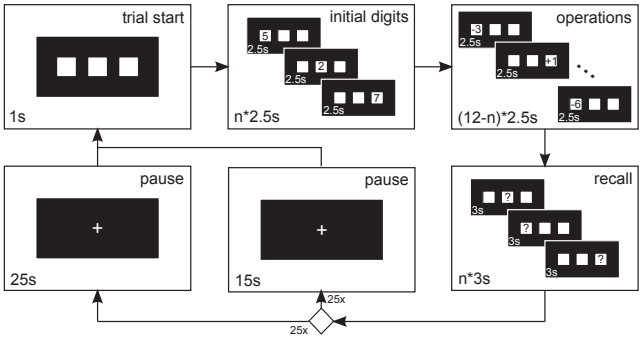


Fig. 1. Experimental design of the memory updating task. The difficulty level n influences the number of boxes per trial. Figure illustrates a 3-box example.

Trials are assigned a label corresponding to the number of boxes shown. In this study, we use the entire 31 seconds of trial data, but exclude pause and recall phases.

B. Data acquisition

During the experiment, we recorded fNIRS signals from 28 sources, emitting near-infrared light and 15 detector optodes located on the forehead to measure hemodynamic activity in the prefrontal cortex. We used a frequency-modulated oximeter (Imagent, ISS Inc.) measuring at two wavelengths (690nm and 830nm). Modulation frequency was set to 110 MHz and the sampling frequency was 19.5 Hz.

In parallel, EEG activity was recorded using three electrodes on the midline at positions Fz, Cz and Pz according to the international 10-20 system. These midline locations were chosen as this region has shown strong activations in previous studies investigating memory load [16]. Additionally, 4 electrodes were placed around the eyes to record electrooculography (EOG). Both modalities are recorded using an ANT amplifier (ANT, Netherlands) and sampled at 256 Hz. All electrodes were referenced to the nose. Ground electrode was placed on the forehead. Source and detector optodes, as well as electrodes are fixed to participants' heads with a rigid custom-made holder. Optode and electrode positions were digitized for each participant individually using an ANT Visor system. The resulting coordinates were used to calculate exact distances for each source and detector pair for the conversion process of raw optical densities to hemoglobin concentration changes.

Figure 2 illustrates optode and electrode positions.

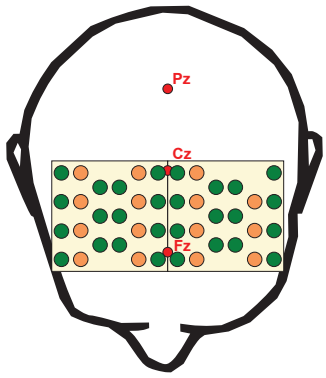


Fig. 2. Optode and electrode montage. Green circles represent sources, orange circles detectors and red circles indicate electrode position.

C. Participants

We recorded 10 healthy participants (3 female) with a mean age of 25.5 years (SD = 0.97). All participants gave written consent and had no history of neurological diseases. They had participated in EEG and fNIRS experiments previously but had no experience with the memory updating task.

D. EEG processing

To reduce eye movement artifacts in the EEG recordings, we applied the EOG regression methods proposed in [20]. We extracted the powerspectrum in 1 Hz wide bins between 4 and 25 Hz using Welch's method. These features were extracted for each of the three electrodes resulting in a 66-dimensional feature space for EEG data. We calculated the mean μ_{train} and standard deviation σ_{train} of each feature in the training set and normalized both training and test sets by subtracting μ_{train} and dividing by σ_{train} (z-normalization).

E. fNIRS processing

We restricted our analysis to the DC component of the measured signal and converted the optical densities to changes in oxygenated and deoxygenated hemoglobin (HbO and HbR) using the HomER package [21]. As HbO and HbR are strongly correlated [22] with responses more pronounced in HbO, we limit our analysis to the HbO signals. As each of the 15 detectors measures the light intensities from each of the 28 sources, a total of 420 channels with different source-detector distances was recorded. To restrict this amount to information bearing channels, and exclude channels with too large source-detector distances, we only consider channels showing a clear pulse artifact, which is expected in clean fNIRS signals. Channels with pulse were identified automatically, if a peak was detected in the log-powerspectrum p frequency band 0.8 to 1.7 Hz:

$$\max\{p(f)|f \in [0.8, 1.7]\} - \text{mean}\{p(f)|f \in [0.8, 1.7]\} > 0.5$$

The log-powerspectrum was extracted with Welch's method and yielded 0.2 Hz wide frequency bins. This procedure reduced the amount of channels from 420 to 13-47 depending on the participant. The data was then lowpass-filtered to attenuate motion artifacts and heart-beat with an elliptic IIR filter with filter-order 6 and cut-off frequency of 0.5 Hz. As a feature, the slope of a line fitted to the data of a trial was extract for each channel using a least-squares approach. The slope has been shown to capture relevant information in single trial fNIRS data in previous studies [14]. This resulted in 13 to 47 features for the fNIRS data. The fNIRS slope features were z-normalized, as described in Section II-D.

F. Evaluation

To evaluate how well 5 classes of memory load can be discriminated, we first investigated the classification of load levels in binary conditions. Each pairing of memory load levels was evaluated for fNIRS and EEG data, resulting in a total of 10 tasks. To assess the possibility of multi-class discrimination, we evaluated the 5-class condition, as well.

Evaluation was done participant-dependently using 10-fold cross-validation. We trained a regularized LDA on the training data, with an optimal shrinkage parameter determined with the analytic methods proposed in [23]. Regularized classifiers have been shown to yield good results in high-dimensional features spaces with small amounts of training samples, as is the case in BCIs with EEG [24] and fNIRS [25], [26]. Multi-class classification was performed using a one-vs-one approach, resulting in 10 classifiers and deciding via majority vote. In addition to the analysis of fNIRS and EEG independently, we evaluated the feature-level fusion of both modalities by combining the 66-dimensional EEG feature space with the 13 to 47 dimensional feature space from fNIRS, resulting in 79 to 113 dimensions for the fused feature space. We denote the fused feature space as FUSION in later section.

III. RESULTS

In the binary classification conditions, high accuracies are achieved with both EEG and fNIRS. As expected, highest accuracies were achieved when discriminating between one-box and 5-box conditions yielding 71% correct on average for fNIRS, 90% for EEG and 93% for FUSION. Accuracies for EEG and FUSION never drop below 58%, while fNIRS only achieves chance accuracies for the discrimination between 1 and 2 box conditions. Classification of one memory load level against another works consistently better for EEG than for fNIRS with a mean over all conditions of 60% accuracy for fNIRS and 74% for EEG. Feature-level FUSION yielded a mean of 74% as well, albeit more results significantly better than chance could be achieved using the feature-level FUSION. Testing for each participant and condition (total of 100 = 10 participants x 10 conditions) individually whether the 10 folds yielded results significantly larger than chance, we found that 26% are significantly better than chance (one-sided t-test, $p < 0.05$) for fNIRS, 61% for EEG and 69% for FUSION.

Results are strongly correlated with the distance between load levels (fNIRS $r = 0.38$, $p < 0.001$; EEG $r = 0.64$, $p < 0.001$; Fusion $r = 0.62$, $p < 0.001$)) meaning that the further the levels are apart, the better the classification. All classification results for the binary conditions can be found in Figure 3 (a).

In addition to the binary conditions, we evaluated how well five levels of memory load could be discriminated 3 (b). Discrimination works significantly better than chance for fNIRS (29%), EEG (42%) and the FUSION approach (44%). Figure 4 illustrates the confusion matrix for the FUSION approach. It is clearly visible that confusions occurred most often with adjacent load levels and that confusions with farther away levels are very rare.

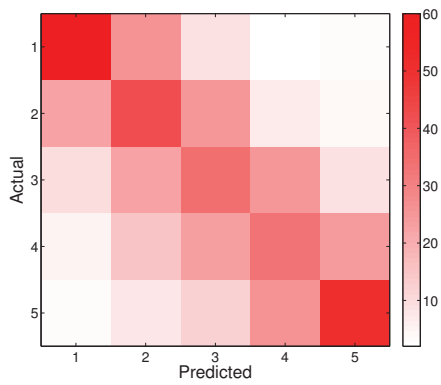


Fig. 4. Confusion matrix in the five class experiment using the fused feature space.

IV. CONCLUSIONS

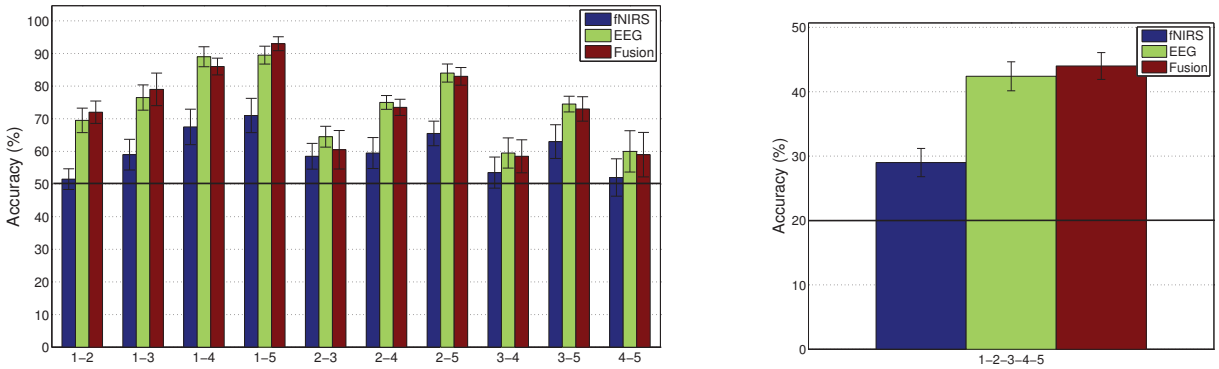
In this paper, we have shown that 5 levels of induced memory load can be robustly discriminated using fNIRS and EEG. This shows the potential for both modalities to be used for more than just the identification of high and low levels of workload or memory load.

A combination of both modalities increased classification results significantly when discriminating between 5 levels of memory load and increased the robustness in binary classification conditions. Even though the combination of both modalities increases robustness of classification, it has to be decided whether this increase justifies the usage of an additional sensor.

Our study was conducted in a lab environment and does thus not face challenges that would be posed by a real-life scenario. Further evaluations are needed to determine whether the window length of 31 seconds can be reduced while maintaining high accuracies to meet the requirements of real-life application.

ACKNOWLEDGMENT

The authors would like to thank Professor Trevor Penney of the National University of Singapore for productive co-operation, providing access to the fNIRS system and useful discussion.



(a) (b)

Fig. 3. Mean classification results for the 10 binary conditions (a) and the five-class condition (b). Whiskers indicate standard errors. Solid lines denotes naive classification rate.

References

[1] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1, pp. 133–145, 2009.

[2] J. Frey, C. Mühl, F. Lotte, M. Hachet, *et al.*, "Review of the use of electroencephalography as an evaluation method for human-computer interaction," in *PhyCS 2014-International Conference on Physiological Computing Systems*, 2014.

[3] M. A. Hogervorst, A.-M. Brouwer, and J. B. Van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Frontiers in neuroscience*, vol. 8, 2014.

[4] T. O. Zander and C. Kothe, "Towards passive braincomputer interfaces: applying braincomputer interface technology to humanmachine systems in general," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025005, 2011.

[5] C. Kothe and S. Makeig, "Estimation of task workload from eeg data: New and current tools and perspectives," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 6547–6551.

[6] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Affective Computing and Intelligent Interaction*. Springer, 2013, pp. 436–446.

[7] C. Herff, D. Heger, F. Putze, J. Hennrich, O. Fortmann, and T. Schultz, "Classification of mental tasks in the prefrontal cortex using fnirs," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 2160–2163.

[8] H. Ayaz, B. Willems, B. Bunce, P. A. Shewokis, K. Izzetoglu, S. Hah, A. Deshmukh, and B. Onaral, "Cognitive workload assessment of air traffic controllers using optical brain imaging sensors," *Advances in understanding human performance: neuroergonomics, human factors design, and special populations*, pp. 21–31, 2010.

[9] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application," *Neuroimage*, vol. 63, no. 2, pp. 921–935, 2012.

[10] M. Strait and M. Scheutz, "What we can and cannot (yet) do with functional near infrared spectroscopy," *Frontiers in Neuroscience*, vol. 8, no. 117, 2014.

[11] G. Pfurtscheller, B. Z. Allison, C. Brunner, G. Bauernfeind, T. Solis-Escalante, R. Scherer, T. O. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer, "The hybrid bci," *Frontiers in neuroscience*, vol. 4, 2010.

[12] F. Putze, S. Hesslinger, C.-Y. Tse, Y. Huang, C. Herff, C. Guan, and T. Schultz, "Hybrid fnirs-eeg based classification of auditory and visual perception processes," *Frontiers in Neuroscience*, vol. 8, no. 373, 2014.

[13] S. Fazli, J. Mehnert, J. Steinbrink, G. Curio, A. Villringer, K.-R. Müller, and B. Blankertz, "Enhanced performance by a hybrid nirs–eeg brain computer interface," *Neuroimage*, vol. 59, no. 1, pp. 519–529, 2012.

[14] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task - quantified in the prefrontal cortex using fnirs," *Frontiers in Human Neuroscience*, vol. 7, no. 935, 2014.

[15] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using eeg spectral power and erps in the n-back task," *Journal of Neural Engineering*, vol. 9, no. 4, p. 045008, 2012.

[16] E. B. Coffey, A.-M. Brouwer, and J. B. van Erp, "Measuring workload using a combination of electroencephalography and near infrared spectroscopy," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1. SAGE Publications, 2012, pp. 1822–1826.

[17] T. A. Salthouse, R. L. Babcock, and R. J. Shaw, "Effects of adult age on structural and operational capacities in working memory." *Psychology and aging*, vol. 6, no. 1, p. 118, 1991.

[18] K. Oberauer, H.-M. Süß, R. Schulze, O. Wilhelm, and W. Wittmann, "Working memory capacityfacets of a cognitive ability construct," *Personality and Individual Differences*, vol. 29, no. 6, pp. 1017–1045, 2000.

[19] S. Lewandowsky, K. Oberauer, L.-X. Yang, and U. K. Ecker, "A working memory test battery for matlab," *Behavior Research Methods*, vol. 42, no. 2, pp. 571–585, 2010.

[20] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of eog artifacts in eeg recordings," *Clinical neurophysiology*, vol. 118, no. 1, pp. 98–104, 2007.

[21] T. J. Huppert, S. G. Diamond, M. A. Franceschini, and D. A. Boas, "Homer: a review of time-series analysis methods for near-infrared spectroscopy of the brain," *Applied optics*, vol. 48, no. 10, pp. D280–D298, 2009.

[22] X. Cui, S. Bray, and A. Reiss, "Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics," *NeuroImage*, vol. 49, no. 4, pp. 3039–46, Feb. 2010.

[23] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[24] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of erp componentsa tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[25] G. Bauernfeind, D. Steyrl, C. Brunner, and G. R. Muller-Putz, "Single trial classification of fnirs-based brain-computer interface mental arithmetic data: A comparison between different classifiers," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 2004–2007.

[26] D. Heger, C. Herff, and T. Schultz, "Combining feature extraction and classification for fnirs bcis by regularized least squares optimization," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 2012–2015.

Investigating Deep Learning for fNIRS based BCI

Johannes Hennrich, Christian Herff, Dominic Heger and Tanja Schultz

Abstract— **Functional Near infrared Spectroscopy (fNIRS) is a relatively young modality for measuring brain activity which has recently shown promising results for building Brain Computer Interfaces (BCI). Due to its infancy, there are still no standard approaches for meaningful features and classifiers for single trial analysis of fNIRS. Most studies are limited to established classifiers from EEG-based BCIs and very simple features. The feasibility of more complex and powerful classification approaches like Deep Neural Networks has, to the best of our knowledge, not been investigated for fNIRS based BCI. These networks have recently become increasingly popular, as they outperformed conventional machine learning methods for a variety of tasks, due in part to advances in training methods for neural networks. In this paper, we show how Deep Neural Networks can be used to classify brain activation patterns measured by fNIRS and compare them with previously used methods.**

I. Introduction

In the last few years, functional Near Infrared Spectroscopy (fNIRS) has become an emerging technology for optical brain activity measurement that can be used in non-invasiv Brain-Computer Interfaces (BCIs). However, there are still no common standards for feature extraction and classification in single trial fNIRS analysis. Often, rather simple methods are used for feature extraction, such as calculating the mean of the measured hemoglobin concentrations within a given time window [1]. Just as for feature extraction, classification methods used are comparably simple. Recently, Bauernfeind et al. [2] compared different classifiers for fNIRS BCIs and recommended using shrinkage LDA. However, their evaluation did only include classifiers that are well established in Brain Computer Interface research, but did not investigate Deep Neural Networks. Therefore, it is still unclear, whether more complex classification schemes, such as Deep Neural Networks (DNN), can be used to exploit additional information that may be hidden in the non linear dynamics of the hemodynamic responses that typically occur in fNIRS data.

In this paper, we investigate the suitability of deep learning, i.e. artificial neural networks with many layers, for fNIRS classification. Deep learning methods have lately regained popularity, as they have shown impressive results for many different classification problems. However, to the best of our knowledge, no studies have used deep neural networks for the classification of fNIRS data, before. We discuss on how to design deep neural networks for fNIRS and evaluate their classification performance in comparison to traditional approaches.

All authors are with the Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany johannes.hennrich@student.kit.edu

A. Functional Near-Infrared Spectroscopy

Functional Near-Infrared Spectroscopy (fNIRS) is a relatively new non-invasive method to capture hemodynamic responses to cortical activity. These responses can be measured with optical sensors which are cheap and portable and allow for high spatial resolution. The underlying metabolic effects fNIRS is based on are the same effects fMRI uses. fNIRS makes use of the fact that oxygenated hemoglobin absorbs near infrared light differently than deoxygenated hemoglobin. This is achieved by using two different wavelengths in the near infrared part of the electromagnetic spectrum to measure concentration levels of oxygenated and deoxygenated hemoglobin and derive the hemodynamic responses to cortical activity. The by far most prominent modality for non-invasive BCIs is electroencephalography (EEG), which derives cortical activity from electric signals measured on the scalp. In comparison to fNIRS, EEG has a higher temporal resolution because hemodynamic effects are inert and appear with some delay. On the other hand, fNIRS provides higher spatial resolution and does not require electrolyte gel, which makes fNIRS more comfortable to wear and faster to put on, especially on bald regions like the forehead.

Many fNIRS studies aim to discover patterns in the hemodynamic responses by averaging over many trials. For research on online BCIs a single-trial analysis of fNIRS activations is necessary. In [3] and [4] activations of a basic motor imagery task were classified online and on single-trial using brain activity produced by motor imagery on the motor cortex. In contrast to these BCIs for direct control, a passive investigation of memory load in fNIRS applying the n-back paradigm was conducted in [5].

B. Deep Learning

To classify the activity captured by fNIRS, a variety of machine learning approaches have been evaluated. In [4] both Support Vector Machines (SVM) and Hidden Markov Models (HMM) yielded significant classification accuracies for motor imagery. In [6] Support Vector Machines were used for a four-class classification task using the mean, median, range and slope of a trial as features. The slope was also used in [7], in combination with a Linear Discriminant Analysis (LDA). [8] and [5] also used LDA classifiers in combination with very simple features.

While all the mentioned classifiers have successfully been used for a long time, Deep Neural Networks have only recently gained popularity as highly efficient training procedures were developed. Training these deep architectures is complicated, as the process gets slower the more layers are

added. To make training feasible, [9] presented a greedy, unsupervised algorithm which uses Restricted Boltzmann Machines (RBM) to pre-train the network and find good initial weights that speed up the actual training. The pre-training is done layer by layer which allows it to be fast even for very deep networks ([10], [11]). Deep neural networks pre-trained with RBMs were used for the classification of phones in automated speech recognition in [12] and [13] and outperformed all previous methods. With the same pre-training, a generative neural network yielded impressive results by learning and reconstructing pictures of handwritten images ([14]). In 2012 [15] showed how these handwritten images can be classified using a deep neural network and achieved a relative improvement of accuracy of 41% compared to conventional methods. Here, we investigate the performance of deep neural networks as classifiers for fNIRS based BCI.

II. Materials and Methods

A. Data Corpus

The data corpus we used for evaluation is from a BCI experiment conducted in [8]. The experiment consisted of three different mental tasks, namely mental arithmetics (MA), word generation (WG) and mental rotation (MR). In mental arithmetics, the subjects were asked to continuously subtract a given number from another given number, in word generation, they had to find words starting with the specified letter and in mental rotation, 3D objects were displayed which the participants had to imagine to rotate around an axis. Tasks were displayed for 10 seconds and followed by a rest interval of 15 seconds. After 30 randomly chosen rest intervals, 10 second long relax trials (R) were included.

To measure the brain activity in a subject's prefrontal cortex, we used a 8-channel fNIRS headset (Oxymon Mark III, Artinis) attached to the forehead.

In fNIRS, each channel produces one value for oxygenated and deoxygenated hemoglobin concentrations, the recorded signal is thus 16-dimensional. At a sampling rate of 10 Hz this resulted in 100 samples per trial for each of the 16 dimensions. A total of 30 trials for each of the three mental tasks and the relax task were recorded. Trials are extracted based on experiment timing and are labeled corresponding to the type of mental task.

B. Preprocessing

To remove common biological and technical artifacts from the measured signal and optimize it for neural networks, there are some important preprocessing steps. To remove slow trends, we subtracted the mean of the surrounding 240s window from each sample in every channel, this has been successfully applied to fNIRS signals in [5]. To attenuate spikes and high-frequency artefacts, in particular the subject's heartbeat, we lowpass filtered with a cutoff frequency of 0.5 Hz [5] using an elliptic IIR filter with filter order 6. After downsampling to 1 Hz to reduce the dimensionality of the data, the 16 channels were stacked so each trial formed a 160 dimensional vector. Since previous studies ([16], [12]) show that z-normalization reduce neural network training duration and increase robustness, we normalized the data to zero mean and unit variance.

C. Deep Neural Network

We used deep artificial neural networks to classify the data.

As deep neural networks have shown great potential at learning relevant features from raw data ([17], [11], [14]) we refrained from using a tailor-made feature extraction and trained the network directly on the preprocessed data.

Instead of initializing the network with random weights, we pre-trained the network in a layer wise, unsupervised manner using restricted boltzmann machines which can speed up training and lead to better generalization [14]. The pre-trained network was then fine tuned by minimizing the cross-entropy error with the method of conjugate gradients (CG). Using CG to train neural networks has been proposed in many publications and has several advantages over standard backpropagation, including a faster convergence and automated estimation of the learning rate ([18], [19]). For our particular problem, CG training was faster, more stable and resulted in higher classification accuracies than backpropagation.

The activation functions we used are linear functions for the input layer and softmax for the output. For the hidden layers, we decided to use logistic activation functions as they are the de facto standard for nonlinear neural networks and yield solid results for most purposes. Choosing the amount of hidden layers and units is a more difficult task. As the optimal network topology is highly dependent on problem type and the training data distribution, there is no universal procedure to derive these numbers. In our case, we only have 27 training samples per class if evaluated in a 10-fold cross-validation. Classic machine learning theory suggests using very small models with few parameters if there is such little training data available. However [20] showed that deep neural networks are able to learn models with much more parameters than available training samples. A common problem that arises when large models are trained with few training samples is overfitting. Overfitting is present if bad generalization reduces the test accuracy, however there are different solutions that address this problem ([21]). To settle on a network topology we ultimately ran a grid search which estimated the optimal size and amount of hidden layers. This way, we found that two hidden layers with 300 and 40 units, respectively, yielded the best results. The amount of units in the input- and classification layer is determined by the dimensionality of the input data and the amount of classes and therefore does not require tuning. To find out how the depth of the network affects its performance we employed two additional networks with one and three hidden layers.

D. Evaluation

For the evaluation of its classification performance, we trained and tested the networks in a 10-fold cross-validation. This procedure was repeated for all 10 subjects and classification tasks. The classes we tried to discriminate were

the three mental tasks (MA, WG, MR) against relax (R) and the mental tasks against each other, which adds up to a total amount of 6 binary classification problems per subject. We compared our classification results with those from [8] which were achieved using a standard LDA classifier. It is important to note that the LDA classifier was trained on custom-built features which require expert knowledge about fNIRS signal and the problem domain. The deep neural network on the other hand runs directly on the preprocessed data, thus providing a more generic solution.

For a more appropriate comparison we cross-validated a shrinkage-LDA on the same data as the deep neural network as a third method. It has been shown in [22] that regularized LDA classifiers perform well on classification tasks where only little training data is available for high dimensional feature spaces. An optimal shrinkage parameter was estimated using the analytic method proposed in [23]. This method uses the same feature space as the neural networks and does not require expert knowledge either.

III. Results

A. Classification Results

All classification accuracies achieved by the three deep neural networks, the LDA with feature extraction and the shrinkage-LDA are presented in Figure 1. Each bar represents the classification of one mental task against relax or another mental task using one of the five methods. All

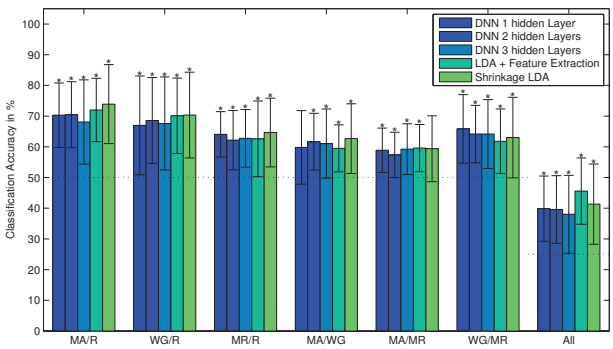


Fig. 1. Classification accuracies of different tasks averaged over the 10 subjects. Whiskers denote standard deviations. Results marked with $*$ are significantly better ($p < 0.01$) than naive classification (dotted lines).

classification results were found to be significantly better than chance level ($p < 0.05$).

The overall classification accuracies for the different methods averaged over subjects and tasks are shown in Table I. Paired t-tests showed no significant differences between classification accuracies of the methods for all classification conditions ($p > 0.05$), except for when comparing the shrinkage-LDA with the 1- and 3-hidden layer DNN ($p = 0.02$ and $p = 0.005$).

To investigate the variances of the classification accuracies, we searched for systematic differences between the subjects. We found a high correlation for each subjects classification rates across the three different approaches

TABLE I

Overall accuracies for different classification methods.

DNN (1 hid.)	DNN (2 hid.)	DNN (3 hid.)	LDA Feat.Extr.	Shrinkage LDA
63.3%	64.1%	62.1%	64.3%	65.7%

(mean ρ: 0.61 for Neural Network vs. LDA + Feature Extraction, 0.89 for Neural Network vs. Shrinkage LDA, 0.70 for LDA + Feature Extraction vs. Shrinkage LDA). This suggests that there are significant variations in the data quality of the different subjects while all three classifiers run at a similar performance. Table II summarizes correlations between methods for a given task.

TABLE II

Correlation coefficients of classification accuracies over the subjects for each combination of classification methods and tasks.

	Neural Network LDA + Feat.Extr.	Neural Network Shrinkage LDA	LDA + Feat.Extr. Shrinkage LDA
MA/R	0.878	0.905	0.790
WG/R	0.869	0.966	0.861
MR/R	0.624	0.816	0.828
MA/WG	0.775	0.929	0.814
MA/MR	0.199	0.912	0.379
WG/MR	0.322	0.852	0.527

We also used the neural networks to classify all three mental tasks and relax against each other. In this 4-class task we achieved a classification accuracy of 40% using the 2-hidden layer DNN, which is significantly better ($p = 0.0012$) than naive classification (25%). With 45% for LDA with feature extraction and 41% for shrinkage LDA, the other methods yield higher, albeit not significantly better results than the neural network ($p = 0.24$ and $p = 0.75$). With the 1- and 3-hidden layer DNN we achieved accuracies of 39% and 38%.

Overall LDA-based methods outperformed all configurations of Deep Neural Networks.

B. Test for Overfitting

A simple but powerful tool to visualize the fine tuning training procedure and discover issues like overfitting is to test the classification accuracy of the network after each epoch of training. Testing is done twice after each epoch, once on the training set and on the test set, resulting in two time lines which can be plotted as a function of the amount of training epochs. Figure 2 shows classification accuracies depending on training epochs for subject 6 for mental arithmetics against relax. Comparable behavior was observed for all subjects. One can see that after several epochs the classification accuracy on the training set saturates on 100%, which denotes that all training samples are getting classified correctly. Similarly the classification accuracy on the test set rises and after several epochs saturates on a value of about 84%, which equals the final test accuracy of the network.

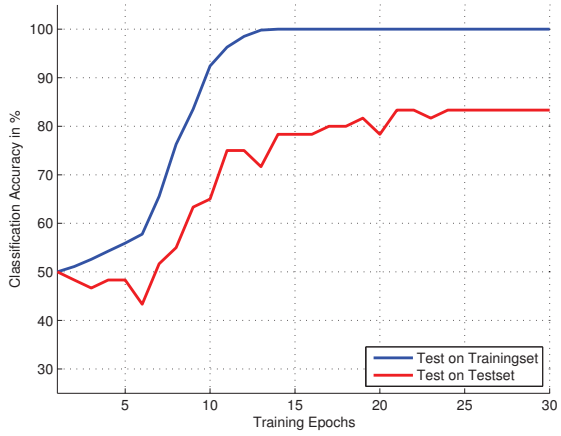


Fig. 2. Classification accuracies for subject 6 for mental arithmetics against relax when evaluated on training- and test set as a function of the amount of training epochs. Comparable behavior was observed for all subjects.

If after reaching a local maximum the test error would start to drop again, the plot would be a clear indicator for overfitting. In that case one should either reduce the complexity of the model by lowering the amount of hidden units or use a stopping criterion which monitors the training process and terminates it after the optimal amount of epochs as explained in [21]. Other possible scenario is that the training accuracy does not saturate, which can be caused by poor initial weights or too few training epochs. Further, it is possible that the training accuracy does saturate, but at a value lower than 100%. This is not necessarily a problem but can indicate that the model is too small and more hidden neurons might improve both training- and test accuracy.

The fact that we do not see any signs of overfitting, but a perfect classification of the training data suggests that far too few training samples are present.

IV. Conclusions

In this study, we showed how deep learning methods can be successfully used for building BCIs based on fNIRS. We achieved classification accuracies for the discrimination of different mental tasks that are comparable to those produced by conventional methods.

Even though the neural network did not yield higher classification rates, it is an promising approach as it does not require a tailor-made feature extraction and expert knowledge about the problem domain. Comparing the neural network with an optimally regularized LDA, also operating on the raw data, we found the shrinkage LDA to yield superior, albeit not significantly better results. By using networks with different amounts of hidden layers we showed that deeper networks do not perform better on this particular task. A possible reason for this is, presumably, the limited training data. Deep neural networks are known to require large training sets for successful learning. We thus recommend to use regularized classifiers if little training data is available.

References

[1] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous Recognition of Affective States by Functional Near Infrared Spectroscopy Signals," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 832–837, Sept. 2013.

[2] G. Bauernfeind, D. Steyrl, C. Brunner, and G. R. Muller-Putz, "Single trial classification of fnirs-based brain-computer interface mental arithmetic data: A comparison between different classifiers," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014.

[3] S. M. Coyle, T. E. Ward, and C. M. Markham, "Brain–computer interface using a simplified functional near-infrared spectroscopy system," *Journal of neural engineering*, vol. 4, no. 3, p. 219, 2007.

[4] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer, "Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface," *NeuroImage*, vol. 34, no. 4, 2007.

[5] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS." *Frontiers in human neuroscience*, vol. 7, no. January, p. 935, Jan. 2014.

[6] A. M. Batula, H. Ayaz, and Y. E. Kim, "Evaluating a four-class motor-imagery-based optical brain-computer interface," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 2000–2003.

[7] S. D. Power, A. Kushki, and T. Chau, "Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by nirs," *PloS one*, vol. 7, no. 7, p. e37791, 2012.

[8] C. Herff, D. Heger, F. Putze, J. Hennrich, O. Fortmann, and T. Schultz, "Classification of mental tasks in the prefrontal cortex using fNIRS." *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 2160–3, 2013.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets." *Neural computation*, vol. 18, no. 7, pp. 1527–54, July 2006.

[10] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," no. 1, 2007.

[11] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] A.-r. Mohamed, G. Dahl, and G. Hinton, "Deep Belief Networks for phone recognition," pp. 1–9, 2009.

[13] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," no. c, pp. 1–10, 2010.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 504–7, July 2006.

[15] D. Cirean, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," no. February, p. 20, Feb. 2012.

[16] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription," *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 24–29, Dec. 2011.

[17] Y. Bengio and N. Y. Georgios, "Learning Deep Physiological Models of Affect," no. April, pp. 20–33, 2013.

[18] J. L. Blue and P. J. Grother, "Training feed-forward neural networks using conjugate gradients," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 179–190.

[19] J. Martens, "Deep learning via Hessian-free optimization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 735–742.

[20] G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," 2010.

[21] W. S. Sarle, "Stopped training and other remedies for overfitting," in *Proc. of the 27th symposium on the interface of computing science and statistics*, 1995, pp. 352–360.

[22] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components–a tutorial." *NeuroImage*, vol. 56, no. 2, pp. 814–25, May 2011.

[23] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, Feb. 2004.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 12 June 2015
doi: 10.3389/fnins.2015.00217

Brain-to-text: decoding spoken phrases from phone representations in the brain

Christian Herff[1†], Dominic Heger[1*†], Adriana de Pesters[2,3], Dominic Telaar[1], Peter Brunner[2,4], Gerwin Schalk[2,3,4] and Tanja Schultz[1]*

[1] Cognitive Systems Lab, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany , [2] New York State Department of Health, National Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, NY, USA , [3] Department of Biomedical Sciences, State University of New York at Albany, Albany, NY, USA, [4] Department of Neurology, Albany Medical College, Albany, NY, USA

OPEN ACCESS

Edited by:
Giovanni Mirabella,
Sapienza University, Italy

Reviewed by:
Christoph Guger,
Guger Technologies OEG, G.tec Medical Engineering GmbH, G.tec Neurotechnology USA Inc., Austria
Damien Coyle,
University of Ulster, UK

***Correspondence:**
Christian Herff and Dominic Heger,
Cognitive Systems Lab, Institute for Anthropomatics and Robotics,
Karlsruhe Institute of Technology,
Adenauerring 4, 76131 Karlsruhe,
Germany
christian.herff@kit.edu;
dominic.heger@kit.edu

† These authors have contributed equally to this work.

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

Received: 09 April 2015
Accepted: 18 May 2015
Published: 12 June 2015

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

It has long been speculated whether communication between humans and machines based on natural speech related cortical activity is possible. Over the past decade, studies have suggested that it is feasible to recognize isolated aspects of speech from neural signals, such as auditory features, phones or one of a few isolated words. However, until now it remained an unsolved challenge to decode continuously spoken speech from the neural substrate associated with speech and language processing. Here, we show for the first time that continuously spoken speech can be decoded into the expressed words from intracranial electrocorticographic (ECoG) recordings.Specifically, we implemented a system, which we call *Brain-To-Text* that models single phones, employs techniques from automatic speech recognition (ASR), and thereby transforms brain activity while speaking into the corresponding textual representation. Our results demonstrate that our system can achieve word error rates as low as 25% and phone error rates below 50%. Additionally, our approach contributes to the current understanding of the neural basis of continuous speech production by identifying those cortical regions that hold substantial information about individual phones. In conclusion, the Brain-To-Text system described in this paper represents an important step toward human-machine communication based on imagined speech.

Keywords: electrocorticography, ECoG, speech production, automatic speech recognition, brain-computer interface, speech decoding, pattern recognition, broadband gamma

1. Introduction

Communication with computers or humans by thought alone, is a fascinating concept and has long been a goal of the brain-computer interface (BCI) community (Wolpaw et al., 2002). Traditional BCIs use motor imagery (McFarland et al., 2000) to control a cursor or to choose between a selected number of options. Others use event-related potentials (ERPs) (Farwell and Donchin, 1988) or steady-state evoked potentials (Sutter, 1992) to spell out texts. These interfaces have made remarkable progress in the last years, but are still relatively slow and unintuitive. The possibility of using covert speech, i.e., imagined continuous speech processes recorded from the brain for human-computer communication may improve BCI communication speed and also increase their usability. Numerous members of the scientific community, including linguists, speech processing

technologists, and computational neuroscientists have studied the basic principles of speech and analyzed its fundamental building blocks. However, the high complexity and agile dynamics in the brain make it challenging to investigate speech production with traditional neuroimaging techniques. Thus, previous work has mostly focused on isolated aspects of speech in the brain.

Several recent studies have begun to take advantage of the high spatial resolution, high temporal resolution and high signal-to-noise ratio of signals recorded directly from the brain [electrocorticography (ECoG)]. Several studies used ECoG to investigate the temporal and spatial dynamics of speech perception (Canolty et al., 2007; Kubanek et al., 2013). Other studies highlighted the differences between receptive and expressive speech areas (Towle et al., 2008; Fukuda et al., 2010). Further insights into the isolated repetition of phones and words has been provided in Leuthardt et al. (2011b); Pei et al. (2011b). Pasley et al. (2012) showed that auditory features of perceived speech could be reconstructed from brain signals. In a study with a completely paralyzed subject, Guenther et al. (2009) showed that brain signals from speech-related regions could be used to synthesize vowel formants. Following up on these results, Martin et al. (2014) decoded spectrotemporal features of overt and covert speech from ECoG recordings. Evidence for a neural representation of phones and phonetic features during speech perception was provided in Chang et al. (2010) and Mesgarani et al. (2014), but these studies did not investigate continuous speech production. Other studies investigated the dynamics of the general speech production process (Crone et al., 2001a,b). A large number of studies have classified isolated aspects of speech processes for communication with or control of computers. Deng et al. (2010) decoded three different rhythms of imagined syllables. Neural activity during the production of isolated phones was used to control a one-dimensional cursor accurately (Leuthardt et al., 2011a). Formisano et al. (2008) decoded isolated phones using functional magnetic resonance imaging (fMRI). Vowels and consonants were successfully discriminated in limited pairings in Pei et al. (2011a). Blakely et al. (2008) showed robust classification of four different phonemes. Other ECoG studies classified syllables (Bouchard and Chang, 2014) or a limited set of words (Kellis et al., 2010). Extending this idea, the imagined production of isolated phones was classified in Brumberg et al. (2011). Recently, Mugler et al. (2014b) demonstrated the classification of a full set of phones within manually segmented boundaries during isolated word production.

To make use of these promising results for BCIs based on continuous speech processes, the analysis and decoding of isolated aspects of speech production has to be extended to continuous and fluent speech processes. While relying on isolated phones or words for communication with interfaces would improve current BCIs drastically, communication would still not be as natural and intuitive as continuous speech. Furthermore, to process the content of the spoken phrases, a textual representation has to be extracted instead of a reconstruction of acoustic features. In our present study, we address these issues by analyzing and decoding brain signals during continuously produced overt speech. This enables us to reconstruct continuous speech into a sequence of words in textual form, which is a necessary step toward human-computer communication using the full repertoire of imagined speech. We refer to our procedure that implements this process as *Brain-to-Text*. Brain-to-Text implements and combines understanding from neuroscience and neurophysiology (suggesting the locations and brain signal features that should be utilized), linguistics (phone and language model concepts), and statistical signal processing and machine learning. Our results suggest that the brain encodes a repertoire of phonetic representations that can be decoded continuously during speech production. At the same time, the neural pathways represented within our model offer a glimpse into the complex dynamics of the brain's fundamental building blocks during speech production.

2. Materials and Methods

2.1. Subjects

Seven epileptic patients at Albany Medical Center (Albany, New York, USA) participated in this study. All subjects gave informed consent to participate in the study, which was approved by the Institutional Review Board of Albany Medical College and the Human Research Protections Office of the US Army Medical Research and Materiel Command. Relevant patient information is given in **Figure 1**.

2.2. Electrode Placement

Electrode placement was solely based on clinical needs of the patients. All subjects had electrodes implanted on the left hemisphere and covered relevant areas of the frontal and temporal lobes. Electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) were composed of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) embedded in silicon with an inter-electrode distance of 0.6-1 cm. Electrode positions were registered in a post-operative CT scan and co-registered with a pre-operative MRI scan. **Figure 1** shows electrode positions of all 7 subjects and the combined electrode positions. To compare average activation patterns across subjects, we co-registered all electrode positions in common Talairach space. We rendered activation maps using the NeuralAct software package (Kubanek and Schalk, 2014).

2.3. Experiment

We recorded brain activity during speech production of seven subjects using electrocorticographic (ECoG) grids that had been implanted as part of presurgical producedures preparatory to epilepsy surgery. ECoG provides electrical potentials measured directly on the brain surface at a high spatial and temporal resolution, unfiltered by skull and scalp. ECoG signals were recorded by BCI2000 (Schalk et al., 2004) using eight 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria). In addition to the electrical brain activity measurements, we recorded the acoustic waveform of the subjects' speech. Participant's voice data was recorded with a dynamic microphone (Samson R21s) and digitized using a dedicated g.USBamp in sync with the

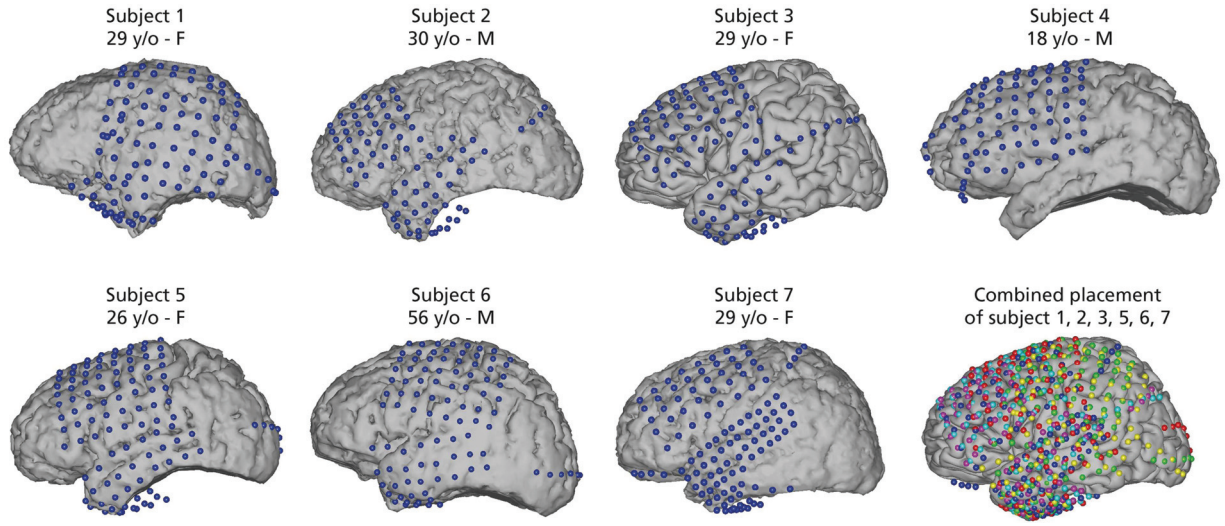


FIGURE 1 | Electrode positions for all seven subjects. Captions include age [years old (y/o)] and sex of subjects. Electrode locations were identified in a post-operative CT and co-registered to preoperative MRI. Electrodes for subject 3 are on an average Talairach brain. Combined electrode placement in joint Talairach space for comparison of all subjects. Participant 1 (yellow), subject 2 (magenta), subject 3 (cyan), subject 5 (red), subject 6 (green), and subject 7 (blue). Participant 4 was excluded from joint analysis as the data did not yield sufficient activations related to speech activity (see Section 2.4).

ECoG signals. The ECoG and acoustic signals were digitized at a sampling rate of 9600 Hz.

During the experiment, text excerpts from historical political speeches (i.e., Gettysburg Address, Roy and Basler, 1955), JFK's Inaugural Address (Kennedy, 1989), a childrens' story (Crane et al., 1867) or *Charmed* fan-fiction (Unknown, 2009) were displayed on a screen in about 1 m distance from the subject. The texts scrolled across the screen from right to left at a constant rate. This rate was adjusted to be comfortable for the subject prior to the recordings (rate of scrolling text: 42–76 words/min). During this procedure, subjects were familiarized with the task.

Each subject was instructed to read the text aloud as it appeared on the screen. A session was repeated 2–3 times depending on the mental and physical condition of the subjects. **Table 1** summarizes data recording details for every session. Since the amount of data of the individual sessions of subject 2 is very small, we combined all three sessions of this subject in the analysis.

We cut the read-out texts of all subjects into 21–49 phrases, depending on the session length, along pauses in the audio recording. The audio recordings were phone-labeled using our in-house speech recognition toolkit BioKIT Telaar et al., 2014 (see Section 2.5). Because the audio and ECoG data were recorded in synchronization (see **Figure 2**), this procedure allowed us to identify the ECoG signals that were produced at the time of any given phones. **Figure 2** shows the experimental setup and the phone labeling.

TABLE 1 | Data recording details for every session.

Participant	Session	Text	Number of phrases	Total recording length (s)
1	1	Gettysburg address	36	279.87
	2	JFK inaugural	38	326.90
2	1	Humpty dumpty	21	129.87
	2	Humpty dumpty	21	129.07
	3	Humpty dumpty	21	126.37
3	1	Charmed fan-fiction	42	310.27
	2	Charmed fan-fiction	40	310.93
	3	Charmed fan-fiction	41	307.50
4	1	Gettysburg address	38	299.67
	2	Gettysburg address	38	311.97
5	1	JFK inaugural	49	341.77
	2	Gettysburg address	39	222.57
6	1	Gettysburg address	38	302.83
7	1	JFK inaugural	48	590.10
	2	Gettysburg address	38	391.43

2.4. Data Pre-Selection

In an initial data pre-selection, we tested whether speech activity segments could be distinguished from those with no speech activity in ECoG data. For this purpose, we fitted a multivariate normal distribution to all feature vectors (see Section 2.6 for a description of the feature extraction) containing speech activity derived from the acoustic data and one to feature vectors when the subject was not speaking. We then determined whether these models could be used to classify general speech activity above chance level, applying a leave-one-phrase-out validation.

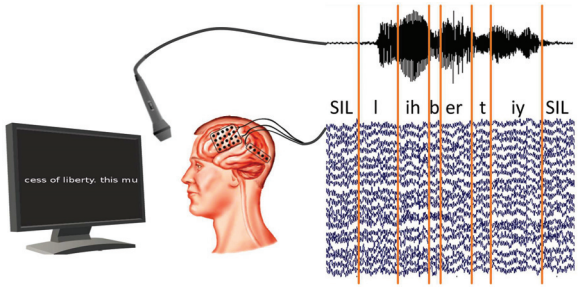


FIGURE 2 | Synchronized recording of ECoG and acoustic data. Acoustic data are labeled using our in-house decoder BioKIT, i.e., the acoustic data samples are assigned to corresponding phones. These phone labels are then imposed on the neural data.

Based on this analysis, both sessions of subject 4 and session 2 of subject 5 were rejected, as they did not show speech related activations that could be classified significantly better than chance (t-test, $p > 0.05$). To compare against random activations without speech production, we employed the same randomization approach as described in Section 2.11.

2.5. Phone Labeling

Phone labels of the acoustic recordings were created in a three-step process using an English automatic speech recognition (ASR) system trained on broadcast news. First, we calculated a Viterbi forced alignment (Huang et al., 2001), which is the most likely sequence of phones for the acoustic data samples given the words in the transcribed text and the acoustic models of the ASR system. In a second step, we adapted the Gaussian mixture model (GMM)-based acoustic models using maximum likelihood linear regression (MLLR) (Gales, 1998). This adaptation was performed separately for each session to obtain session-dependent acoustic models specialized to the signal and speaker characteristics, which is known to increase ASR performance. We estimated a MLLR transformation from the phone sequence computed in step one and used only those segments which had a high confidence score that the segment was emitted by the model attributed to them. Third, we repeated the Viterbi forced alignment using each session's adapted acoustic models yielding the final phone alignments. The phone labels calculated on the acoustic data are then imposed on the ECoG data.

Due to the very limited amount of training data for the neural models, we reduced the amount of distinct phone types and grouped similar phones together for the ECoG models. The grouping was based on phonetic features of the phones. See **Table 2** for the grouping of phones.

2.6. Feature Extraction

We segmented the neural signal data continuously into 50 ms intervals with an overlap of 25 ms, which enabled us to capture the fast cortical processes underlying phones, while being long enough to extract broadband (70–170 Hz) gamma activity reliably. Each of the 50 ms intervals was labeled with the corresponding phone obtained from the audio phone labeling.

TABLE 2 | Grouping of phones.

Grouped phone	IPA phones
aa	ɑ æʌ
b	b
ch	tʃ ʃ ʒ
eh	ɛ ɝ eɪ
f	f
hh	h
ih	i ɪ
jh	dʒ g j
k	k
l	ɫ
m	m
n	n ŋ
ow	oʊ ɔ
p	p
r	r
s	s z ð θ
t	t d
uw	u ʊ
v	v
w	w
Diphtongs	
ow ih	ɔɪ
aa ih	aɪ
aa ow	aʊ

English phones are based on the International Phonetic Alphabet (IPA).

We extracted broadband-gamma activations as they are known to be highly task-related for motor tasks (Miller et al., 2007), music perception (Potes et al., 2012), auditory processes (Pei et al., 2011b; Pasley et al., 2012) and word repetition (Leuthardt et al., 2011b). Broadband-gamma features were extracted from the ECoG electrical potentials as follows: linear trends in the raw signals were removed from each channel. The signals were down-sampled from 9600 to 600 Hz sampling rate. Channels strongly affected by noise were identified and excluded from further processing. Specifically, we calculated the energy in the frequency band 58–62 Hz (line noise) and removed channels with more noise energy than two interquartile ranges above the third quartile of the energy of all channels in the data set. This way, an average of 7.0 (std 6.5) channels were removed per subject.

The remaining channels were re-referenced to a common average (i.e., CAR filtering). Elliptic IIR low-pass and high-pass filters were applied to represent broadband gamma activity in the signals. An elliptic IIR notch filter (118–122 Hz, filter order 13) was applied to attenuate the first harmonic of 60 Hz line noise, which is within the broadband gamma frequency range.

Resulting 50 ms intervals are denoted as $X_{i,c}(t)$ and consist of n samples ($t \in [1, \ldots, n]$). For each interval i and channel c, the signal energy $E_{i,c}$ was calculated and the logarithm was applied to make the distribution of the energy features approximately Gaussian: $E_{i,c} = log(\frac{1}{n}\sum_{t=1}^{n} X_{i,c}(t)^2)$. The

logarithmic broadband gamma power of all channels were concatenated into one feature vector $E_i = [E_{i,1}, \ldots, E_{i,d}]$. To integrate context information and temporal dynamics of the neural activity for each interval, we included neighboring intervals up to 200 ms prior to and after the current interval, similar context sizes have been found relevant in speech perception studies Sahin et al., 2009. Therefore, each feature vector was stacked with four feature vectors in the past and four feature vectors in the future. Stacked feature vectors $F_i = [E_{i-4}, \ldots, E_i, \ldots, E_{i+4}]^\top$ were extracted every 25 ms over the course of the recording sessions and the fitting phone label (ground truth from acoustic phone labeling) was associated.

2.7. Identification of Discriminability

The high temporal and spatial resolution of ECoG recordings allowed us to trace the temporal dynamics of speech production through the areas in the brain relevant for continuous natural speech production. To investigate such cortical regions of high relevance, we calculated the mean symmetrized Kullback-Leibler divergence (KL-div) among the phone models for each electrode position and at every time interval.

The Kullback-Leibler divergence (KL-div) is a measure of the difference between two distributions P and Q. It can be interpreted as the amount of discriminability between the neural activity models in bits. It is non-symmetric and does not satisfy the triangle inequality. The KL-div can be interpreted as the amount of extra bits needed to code samples from P when using Q to estimate P. When both distributions P and Q are normal distributions with means μ_0 and μ_1 and covariances Σ_0 and Σ_1, respectively, the KL-div can be easily calculated as

$$D_{KL}(N_0||N_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - d - log_2(\frac{det(\Sigma_0)}{det(\Sigma_1)})) \quad (1)$$

with d being the dimensionality of the distributions. The closed-form of the KL-div enables us to calculate the difference between two phone models. To estimate the discriminability of a feature $E_{i,c}$ (log broadband gamma power of a particular channel and time interval) for the classification of phones, we calculate the mean KL-div between all pairs of phones for this particular feature. The mean between all divergences symmetrizes the KL-div and yields one number in bits as the estimation of the discriminability of this particular feature $E_{i,c}$.

2.8. Feature Selection

We selected features with the largest average distance between phone models based on the mean KL-div (cf. previous section) in the training data during each run of the leave-one-phrase-out validation. The number of features selected was automatically determined based on the distribution of KL-div for this specific run as follows: We normalized the mean KL-div values d_k for every feature k by their average ($\hat{d}_k = \frac{d_k}{\sum_k d_k}$). Then, we sorted the values in descending order and selected features with large normalized mean KL-div until the sorted sequence did not decline more than a threshold $t = -0.05$: $\arg\max_l \; sort(\hat{d}_k)_l - -sort(\hat{d}_k)_{l+1} < t$. The threshold value $t = -0.05$ corresponds to a very low decline in KL-div and thus reflected the point after which little additional information was present. This way, only the l most relevant features are selected to limit the feature space.

Note that features are selected solely based on the Kullback-Leibler divergence in the training data and do not include any prior assumptions on the suitability of specific regions for phone discrimination. We further reduced the feature space dimensionality by linear discriminant analysis (LDA) (Haeb-Umbach and Ney, 1992) using the phone labels on the training data.

2.9. ECoG Phone Model Training

Each phone was modeled in the extracted feature space by a normal distribution. Thus, models characterized the mean contribution and variance of the neural activity measured at each electrode. We represented the stacked cortical activity feature vectors F_i of each phone j by a model λ_j as a multivariate Gaussian probability density function $p(F_i|\lambda_j) \sim \mathcal{N}(\mu_j, \Sigma_j)$ determined by the mean feature vectors μ_j and their diagonal variance matrix Σ_j calculated from training data. Gaussian models were chosen as they represent the underlying feature distribution suitably well. Furthermore, Gaussian models can be robustly calculated from a small amount of data, they are computationally very efficient and allow a closed form calculation of the Kullback-Leibler-Divergence.

2.10. Decoding Approach

Following a common idea of modern speech recognition technology (Rabiner, 1989; Schultz and Kirchhoff, 2006), we combined the information about the observed neural activity with statistical language information during the decoding process by Bayesian updating (Rabiner, 1989). Simplified, the process can be understood (Gales and Young, 2008) as finding the sequence of words $W = w_1 \ldots w_L$ which is most likely given the observed ECoG feature segments $X = F_1 \ldots F_T$. This probability $P(W|X)$ can be transformed using Bayes' rule:

$$\hat{W} = \arg\max_W \{P(W|X)\} = \arg\max_W \{p(X|W)P(W)\} \quad (2)$$

Here, the likelihood $p(X|W)$ is given by the ECoG phone models and $P(W)$ is calculated using a language model. The likelihood of ECoG phone models $p(X|W)$ given a word W is calculated by concatenating ECoG phone models to form words as defined in a pronunciation dictionary. Specifically, we employed a pronunciation dictionary containing the mapping of phone sequences to words, for example, describing that the word "liberty" comprises of the phone sequence "/l/ /ih/ /b/ /er/ /t/ /iy/." We constructed a minimized and determinized search graph consisting of the phone sequences for each recognizable word. To capture important syntactic and semantic information of language, we used a statistical language model (Jelinek, 1997; Stolcke, 2002) that predicts the next word given the preceding words. In N-gram language modeling, this is done by calculating probabilities of single words and probabilities for predicting words given the $n-1$ previous words. Probabilities

for single word occurrence ($n = 1$) are called uni-grams. Probabilities for the co-occurrence of two words ($n = 2$) are called bi-grams. For the *Brain-to-Text* system, we estimate bi-grams on the texts read by the subjects. It is important to note that even though this results in very specialized models, the correctness of our results is still assured, as the same language models are utilized for both the real as well as for the control analyses.

Finally, the decoding of spoken phrases from neural data X is performed by finding the word sequence $\hat{W}$ in the search graph that has the highest likelihood for producing the neural data with respect to the ECoG phone models and language information given by pronunciation dictionary and language model.

Figure 3 illustrates the different steps of decoding continuously spoken phrases from neural data. *ECoG signals over time* are recorded at every electrode and divided into 50 ms segments. For each 50 ms interval of recorded *broadband gamma activity*, stacked feature vectors are calculated (*Signal processing*). For each *ECoG phone model* calculated on the training data, the likelihood that this model emitted a segment of ECoG features can be calculated, resulting in *phone likelihoods over time*. Combining these Gaussian *ECoG phone models* with language information in the form of a *dictionary* and an n-gram *language model*, the *Viterbi* algorithm calculates the *most likely word sequence* and corresponding *phone sequence*. To visualize the decoding path, the *most likely phone sequence* can be shown in the *phone likelihoods over time* (red marked areas). The system outputs the decoded word sequence. Overall, the system produces a textual representation from the measured brain activity (see also Supplementary Video).

2.11. Evaluation

For the evaluation of our *Brain-to-Text* system, we trained neural phone models using all but one phrase of a recording session and decoded the remaining phrase. This evaluation process was repeated for each phrase in the session. Through this leave-one-phrase-out validation, we make sure that all feature selection, dimensionality reduction and training steps are only performed on the training data while the test data remains completely unseen. For comparison, we performed the decoding with randomized phone models. This is a baseline that quantifies how well the language model and dictionary decode phrases without any neural information. To obtain an estimate for chance levels in our approach, we shifted the training data by half its length in each iteration of the leave-one-phrase-out validation while the corresponding labels remained unchanged. This way, the data for the random comparison models still have the typical properties of ECoG broadband gamma activity, but do not correspond to the underlying labels. Furthermore, as the labels are not changed, prior probabilities remain the same for the random and the actual model case. As the shifting point is different for all iterations of the specific session, we get an estimate of the chance level performance for every phrase. The mean over all these results thus allows a robust estimation of the true chance level (randomization test).

It is also important to bear in mind that *Brain-to-Text* is still at a disadvantage compared to traditional speech recognition systems as our data contained only several minutes of ECoG signals for each subject. This limited model complexity compared to traditional speech recognition systems, which are usually

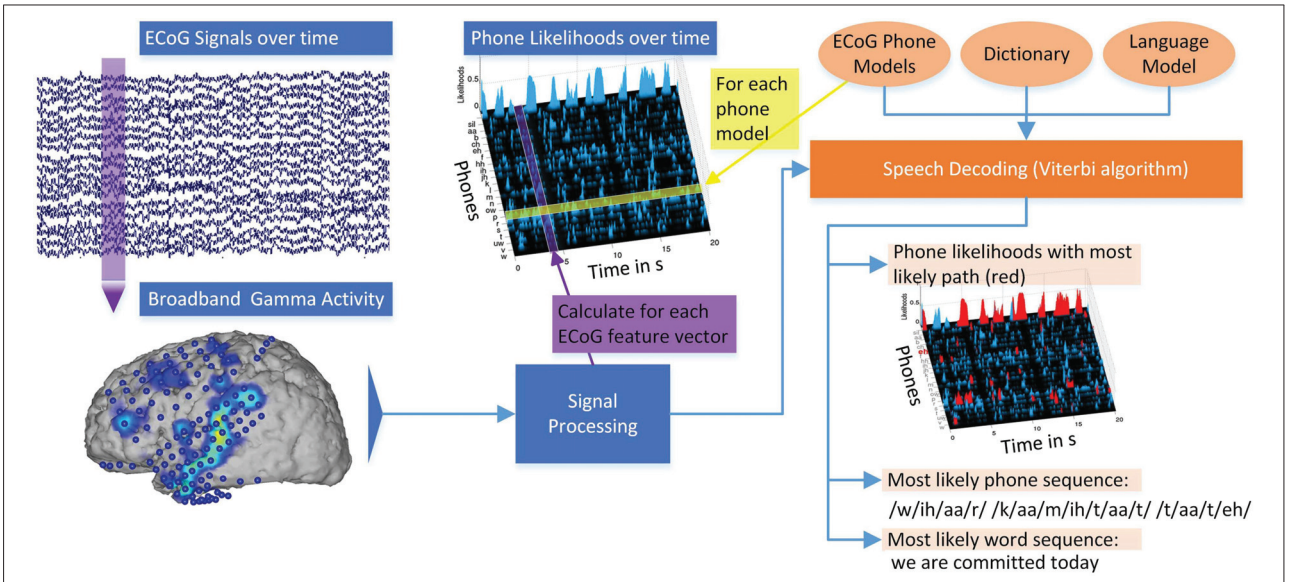


FIGURE 3 | Overview of the *Brain-to-Text* system: ECoG broadband gamma activities (50 ms segments) for every electrode are recorded. Stacked broadband gamma features are calculated (Signal processing). Phone likelihoods over time can be calculated by evaluating all Gaussian ECoG phone models for every segment of ECoG features. Using ECoG phone models, a dictionary and an n-gram language model, phrases are decoded using the Viterbi algorithm. The most likely word sequence and corresponding phone sequence are calculated and the phone likelihoods over time can be displayed. Red marked areas in the phone likelihoods show most likely phone path. See also Supplementary Video.

trained on thousands of hours of acoustic data and billions of words for language model training.

We evaluated the performance of our *Brain-to-Text* system with different dictionary sizes. For this purpose, we created new dictionaries for every test phrase including the words that were actually spoken plus a set of randomized set of words from the full dictionary. Created dictionaries were the same for *Brain-To-Text* and randomized models to ensure that the words chosen had no influence on the comparison between models. The language model was limited to the words in the dictionary accordingly. This approach allowed us to perpetually increase the dictionary size.

3. Results

3.1. Regions of Discriminability

Figure 4 illustrates the spatio-temporal dynamics of the mean KL-div between the phone models on a joint brain surface (Talairach model, Talairach and Tournoux, 1988) for nine temporal intervals with co-registered electrodes of all subjects. KL-div values plotted in **Figure 4** exceed 99% of the KL-div values with a randomized phone-alignment (data shifted by half its length while the labels remain the same).

Starting 200 ms before the actual phone production, we see high KL-div values in diverse areas including Broca's area, which is generally associated with speech planning (Sahin et al., 2009). 150 ms prior to the phone production, Broca's area still has high KL-div scores, but additionally sensorimotor areas and regions in the superior temporal gyrus associated with auditory and language function show increasing discriminability. Subsequently, activations in Broca's area vanish and motor area discriminability increases until peaking at the interval between 0 and 50 ms (which corresponds to the average length of phones). Discriminability increases in auditory regions until approximately 150 ms after phone production.

3.2. Decoding Results

For each phrase to be decoded, the most likely phone-path can be efficiently calculated using Viterbi decoding (Rabiner, 1989). Comparing the extracted phone labels for each feature vector with the baseline labels from the audio alignment, we calculate single-frame accuracies for the decoding of phones from continuous speech production. Reducing the size of

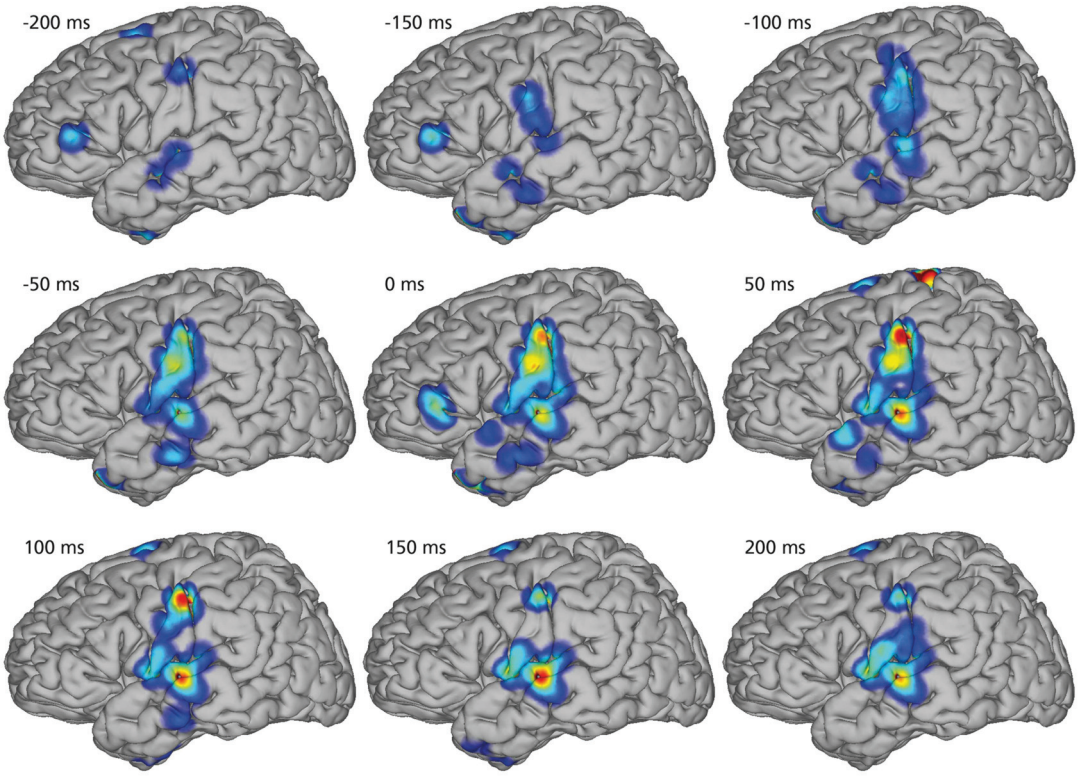


FIGURE 4 | Mean Kullback-Leibler Divergences between models for every electrode position of every subject. Combined electrode montage of all subjects except subject 4 in common Talairach space. Heat maps on rendered average brain shows regions of high discriminability (red). All shown discriminability exceeds chance level (larger than 99% of randomized discriminabilities). The temporal course of regions with high discriminability between phone models shows early differences in diverse areas up to 200 ms before the actual phone production. Phone models show high discriminability in sensorimotor cortex 50 ms before production and yield different models in auditory regions of the superior temporal gyrus 100 ms after production.

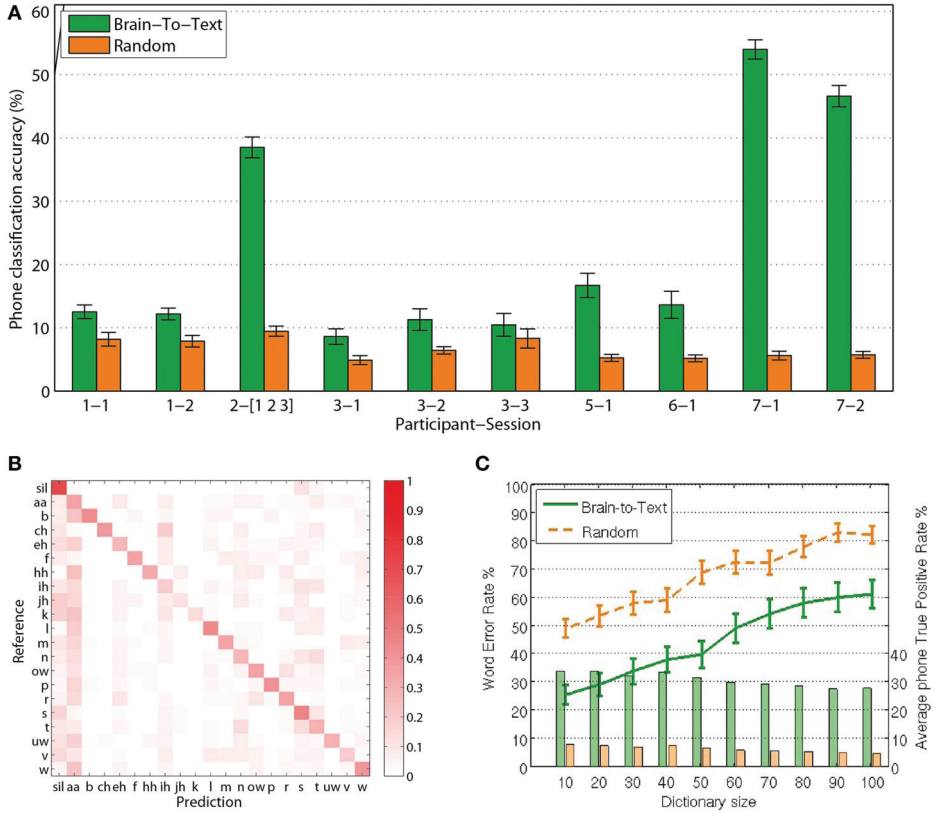


FIGURE 5 | Results: (A) Frame-wise accuracy for all sessions. All sessions of all subjects show significantly higher true positive rates for *Brain-To-Text* (green bars) than for the randomized models (orange bars). **(B)** Confusion matrix for subject 7, session 1. The clearly visible diagonal indicates that all phones are decoded reliably. **(C)** Word Error Rates depending on dictionary size (lines). Word error rates for *Brain-To-Text* (green line) are lower than the randomized models for all dictionary sizes. Average true-positive rates across phones depending on dictionary size (bars) for subject 7, session 1. Phone true positive rates remain relatively stable for all dictionary sizes and are always much higher for *Brain-To-Text* than for the randomized models.

the dictionary to 10 words, including those that are to be evaluated, *Brain-to-Text* yielded significantly higher accuracies (two-sided t-test, $p < 0.05$ for all sessions) for single phone decoding in all sessions compared to random models. **Figure 5A** shows average phone recognition accuracies (green) and average random recognition accuracies (orange) for each session. The best session resulted in average accuracies above 50% for the correct classification of 20 phones plus SILENCE. While all sessions resulted in significantly higher accuracies than random models, the results of subject 2 and subject 7 clearly outperform those of all other subjects. The outstanding performance of subject 7 might be explained by the high-density grid on the superior temporal gyrus. We further investigate the results of subject 7, session 1 (results for all other subjects and sessions can be found in the Supplementary Material) by investigating the confusion matrix (**Figure 5B**) that shows which phones in the reference corresponded to which phones in the predicted phrase. The clearly visible diagonal in this confusion matrix illustrates that our approach reliably decodes the complete set of phones.

In *Brain-to-Text*, we decode entire word sequences of each test phrase. Even with a small dictionary size, a large number of different phrases can be produced, as the number of words may vary and words can be arbitrarily combined. Therefore, we utilize the Word Error Rate (WER) to measure the quality of a decoded phrase. The word error rate (WER) between a predicted phrase and the corresponding reference phrase consists of the number of editing steps in terms of substitutions, deletions and insertions of words necessary to produce the predicted phrase from the reference, divided by the amount of words in the reference.

Figure 5C shows the average WER depending on dictionary size (green line). For all dictionary sizes, the performance is significantly better than randomized results (orange line). Significance was analyzed using paired t-tests between the Word Error Rates of *Brain-To-Text* and the randomized models ($p < 0.001$, one-sided paired t-test). With 10 words in the dictionary, 75% of all words are recognized correctly. The approach scales well for increasing dictionary sizes. Average phone true positive rates remain rather stable even when dictionary sizes increase (bars in **Figure 5C**).

4. Discussion

4.1. ECoG Phone Models

Gaussian models as a generative statistical representation for log-transformed broadband gamma power have been found well-suited for the observed cortical activity (e.g., Gasser et al., 1982; Crone et al., 2001b). These models facilitate the analysis of the spatial and temporal characteristics of each phone model within its 450 ms context. Note that the modeling of phones does not contradict recent findings of articulatory features in neural recordings during speech perception (Pulvermüller et al., 2006; Mesgarani et al., 2014) and production (Bouchard et al., 2013; Lotte et al., 2015), since multiple representations of the same acoustic phenomenon are likely.

Note that only one context-independent model is trained for each phone, i.e., without consideration of preceding or succeeding phones due to the limited amount of data, even though effects of context have been shown in neural data (Mugler et al., 2014a). While context dependent modeling is very common in acoustic speech recognition (Lee, 1990) and known to significantly improve recognition performance, it requires substantially more training data than available in our ECoG setting.

4.2. Regions of Discriminability

In our approach, the phone representation through Gaussian models allows for detailed analysis of cortical regions, which have high discriminability among the different phones over time. The cortical locations identified using the KL-div criterion are in agreement with those that have been identified during speech production and perception in isolated phoneme or word experiments (Canolty et al., 2007; Leuthardt et al., 2011a). These findings extend the state-of-the-art by showing for the first time the dynamics for single phone discriminability and decoding during continuous speech production.

As our experiments demand overt speech production from prompted texts, it is evident that multiple processes are present in the recorded neural data, including speech production, motor actions, auditory processing, and language understanding. By demonstrating that phones can be discriminated from each other, we show that such a phone-based representation is indeed a viable form of modeling cortical activity of continuous speech in this mixture of activation patterns.

4.3. Decoding Results

The reported phone decoding accuracies are significantly higher for *Brain-to-Text* than for randomized models in all subjects, which shows that continuous speech production can be modeled based on phone representations. The clearly visible diagonal in the confusion matrix **Figure 5B** emphasizes that the decoding performance is based on a reliable detection of all phones and not only on a selected subset.

Different conditions, such as varying task performance of the subjects, and different positions and densities of the electrode grids, yielded highly variable decoding performances for the different subjects, however the low WER (see Supplementary Material) and phone true positive rates for subject 1,2, and 7 imply the potential of *Brain-to-Text* for brain-computer interfaces.

4.4. Conclusion

Decoding overt speech production is a necessary first step toward human-computer interaction through imagined speech processes. Our results show that with a limited set of words in the dictionary, *Brain-to-Text* reconstructs spoken phrases from neural data. The computational phone models in combination with language information make it possible to reconstruct words in unseen spoken utterances solely based on neural signals (see Supplementary Video). Despite the fact that the evaluations in this article have been performed offline, all processing steps of *Brain-to-Text* and the decoding approach are well suited for eventual real-time online application on desktop computers. The approach introduced here may have important implications for the design of novel brain-computer interfaces, because it may eventually allow people to communicate solely based on brain signals associated with natural language function and with scalable vocabularies.

Funding

This work was supported by the NIH (EB00856, EB006356, and EB018783), the US Army Research Office (W911NF-08-1-0216, W911NF-12-1-0109, W911NF-14-1-0440) and Fondazione Neuron, and received support by the International Excellence Fund of Karlsruhe Institute of Technology. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

Acknowledgments

We thank Dr. Anthony Ritaccio for patient interactions, Dr. Aysegul Gunduz for help with data recording and Dr. Cuntai Guan for valuable discussions.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2015.00217/abstract

References

Unknown. (2009). *"Traitor among us" and "Split Feelings"*. Available online at: https://www.fanfiction.net/

Blakely, T., Miller, K. J., Rao, R. P., Holmes, M. D., and Ojemann, J. G. (2008). "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids," in *Engineering in Medicine and Biology Society, 2008. EMBC 2008. 30th Annual International Conference of the IEEE* (Vancouver, BC: IEEE), 4964–4967.

Bouchard, K., and Chang, E. (2014). "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *Engineering in Medicine and Biology Society, 2014. EMBC 2014. 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE).

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495, 327–332. doi: 10.1038/nature11911

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

Canolty, R. T., Soltani, M., Dalal, S. S., Edwards, E., Dronkers, N. F., Nagarajan, S. S., et al. (2007). Spatiotemporal dynamics of word processing in the human brain. *Front. Neurosci.* 1:14. doi: 10.3389/neuro.01.1.1.014.2007

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Crane, W., Gilbert, J. S., McConnell, W., Tenniel, J. S., Weir, H., and Zwecker, J. B. (1867). *Mother Gooses Nursery Rhymes. A Collection of Alphabets, Rhymes, Tales and Jingles*. London: George Routledge and Sons.

Crone, N., Hao, L., Hart, J., Boatman, D., Lesser, R., Irizarry, R., et al. (2001a). Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology* 57, 2045–2053. doi: 10.1212/WNL.57.11.2045

Crone, N. E., Boatman, D., Gordon, B., and Hao, L. (2001b). Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* 112, 565–582. doi: 10.1016/S1388-2457(00)00545-9

Deng, S., Srinivasan, R., Lappas, T., and D'Zmura, M. (2010). EEG classification of imagined syllable rhythm using hilbert spectrum methods. *J. Neural Eng.* 7:046006. doi: 10.1088/1741-2560/7/4/046006

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "who" is saying "what"? brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Fukuda, M., Rothermel, R., Juhász, C., Nishida, M., Sood, S., and Asano, E. (2010). Cortical gamma-oscillations modulated by listening and overt repetition of phonemes. *Neuroimage* 49, 2735–2745. doi: 10.1016/j.neuroimage.2009.10.047

Gales, M., and Young, S. (2008). The application of hidden markov models in speech recognition. *Found. Trends Signal Process.* 1, 195–304. doi: 10.1561/2000000004

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Comput. Speech Lang.* 12, 75–98. doi: 10.1006/csla.1998.0043

Gasser, T., Bächer, P., and Möcks, J. (1982). Transformations towards the normal distribution of broad band spectral parameters of the eeg. *Electroencephalogr. Clin. Neurophysiol.* 53, 119–124. doi: 10.1016/0013-4694(82)90112-2

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., et al. (2009). A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE* 4:e8218. doi: 10.1371/journal.pone.0008218

Haeb-Umbach, R., and Ney, H. (1992). "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Vol. 1 (San Francisco, CA), 13–16.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice Hall PTR.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT press.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., and Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* 7:056007. doi: 10.1088/1741-2560/7/5/056007

Kennedy, J. F. (1989). *Inaugural Addresses of the Presidents of the United States*. Washington, DC. Available online at: www.bartleby.com/124/

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PLoS ONE* 8:e53398. doi: 10.1371/journal.pone.0053398

Kubanek, J., and Schalk, G. (2014). NeuralAct: a tool to visualize electrocortical (ECoG) activity on a three-dimensional model of the cortex. *Neuroinformatics* 13, 167–174. doi: 10.1007/s12021-014-9252-3

Lee, K.-F. (1990). Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. *IEEE Trans. Acoust. Speech Signal Process.* 38, 599–609. doi: 10.1109/29.52701

Leuthardt, E. C., Gaona, C., Sharma, M., Szrama, N., Roland, J., Freudenberg, Z., et al. (2011a). Using the electrocorticographic speech network to control a brain–computer interface in humans. *J. Neural Eng.* 8:036004. doi: 10.1088/1741-2560/8/3/036004

Leuthardt, E. C., Pei, X.-M., Breshears, J., Gaona, C., Sharma, M., Freudenberg, Z., et al. (2011b). Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. *Front. Hum. Neurosci.* 6:99. doi: 10.3389/fnhum.2012.00099

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., et al. (2015). Electrocorticographic representations of segmental features in continuous speech. *Front. Hum. Neurosci.* 9:97. doi: 10.3389/fnhum.2015.00097

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., et al. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* 7:14. doi: 10.3389/fneng.2014.00014

McFarland, D. J., Miner, L. A., Vaughan, T. M., and Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topogr.* 12, 177–186. doi: 10.1023/A:1023437823106

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*. 343, 1006–1010. doi: 10.1126/science.1245994

Miller, K. J., Leuthardt, E. C., Schalk, G., Rao, R. P., Anderson, N. R., Moran, D. W., et al. (2007). Spectral changes in cortical surface potentials during motor movement. *J. Neurosci.* 27, 2424–2432. doi: 10.1523/JNEUROSCI.3886-06.2007

Mugler, E., Goldrick, M., and Slutzky, M. (2014a). "Cortical encoding of phonemic context during word production," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE).

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., et al. (2014b). Direct classification of all american english phonemes using signals from functional speech motor cortex. *J. Neural Eng.* 11:035015. doi: 10.1088/1741-2560/11/3/035015

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. (2011a). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *J. Neural Eng.* 8:046028. doi: 10.1088/1741-2560/8/4/046028

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., and Schalk, G. (2011b). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage* 54, 2960–2972. doi: 10.1016/j.neuroimage.2010.10.029

Potes, C., Gunduz, A., Brunner, P., and Schalk, G. (2012). Dynamics of electrocorticographic (ecog) activity in human temporal and frontal cortical areas during music listening. *Neuroimage* 61, 841–848. doi: 10.1016/j.neuroimage.2012.04.022

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Roy, E., and Basler, P. (1955). *The Gettysburg Address, in The Collected Works of Abraham Lincoln*. New Brunswick, NJ: Rutgers University Press.

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Brocas area. *Science* 326, 445–449. doi: 10.1126/science.1174481

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* 51, 1034–1043. doi: 10.1109/TBME.2004.827072

Schultz, T., and Kirchhoff, K. (2006). *Multilingual Speech Processing*. Burlington, MA: Elsevier, Academic Press.

Stolcke, A. (2002). "SRILM - An extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)* (Denver, CO).

Sutter, E. E. (1992). The brain response interface: communication through visually-induced electrical brain responses. *J. Microcomput. Appl.* 15, 31–45. doi: 10.1016/0745-7138(92)90045-7

Talairach, J., and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. Thieme.

Telaar, D., Wand, M., Gehrig, D., Putze, F., Amma, C., Heger, D., et al. (2014). "BioKIT - real-time decoder for biosignal processing," in *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)* (Singapore).

Towle, V. L., Yoon, H.-A., Castelle, M., Edgar, J. C., Biassou, N. M., Frim, D. M., et al. (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain* 131, 2013–2027. doi: 10.1093/brain/awn147

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Music rhythm reconstruction from ECoG

C. Herff[1*], G. Johnson[2], J. Shih[3], T. Schultz[1], D. Krusienski[2]
[1]University of Bremen, Bremen, Germany; [2]Old Dominion University, Norfolk, VA, USA; [3]Mayo Clinic, Jacksonville, FL, USA

*Enrique-Schmidt-Str. 5, 28309 Bremen, Germany. E-mail: christian.herff@uni-bremen.de

Introduction: Imagine being able to record the music you're humming in your head and play it back to others. Prior neuroscientific studies have highlighted auditory processing in the brain in relation to speech [1,2,3] and demonstrated that envelope and spectral properties of perceived speech could be reliably reconstructed. Other studies showed successful reconstruction of acoustic properties from speech production [4] and automatic speech recognition of entire phrases [5]. Perception of melodies and rhythm is generally ubiquitous in humans across age and culture and should also lead to robust representations in neural data. However, only a few studies have investigated neural responses to complex musical stimuli in electrocorticography (ECoG) [6] and have not investigated rhythms and melodies indepentently. Numerous studies exist investigating different aspects of music perception using functional Magnetic Resonance Imaging (fMRI) or the scalp electroencephalogram (EEG), but both of these technologies have limitations in either spatial or spectral resolution, which are necessary for the investigation of the fast processes underlying music perception and production. ECoG, on the other hand, measures high spatial and temporal resolution electrical potentials unfiltered by the skull and scalp, which is more ideal for the investigation of music. Here we investigate cortical responses to perceived drum rhythms and demonstrate reconstruction of the perceived rhythm sound intensities. The investigated drum rhythms lack the rich melodic and harmonic information present in previous studies. Reconstruction of this very simple musical stimuli therefore illustrate basic rhythm perception.

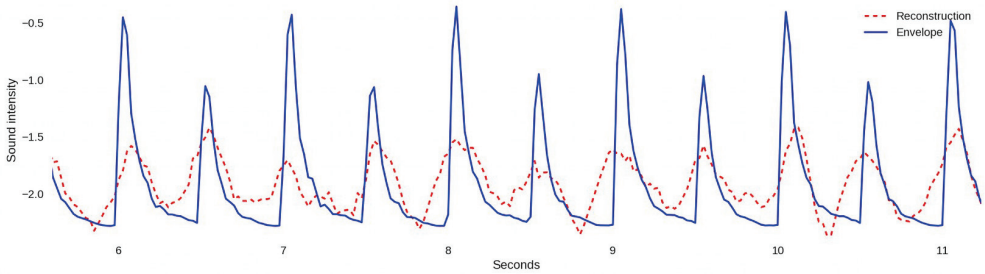


Figure 1. *Drum envelope (blue) and reconstructed (red-dotted) envelope based on high-gamma activity in ECoG electrodes.*

Material and Methods: We presented a simple drum rhythm to seven participants undergoing surgery for intractable epilepsy. Subjects had between 34 and 96 subdural ECoG electrodes implanted (3 left, 4 right hemisphere, frontal; parietal and/or temporal areas covered), based on the clinical need. The sound envelope was extracted using the Hilbert transform in 50 ms windows. We extracted broadband-gamma (70-170 Hz) power in 50 ms windows and time-aligned the ECoG activity to the presented sound stimuli.

Results: We evaluated the possibility to reconstruct perceived sound intensity from the gamma power features across spatial channels using Lasso regression, and evaluated the correlation coefficients (Spearman's r) between actual sound intensity and reconstructed envelope. Statistically significant ($p<0.01$) correlations could be achieved for all subjects with correlations coefficients up to 0.45 (mean 0.15). Figure 1 illustrates the drum envelope (blue) and the reconstruction (red) from ECoG signals.

Discussion: We show that neural data measured directly from the cortex using ECoG can be used to accurately reconstruct the intensity of a repetitive drum stimulus.

Significance: This is a first step towards synthesizing musical rhythms from mental imagery using intracranial signals by reconstruction very basic musical phenomenon.

References

[1] Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. PLoS ONE, 8(1).
[2] Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., ... & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. PLoS-Biology, 10(1), 175.
[3] Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174), 1006-1010.
[4] Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., ... & Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. Frontiers in neuroengineering, 7.
[5] Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. Frontiers in Neuroscience, 9.
[6] Potes, C., Gunduz, A., Brunner, P., & Schalk, G. (2012). Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. Neuroimage, 61(4), 841-848.

Towards direct speech synthesis from ECoG: A pilot study

Christian Herff[1], Garett Johnson[2], Lorenz Diener[1], Jerry Shih[3], Dean Krusienski[2] and Tanja Schultz[1]

Abstract— **Most current Brain-Computer Interfaces (BCIs) achieve high information transfer rates using spelling paradigms based on stimulus-evoked potentials. Despite the success of this interfaces, this mode of communication can be cumbersome and unnatural. Direct synthesis of speech from neural activity represents a more natural mode of communication that would enable users to convey verbal messages in real-time.**

In this pilot study with one participant, we demonstrate that electrocoticography (ECoG) intracranial activity from temporal areas can be used to resynthesize speech in real-time. This is accomplished by reconstructing the audio magnitude spectrogram from neural activity and subsequently creating the audio waveform from these reconstructed spectrograms. We show that significant correlations between the original and reconstructed spectrograms and temporal waveforms can be achieved. While this pilot study uses audibly spoken speech for the models, it represents a first step towards speech synthesis from speech imagery.

I. Introduction

Brain-Computer Interface (BCI) research has made tremendous advances in the last several decades. The most prominent BCIs for communication rely on stimulus-evoked potentials in conjunction with spelling paradigms to type a single letter at a time [1], [2]. Because these systems operate in a stimulus-locked fashion, users can only communicate in predefined intervals. While speech can be synthesized via text-to-speech methods, these systems cannot operate in real-time. Direct synthesis of speech from neural activity represents a more natural mode of communication that would enable users to convey verbal messages in real-time. Additionally, such systems could convey other important aspects of speech communication such as accentuation and prosody.

Neuroscientific investigations show detailed insights into the production of speech [3], [4] and show that speech production and perception are processed very differently in motor areas [5]. Studies have shown that a complete set of English phonemes can be classified from electrocorticography (ECoG) [6], [7]. Others showed that speech recognition technology can be used to reconstruct a textual representation of spoken phrases using ECoG [8], [9]. Despite their innovative direction, these approaches suffer from the same limitations as typing approaches, as additional information of the spoken phrases is lost.

Pasely et al. were able to reconstruct perceived speech from neural activity [10] and Martin et al [11] showed reconstruction of low dimensional spectral representations from audible and imagined speech. We extend on these ideas by reconstructing a complete spectrogram from neural activity. We then use these reconstructed spectrograms to synthesize a waveform of the speech signal. This approach enables users to not only convey a message, but also add extra information such as accentuation, prosody and accent.

In this pilot study, we recorded audible speech and ECoG activity simultaneously from one participant and showed that the speech spectrogram can be reconstructed with promising correlations in an offline analysis. Furthermore, we show that this scheme is fast enough for real-time, online synthesis of speech from the neural signals.

II. Materials and Methods

A. Data Acquisition

Data were collected from a 42 year-old female patient with medically intractable epilepsy who underwent clinical evaluation to localize the epileptogenic zone prior to surgical resection. The patient consented to participate in the study as approved by the IRB of both Mayo Clinic and Old Dominion University. The patient had temporary placement of bilateral temporal depth electrodes (8 contacts apiece, 5 mm spacing), as well as three additional subdural strips placed on the cortex of the left temporal lobe (6 contacts apiece, 1 cm spacing). Electrode (Ad-Tech Medical Instrument Corporation, Wisconsin) placement and duration of intracranial monitoring were solely based on clinical evaluation, with no consideration given to this study. Electrode placements were verified using a postoperative CT. Figure 1 illustrates locations of subdural electrodes.

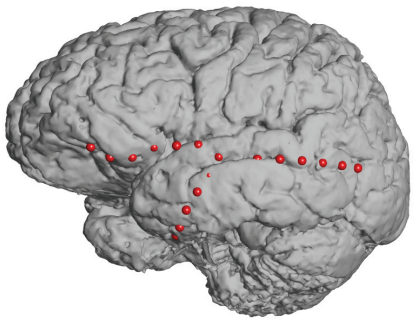


Fig. 1. Electrode positions for the pilot study participant.

The ECoG data were bandpass filtered between 0.5-500 Hz, digitized, and recorded using two 16-channel g.USB amplifiers (Guger Technologies, Austria) at a 1200 Hz

[1] C.H., L.D. and T.S. are with the Cognitive Systems Lab, University of Bremen, Bremen, Germany
[2] G.J. and D.K. are with the Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Lab, Old Dominion University, VA, USA
[3] J.S. is with Mayo Clinic, Jacksonville, FL, USA christian.herff@uni-bremen.de

sampling rate. Simultaneously, a Snowball iCE microphone (Blue Microphones, California) sampled the voice data at 48 kHz. The data recordings were synchronized using the general-purpose BCI system BCI2000 [12].

B. Experiment

For this study, a sentence was presented to the participant visually and aurally for 4 seconds. Subsequently, the participant had 4 seconds to recite the phrase aloud from memory. Sentences from the Harvard Sentence corpus [13] were used. A total of 50 sentences were recited, which resulted in a total of 200 seconds of data.

C. Feature Extraction

The recorded ECoG data were segmented into 50 ms intervals with 25 ms overlap. This duration is short enough to capture the cortical processes associated with speech production and are long enough to extract broadband gamma (70-170 Hz) activity, which is known to be highly task-related [14], [15].

To extract broadband-gamma, linear trends were first removed and data were subsequently downsampled to 600 Hz. The first harmonic of 60 Hz line noise was attenuated using an elliptic IIR notch filter. Elliptic IIR low-pass and high-pass filters were used to isolate the gamma band. Signal energy was then calculated on the filtered signal. A logarithm was applied to the energy estimates to give the power features a more Gaussian distribution.

Context information was included by concatenating 4 neighboring feature vectors up to 200 ms before and after the current interval. This resulted in a total of $18 \cdot 9 = 162$ features in each feature vector x_n for a time interval n.

The audio data was downsampled to 12 kHz to reduce the total spectrogram size. The audio spectrogram is calculated by taking the Short-Time Fourier Transform (STFT) in 50 ms intervals with 25 ms overlap, windowed using Hanning windows. This results in 301 frequency bins per interval. Only the magnitude of the STFT was utilized, as phase information can not be reconstructed from neural signals. The spectral information of a time interval n is denoted as f_n. As ECoG data and audio data are recorded simultaneously, each ECoG feature vector x_n can be assigned a corresponding audio spectrum f_n.

With the phase information missing, the audio signal can not be trivially reconstructed anymore and an approximation method as described in Section II-E is needed.

D. Spectrogram Reconstruction

A linear mapping between ECoG features and log power is estimated in a specific frequency bin. This mapping is obtained using a Lasso regression [16]. The optimal regularization weight α was determined using a nested 10-fold cross-validation. This results in a weight-vector v_i for each spectral bin i and a scalar intercept b_i. Once the models are trained for all spectral bins, all weight-vectors and intercepts can be combined to form a mapping matrix v and an intercept vector b. Using this combined representation, a new frame x_n of ECoG activity can be transfered to the spectral power representation f_n of the audio by simply calculating

$$f_n = v * x_n + b \tag{1}$$

Using a simple linear model for ECoG to speech mapping might not be optimal. Spectral reconstruction methods using deep learning methods have achieved great results in the past [17], but are usually orders of magnitude slower in training and require more time for reconstruction of each spectrum than the simple matrix multiplication needed in our approach. Since this is a pilot study, a linear model was used knowing that more complex methods should be investigated in the future.

E. Speech Synthesis

Given the spectrogram reconstructed from the measured ECoG activity f, one can reconstruct an audio waveform by iteratively modifying the spectral coefficients of a signal initialized with noise. Griffin and Lim [18] proposed Algorithm 1 to reconstruct the waveform from the spectrogram. With

Algorithm 1: Waveform reconstruction

Data: Spectrogram f
Result: Waveform w
$w \leftarrow$ noise;
for $i \leftarrow 1$ **to** l **do**
  $X \leftarrow \text{STFT}(w)$;
  $Z \leftarrow f \exp(i \angle X)$;
  $w \leftarrow \text{ISTFT}(Z)$;

STFT & ISTFT being the Short-Term Fourier Transform and the Inverse Short-Term Fourier Transform, respectively. This allows the reconstruction of a complete audio waveform from the reconstructed spectrograms. Generally, only few iterations l of this procedure are necessary to yield sufficient audio quality. A value of $l = 8$ was chosen as no improvements could be seen with more iterations and processing was still very fast for 8 iterations. This algorithm can be used either on the complete reconstructed spectrogram in offline-analyses, or on each individual spectrum for online-synthesis. In this study, waveform reconstruction was performed on the entire reconstructed spectrogram.

III. Results

A. Spectrogram and Waveform Reconstructions

Figure 2 illustrates an original and reconstructed (log) spectrogram. Figure 3 shows an example of original and reconstructed speech waveforms.

B. Computation Time

To assess the feasibility of our approach for online synthesis of speech from neural signals, all involved components were evaluated in terms of computational time and the thus induced time lag. As hardware offsets induced by data recording and audio output are not within the scope of this analysis, they have not been included. All calculations are

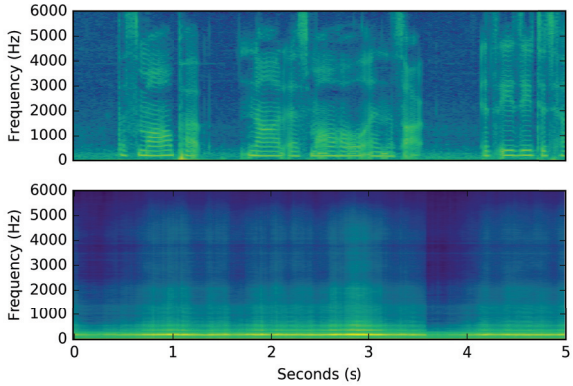


Fig. 2. Original (top) and reconstructed (bottom) spectrograms.

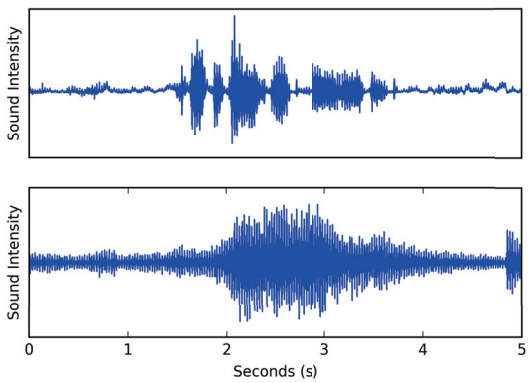


Fig. 3. 5 seconds of original (top) and reconstructed (bottom) waveform. Only very broad characteristics of the waveform can be seen in the reconstructed waveform.

performed on an Intel Core i7 processor running at 3.6 GHz . The time needed for data filtering, feature calculation, spectrogram reconstruction using the linear filter described in Section II-D and the waveform reconstruction described in Section II-E of one frame x_n of ECoG features resulting in 50 ms of audio were measured.

As can be seen in Table I, all operations can be performed in under 1 ms resulting in a total offset far smaller than the 50 ms interval length. Speech synthesis from neural signals can thus be performed in real time.

TABLE I

TIME NEEDED FOR COMPONENTS.

Operation	Computation time
Data filtering	<1 ms
Feature calculation	<1 ms
Spectrogram reconstruction	<1 ms
Waveform synthesis	<1 ms

C. Reconstruction Quality

All evaluations were performed using a 10-fold cross-validation: The Lasso regression models were trained on 90% of the data and were used to reconstruct spectrograms for the remaining 10%. This procedure was repeated 10 times, so that all data were used for testing once. The Lasso regularization parameter α was optimized using a nested 10-fold cross-validation on the training data. The models need approximately 1.5 seconds to be trained for each frequency bin. This would result in a total training time of about 450 seconds for the complete model.

We calculated the Spearman correlation coefficient ρ between the original and reconstructed spectrogram for each frequency bin to assess which parts of the spectral information can be robustly reconstructed. Figure 4 illustrates correlation coefficients over frequency bin. The mean overall correlation over all frequency bins is $\rho = 0.36$. Correlations below 200 Hz are around chance level as no speech information is present in this frequency range. From 200 Hz onwards, rank correlation coefficients increase until reaching a level of approximately 0.4 at around 300 Hz. As the first formant of vowel production usually starts around 300 Hz, high correlation in these frequency ranges is especially important. Rank correlations remain stable up to approximately 5 kHz, after which only little speech information is left in the spectrogram and correlation coefficients deteriorate rapidly in our evaluations. Despite these very promising results, it is evident from the short excerpt in Figure 2 that only very broad aspects of the spectrogram are reconstructed and improvements are still necessary to capture all delicate processes in the speech spectrogram. The achieved correlation coefficients are similar to those achieved by average subjects in [11].

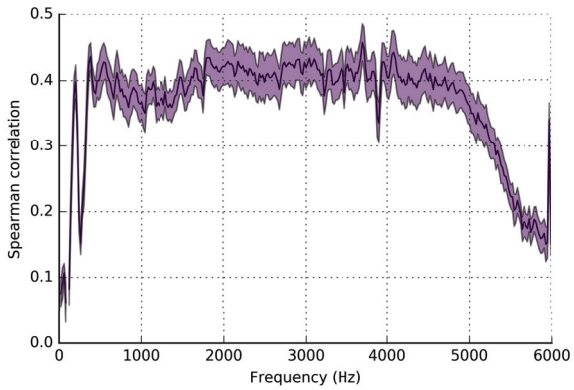


Fig. 4. Spearman correlation coefficients between original and reconstructed spectrograms for different frequency ranges. Purple shaded region denotes standard error of the mean over folds. Reconstruction remains relatively stable between 500 and 5000 Hertz.

To evaluate the synthesized waveform, Spearman correlations between the mean absolute Hilbert envelope in 50 ms intervals of the original and reconstructed waveforms were calculated. This yielded a Spearman correlation of $\rho = 0.41$,

which is significantly better than chance (Randomization Tests, $p < 0.001$). As can be seen in Figure 3, the reconstructed waveform broadly captures the envelope of speech activity, but no detailed resemblance can be observed. Unsurprisingly, the reconstructed waveforms are not intelligible for our pilot study participant. We hypothesize that this might be due to the suboptimal electrode montage only covering areas in the temporal lobe with low density and thus not providing any information from motor areas which have been found to contain a lot of relevant information about speech production [5], [7], [8].

D. Interpretation of Regression Models

To visualize which neural activity is used to reconstruct the spectrogram, the corresponding forward models to the Lasso backward models v were estimated. This is done using the method described by Haufe et al. [19]. Figure 5 visualizes the mean forward model over all frequency bins for the pilot participant. Highest model weights are on regions in the auditory cortex. Activations are rendered using the NeuralAct Software package [20].

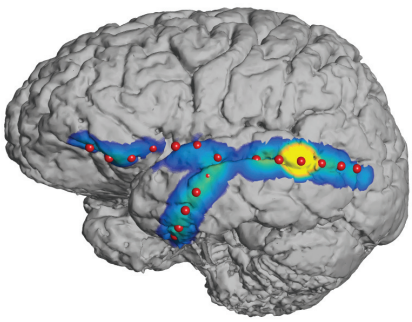


Fig. 5. Average activation pattern of regression models for spectrogram reconstruction.

IV. Conclusions

In a pilot study with one participant, we have shown that intracranial ECoG recordings can be used to synthesize speech. This is achieved by mapping the neural activity directly to magnitude spectrograms which allow for a reconstruction of a speech waveform. Our method yields reconstruction similar to previously reported spectral reconstructions despite a suboptimal electrode montage. Even though the reconstructed waveforms in this pilot study are not intelligible, performance is expected to improve with better coverage of more relevant brain areas. Most significantly, we verified that our approach is capable of achieving real-time online synthesis of speech from neural recordings, which is key in the development of future speech neuroprosthetics.

References

[1] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a p300-based brain-computer interface," *Rehabilitation Engineering, IEEE Transactions on*, vol. 8, no. 2, pp. 174–179, 2000.

[2] G. R. Müller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller, "Steady-state visual evoked potential (ssvep)-based communication: impact of harmonic frequency components." *Journal of neural engineering*, vol. 2, no. 4, pp. 123–130, 2005.

[3] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, pp. 327–332, 2013.

[4] K. Bouchard and E. Chang, "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE, 2014.

[5] C. Cheung, L. S. Hamiton, K. Johnson, and E. F. Chang, "The auditory representation of speech sounds in human motor cortex," *eLife*, vol. 5, p. e12577, mar 2016. [Online]. Available: https://dx.doi.org/10.7554/eLife.12577

[6] E. Mugler, M. Goldrick, and M. Slutzky, "Cortical encoding of phonemic context during word production," in *Engineering in Medicine and Biology Society, 2014. EMBS 2014. 36th Annual International Conference of the IEEE*. IEEE, 2014.

[7] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex," *Journal of Neural Engineering*, vol. 11, no. 3, p. 035015, 2014.

[8] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, no. 217, 2015.

[9] D. Heger, C. Herff, A. d. Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Continuous speech recognition from ecog," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[11] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in Neuroengineering*, vol. 7, no. 14, 2014.

[12] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, 2004.

[13] IEEE, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[14] E. C. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenberg, D. Barbour, and G. Schalk, "Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task." *Frontiers in human neuroscience*, vol. 6, pp. 99–99, 2011.

[15] K. J. Miller, E. C. Leuthardt, G. Schalk, R. P. Rao, N. R. Anderson, D. W. Moran, J. W. Miller, and J. G. Ojemann, "Spectral changes in cortical surface potentials during motor movement," *The Journal of neuroscience*, vol. 27, no. 9, pp. 2424–2432, 2007.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[17] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–7.

[18] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, 1984.

[19] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, 2014.

[20] J. Kubanek and G. Schalk, "Neuralact: A tool to visualize electrocortical (ecog) activity on a three-dimensional model of the cortex," *Neuroinformatics*, pp. 1–8, 2014.

Cross-subject classification of speaking modes using fNIRS

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

[1] Cognitive Systems Lab (CSL),
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{christian.herff,dominic.heger,felix.putze,tanja.schultz}@kit.edu
[2] Institute for Infocomm Research,
Agency for Science, Technology and Research (A*STAR), Singapore

Abstract. In Brain-Computer Interface (BCI) research, subject and session specific training data is usually used to ensure satisfying classification results. In this paper, we show that neural responses to different speaking tasks recorded with functional Near Infrared spectroscopy (fNIRS) are consistent enough across speakers to robustly classify speaking modes with models trained exclusively on other subjects. Our study thereby suggests that future fNIRS-based BCIs can be designed without time-consuming training, which, besides being cumbersome, might be impossible for users with disabilities. Accuracies of 71% and 61% were achieved in distinguishing segments containing overt speech and silent speech from segments in which subjects were not speaking, without using any of the subject's data for training. To rule out artifact contamination, we filtered the data rigorously.
To the best of our knowledge, there are no previous studies showing the zero training capability of fNIRS based BCIs.

Keywords: fNIRS, BCI, speech imagery, cross-subject, session-transfer

1 Introduction

1.1 Motivation

A Brain-Computer Interface (BCI) is a communication channel between a user and a machine. Typical BCI applications target users with disabilities for whom standard input mechanisms are not feasible, due to motor limitations caused by brain stem stroke, cancer or amyotrophic lateral sclerosis, to name a few examples.
Functional Near Infrared Spectroscopy (fNIRS) provides robust measurement of

* 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.

hemodynamic responses in the brain, which are related to neural activity. It is less affected by artifacts caused by movements of the subjects than the de-facto standard modality in BCI, namely electroencephalography (EEG). Compared to functional magnetic resonance imaging (fMRI), which is based on the same hemodynamic effects, fNIRS is far cheaper and more portable. Even though fNIRS is a relatively new brain imaging modality, its feasibility for BCI has been shown in a number of papers [2, 4].

Traditionally, BCIs rely on motor imagery for control, requiring the users to imagine movement of certain parts of their body. Naito et al. [11] first showed the usage of speech related activations, in the form of singing, with a very simple fNIRS sensor. In a very recent study [6], we showed that overt as well as imagined speech is a very reliable and promising paradigm for fNIRS-based human-machine interaction.

As brain signals are non-stationary and user-specific, i.e. they vary significantly over time and, even more so, between users, BCIs usually rely on training intervals from the same session to calibrate the system. Especially in applications for motion impaired users, a training procedure is cumbersome and reduces the time of actual interaction with the system. Recent advances in EEG-based BCI have shown that the usage of data from other subjects and sessions can reduce the time needed to calibrate the system [7, 10] without compromising the system performance. With large numbers of sessions available for each subject, calibration time can completely be rendered obsolete [8].

In this study, we show that by using fNIRS data from other subjects, we can robustly distinguish between different speaking modes without any calibration data of the current user. Consequently, we do not require multiple sessions per user, but rely on only on a very limited dataset of 5 subjects in total. Furthermore, our strict filtering assures that hemodynamic responses are used for classification while all artifacts are removed. The results achieved in this setup indicate the huge potential of fNIRS for BCIs, which are immediately usable without calibration time.

For this study, we investigated the following speaking modes in classification tasks: Normal audible speech (AUD_{Speech}), silently uttered speech, for which the subjects moved their articulatory muscles as if speaking but not producing any sounds (SIL_{Speech}), and speech imagery (IMG_{Speech}), for which the subjects had to to imagine themselves of speaking, including imagining to move their articulatory muscles.

1.2 Functional Near Infrared Spectroscopy

fNIRS measures the changes in oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR), which are triggered by changes in blood volume due to neural activity in the brain's cortical areas. Using light-sources and detector-optodes, which are fixated to the subjects' heads, these hemodynamic responses can be measured. Light in the near infrared range (620 - 1000 nm) disperses through biological tissue, such as scalp, skull and cortical areas of the brain, but is absorbed by hemoglobin. The modified Beer-Lambert law [12] can be applied to transfer raw

optical densities (ΔOD) into changes in HbO and HbR, denoted as ΔHbO and ΔHbR, respectively:

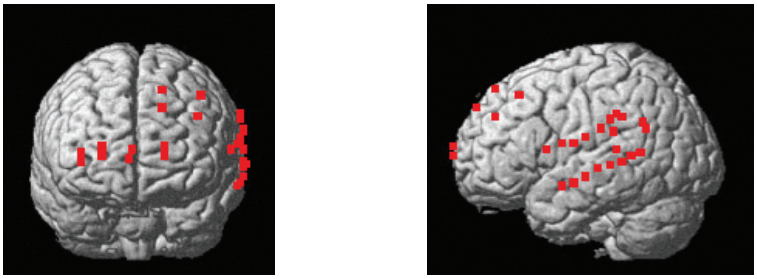
$$\Delta HbO = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbO}} \qquad \Delta HbR = \frac{\Delta OD}{b \cdot l \cdot \alpha_{HbR}} \tag{1}$$

with source-detector distance l, photon path length b and absorption coefficients α_{HbO} and α_{HbR} for HbO and HbR.
A typical hemodynamic response triggered by cortical activity increases on stimulus onset for HbO and decreases for HbR. After the end of the activation, the levels are expected to return to baseline.

2 Experiment

2.1 Setup

To record fNIRS data, we used a Dynot232 system by NIRX Medical Technologies equipped with 32 optodes, sampling at 1.81 Hz. All optodes were used as sources and detectors simultaneously. We used infrared wavelengths of 760 and 830 nm in this study. For every source-detector pair, the system outputs raw optical densities. We limited these to pairs with distances ranging from 2.5 to 4.5 cm, resulting in 252 channels of raw optical densities.
To measure neural activity in the relevant areas, four optodes were placed on Broca's area, 10 on Wernicke's area, both on the left hemisphere. The prefrontal cortex was covered with 12 optodes and six optodes were placed on the lower left motor cortex. Exact optode positions were registered with an ANT Visor infrared camera system[1] and plotted on a brain surface image using the NIRS-SPM software [13]. Figure 1 illustrates exact optode positions in our experiment.



(a) (b)

Fig. 1. (a) Optode positions frontal view. (b) Optode positions left lateral view. Created with [13].

[1] http://www.ant-neuro.com/products/visor/

2.2 Data Acquisition

Five male subjects participated in this study. All of them were right-handed and had a mean age of 27.6 years. Subjects had the 32 NIRS-optodes fixated to their heads by a helmet. Ten sentences in English from the broadcast-news domain were used for the experiment. Only subject 1's mother tongue was English, but all subjects spoke English fluently.
In the experiment, subjects produced utterances in the three modes AUD, SIL and IMG, where each utterance was separated by pauses. Sentences were prompted by displaying them on a screen placed 50cm away from the subjects. Trials are labeled according to the respective modes, i.e. AUD_{Speech}, SIL_{Speech}, IMG_{Speech} and AUD_{Pause}, SIL_{Pause}, IMG_{Pause}. In every mode, each sentence was repeated three times, resulting in a total of 30 trials per mode and per subject. Every utterance of a sentence and every subsequent pause are denoted as separate trials. Two subjects terminated the recordings prematurely resulting in fewer than 30 trials per mode. See Table 1 for full corpus characteristics.
The experimental design is described in more detail in our previous analysis [6].

Table 1. Corpus characteristics

Subject-ID	1	2	3	4	5
Mother tongue	English	German	Sinhala	German	Farsi
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

3 Methods

3.1 Signal Preprocessing

The HomER package[2] was used to transfer the 252 channels of raw optical densities into ΔHbO and ΔHbR values. After linear detrending the channels, trials were extracted based on the experiment time information. Each trial was assigned a class label, which correspond to the *Speech* or *Pause* categories.
Cui et al. [5] showed that NIRS channels containing artifacts can be identified using the correlation between HbO and HbR. Usually, HbO and HbR should be strongly negatively correlated, but motion induced artifacts lead to positive

[2] http://www.nmr.mgh.harvard.edu/PMI/resources/homer/home.htm

correlations, as both values will spike when the optodes are shifted or are lifted off the scalp. To clean the data from artifacts, all channels which were not negatively correlated ($r > -0.3$) for every subject were removed from the dataset. This way, the initial 252 channels were reduced to 60 channels that do not contain artifacts for any of the subjects. Almost all channels on the forehead are removed through this procedure, as they are most vulnerable to movement induced artifacts.

3.2 Feature Extraction

Following Leamy et al. [9], we assume an idealized hemodynamic response for feature extraction. A rise in *HbO* is expected during speech activity and levels should return to baseline for the subsequent *Pause* trials (and vice-versa for *HbR*). To make use of this observation, the mean μ of samples 9 to 15 (corresponding to roughly 4 seconds) is subtracted from the mean of the first 7 samples ($\sim$ 4 seconds) in every trial t for ΔHbO and ΔHbR for every channel i.

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

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

Given this feature extraction, we extract 120 features in total per trial. The features were normalized to zero mean and unit standard deviation (z-normalization).

3.3 Feature Selection

Ang et al. [1] presented the *Mutual Information based Best Individual Feature (MIBIF)* algorithm, a feature selection approach based on a high relevance criterion to reduce the feature space dimensionality. It has proven highly effective for BCI data [1] and is orders of magnitude faster than more complex *Mutual Information* based approaches which try to incorporate redundancy measures [3]. The Mutual Information $I(X;Y)$ can be understood as the amount of information shared by two random variables X and Y. Therefore, a feature containing highly relevant information should have a high Mutual Information with the class labels. *MIBIF* selects the k features with highest Mutual Information with the class labels. Assuming that the training data is representative of the test data, such selected features should increase the classification accuracy.
We set $k = 5$ after studying the distributions of Mutual Information of features with the class labels. See Figure 2 for the distribution of the Mutual Information when selecting features on four subjects for classification on the remaining fifth. Features are sorted decreasingly by their Mutual Information. It can be easily seen that the largest portion of the Mutual Information is explained by the first $k = 5$ features while the remaining 115 contribute only very little information. Selected features were very consistent across the different folds, but varied in between tasks.

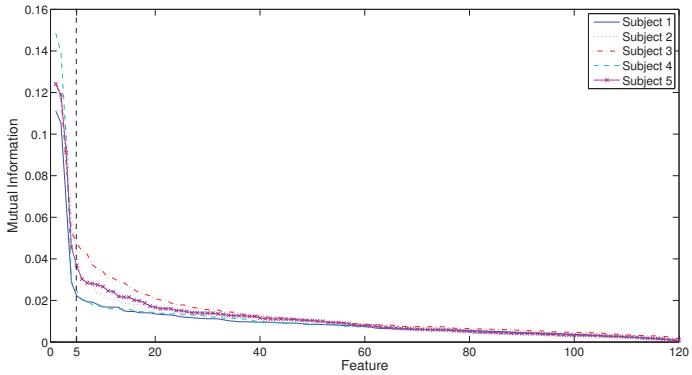


Fig. 2. Mutual Information over number of features for each subject when selecting features on the remaining four subjects for the AUD_{Speech} versus AUD_{Pause} task. The dotted line indicates the five selected features.

3.4 Classification and Evaluation

To evaluate our system, we applied a leave-one-speaker-out cross validation. A Linear Discriminant Analysis (LDA) classifier was trained on the 5-dimensional feature set S, determined with *MIBIF*. The LDA was trained on 4 subjects and tested on the remaining subject in a round-robin manner. Presented results were then averaged over all 5 rounds.
In a first experiment, all three *Speech* modes were combined and tested against all three combined *Pause* modes to discriminate speech activity from inactivity. Subsequently, every mode was classified from its respective *Pause* trials in binary classification experiments. Additionally, the three *Speech* modes were discriminated from each other.

4 Results

All classification results are presented in Figure 3. Differentiating between combined *Speech* (build from AUD_{Speech}, SIL_{Speech}, and IMG_{Speech}) and combined *Pause* worked reasonably well with an average accuracy of 58%. Subsequently, every *Speech* mode was tested individually against its respective *Pause* mode. Audible speech yielded best results with 71% average classification accuracy. This was expected, as neural activity from speech production, speech planning and auditory activity should be observed. Results for silent speech (SIL_{Speech}) are slightly lower (61%), which is explicable by the lack of auditory activity in the fNIRS signals. Discriminating IMG_{Speech} from IMG_{Pause}, when only speech planning activity is present, did not yield results better than chance level. This can be explained by the large variability in speech imagery across subjects, as their might be a lack of a consistent form of imagined speech, even though all speakers were instructed to imagine reading the sentences out loud. Our

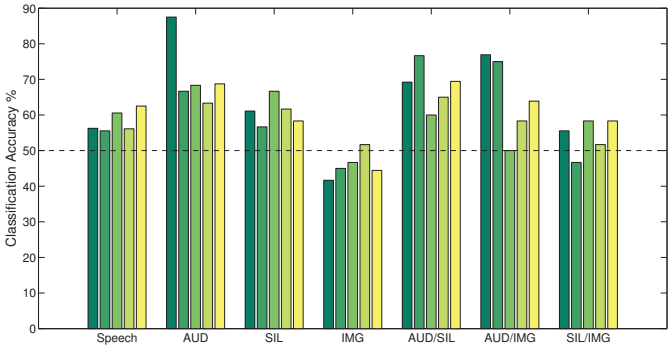


Fig. 3. Classification results for binary classification experiments *Speech* against *Pause* in all modes and between *Speech* of different speaking modes. Each color represents one subject. Dotted line stands for naive classification accuracies.

dataset is small and contains subjects from very different backgrounds (4 different mother tongues), thus the absence of a uniform activation pattern across subjects for speech imagery, for which neither muscle control, nor speech production or acoustic feedback are present is not too surprising.
Differentiating between the different speaking modes worked reliably as well. Classification between AUD_{Speech} and SIL_{Speech} worked best with 68% accuracy. We were able to distinguish between AUD_{Speech} and IMG_{Speech} with 65% accuracy and our setup achieved 55% for SIL_{Speech} versus IMG_{Speech}.
In addition to the classification accuracies, we conducted t-tests to reject the null hypothesis that classification results were equal to naive classification. All experiments, except for IMG_{Speech} versus IMG_{Pause}, were significantly ($p < 0.05$) better than naive classification.
A summary of all classification results can be found in Table 2. These high results, which were achieved with the small dataset of just 5 subjects and which are rigorously artifact cleaned, show that fNIRS has huge potential for cross-subject classification in BCI.

Table 2. Average classification results and standard deviations in %.

	Speech/*Pause*	AUD	SIL	IMG	AUD/SIL	AUD/IMG	SIL/IMG
Accuracy	58	71	61	46	68	65	54
Standard deviations	3.1	9.5	3.8	3.7	6.2	11.3	5.0

5 Conclusion

We have shown that fNIRS signals from speech related tasks produce brain activity that is consistent across multiple subjects. By selecting only the five

most relevant features that are reliable across all subjects, we are able to classify speaking modes solely based on training data from other subjects and thus make user specific training obsolete. Our rigorous filtering for artifacts and the significant results further support the argument that fNIRS signals from speech tasks have huge potential for future BCI applications, as they potentially reduce the amount of training needed in future experiments.

References

1. Ang, K.K., Chin, Z.Y., Zhang, H., Guan, C.: Filter bank common spatial pattern (FBCSP) in brain-computer interface. In: IEEE International Joint Conference on Neural Networks. IJCNN. pp. 2390–2397. IEEE (2008)
2. Ang, K.K., Guan, C., Lee, K., Lee, J.Q., Nioka, S., Chance, B.: A Brain-Computer Interface for mental arithmetic task from single-trial near-infrared spectroscopy brain signals. 20th International Conference on Pattern Recognition pp. 3764–3767 (2010)
3. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE transactions on neural networks pp. 537–550 (1994)
4. Coyle, S.M., Ward, T.E., Markham, C.M.: Brain-computer interface using a simplified functional near-infrared spectroscopy system. Journal of Neural Engineering pp. 219–226 (2007)
5. Cui, X., Bray, S., Reiss, A.L.: Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. NeuroImage pp. 3039–3046 (2010)
6. Herff, C., Putze, F., Heger, D., Guan, C., Schultz, T.: Speaking mode recognition from functional near infrared spectroscopy. In: International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). p. To appear (2012)
7. Krauledat, M., Schröder, M., Blankertz, B., Müller, K.R.: Reducing calibration time for brain-computer interfaces: A clustering approach. Advances in Neural Information Processing Systems pp. 753–760 (2007)
8. Krauledat, M., Tangermann, M., Blankertz, B.: Towards zero training for brain-computer interfacing. PLoS One p. e2967 (2008)
9. Leamy, D.J., Collins, R., Ward, T.: Combining fNIRS and EEG to improve motor cortex activity classification during an imagined movement-based task. In: HCI (20). pp. 177–185 (2011)
10. Lotte, F., Guan, C.: Learning from other subjects helps reducing Brain-Computer Interface calibration time. IEEE International Conference on Acoustics Speech and Signal Processing pp. 614–617 (2010)
11. Naito, M., Michioka, Y., Ozawa, K., Ito, Y., Kiguchi, M., Kanazawa, T.: 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. pp. 1028–1037 (2007)
12. Sassaroli, A., Fantini, S.: Comment on the modified Beer-Lambert law for scattering media. Physics in Medicine and Biology pp. N255–N257 (2004)
13. Ye, J.C., Tak, S., Jang, K.E., Jung, J., Jang, J.: NIRS-SPM : Statistical parametric mapping for near-infrared spectroscopy. NeuroImage pp. 428–447 (2009)

Towards continuous speech recognition for BCI

Christian Herff[1], Adriana de Pesters[2], Dominic Heger[1], Peter Brunner[2,3], Gerwin Schalk[2,3], and Tanja Schultz[1]

[1]Cognitive Systems Lab, University of Bremen (formerly at Karlsruhe Institute of Technology)
Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
christian.herff@uni-bremen.de
http://csl.uni-bremen.de
[2]National Resource Center for Adaptive Neurotechnologies, Wadsworth Center, New York State Department of Health, Albany, USA
[3]Department of Neurology, Albany Medical College, Albany, USA

Abstract. For the last two decades, brain-computer interface (BCI) research has worked towards practical and useful applications for communication and control. Yet, many BCI communication approaches suffer from unnatural interaction or time-consuming user training. As continuous speech provides a very natural communication approach, it has been a long standing question whether it is possible to develop BCIs that perform speech recognition from cortical activity. Imagined speech as a BCI paradigm for locked-in patients would mean a large improvement in communication speed and usability without the need for cumbersome spelling using individual letters.
We showed for the first time that automatic speech recognition from neural signals is possible. Here, we evaluate the feasibility of speech recognition from neural signals using only temporal offsets associated with speech production and omitting information from speech perception. This analysis provides first insights into the potential usage of imagined speech processes for speech recognition, for which no perceptive activity is present.

Keywords: speech, BCI, Automatic Speech Recognition, ASR, Brain-Computer Interface

1 Introduction

Previous neuroscientific studies provided evidence for neural representations of speech, such as phones and phonetic features during speech perception [3, 12, 9]. Other studies classified [10, 1, 8] or investigated the production [18, 4] of limited sets of phones, syllables, and words. A complete set of manually labeled phones was classified in single word production in [13]. However, it was unclear whether the brain encodes a complete repertoire of phonetic representations during the production of continuous speech that allows the decoding of words and phrases.

In a study with 7 participants [6], we showed for the first time that continuously spoken speech is represented in the brain as a sequence of phones.

These phones can be decoded from electrocorticographic (ECoG) recordings and allow the composition of the spoken words, which we call *Brain-to-Text*. All participants were undergoing surgery for intractable epilepsy and agreed to participate in our experiment. Electrode locations were determined based solely on clinical needs of the patients. We used electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) with inter-electrode distances of $0.6 - 1$ cm. BCI2000 [16] was used to record ECoG signals from eight 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria).

In our experiment, we recorded ECoG activity and the acoustic waveform simultaneously, while participants read aloud different texts consisting of childrens' literature, fan fiction or political speeches. We time-aligned the neural data to a phone labeling obtained from the acoustic data using our in-house speech recognition toolkit BioKIT [17]. This allowed us to identify the neural activity corresponding to the production of each phone. See Figure 1 for data recording in our experiment and aligning of ECoG and acoustic data. We segmented the texts into phrases and used the recorded ECoG data of all but one phrase for feature selection and training, then evaluated our approach on the ECoG data of the remaining phrase in a round-robin manner (leave-one-phrase-out validation). We compared the results from temporal offsets associated with speech production to productive and perceptive temporal offsets to analyze the feasibility of continuous speech recognition from imagined speech processes, as perceptive activity is only present when participants hear their own voice.

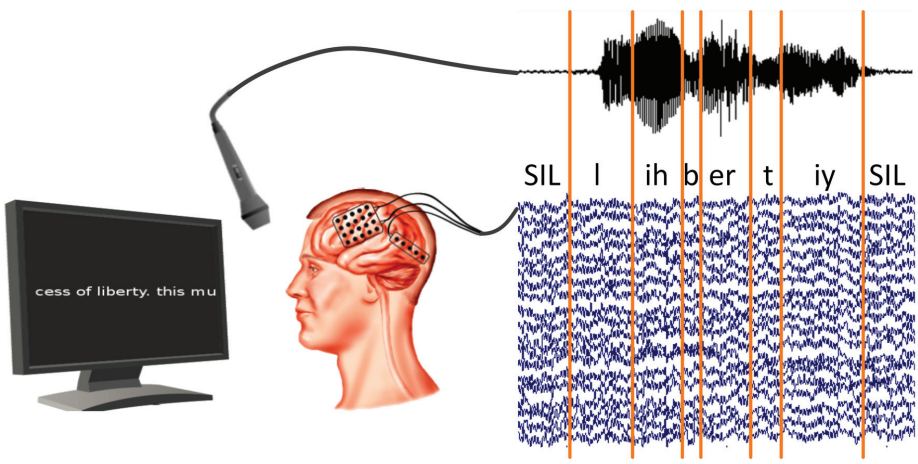


Fig. 1. Synchronized data recording of ECoG data and the acoustic stream.

2 Phone modeling in ECoG

To model phones in ECoG data, we extracted broadband-gamma (70-170 Hz) activity in 50 ms windows for each channel. The temporal dynamics of speech production were captured by including the features of the four neighboring windows before and after each window in the feature vector, i.e. representing a context of 450 ms length. We modeled each phone with a multivariate Gaussian distribution representing the mean broadband-gamma activity and the corresponding variance for all locations and time lags. We analyzed the discriminability between the different phone models by employing their Kullback-Leibler divergences (KL-div) for every electrode position and time lag. The spatio-temporal distributions of KL-div results give interesting insights into the spatio-temporal dynamics of cortical activity during continuously spoken speech. Figure 2 illustrates discriminability between phones for cortical locations and time offsets on a combined electrode montage of all participants. Phone discriminability can be observed 200 ms prior to phone production in prefrontal areas associated with speech planning (Broca's area). 100 ms prior to phone production, discriminability increases in motor areas and auditory cortex and vanishes in previously observed regions. At phone onset, discriminability peaks in motor cortex, while discriminability is largest in auditory cortex 100 ms after phone production. 200 ms after phone production, phone models can be discriminated in auditory cortex. The activations after the actual phone production are presumably triggered by the participants' perception of their own voice.

We also use the KL-div values to automatically select the best ECoG features for our *Brain-To-Text* system.

To evaluate the feasibility of our system for realistic brain-computer interfaces based on imagined speech production, we performed an analysis that focuses on activity prior to phone onset. By only keeping the temporal offsets between -200 ms to 0 ms (see Figure 2), no perceptive activity from hearing one's own voice should remain in the data. This restriction to productive temporal offsets is a first simulation of imagined speech, in which no perceptive activity is present, as participants do not hear their own voice. We refer to these results as *Production* and compare them to those obtained with all temporal offsets, refereed to as *Production & Perception*. This analysis therefore provides a first insight into the feasibility of our system for imagined speech.

3 Automatic Speech Recognition for BCI

We combined the phone-based speech representations of cortical activity with language information using automatic speech recognition technology to reconstruct the words in unseen spoken phrases. Language information is included into the decoding process through a language model and a pronunciation dictionary. The pronunciation dictionary contains the mapping of phone sequences to words. The language model statistically models syntactic and semantic information by predicting the next words given the preceding words [7].

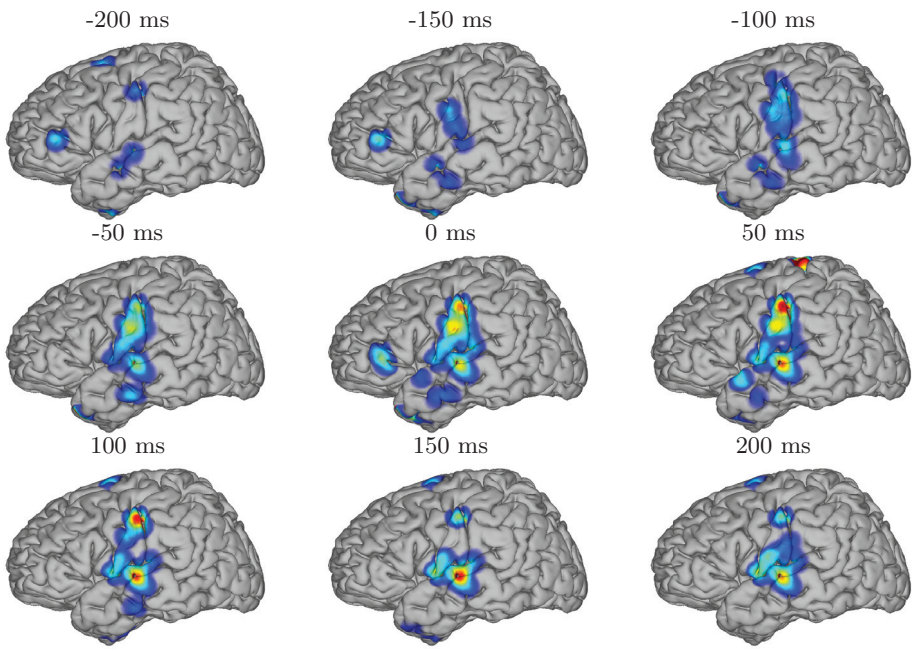


Fig. 2. Discriminability (Mean Kullback-Leibler Divergences) between phones for electrode position of all participants. Color overlays on the rendered average brain show regions of high discriminability (red) to lower discriminability (blue), all overlays are larger than random discriminability. Early differences can be observed in diverse areas up to 200 ms before phone production. Sensorimotor cortex shows high discriminability 50 ms before production, while discriminability in auditory regions of the superior temporal gyrus peaks 100 ms after production.

Our results show that, with a limited set of words in the dictionary, *Brain-to-Text* is able to reconstruct full sentences. Figure 3 illustrates the different steps of decoding continuously spoken phrases from neural data. *ECoG signals over time* are recorded at every electrode and divided into 50ms segments. For each 50 ms interval of recorded *broadband gamma activity*, stacked feature vectors are calculated (*Signal processing*). For each *ECoG phone model* calculated on the training data, the likelihood that this model emitted a segment of ECoG features can be calculated, resulting in *phone likelihoods over time*. Combining these Gaussian *ECoG phone models* with language information in the form of a *dictionary* and an n-gram *language model*, the *Viterbi* algorithm calculates the *most likely word sequence* and corresponding *phone sequence*. To visualize the decoding path, the *most likely phone sequence* can be shown in the *phone likelihoods over time* (red marked areas). The system outputs the decoded word sequence. Once the ECoG phone models are trained, phrases can be decoded in real-time.

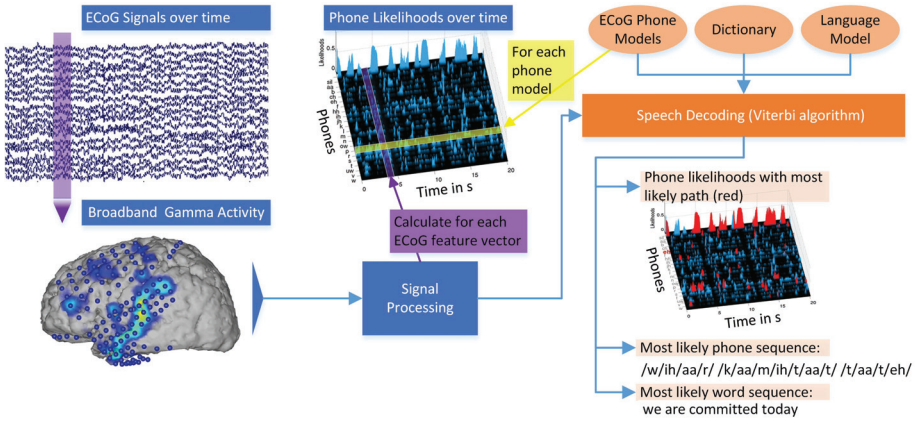


Fig. 3. Overview of the *Brain-to-Text* decoding process

4 Results

To evaluate the performance of *Brain-to-Text*, we compared the decoding results of our approach to randomized models (randomization test by shifting the labels of the training data by half the session length). The randomized results illustrate the impact that the language model and dictionary have when no usable neural information is present. Figure 4 shows phone classification accuracies for all participants and sessions. Classification accuracies for combined productive and perceptive areas (purple bars) are better than accuracies achieved with randomized models (yellow bars) for all sessions of all participants. To estimate how well a hypothetical device based on imagined speech production might be, we evaluated our approach only based on productive areas, by excluding all activations from time offsets after phone onset. As the participants cannot hear their own voice prior to the onset of the phone, this ensures that no perceptive activity should be used in this evaluation. Results on productive areas only (turquoise bars) outperform the randomized models for all sessions, but are usually worse than accuracies achieved when using all neural activity.

As *Brain-to-Text* outputs word sequences, we evaluated the Word Error Rate between our predicted word sequence and the reference phrase. One of the major limitations in our study is the small amount of training data per session, with only a few minutes of data. For comparison, speech recognition systems based on acoustic speech are usually trained on thousands of hours of data. To account for the limited amount of training data, we restricted the amount of recognizable words in the dictionary to a range of 10 to 100 words. We were able to achieve Word Error Rates as low as 25% when using a dictionary of 10 words. Word Error Rates depending on dictionary size for the best performing participant are shown in Figure 5. Word Error Rates are lowest (between 25% and just over 60%) when using perceptive and productive (purple line) time offsets. Neural activity only resulting from speech production yields slightly higher Word Error

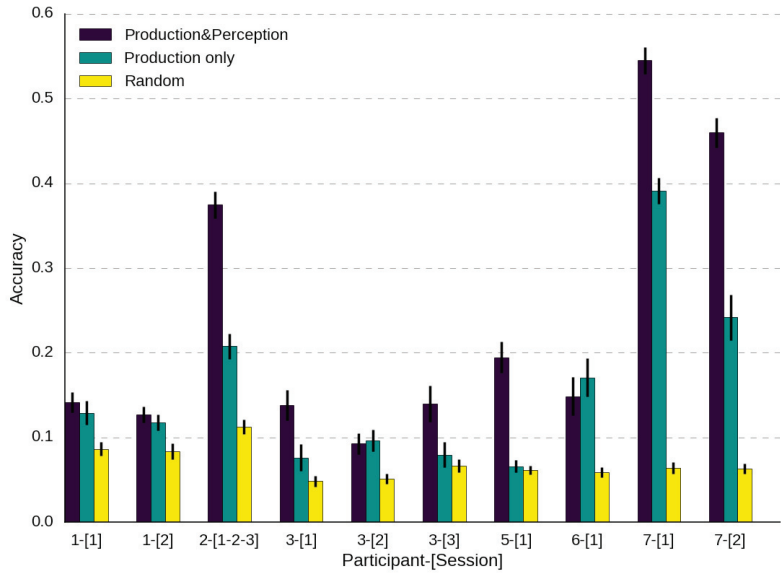


Fig. 4. Phone classification accuracies for all participants and sessions. Error bars depict standard errors. Our system shows significantly better accuracies than random models (yellow bars) when using all information (purple bars) and when only using productive temporal offsets (turquoise bars).

Rates (turquoise line) than perceptive & productive activity, but still outperforms randomized models (yellow line) for all dictionary sizes. Using productive activity only, more than 60% of words are recognized correctly for a dictionary of 10 words.

To ensure that word recognition is not based on the robust recognition of a small subset of phones, we also analyzed average phone true positive rates. For this analysis, we obtained the ground truth of phone timings from the audio alignment described earlier. Bars in Figure 5 show true positive rates averaged across all phones on window-level. Again, productive and perceptive time offsets (purple) combined yield the best results, but using only productive neural activity (turquoise) still yields high average phone true positive rates above 20%. Both systems using neural activity outperform random true positive rates (yellow). Average phone true positive rates remain rather stable even when dictionary sizes increase.

Even though detailed results are only shown for the participant which gave the best recognition results, we found significantly better results than random models in Word Error Rate and single phone true positive rates for all sessions in this study.

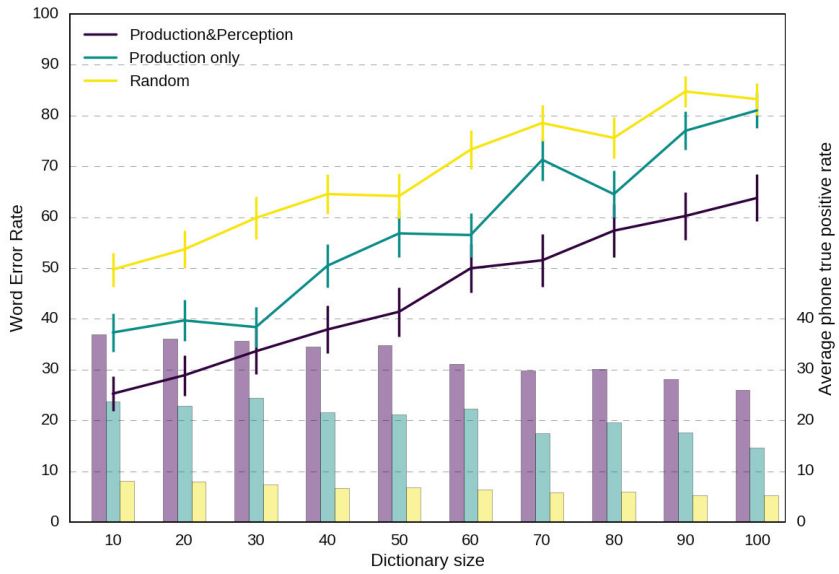


Fig. 5. Word Error Rates over dictionary size (lines); average true positive rates across phones depending on dictionary size (bars). Error bars depict standard errors. While the full set of temporal offsets performs best (purple), information from productive time offsets (turquoise) still outperforms random models (yellow) for all dictionary sizes both in Word Error Rates and true positive rates.

5 Conclusion

In summary, our results support the hypothesis that *Brain-to-Text* may eventually allow people to communicate using brain signals associated with continuous spoken language, i.e. without the current limitations of a restricted set of commands or unnatural selection procedures. We showed that participants' neural activity could be used to decode continuously spoken phrases into a textual representation, even when omitting neural activity associated with the perception of their own voice. This illustrates the feasibility of speech recognition from neural activity when participants only imagine to speak. Thus, using continuous speech production for BCIs has the potential to increase naturalness and information transfer rates and the practical utility of current BCI communication approaches. Ultimately, speech processes for BCIs might lead to information transfer rates similar to that of continuous speech while being more natural to the user.

While the generative models used in this study allow for a good illustration and fast training of phone models, we have shown that more advanced discriminative models can improve results [5].

Recent advances in the modeling of imagined phones [2], reconstruction of imagined speech spectra [11] and investigations in silent reading [14, 15], suggest

that covert and overt speech share a neural substrate. Our presented results suggest that neural activity from productive temporal offsets allows reconstruction of a textual form, without the need for perceptive information. These findings highlight the potential of *Brain-to-Text* to be used on imagined continuous speech in the future.

References

1. Timothy Blakely, Kai J Miller, Rajesh PN Rao, Mark D Holmes, and Jeffrey G Ojemann. Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4964–4967. IEEE, 2008.
2. Jonathan S Brumberg, E Joe Wright, Dinal S Andreasen, Frank H Guenther, and Philip R Kennedy. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in neuroscience*, 5, 2011.
3. Edward F Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432, 2010.
4. Miho Fukuda, Robert Rothermel, Csaba Juhász, Masaaki Nishida, Sandeep Sood, and Eishi Asano. Cortical gamma-oscillations modulated by listening and overt repetition of phonemes. *Neuroimage*, 49(3):2735–2745, 2010.
5. Dominic Heger, Christian Herff, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Continuous speech recognition from ecog. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
6. Christian Herff, Dominic Heger, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9(217), 2015.
7. Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.
8. Spencer Kellis, Kai Miller, Kyle Thomson, Richard Brown, Paul House, and Bradley Greger. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7(5):056007, 2010.
9. Jan Kubanek, Peter Brunner, Aysegul Gunduz, David Poeppel, and Gerwin Schalk. The tracking of speech envelope in the human cortex. *PloS one*, 8(1):e53398, 2013.
10. Eric C Leuthardt, Charles Gaona, Mohit Sharma, Nicholas Szrama, Jarod Roland, Zac Freudenberg, Jamie Solis, Jonathan Breshears, and Gerwin Schalk. Using the electrocorticographic speech network to control a brain–computer interface in humans. *Journal of neural engineering*, 8(3):036004, 2011.
11. Stephanie Martin, Peter Brunner, Christopher Holdgraf, Hans-Jochen Heinze, Nathan Earl Crone, Jochem Rieger, Gerwin Schalk, Robert Thomas Knight, and Brian Pasley. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering*, 7(14), 2014.
12. Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, page 1245994, 2014.
13. Emily M Mugler, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele, Joshua Rosenow, Jerry J Shih, Dean J Krusienski, and Marc W Slutzky. Direct classification of all american english phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, 11(3):035015, 2014.

14. Marcela Perrone-Bertolotti, Jan Kujala, Juan R Vidal, Carlos M Hamame, Tomas Ossandon, Olivier Bertrand, Lorella Minotti, Philippe Kahane, Karim Jerbi, and Jean-Philippe Lachaux. How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. *The Journal of Neuroscience*, 32(49):17554–17562, 2012.
15. Christopher I Petkov and Pascal Belin. Silent reading: Does the brain hearboth speech and voices? *Current Biology*, 23(4):R155–R156, 2013.
16. Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *Biomedical Engineering, IEEE Transactions on*, 51(6):1034–1043, 2004.
17. Dominic Telaar, Michael Wand, Dirk Gehrig, Felix Putze, Christoph Amma, Dominic Heger, Ngoc Thang Vu, Mark Erhardt, Tim Schlippe, Matthias Janke, et al. BioKIT - real-time decoder for biosignal processing. In *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*, 2014.
18. Vernon L Towle, Hyun-Ah Yoon, Michael Castelle, J Christopher Edgar, Nadia M Biassou, David M Frim, Jean-Paul Spire, and Michael H Kohrman. Ecog gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain*, 131(8):2013–2027, 2008.