CARL VON OSSIETZKY universität OLDENBURG

Speech-related brain responses as a basis for auditory brain-computer interfaces.

Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines Doktors der Naturwissenschaften (Dr. rer. nat.) der Physik angenommene Dissertation.

von Herrn Carlos Filipe da Silva Souto
geboren am 29.06.1984 in Porto

Gutachter: Prof. Dr. Dr. Kollmeier
Prof. Dr. Verhey

Tag der Disputation: 21. Oktober 2019

Abstract

Hearing aid users have strong difficulties to follow a conversation in a complex acoustic situation with several competing speakers inside a reverberant room with background noise. Although hearing aid algorithms are able to segregate different sources and mask noise, they possess no information of the users' intention - especially when the posture of the head of the user does not provide usable information (e.g. several competing speakers are standing close to each other). In the future, auditory-based Brain Computer Interfaces (BCI) could provide a possible solution by controlling hearing aid functions, like detecting the speaker of interest and pointing a spatial filter to the respective direction or selecting a specific noise filter according to the ability of the listener to understand the attended speaker. Such an auditory BCI approach would optimally rely on brain responses related to features of speech (promising features are rhythm, direction and context).

The current thesis aims at a related auditory BCI paradigm, able to decode the subjects' intention, using natural speech sounds in an everyday life scenario. The goal is to characterize the relation between the features of the presented speech sounds and the respective evoked EEG responses in an attended or non-attended case as the basis for a simple and computational efficient automatic classifier. In the first part of the thesis, a complex acoustic situation with competing speakers that talk with different speech rhythms was simulated to investigate the effect of spatial attention on natural speech stimuli. Two 20 s long competing streams of spoken syllables in rates of 2.3 (female) and 3.1 Hz (male) were presented from two directions simultaneously, while the auditory BCI approach employed the auditory steady-state response (ASSR) to automatically detect which stream a listener selectively attends to. The second part of the thesis investigated the envelope of spoken sentences (reflecting the speech-rhythm) as a feature for a BCI. A computational simple classification approach was tested based on correlation of envelope and evoked single-trial EEG to segregate different speakers and sentences. This was done using a short analysis window of 1.6 s, which is fairly corresponding to the duration of a short spoken sentence. To provide a first test for the robustness of the BCI against noise as well as to obtain some insight into which features of the stimuli are most important for classification additional test conditions were added. Therefore, a speech-in-noise stimulus with a flat envelope and a non-speech stimulus, mimicking the envelope and the overall spectrum of one sentence, were segregated from either a single sentence spoken from a male or female speaker in quiet. The third part of the thesis used the signal processing and classification methods of the BCI approach presented in the second part of the thesis to test its ability to generalize to speech in noise and further to observe a possible relation between its performance and the human speech intelligibility. Therefore, the performance of this approach was tested in a speech-in-babble-noise scenario classifying two different sentences at three different signal-to-noise ratios using an analysis window of 1.2 s (according to length of shortest used sentence), while the subjects performed a concurrent psychoacoustic attentive task.

Speech rhythm proved to be a feasible feature for an auditory based classification approach. It was possible to classify ASSR evoked from streams of spoken syllables and further to segregate sentences or speakers by classifying correlation between speech envelopes and evoked single-trial EEG, even under varying levels of speech-simulating noise and the use of short analysis windows. The performance of the correlation-based BCI approaches are comparable to recent published BCI approaches. Further, a relation between the BCIs performance and the human speech intelligibility was found. The computational simple auditory BCI approaches shown in this thesis point towards the development of a future BCI-controlled hearing aid.

Zusammenfassung

In einer komplexen akustischen Situation mit mehreren konkurrierenden Sprechern, die sich in einem hallenden Raum mit Hintergrundgeräuschen befinden, haben Hörgeräteträger große Schwierigkeiten ein Gespräch zu verfolgen. Obwohl Hörgeräte-Algorithmen in der Lage sind verschiedene Schallquellen zu trennen und Störgeräusche zu unterdrücken, besitzen sie jedoch keine Informationen über die Absicht des Trägers. Dies gilt insbesondere, wenn die Position des Kopfes des Hörgeräteträgers keine brauchbare Information liefert (z.B. mehrere konkurrierende Sprecher befinden sich nahe beieinander). In der Zukunft könnte die Steuerung von Hörgerätefunktionen durch auditorisch basierte Gehirn Computer Steuerungen (engl. Brain Computer Interfaces; BCI) eine mögliche Lösung für dieses Problem darstellen. Das BCI könnte z.B. einen attendierten Sprecher detektieren und einen räumlichen Filter in die entsprechende Richtung richten oder die Wahl eines spezifischen Rauschfilters treffen, entsprechend der Fähigkeit des Trägers den attendierten Sprecher zu verstehen. Ein solcher BCI-Ansatz würde optimaler Weise Gehirnantworten, die durch Sprachmerkmale evoziert wurden nutzen (vielversprechende Merkmale hierbei sind Rhythmus, Richtung und Kontext).

Ziel der vorliegenden Arbeit ist es auf ein geeignetes auditorisches BCI-Paradigma hinzuarbeiten, welches natürliche Sprachlaute verwendet, um die Intention des Probanden zu entschlüsseln. Zu diesem Zweck soll der Zusammenhang zwischen den Merkmalen der präsentierten Sprachlaute und der entsprechend evozierten EEG-Antworten charakterisiert werden - in einem attendierten oder nicht-attendierten Fall - und dies als Grundlage für eine simple und recheneffiziente automatische Klassifikation genutzt werden. Im ersten Teil der Arbeit wurde eine komplexe akustische Situation mit konkurrierenden Sprechern - die mit unterschiedlichen Rhythmen sprechen - simuliert, um die Wirkung der räumlichen Aufmerksamkeit auf natürliche Sprachreize zu untersuchen. Zwei 20 s lange konkurrierende Folgen von gesprochenen Silben wurden in Raten von 2,3 (weiblicher Sprecher) und 3,1 Hz (männlicher Sprecher) gleichzeitig aus zwei Richtungen präsentiert, während der BCI-Ansatz die stationäre neuronale Dauerantworten (engl. steady-state response; ASSR) verwendete, um die von einem Probanden attendierte Silbenfolge automatisch zu erkennen. Der zweite Teil der Arbeit testet die Einhüllenden von gesprochenen Sätzen - welche den Sprachrhythmus widerspiegeln - als Grundlage für ein BCI. Ein einfacher Klassifikationsansatz mit geringem Rechenaufwand konnte verschiedene Sprecher und Sätze trennen, durch die Klassifikation der Korrelation von Spracheinhüllenden und einzelner Epochen des evozierten EEGs. Es wurde dafür ein kurzes Analysefenster von 1,6 s verwendet, welches in etwa der Dauer eines kurzen gesprochenen Satzes entspricht. Um die Robustheit des BCIs gegenüber Störgeräuschen erstmals zu testen und einen Einblick darüber zu erhalten, welche Merkmale der präsentierten Reize für die Klassifizierung am wichtigsten sind, wurde ein von Störgeräuschen maskierter Sprachreiz mit flacher Einhüllenden und ein Nicht-Sprachreiz (dieser künstliche Reiz ahmt die Einhüllende und das Gesamtspektrum eines Satzes nach) eingeführt. Das BCI sollte diese von einem einzelnen Satz trennen, der entweder von einem männlichen oder weiblichen

Sprecher in Ruhe gesprochen wurde. In dem dritten Teil der Arbeit werden die Signalverarbeitungs- und Klassifizierungsmethoden des zweiten Teils dieser Arbeit verwendet, um die Anpassungsfähigkeit des BCIs an eine Störgeräuschsituation und weiter eine mögliche Abhängigkeit der Leistung des BCIs von der menschlichen Sprachverständlichkeit zu untersuchen. Aus diesem Grund wurde die Leistung dieses BCI-Ansatzes in einem Geplapper Störgeräusch-Szenario getestet, während die Probanden gleichzeitig eine psychoakustische Aufmerksamkeitsaufgabe durchführten. Das BCI konnte unter Verwendung eines Analysefensters von 1,2 s Länge (entsprechend der Länge des kürzesten verwendeten Satzes) in drei verschiedenen Signal-zu-Rausch-Verhältnis-Konditionen zwei unterschiedliche Sätze trennen.

Der Rhythmus von Sprache erwies sich als praktikables Merkmal für einen auditorisch basierten Klassifikationsansatz. Es war möglich ASSR auf gesprochene Silbenfolgen zu klassifizieren und weiter durch Klassifikation von Korrelation zwischen Spracheinhüllenden und einzelnen Epochen des evozierten EEGs Sätze oder Sprecher zu trennen, selbst unter Einfluss von unterschiedlich starkem sprachsimulierenden Störgeräuschen und der Verwendung von kurzen Analysefenstern. Die Leistung der korrelationsbasierten BCI-Ansätze ist mit anderen kürzlich veröffentlichten BCI-Ansätzen vergleichbar. Weiterhin wurde ein Zusammenhang zwischen der Leistung der BCIs und der menschlichen Sprachverständlichkeit festgestellt. Die in dieser Arbeit vorgestellten simplen auditorischen BCI-Ansätze mit geringem Rechenaufwand bieten einige vielversprechende erste Ansätze in Hinblick auf eine zukünftige BCI Steuerung von Hörgeräten.

Content

1 General Introduction

Brain Computer Interfaces (BCI) detect and classify specific task-related neurophysiological changes to control electronic devices without the need of manual interaction (Vidal, 1973; Brunner et al., 2011). For example, the use of BCI allow communication with severely paralyzed people (Birbaumer et al., 1999; Kübler et al., 2001), BCI controlled prostheses continuously improve the quality of life of physically disabled people (Birbaumer, 2006). Electroencephalography (EEG) is most commonly employed for this purpose, due to its advantages to be noninvasive, highly temporal accurate, cost efficient and the possibility to be portable. Chapter 2 describes EEG and the classification of its respective Data in more detail. The majority of BCI controlled systems are based on visually evoked neuronal responses or use neuronal changes that occur when a specific movement is imagined (e.g. Blankertz et al. 2010). Since 2006 more and more Auditory-based BCI approaches are investigated (e.g., Halder et al., 2010; Kim et al., 2011). These approaches do not rely on the visual abilities or motor skills of the user and therefore provide an additional alternative for visually disabled and severely paralyzed people (Sellers and Donchin, 2006; Halder et al., 2010; Kim et al., 2011). For example, Klobassa et al. (2009) expanded the established visual P300 spelling paradigm of Farwell and Donchin (1988) and Donchin et al. (2000), i.e. a BCI paradigm that makes it possible to spell words or sentences by detecting visual P300 responses of the user to a flashing row or column containing a specific symbol of interest in a six by six matrix with letters and numbers. They presented six environmental sounds in a 6 x 6 P300-Speller. One subject group obtained an adequate visual stimulation. The classification of P300 responses from both groups showed equivalent results. The current thesis aims to test the possibility of a related auditory BCI paradigm based on natural speech sounds that further should be usable in everyday life scenario. The objective is to characterize the relation between the acoustic input signal to the brain (i.e. speech and its most relevant features) and the EEG response that can be recorded in the attended or non-attended case as the basis for an automatic classifier to decode the subjects' attention and intention.

One possible future application field for such auditory-based BCIs could be the control of hearing aids. The target scenario for such an auditory BCI is a difficult acoustic situation, like several speakers in a scenario with background noise and reverberation. The aim is to detect the acoustical object of interest and pointing a spatial filter to the respective direction or selecting of a specific noise filter according to the ability of the listener to understand the acoustic signal of interest. An auditory BCI approach would optimally rely on the potentials that are evoked by signals that naturally accrue in such a situation. Hence, it is necessary to investigate the feasibility of speech-evoked potentials for classification. For this purpose, three promising speech features are (a) meaning or context, (b) direction and (c) rhythm. Since 2013 several auditory BCI approaches were investigated focusing on those speech features (e.g., Nakamura et al., 2013; Hill et al., 2014). Hill et al. (2014) investigated event related potentials to synthesized words as a feature for their BCI approach. In an oddball-paradigm they presented the words "yes" (standard) and "yep" (deviant) from the

left side and “no” (standard) and “nope” (deviant) from the right side. The participants were asked to answer yes- and no-questions by shifting their attention to the stimuli with the respective meaning. The BCI was able to detect the direction of interest out of two with an average accuracy of 77%, i.e. 27% above chance. Other BCI approaches, like the one of Nakamura et al. (2013), are based on classifying auditory steady-state responses (ASSR; Picton et al., 1987) to speech that is artificially modulated, i.e. providing a specific rhythm. Nakamura et al. (2013) classified ASSR to amplitude modulated synthesized sentences presented from two directions with comparable accuracy to the one by Hill et al. (2014). Further studies like Kim et al. (2011) classified ASSR to continuous trains of tone bursts. Kim et al. (2011) presented two trains of pure tone bursts with different presentation rates (37 and 43 Hz) from two different directions simultaneously (left and right) and were able to detect the spatial attention of listener.

In the current thesis the use of BCI to segregate real speech stimuli using simple and computational efficient classifiers will be explored as a prerequisite for future BCI controlled hearing devices, like hearing aids. Most of the BCI approaches so far (Hill et al., 2014) utilized artificial stimuli. As a first step towards real speech, in Chapter 3 the effect of spatial attention on natural speech stimuli is investigated. A complex acoustic situation with competing speakers that talk with different speech-rhythms was simulated presenting two 20 s long streams of spoken syllables in rates of 2.3 (female) and 3.1 Hz (male) from two directions simultaneously. It is aimed to measure and classify the ASSR to both steams and automatically detect the spatial attention of the listener.

One step beyond using a fixed rhythm as needed for ASSR paradigms is to follow the stimulus envelope, which is reflecting the speech-rhythm as a feature for a BCI. Such an approach based on spoken sentences is evaluated in Chapter 4. To provide a rather stable rhythm, the structure of the Oldenburg sentence test (OLSA) was chosen here, which is a well-established psychoacoustic matrix test in German language (Wagener et al., 1999, Kollmeier et al., 2015). This approach is generally in line with more recent auditory BCI paradigms like the ones of O’Sullivan et al. (2014), Ekin et al. (2016) and Biesmans et al. (2017) that use a stimulus-reconstruction method (Rieke et al., 1995; Mesgarani et al., 2009) to estimate the low frequency components of the attended speech envelope from the evoked EEG recordings. All these approaches base essentially on the findings by Aiken and Picton (2008), who discovered that the envelope components below 7 Hz of speech are represented in the evoked EEG of the auditory cortex. O’Sullivan et al. (2014) presented two competing spoken stories from two different talkers and directions simultaneously. Utilizing the correlation of the estimated envelope of the attended speech with the actual envelopes of both speech signals, they were able to detect which story a listener was selectively attending to. Ekin et al. (2016) extended this approach by the addition of a reconstruction filter for the unattended story, which achieved a comparable performance. Biesmans et al. (2017) further enhanced the performance of the BCI approach of O’Sullivan et al. (2014) by solving a single least squares estimation of the attended speech envelope over the entire training data set, decreasing the length of the analysis window and optimizing the stimulus

envelope extraction using a simple model based on a combination of power law relation (loudness model) and gammatone filter bank.

The performance of a BCI depends on the accuracy and the speed of its classifier and can be compared by the bits per class that are generated after the duration of one minute (Information transfer-rate (ITR); Shannon and Weaver, 1964; Besserve et al., 2007; Speier et al., 2013). Although the stimulus-reconstruction method - as described by O'Sullivan et al. (2014), Ekin et al. (2016) and Biesmans et al. (2017) - exhibits comparatively high classification accuracies, it depends on a rather long analysis window of about 30 to 60 s. The ITR of a BCI with an analysis window of 30 to 60 s for the systems mentioned so far range between 0.29 and 0.62 bits/min. This would be too low and its latency too high for the controlling of hearing aids in a natural environment. Furthermore, the limited data processing power of current processors intended for mobile usage (e.g., for smart phones) could be a problem when loading them with state-of-the-art adaptive filter methods or high computational neural network approaches (Kottaimalai et al., 2013). The major objective of Chapter 4 is therefore the better applicability for real-time applications and increase (or at least preservation) in ITR of a speech-driven BCI by exploring (a) a sufficiently short analysis windows (here 1.6 s), providing sufficient low latencies to be useful for hearing aid control and (b) the use of a simple and computational efficient classifier (without extensive computational methods like adaptive filters). The classifier suggested here exploits the correlations between all possible speech envelopes and their corresponding single-trial EEG responses. The presented approach resulted from an optimization search across several approaches tested in respective pilot studies, that have been partially rejected or specifically adapted (see General appendix 8 for an overview).

In Chapter 4 further specific test stimuli were included into the classification to test the feasibility and robustness of the BCI as well as to obtain some insight into which features of the stimuli are most important for classification. The stimuli were (a) a speech-in-noise condition with a flattened stimulus envelope while providing clearly intelligible speech and (b) a non-speech stimulus mimicking the envelope and the overall spectrum of one sentence spoken by a male speaker (amplitude modulated tone complex). The underlying research questions were: (1) - robustness of the BCI against noise - If the original speech envelope is represented in the EEG recording, the classifier (trained with the clear, only slightly modulated envelope) should be able to segregate the underlying sentence, from any other presented sentence with an accuracy comparable to the conditions when both sentences are presented as clean speech. (2) - most important stimuli feature for classification - Does the BCI accuracy depend on the specific characteristics of the respective stimulus envelope, or does the stimulus need to be perceived as intelligible speech explicitly? Therefore, - in condition b) - a non-speech stimulus (mimicking the envelope of a spoken sentence) was segregated from an arbitrary test sentence and its resulting classification accuracy was compared to the accuracy resulted by segregating the respective arbitrary sentence from the original sentence (the non-speech stimulus is mimicking). If the accuracies are comparable, the envelope of the

stimulus would have to be considered as being more important for the classifier then the property to be perceived as intelligible speech.

In Chapter 5 the performance of the BCI paradigm (signal processing and classification methods) described in the Chapter 4 is investigated using variable levels of background speech-noise. This provides a rather realistic test scenario for an auditory BCI approach. EEG and psychoacoustic individual speech intelligibility of the received speech are measured in parallel to investigate the relation between the BCIs performance and the actual speech intelligibility of the user. For this purpose, original speech material from the Oldenburg children sentence test (OLKISA; Wagener and Kollmeier, 2005) is used. This shortened version of the OLSA (only three words instead of five per sentence) is chosen to keep the overall measurement duration for each subject in an acceptable range within a single measurement session, while providing a sufficiently large amount of epochs for classification. As distortion stimulus the stationary Olnoise was used, which is the specific speech-noise of the OLKISA and OLSA test. It provides the same long-term spectrum as the speech material to simulate a realistic scenario for different SNR. Two major questions are asked in Chapter 5. First, to which extent does the BCI generalize to speech in noise during an attentive task when trained on clean speech only? Second, does the BCI still provide a performance above chance level at SNRs at which human subjects already show clearly reduced speech intelligibility? This would be a necessity to use the BCI for controlling a hearing aid, e.g., to select a noise reduction software appropriate for the acoustical situation or even to estimate the current speech reception level of the listener. For this reason, the performance of the simple correlation based auditory BCI approach of Chapter 4 was tested in a speech-in-speech-noise scenario, at three different SNR conditions, classifying two different sentences within a short time window of just 1.2 s duration (according to the shortest used OLKISA sentence).

Chapter 6 gives a general summary and discussion of the three studies presented in Chapter 3-5 and sets them into context of actual BCI research. Finally, in Chapter 7 the major outcomes of this thesis are concluded and the importance for future BCI applications as well as ideas for possible resulting follow-up projects are discussed.

2 EEG Overview

Electroencephalography (EEG) is a noninvasive and cost-saving electrophysiological method to measure neuronal activity. Movements of ions through the membrane of active neurons generate dipoles. The superposition of several dipoles generates an electrical field on the surface of the scalp. The potential changes of this electrical field have a magnitude of µV and can be measured using at least two electrodes. It is possible to measure the neuronal activity with high temporal accuracy of milliseconds. The disadvantages of EEG are a low spatial resolution, a low signal to noise ratio (SNR) and its vulnerability to artifacts. In this Chapter a general overview of EEG is given, providing some insight to the physiological background, the features to acoustic stimulation, a suitable classification method, often occurring problems and respective signal processing solutions.

2.1 Physiological Model

With EEG it is possible to measure the neuronal activity of the brain, in the form of potential changes of the electrical field on the surface of the scalp, using at least two electrodes. Different models do exist that describe the source of EEG. In this chapter the prominent dipole model is described by giving a short résumé of the work of Scherg (1991). Scherg (1991) showed that the changes of the electrical field measured on the scalp can be explained by a superposition of dipole fields, which are generated inside of the head. To further describe the source of the electrical field on the scalp following two models are necessary: (1) a head model describing the propagation of electrical activity through the head and the resulting electrical field on the scalp and (2) a model describing the electrical activity of a neuronal structure.

(1) Scherg (1991) described that the head can be represented by a solid high conductive sphere (brain) that is surrounded by a shell with low conductibility (skull) and a further shell with high conductibility (scalp). Due to the low conductibility of the skull shell in respect to the inner sphere and outer shell, it has capacitary properties. Therefore, the electrical field, which is generated by a superposition of several dipoles located in the head, is weekend and spatial widened when observed on the scalp.

(2) The physiology of an active neuron can be described as follows (as described by Zschocke et al., 2012; and Scherg, 1991). Synapses that are connected to a neuron can act excitatory or inhibitory to the activation of the neuron. In Figure 1 an excitatory activation of a synapse is shown (Zschocke et al., 2012). During an activation of a synapse, transmitters are released into the synaptic gap, which defuse to the subsynaptic membrane of the neuron and get bound to receptors. This leads to the opening of ion-channels and movement of ions through the subsynaptic membrane. Depending on the type of synaptic activation the subsynaptic membrane gets either depolarized (if excitatory) or hyperpolarized (if inhibitory). Therefore, it possesses a different electrical charge relative to the remaining membrane surface of the

neuron dendrites and soma (postsynaptic membrane), which leads to the creation of an electrical dipole alongside the postsynaptic membrane. The respectively generated electrical field is either referred to as the excitatory postsynaptic potential (EPSP, see Figure 1) or inhibitory postsynaptic potential (IPSP). The soma can be attached to several dendrites. If the depolarization at the end of the soma surpasses a specific threshold, an action potential is released leading to an even spreading deposition of the axons' membrane until the neurons' synapse is reached, allowing an information transition between nerve cells.

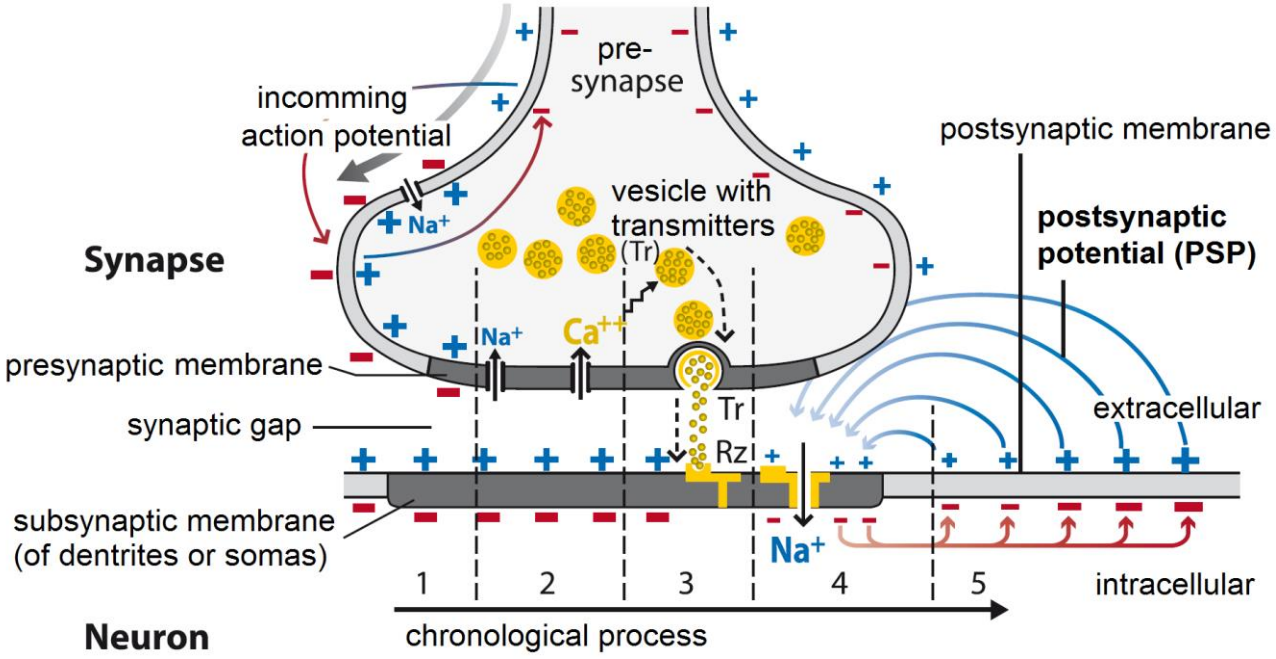

Figure 1: Schematic chronological presentation of neurotransmission and resulting generation of an excitatory postsynaptic potential (Zschocke et al., 2012). 1) Incoming action potential of synapse leads to depolarization of the presynaptic membrane. 2) Resulting influx of Ca^{++}-ions launches the movement of a vesicle with transmitters towards the presynaptic membrane. 3) Vesicle attaches to presynaptic membrane and transmitters are released into the synaptic gap, diffusing to the subsynaptic membrane and bound to receptors. 4) In result ion-channels are opened and Na^{+}-ions pass through the sub synaptic membrane, depolarizing the membrane. 5) The sub-synaptic-membrane possesses a different electrical charge relative to the remaining membrane surface of the neuron dendrites' and soma (postsynaptic membrane), creating an electrical dipole alongside the postsynaptic membrane, generating an electrical field, referred to as excitatory postsynaptic potential (EPSP; Zschocke et al., 2012).

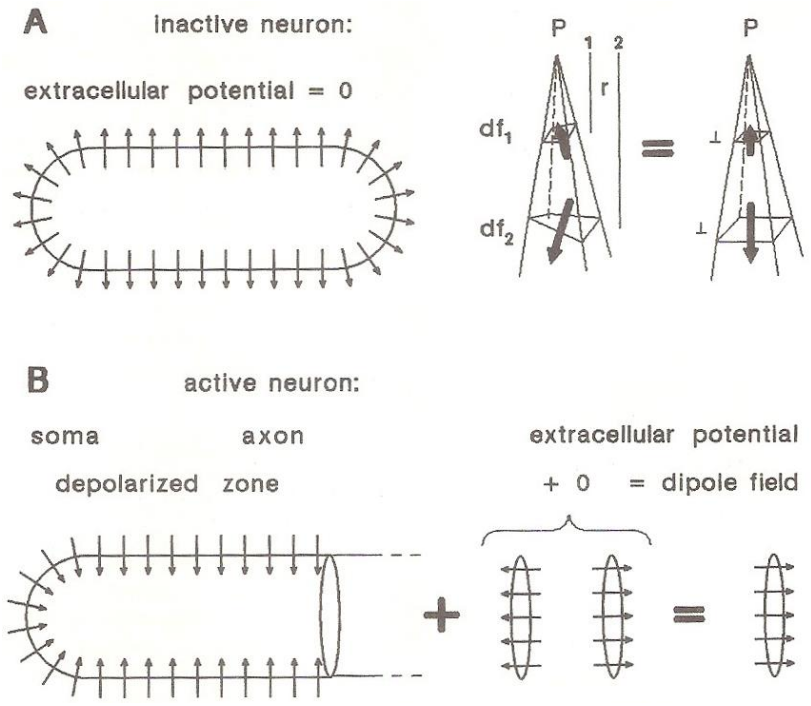

Figure 2: Demonstration of the extracellular potential of an inactive (A) and an active (B) neuron (Scherg, 1991). A) Random extracellular measurement point P shows no measurable electrical field, due to cancelation of equal electrical fields with opposing polarities created by differential surface elements (df_1 and df_2) at the radial distance (r_1 and r_2). B) Dipole field of depolarized zone (soma and initial axon-section) can be described by a dipole disc with equal dipole-density.

To show that the macroscopic electrical field that is generated during the activation processes of a neuron is predominantly determined by the geometry of the neuron, Scherg (1991) introduced following model. In Figure 2A) the resting potential of the membrane of an inactive neuron is characterized (Scherg, 1991). Due to the difference in charge of ions located at both sides of the membrane, small electrical dipoles are generated. The density of dipoles per differential element of the membrane can be defined as dipole-density: $D = \frac{q \cdot d}{dF}$ (d: thickness of membrane, dF: differential element of membrane with extracellular charge $+q$ and intracellular charge $-q$). The extracellular electrical field, here referred to as extracellular potential, of an inactive neuron is equal to zero. This is demonstrated using a random extracellular measurement point P. The size of a surface element (df) of the membrane determined by a solid angle at the measurement point P increases with the quadratic distance (r^2) to P, while the elicited electrical field decreases with the quadratic distance (r^2) to P. At P the electrical fields of the two surface elements (df_1 and df_2) are equal in size with opposed polarity, therefore canceling each other. In Figure 2B) extracellular potential of an active neuron is characterized. An action potential is released, if the depolarization at the end of the soma surpasses a specific threshold. The depolarization of a membrane is caused by the movement of ions through the membrane. Due to the difference in charge between both sides of the membrane, small electrical dipoles are generated. Figure 2B) shows a depolarized zone composed of the

depolarized membrane of the soma and the initial axon-section. The membrane behind the depolarized zone is still in its resting potential. Assuming that the depolarized zone is closed with a disc that possesses no charge, it is possible to replace this disc with two equal dipole discs with opposed polarity, so that the dipole fields of the two discs are canceled by each other. If the dipole-density of the two discs is chosen to be equal to the dipole-density of the depolarized zone, the summated extracellular potential of the depolarized zone and one dipole disc is equal to zero. The electrical field of the remaining dipole disc is now describing the extracellular potential of the active neuron. The effective dipole moment of the disc is proportional to the section area of the axon and further proportional to the dipole-density (D): $p_{e_{ff}} = D \cdot dF$. The dipole-density is only dependent on the amplitude of the action potential.

Scherg (1991) further showed that a continuous distribution of dipole-discs along the axon is needed to appropriately describe the potential of its membrane during an action potential (as described in Figure 3; Scherg, 1991). In Figure 3 (top) the change of the membrane potential at the beginning of the axion (soma side) is shown at five different phases (Figure 3, middle) of a propagating action potential. The dipole-disc at the beginning of the axion is oriented in direction of the axon during the depolarization phase (0-2). During the repolarization phase (3-4) the orientation of this disc points in the direction of the soma. After completion of the repolarization phase (5), the membrane at the beginning of the axon possesses its resting potential (like shown in Figure 2A)) and the respective dipole-disc has no charge. If the equally distributed dipole-discs are observed together, they build a dipole that moves along the axon during the depolarization phase (0-2). During the repolarization phase (3-4) a second combined dipole arises, pointing in opposite direction, increasingly compensating the dipole field of the first one. After completion of the repolarization phase (5) the quadrupole is moving along the axon. In Figure 3 (bottom) the change of the elicited electrical far field is simulated during these five phases. Scherg (1991) notes that after a distance of 20 mm from the center of an active axon the dipole field is dominating the quadrupole field. Hence, as far as action potentials are concerned, electrodes placed on the scalp can only measure dipole fields created at the beginning, end or bends of axons.

Using these models Scherg (1991) concludes that neurons with widely extended cell-bodies or dendritic trees, like e.g. pyramid cells of the cortex, generate dipole fields at the soma (sum of EPSP) that are bigger than the dipole fields generated at the beginning, end or bends of their axons. Here the geometry of the cells' dendritic trees and the position of its synapses is critical for the structure of the elicited dipole field. Therefore, the sum of EPSP generated by the pyramid cells of the cortex might be the source of cortical activity to auditory stimulation measured at the scalp. Since neurons at the beginning auditory pathway located at the brainstem have diameters of less than 40 µm, the superposition of dipole fields generated during the propagation of action potentials may be the main source of brainstem activity to auditory stimulation measured at the scalp.

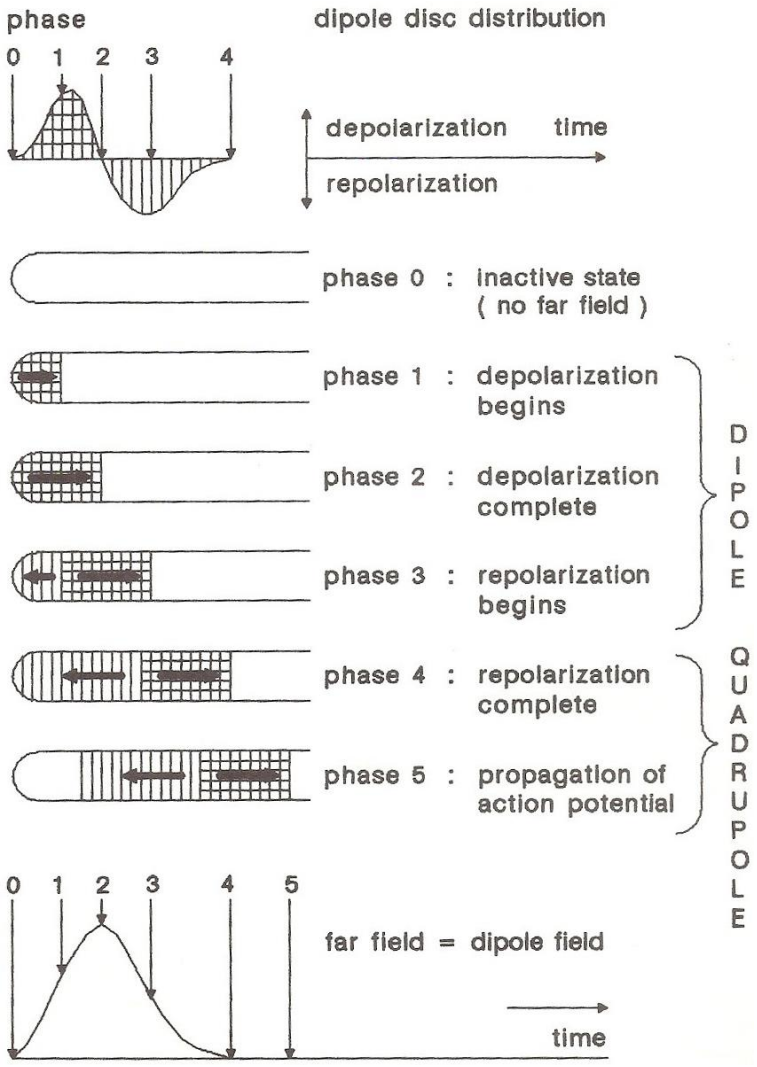

Figure 3: Simulation of the membrane potential (Top) at the beginning of an axon and the elicited electrical far field (Bottom) during propagation of an action potential (Scherg, 1991). Middle: propagating action potential divided in five phases (depolarization: 0-2, repolarization: 3-4, time after completion of repolarization 5).

2.2 EEG Features to auditory stimulation

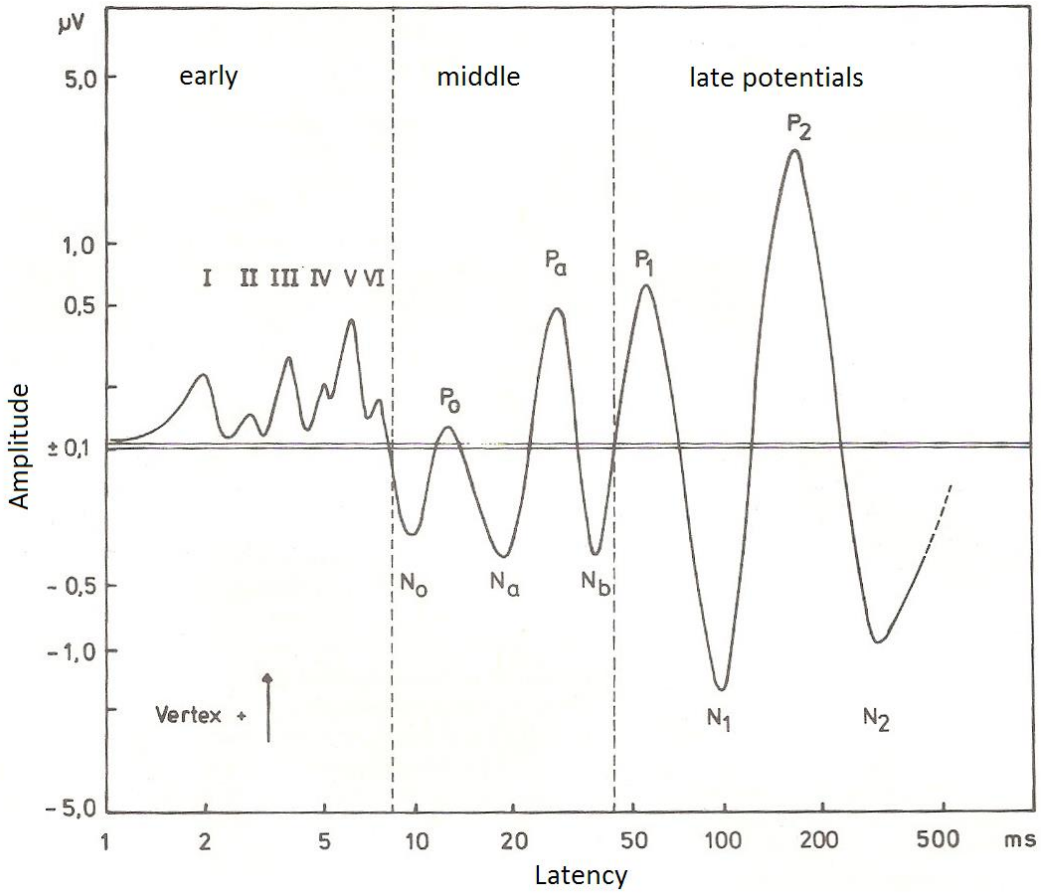

Figure 4: Schematic presentation of the auditory evoked potentials on a logarithmic scale (Scherg, 1991). The potentials are structured in early, middle and late components due to their latency (recorded from vertex in reference to mastoid).

The current thesis addresses different Brain Computer Interface (BCI) approaches based on auditory stimulation. Therefore, the three most prominent groups of EEG features to auditory stimulation are introduced in the following. In Figure 4 the so called auditory evoked potentials (AEP; Picton et al., 1974) are demonstrated (Scherg, 1991). The AEP are generated, whenever an acoustical stimulus is presented or when a permanently presented acoustical stimulus changes its intensity. The AEP are structured in early (1-10 ms), middle (10-50 ms) and late (50-250 ms) components due to their latency (recorded from vertex in reference to mastoid) and therefore can be assigned to different areas on the auditory pathway (early AEP: different stages of the brainstem to midbrain; middle & late AEP: from thalamus to auditory cortex). The magnitude of the different components ranges from 0.1 µV to several µV (Picton et al., 1974).

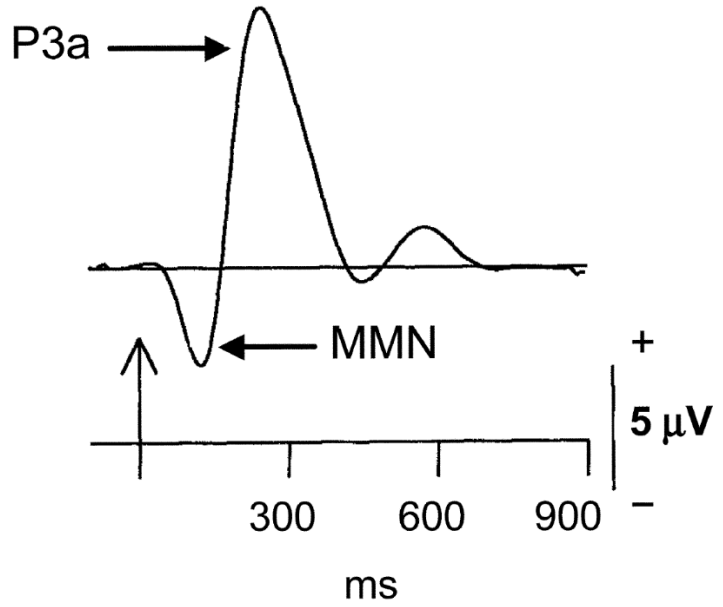

Figure 5 Demonstration of event related potentials that indicate the novelty of a stimulus (Friedman et al., 2001). The mismatch negativity (MMN) and the P3a shown here where calculated by subtracting the averaged recordings of the EEG responses to standard stimuli from the averaged EEG responses to the deviant stimuli of an oddball paradigm. The participants were watching a silent movie during the measurement and were instructed to ignore acoustical stimuli.

Event related potentials (ERP; Friedman et al., 2001) are EEG potentials sensitive to the manipulation of the cognitive context in which the evoking stimuli are incorporated. Therefore, ERP components can be used to detect possible differences of cognitive processing between conditions. The ERP are evoked by deviant events that are presented within a train of homogeneous standard stimuli (a.k.a. oddball-paradigm; Friedman, 2001). In Figure 5 a demonstration of two prominent ERP is given (Friedman et al., 2001). One Example of an ERP is the P300, which is a positive EEG-potential that occurs with a latency of about 300 ms to a stimulus that a listener is expecting or to a stimulus that surprises the listener and therefor unconsciously takes his attention (Pritchard, 1981). The P300 is distinguished into the following two groups: (1) the P3a, which is an indicator for the novelty of an event and is best measurable at the frontal/central area of the scalp. (2) the P3b, which is an indicator for stimulus relevance/attention and is best measurable at the parietal area of the scalp (Comorchero and Polich, 1999). The attention effect of the P3b is reflected in the magnitude of its amplitude. If for example two different deviant stimuli are presented in an oddball-paradigm, the P3b evoked by the attended deviant has a bigger amplitude than the P3b evoked by the ignored deviant stimulus. Another example for a prominent ERP is the mismatch negativity (MMN; Näätänen et al., 2007; Picton et al., 2000), which is represented in the EEG as a negative deflection with a latency of 150-250 ms to a deviant stimulus. The MMN is an indicator for stimulus novelty and is evoked whether the deviant stimulus is attended or ignored. It is sensitive to a variety of stimulus changes, from

simple feature changes to complex context differences. The bigger the difference between standard and deviant, the shorter the latency of the evoked MMN (Näätänen et al., 2007).

The last EEG feature introduced here is the auditory steady-state response (ASSR; Picton et al., 1987; Picton et al., 2003), which represents how the brain follows the modulation of a continuously presented stimulus. The ASSR are evoked by auditory stimuli presented at rates between 1 and 200 Hz or by continuously presented stimuli that are periodically modulated in amplitude or frequency. ASSR can thereupon be revealed in the respective frequency bin of the recorded long-term EEG spectrum (Picton et al., 2003). ASSR evoked by lower modulation rates (40 Hz and lower) are related to interactions between thalamus and the auditory cortices, while the main generators of ASSR evoked by higher modulation rates (80 Hz and higher) are related to the brainstem (Herdman et al., 2002; Picton et al., 2003). Picton et al. (1987) described, that ASSR evoked by modulation rates around 40 Hz are the easiest to measure. ASSR to faster rates are also good measurable, even though smaller in amplitude, due to the lower EEG background noise at higher frequencies. ASSR to slower rates are more difficult to measure, even though higher in amplitude, due to the higher background noise of the EEG at lower frequencies. Lopez et al. (2009) and Kim et al. (2011) demonstrated, that the amplitude of ASSR to modulation rates around 40 Hz can be affected by selective attention. For example, two different sinusoidal amplitude modulated stimuli with carrier frequencies at 1 kHz (S_1) and 2.5 kHz (S_2) and modulation frequencies of 35 Hz (S_1) or 45 Hz (S_2) respectively are presented dichotically to a participant. If the participant attends to S_1 while ignoring S_2, the recorded EEG spectrum would reveal a higher amplitude of the 35 Hz bin than the 45 Hz bin. For ASSR to low amplitude modulation rates, it is not fully clear if the elicited EEG is a superposition of transient responses or a steady state response. ASSR to low amplitude modulation rates are often referred to as envelope following response (EFR), especially when the modulation is varying over time.

2.3 Classification

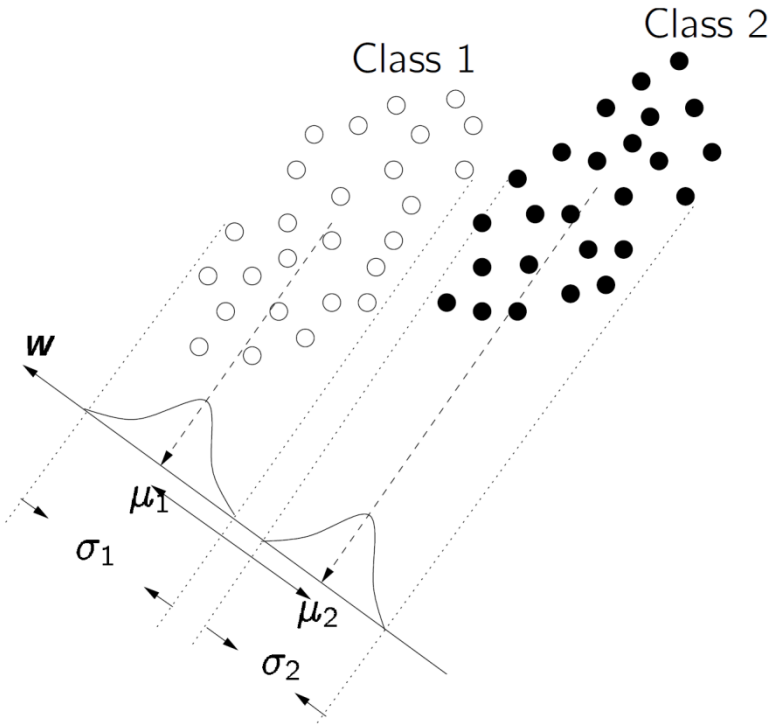

Figure 6: Exemplary presentation of a Linear Discriminant Analysis for two classes (Mika, 2002). The data is projected to a new axes $\boldsymbol{w}$ *that assures the biggest possible separability of the two classes.* $\boldsymbol{w}$ *is found by maximizing the distance between the projected mean of each class (*μ_1*,* μ_2*) while minimizing the projected variation of each class (*σ_1*,* σ_2*).*

An auditory BCI could detect brain responses that are evoked by different sources and use a classification method to segregate these sources or to detect o specific source of interest. For a BCI application it is reasonable to choose a computational efficient and mathematically robust classification method, due to the limited availability of neuronal data samples and the limitation in processing power as well as processing time. In contrast to more powerful, but computational expensive classification methods, like e.g. Neural Networks (NN; Martinez, 2014; Schmidhuber, 2015) or Supportive Vector Machines (SVM; Cristianini et al., 2000), Linear Discriminant Analysis (LDA; Fisher, 1936; Mika et al., 1999) is not dependent on big datasets for training to ensure reasonable results. The LDA searches for a linear combination of variables that best separate the known categories, focusing on maximizing the distance between the mean of each category while minimizing the variance of each category (Fisher criterion; Fisher, 1936). Therefore, a dimension reduction is executed, projecting the data to a set of new axis which assure the biggest possible separability of the categories. In the remaining of section 2.3, an insight into the mathematical background of the LDA is given using a two-class example, based on the description given in the dissertation of Mika (2002). $\boldsymbol{X}_1 = \{x_{1,1}, \dots, x_{1,N}\}$ and $\boldsymbol{X}_2 = \{x_{2,1}, \dots, x_{2,N}\}$ are two different classes with N samples each. The Fisher criterion

$$J(\boldsymbol{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1 + \sigma_2}$$

describes the separability $J(\boldsymbol{w})$ of the two classes in dependence of to the new projection axes $\boldsymbol{w}$. The projected mean of each class (μ_i) and the projected variance of each class (σ_i) can be calculated using following equation:

$$\mu_i = \frac{1}{N_i} \Sigma_{j=1}^{N} \boldsymbol{w}^T x_{i,j} = \boldsymbol{w}^T \boldsymbol{m}_i \qquad \sigma_i = \sum_{j=1}^{N} \left(\boldsymbol{w}^T x_{i,j} - \mu_i\right)^2 .$$

It is distinguishable, that the separability of the two classes $J(\boldsymbol{w})$ is maximal, when the distance between the projected mean of each class (μ_1, μ_2) is maximal and the projected variance of each class (σ_1 ,σ_2) is minimal. An exemplary illustration for a two class LDA is shown in Figure 6 (Mika, 2002). A combination of these three equations and conversion leads to the following equivalent representation of the Fisher criterion:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T S_B \boldsymbol{w}}{\boldsymbol{w}^T S_W \boldsymbol{w}},$$

at which the substitutions S_B (a.k.a. between class scatter matrix) and S_W (a.k.a. within class scatter matrix) are defined as followed:

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \qquad S_W = \Sigma_{i=1}^{2} \Sigma_{j=1}^{N} \left(x_{i,j} - \boldsymbol{m}_i\right)^2 .$$

The Fisher discriminant $\boldsymbol{w}$ has a global solution, that maximizes $J(\boldsymbol{w})$. To find this solution the differentiation of $J(\boldsymbol{w})$ (with respect to $\boldsymbol{w}$) is equal to zero, which leads to:

$$\frac{dJ(\boldsymbol{w})}{d\boldsymbol{w}} = (\boldsymbol{w}^T S_B \boldsymbol{w})\, S_W \boldsymbol{w} - (\boldsymbol{w}^T S_W \boldsymbol{w})\, S_B \boldsymbol{w} = 0 \qquad \langle = \rangle \qquad S_B \boldsymbol{w} = \frac{\boldsymbol{w}^T S_B \boldsymbol{w}}{\boldsymbol{w}^T S_W \boldsymbol{w}} S_W \boldsymbol{w} .$$

The resulting equation is showing an eigenproblem, at which $\boldsymbol{w}$ is the leading eigenvector and the quantity $\frac{\boldsymbol{w}^T S_B \boldsymbol{w}}{\boldsymbol{w}^T S_W \boldsymbol{w}}$ is largest eigenvalue. A possible solution for this eigenproblem is (dropping all scalar factors):

$$\boldsymbol{w} = S_W^{-1} (\boldsymbol{m}_2 - \boldsymbol{m}_1) .$$

Now we assume that the data of each class ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$) is normal distributed and used for training of the LDA. It is possible to make the decision to which of the two class (C_1 or C_2) a test sample x is belonging to, using the Bayes' theorem:

$$P(C_1|q(x)) = \frac{p(q(x)|C_1) \cdot P(C_1)}{p(q(x)|C_1) \cdot P(C_1) + p(q(x)|C_2) \cdot P(C_2)}$$

and likewise, for $P(C_2|q(x))$, while $q(x) = \boldsymbol{w}^T x$ and

$$p(q(x)|C_i) = \left(2\pi\sigma_i^2\right)^{-1/2} exp\left(-\frac{(q(x)-\mu_i)^2}{2\sigma_i^2}\right).$$

The prior class probabilities $P(C_1)$ and $P(C_2)$ can be estimated from the training samples ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$). In this case, LDA would assign the test sample x to class 1, whenever the posterior probability $P(C_1|q(x)) \geq \frac{1}{2}$ and to class 2 otherwise (Mika, 2002).

2.4 Common Artifacts and Solutions

The biggest challenge of EEG signal processing is to compensate for the poor SNR of the measured data. This SNR property is caused by interfering signals that are measured besides the respective evoked brain responses (see 2.2) and by additional changes of measurement conditions over time. In the following, an overview of the five most common measurement artifacts is given with corresponding possible signal processing solutions:

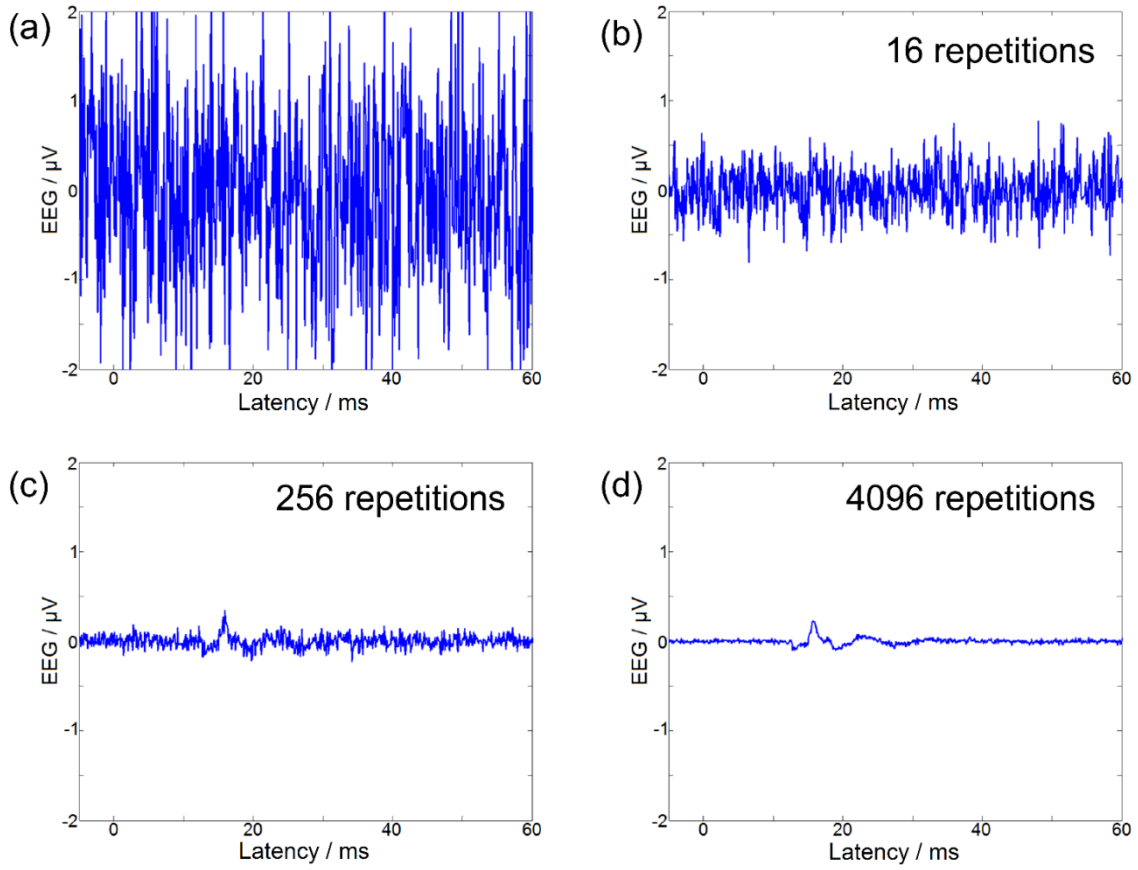

Figure 7: Auditory evoked brainstem response to a chirp stimulus. a) Single trial EEG response masked by spontaneous background activity of the brain. b), c) and d): Data averaged over successively increasing repetitions of presentations (trials) of the same stimulus revealing the correlated brainstem response. (Data recorded and pictures created for demonstration purposes by Helge Lüddemann, Oldenburg 2003).

(1) The most prominent interfering signal is the spontaneous EEG that is permanently generated by the background activity of the brain. The spontaneous EEG is uncorrelated to any external stimulation and is expressed as an overlaying noise floor in the EEG recording. In Figure 7a) an example of a single trial EEG measurement to an auditory chirp stimulus (acoustical stimulus optimized for synchronous excitation along the basilar membrane; Dau et al., 2000) is shown. The low amplitude EEG response to the stimulus is masked by the spontaneous EEG. If a stimulus is repeatedly presented with a sufficient inter-stimulus-interval, it can be assumed, that the brain responds the same way every time. The spontaneous EEG is

normal distributed around zero assuming, that the recorded data is offset corrected. Based on these two assumptions, the amplitude of the spontaneous EEG can be reduced by the factor $\frac{1}{\sqrt{N}}$ by applying an arithmetic mean over multiple EEG recordings (N) to the same presented stimulus. A high amount of repetition is necessary to reveal low amplitude EEG responses, as shown in Figure 7c), after averaging over 256 repetitions the evoked brainstem response becomes visible. Hoke et al. (1984) described that the power of noise can vary from epoch to epoch and therefore, an optimal SNR can be obtained by weighting each epoch with its inverse power of noise during averaging. By the execution of additional iteration steps, using the inverse power of the residual noise (difference of the epoch and the weighted average) derived in the iteration step before as the new weighting for this epoch, Riedel et al (2002) further enhance the SNR obtained by the weighted average method (due to the enhanced noise estimation).

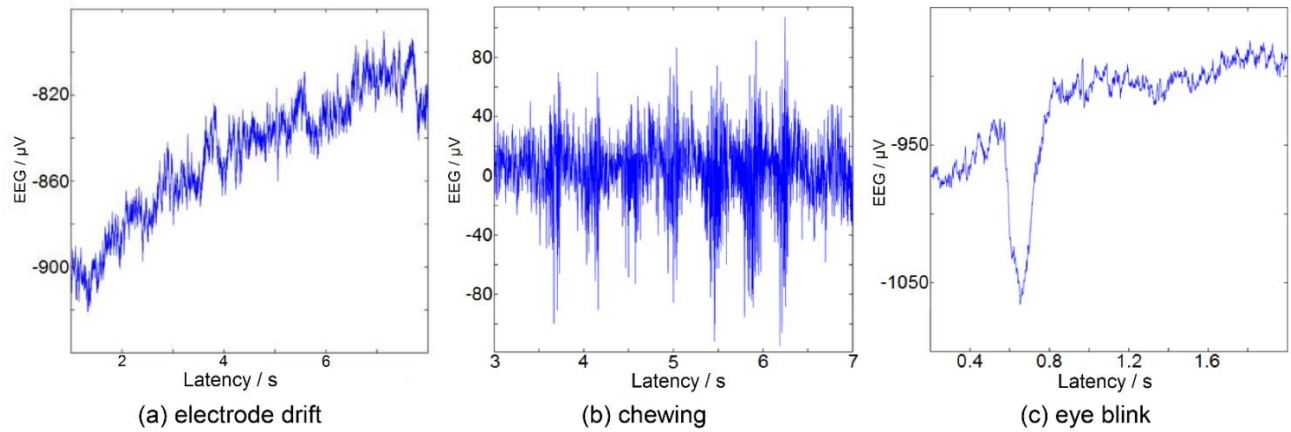

Figure 8: Demonstration of three common EEG artifacts. No external stimulation was introduced during all three recordings and an average reference over the two mastoid channels was used. a) Section of an EEG recording - which lasted over several minutes - at the vertex showing a drift in the data (a.k.a. electrode drift). b) EEG recorded over channels close to the jaw mussels while the subject performed a chewing motion (muscle artifact). c) EEG recorded at frontal channels while the subject performed an eye blink. (Data recorded and pictures created for demonstration purposes by Carlos da Silva Souto, Oldenburg 2011).

(2) The electrode drift is another typical EEG artifact (see Figure 8a)). It is possible that measurement conditions of the electrodes change over time - e.g. the resistance of the skin could change over time due to sweating - this causes a shift of voltage in the recorded data of several hundred µV over the period of several minutes. Since this artifact affects the measured data like a low frequency oscillation, it can be removed by applying a zero-phase high-pass filter with a cut-off frequency of 1 Hz or lower to the data. Alternatively, the electrode drift can be estimated applying an appropriate nonlinear fit to the data to extract the low frequency trends of the data and subtract this from the original data.

(3) External electrical fields can directly influence the measurement and are visible in the long-term spectrum of the EEG. Two examples for external sources of electrical fields are the German electricity

network which runs an alternating current with a frequency of 50 Hz and the German train network which runs an alternating current at 16.7 Hz. It is possible to use a low-pass or notch filter to attenuate the affected frequency band of the data, if the loss of information is acceptable. To avoid this kind of artifacts it is possible to execute the measurement in an appropriately shielded booth. When measuring inside a shielded booth, it is necessary to avoid any electrical AC devices, like e.g. light sources. The electrical field of DC devices does not frequently change its direction and is represented in the EEG as an offset. The DC-offset of an epoch can be corrected by averaging over EEG samples recorded during silence and reducing this estimate from the whole epoch. If an AC device is necessary for stimulation, e.g. headphones for acoustical stimulation, it is necessary to additionally shield these artifact sources properly and position them as far as possible away from the measurement electrodes (for this example insert tube-headphones with additional shielding would be appropriate).

(4) Another group of prominent artifacts are created by muscle movement. Active groups of muscles that are located close the electrodes, e.g. jaw or neck muscles, produce measurable potentials. In Figure 8b) the effect of a chewing movement on the EEG is demonstrated. The features of this artifacts are a high amplitude and a broad frequency spectrum. Therefore, they can hardly be removed neither by filter methods nor by averaging over multiple measurement repetitions. However, epochs affected by this artifact can easily be detected by setting a peak to peak artifact threshold of 30 to 100 µV (depending on the general noise level of the recorded data) and subsequently, these recording epochs can be removed. Since data is lost this way it is important to inform the participant ahead of the measurement about this problem and to ensure that the participant stays relaxed and calm during the measurement.

(5) The last EEG-artifacts addressed here are created by eye movement and effect predominantly channels located at the forehead. The eyes generate an electric dipole field, due to the ionized fluid inside the eyeball (Durban, 2006). Therefore, eye movements like horizontal and vertical shifts lead to shifts of up to 100 µV in the EEG recording. An eye blink generates a specific artifact shown in Figure 8c), due to the 90-degree reflex-rotation of the eyeball upwards and beck to the front. Like muscle artifacts (see above), affected epochs can be detected and removed using an artifact threshold. If the amount of recorded EEG epochs is low, it is advisable to use advanced statistical methods like independent-component-analysis to statistically estimate the artifact that is caused by eye movement and reduce it from the data (potentially rising the general noise floor of the data due to estimation errors). To prevent eye artifacts in the EEG recording, it is important to instruct the participant to stay calm during the measurement, to avoid excessive eye blinking and to watch at a visual fixation point, e.g. a fixation-cross on a wall, to avoid eye movement. If the eyes of the participant are closed during the measurement or the participant is very tired, a relaxation of the visual cortex occurs which often leads to a raised activity in the alpha-band (8-12 Hz) of the EEG (Hughes and Crunelli, 2005). The resulting so-called alpha-waves are visible in the recorded EEG in the form of sinusoidal waves that obscure the underlying recordings.

3 Influence of attention on speech-rhythm evoked potentials: first steps towards an auditory brain-computer-interface driven by speech.[1]

3.1 Abstract

A Brain-computer interface (BCI) uses neuronal responses to control external systems. The majority of BCI systems are based on visual stimuli, only few apply auditory input. Because auditory-based BCI do not rely on visual skills or mobility of the body, they could be an alternative for visually or physically disabled people.

This study investigates the performance of an auditory paradigm using two competing streams of repeatedly presented speech syllables. The streams had different repetition rates of 2.3 and 3.1 Hz. Our auditory BCI approach uses the auditory steady-state response (ASSR) to automatically detect which stream a listener selectively attends to.

In a single trial classification ten healthy volunteers achieved an accuracy significantly above chance of 61 % and an information transfer-rate (ITR) of 0.2 bit/min. The use of the average over six random trials improved the average classification accuracy to 79 % while keeping the ITR comparable.

In conclusion it is possible to classify ASSR evoked from streams of spoken syllables. For a real life application it is necessary to improve the performance of this auditory BCI, but it is a step towards the long term goal of using BCI on natural speech features and eventually controlling the processing of hearing devices.

[1] This Chapter is published as:

da Silva Souto, C., Lüddemann, H., Lipski, S., Dietz, M., & Kollmeier, B. (2016). Influence of attention on speech-rhythm evoked potentials: first steps towards an auditory brain-computer interface driven by speech. *Biomedical Physics & Engineering Express*, *2*(6), 065009.

3.2 Introduction

A Brain Computer Interface (BCI) uses brain activity to control electronic devices. Task-related neurophysiological changes are detected, classified, and finally activate external systems. In this way, BCI can be used to control medical devices without manual interaction. For example, BCI controlled prostheses are a great benefit for physically disabled people and can improve their quality of life permanently (Birbaumer, 2006).

The majority of BCI systems are based on neuronal responses to visual stimuli or use neuronal activation patterns that occur while a certain stimulus is perceived or while a specific action is imagined. Only a few studies on BCI apply auditory evoked potentials (e.g., Halder et al., 2010; Kim et al., 2011). Because auditory-based BCI are not dependent on visual skills, mobility or posture of the head and eyes, they could be an alternative for people that are visually disabled or completely paralyzed (Halder et al., 2010; Kim et al., 2011).

Some auditory BCI are based on artificial stimuli or noises in oddball-paradigms to expand established visual BCI-paradigms like the visual P300-Speller (Farwell and Donchin, 1988; Donchin et al., 2000) to auditory stimulation (Klobassa et al., 2009). Klobassa et al. (2009) for instance presented six environmental sounds in a 6 x 6 P300-Speller paradigm to two groups of participants. Additional visual stimuli were presented to one of the groups. They were able to measure and classify P300 responses from both groups with equivalent accuracy.

Recently some studies investigated on synthetic speech stimuli. Hill et al. (2014) realized an auditory BCI based on streams of synthesized words in an oddball-paradigm. Presenting the stimuli from two different directions simultaneously, they were able to classify event related potentials to detect the direction of interest with an average accuracy of 77%. Nakamura et al. (2013) showed a BCI classifying auditory steady-state responses (ASSR; Picton et al., 1987) to amplitude modulated synthesized sentences presented from two directions. Their BCI approach performed comparable to the one by Hill et al. in terms of classification accuracy.

Further studies like the one of O'Sullivan et al. (2014) use regression methods to estimate the envelope of a presented speech signal from the evoked EEG recordings (aka. stimulus-reconstruction method; Rieke et al., 1995; Mesgarani et al., 2009). O'Sullivan et al. simulated a competing speaker situation by presenting two spoken stories from two different talkers and directions simultaneously. Their BCI approach utilized the correlation of the estimated signal with the actual envelopes of both speech signals. With this method they were able to detect which story a listener was selectively attending to.

Other studies like Kim et al. (2011) classified steady state responses to continuous trains of tone bursts or frequency-/amplitude modulated signals. Doing this, Kim et al. (2011) tested an auditory BCI that could detect which one of two sound sources a listener selective attends to. They presented two trains of pure

tone bursts with different presentation rates (37 and 43 Hz) from two different directions simultaneously (left and right) and were able to measure and classify ASSR that showed an effect of attention.

This study investigates the performance of an auditory paradigm using speech syllables. The ASSR to streams of syllables is used to automatically detect which one of two speakers a listener selectively attends to. Prospectively, auditory BCI could be applied to control beam-formers in hearing aids to point to a certain direction of interest, so that a difficult acoustical situation for an aided hearing impaired person may be enhanced, e.g., a cocktail party with multiple speakers and diffuse noise in the background, might be enhanced.

The aim of our study was to find new approaches for auditory BCI that can control hearing aids. Different from Kim et al. (2011), we are trying to use more natural stimuli like speech. Unmodified speech stimuli sound natural for the listener and should make the usage of BCI controlled systems more intuitive. Therefore it is reasonable to focus on the features of speech itself like rhythm, relevance and direction. To take a first step into the direction of using real spoken sentences, we use two streams of spoken syllables and present them in rates of 2.3 and 3.1 Hz to simulate a situation with two competing speakers that talk with different speech-rhythms. By measuring the electroencephalogram (EEG) of participants who selectively attend to one of the two speakers, we aimed to measure and classify the ASSR to both steams and detect to which speaker the listener was attending.

3.3 Methods

3.3.1 Participants

Initially twelve healthy student volunteers were recruited (5 female, mean age: 25, range: 15-35 years). Prior to the experiment, the hearing status was assessed with a tone-audiogram (125 Hz - 10 kHz). Thresholds were better than 15 dB HL at any tested frequency for 11 participants, one volunteer had a peak of 25 dB HL at 750 Hz on the right side. All volunteers were right-handed according to the questionnaire by Oldfield (1970). None reported any present or previous psychiatric or neurological disorder. Written informed consent was obtained from all adult volunteers and from the parents for minors prior to any testing. Participants were paid for participation. The study was carried out in accordance with the Declaration of Helsinki, 2008.

3.3.2 Stimuli

The following spoken syllables were used as stimuli: /te/ and /ti/ produced by a female speaker were presented repetitively at a rate of 2.3 Hz; /ka/ and /ku/ spoken by a male speaker were presented repetitively at a rate of 3.1 Hz (recorded: Tascam DR-40, 16 bit, 48 kHz). All syllables were matched in terms

of intensity and subjectively perceived loudness using Praat software. The two streams, 2.3 Hz from the female speaker (at 66.5 dB SPL averaged over 20 s) and 3.1 Hz from the male speaker (at 68.5 dB SPL averaged over 20 s) were presented simultaneously to the left and the right ear respectively (e.g. one Trial: female speaker from the left and male speaker from the right side; shown in Figure 9). The syllables /te/ and /ka/ were frequently presented, the syllables /ti/ in the female speaker's stream and /ku/ in the male speaker's stream were interspersed irregularly (pseudo-random position). This served as control for attention: participants had to count the irregular deviant syllables in the stream they attended to. A male and a female speaker were chosen in order to ease the separation of the auditory streams. For the same reason the consonant in each stream was kept constant. The specific acoustical characteristics of the syllables are shown in Table 1.

The syllable rate of spoken German is usually higher than the rates applied in the present study. Since German can be characterized as a stress-timed language (Grabe & Low, 2002) and stressed syllables are perceived as the "beat", presenting homogeneously accented syllables as sequences at a rate of 2-3 Hz is perceived as a normal speech rate.

Table 1: Specification of the acoustical characteristics of the used syllables.

Syllables	Spectrum (Hz)			Duration (ms)		
	Formant 1	Formant 2	Fundamental Frequency Vowel	Voice-Onset	Vowel	Total
/te/	440	2441	start 270 end 260	40	115	155
/ti/	385	2519		40	116	155
/ka/	659	1353	start 130 end 120	52	141	193
/ku/	372	947		52	142	193

Stimulus presentation was controlled using Matlab (7.3.0 R2006b), sounds were presented using a RME DIGI 96-8 PAD soundcard and converted via digital-analog converter (RME ADI-8 DS, sampling rate: 48 kHz). The analog stimuli were attenuated (Trucker-Davis, PA5) and presented over insert tube-headphones (Trucker-Davis, HB7 headphone amplifier / Etymotic Research, ER-2).

3.3.3 Task and Experimental design

The participants sat in a comfortable chair in an electrically and acoustically shielded booth. They were instructed to selectively attend to a specific stream and count the deviant syllables (/ti/ or /ku/) in that stream while fixating a visual cross on a wall in front of them. Each stimulation trial started with a brief tone-signal at the left or right ear which indicated the direction of selective attention (5 pulses of pure

tone, 440 Hz, ISI: 0.1 s). After a 1 s pause the syllable sequences started simultaneously on both sides (see Figure 9). Within short breaks of 4 to 8 s between trials participants wrote down the number of deviant stimuli that they had counted in the previous trial. Figure 9 shows a schematic example of one trial. This paradigm is similar to the one by Kim et al. (2011), except that here spoken syllables were presented at rates of 2-4 Hz instead of artificial stimuli at rates of approximately 40 Hz.

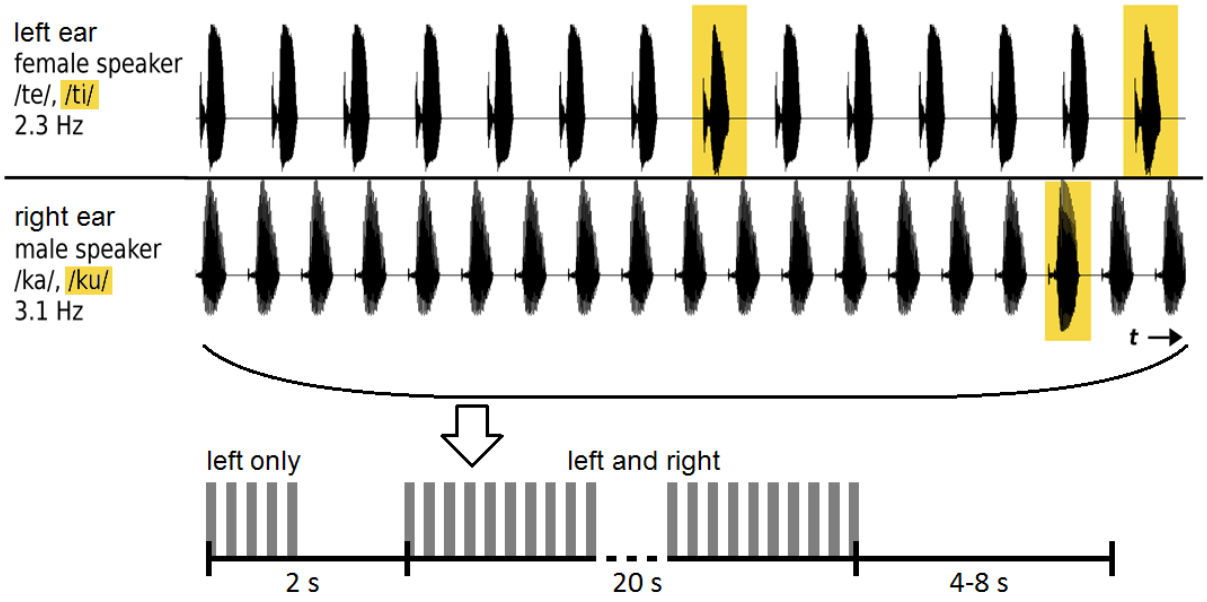

Figure 9: Top: Syllable streams were presented simultaneously to the left and the right ear at 2.3 Hz and 3.1 Hz, respectively. Participants selectively attended to one side and counted deviating syllables (yellow). Bottom: Stimulation layout: At the onset of each trial an acoustic signal indicated the direction of attention. Syllables were presented after a delay of 1 s in a 20 s syllable train. Each trial ended with a pause of 4-8 s during which participants wrote down the number of counted deviants.

Each participant took part in two 90 minute sessions on two days. The first session consisted of training, the second session was the actual measurement. Seventy-two trials with a duration of 26 to 30 s each were presented in each session consisting of 1656 syllables from the female speaker at 2.3 Hz (8.7% deviants) and 2232 syllables from the male speaker at 3.1 Hz (6.5% deviants, 3-5 deviant-syllables per trial). Sessions were divided into 4 blocks (18 trials each) with 5 to 10 minute breaks between blocks. The direction of attention switched after each trial and the direction of speakers changed after each block.

3.3.4 EEG Recording

EEG data were recorded with the Biosemi ActiveTwo system (speed mode 6, 11 channels, Electro-Capm International 10-20 system, Parker contact gel). Before data recording all electrode offsets were set not higher than 10 mV. Data was collected with the Biosemi software ActiView (6.03) and digitized at 1024 Hz without any additional filtering.

3.3.5 Data processing

The data set was referenced to the averaged mastoid-channels ((A1+A2)/2) and the channels were reduced to a nine-electrode-cluster (FC1, Fcz, FC2, C1, Cz, C2, CP1, Cpz, CP2). A band-pass filter with lower and higher cutoff frequency of 1 and 10 Hz was applied to reduce electrode drifts and high frequency noise. This cutoff furthermore allows for better power of noise estimation of the relevant frequency area. The data was separated into epochs of 20 s according to triggers, after filtering. For baseline correction the DC-offset was estimated, by averaging the samples of the 500 ms window before each epoch, and subtracted. In the last step data was sorted according to events, averaged over epochs and channels and fast-Fourier transformed (FFT).

To receive an optimal signal to noise ratio (SNR) during averaging, every epoch was weighted with its inverse power of noise (Hoke et al., 1984). For a better estimation of the noise, two iteration steps were executed using the residual noise of the step before (Riedel et al., 2002).

3.3.6 Statistical analysis and classification

Statistical significance of ASSR-amplitudes was tested with a (2x2x2) repeated analysis of variance (rANOVA). To test for sphericity of the data a Mauchly-Test was applied. The assumption of sphericity was not violated in any of the conducted analyses. Within subject factors were attention (attended, unattended), syllable rate (2.3 Hz, 3.1 Hz), corresponding to speaker gender (female, male) and harmonic (no: ground frequency [2.3 Hz, 3.1 Hz], yes: first harmonic [4.6 Hz, 6.2 Hz]).

To classify the EEG data and to show the quality of the BCI, a linear discriminant analysis (LDA) was employed. In two tests the absolute amplitude of the corresponding frequency bin and its first harmonic (female: 2.3 Hz, 4.6 Hz; male: 3.1 Hz, 6.2 Hz) were used for classification. Epochs of the following conditions were chosen randomly:

1. attention female speaker vs no attention male speaker (A: female vs N: male)
2. attention male speaker vs no attention female speaker (N: female vs A: male)

Two-third of the epochs were used for training and 1/3 for testing. The tests were made for single trials, weighted averages of 3 or 6 randomly chosen epochs of the specific condition. For a more accurate classification each LDA was repeated 1000 times and their results were averaged. Information transfer-rates (ITR) were calculated using the following equation (Besserve et al., 2007; N: amount of classes, p: recognition rate, T: duration of epoch: single trial (T=20 s), 3 epochs (T=60 s), 6 epochs (T=120 s)):

$$\mathrm{ITR} = \frac{60}{\mathrm{T}}\left[\log_2 \mathrm{N} + \mathrm{p}\cdot\log_2 \mathrm{p} + (1-\mathrm{p})\cdot\log_2\left(\frac{1-\mathrm{p}}{\mathrm{N}-1}\right)\right],$$

$$\Delta\mathrm{ITR} = \frac{60}{\mathrm{T}}\cdot\log_2\left(\frac{\mathrm{p}\cdot(\mathrm{N}-1)}{1-\mathrm{p}}\right)\cdot\Delta\mathrm{p}.$$

3.4 Results

Data from 10 volunteers entered the analysis. Data from 2 male participants were excluded due to excessive head movement during the measurement.

Figure 10 shows the spectral density of the EEG signal averaged over all subjects. ASSR to the presented syllable rates and their first harmonics are visible and modulated by attention. The rANOVA reveals a nearly significant effect of attention for ASSR to syllable rates and their first harmonics combined ($F(1,9)=5.1$; $p=0.051$). This effect remains nearly significant for ASSR to syllable rates stand-alone ($F(1,9)=4.8$; $p=0.055$), but vanishes for responses to first harmonics ($F(1,9)=0.9$; $p=0.36$). Syllable rate in general shows no significant effect ($F(1,9)=3.0$; $p=0.12$).

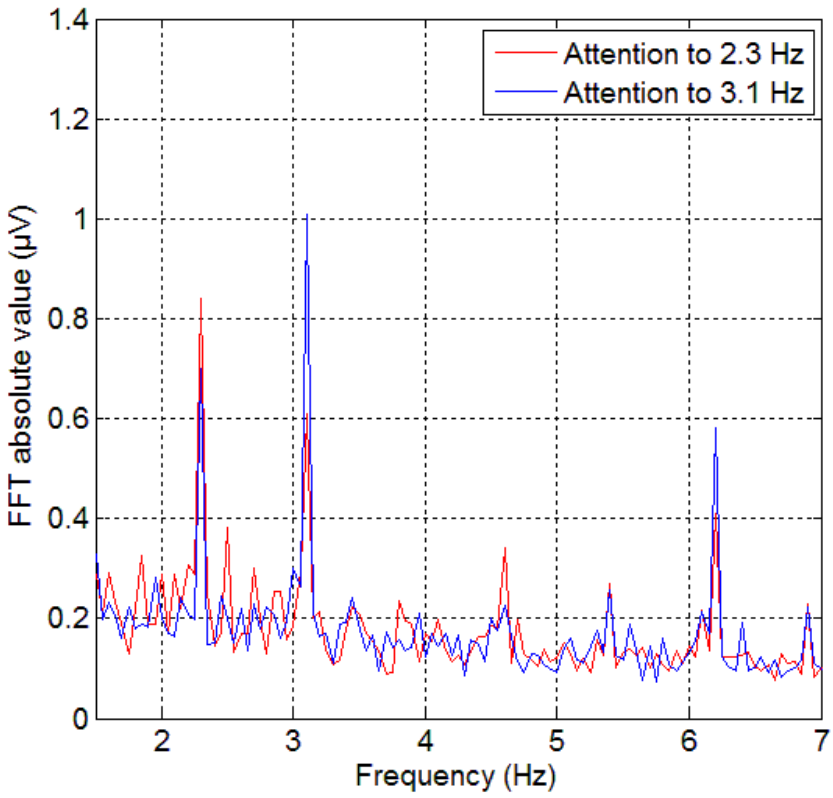

Figure 10: Spectral density of the EEG signal averaged over all subjects for two different measurement conditions: (red line) Attention to syllables /te/ and /ti/ (female speaker) presented in 2.3 Hz rates; (blue line) Attention to syllables /ka/ and /ku/ (male speaker) presented in 3.1 Hz rates.

Figure 11 shows the ASSR absolute amplitudes for each participant and classification condition with corresponding mean noise for error estimation (root-mean square (RMS) of frequency bins ± 0.5 Hz). For example in condition "A: female vs N: male" the 2.3 Hz response corresponding to the attended female speaker is classified against the 3.1 Hz response corresponding to the unattended male, and parallel the 4.6 Hz response against the 6.2 Hz response. Figure 11 illustrates that in about 70% of the cases, the attended amplitude is bigger than the unattended. Participant 2 (lowest age) shows the biggest responses overall.

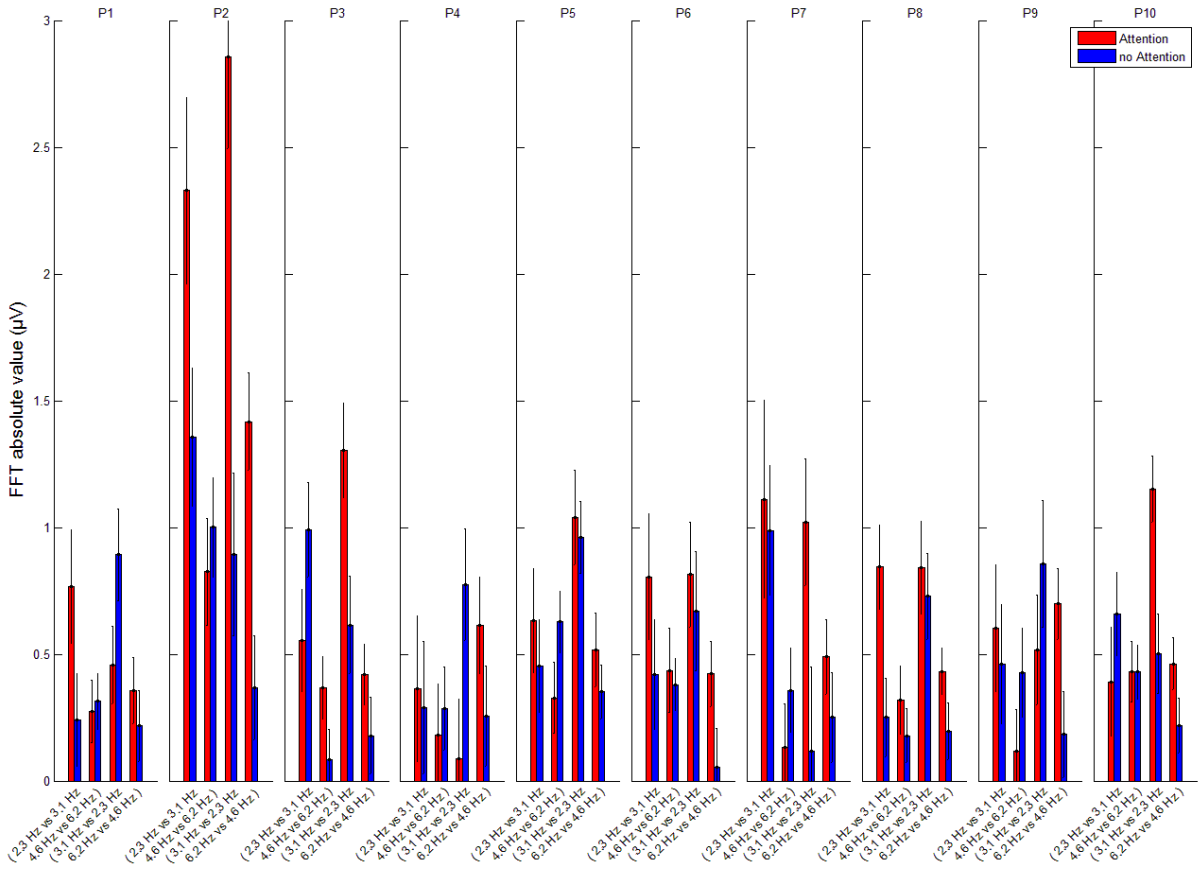

Figure 11: ASSR absolute amplitudes (2.3, 3.1 Hz) and their first harmonics (4.6, 6.2 Hz) with corresponding mean noise (RMS of ± 0.5 Hz) are presented for all participants (P1-P10). Red bars show ASSR to syllables that were attended and blue bars to syllables that were ignored. Attended and unattended responses are grouped to represent later classification, in doing so two groups in round brackets are classified parallel and represent one condition (see sec. 2.6).

Table 2: Single trial classification accuracies (p) and corresponding ITR averaged over all participants.

condition	p (%)	ITR (bit/min)
A: female vs N: male	60.14 ± 5.88	0.121 ± 0.098
N: female vs A: male	62.79 ± 12.27	0.290 ± 0.384
average	61.46 ± 9.71	0.206 ± 0.293

LDA results and corresponding ITR averaged over participants are shown in Table 2. Single trial classification revealed an average classification accuracy of 61 % with an ITR of 0.2 bit/min. For the single trial analysis (AV 1), the classification accuracy did not differ significantly between testing conditions (AV 1: $F(1,9)=0.25$; $p=0.63$). To see how valid this statement is for classifications of signals with a higher SNR we increase the time-window of the classifier by averaging over three (AV 3) and six (AV 6) random epochs.

Doing so the mean accuracy increases to 70 % (AV 3) and 79 % (AV 6) respectively (AV 3: ITR=0.191 ± 0.206 bit/min; AV 6: ITR=0.176 ± 0.153 bit/min). By increasing the time-window the classification accuracy starts being significantly higher for the testing condition “N: female vs A: male” (AV 3: F(1,9)=4.6; p=0.06; AV 6: F(1,9)=8.1; p=0.02). Figure 12 shows the density of classification accuracies averaged across participants for the two test conditions and three different analysis time-windows. The results revealed a broad variation of about 40 % in classification accuracy depending on subjects. It can be seen that the attention to the male speaker is classified with a higher median accuracy than the attention to the female speaker with rising amount of averaged epochs.

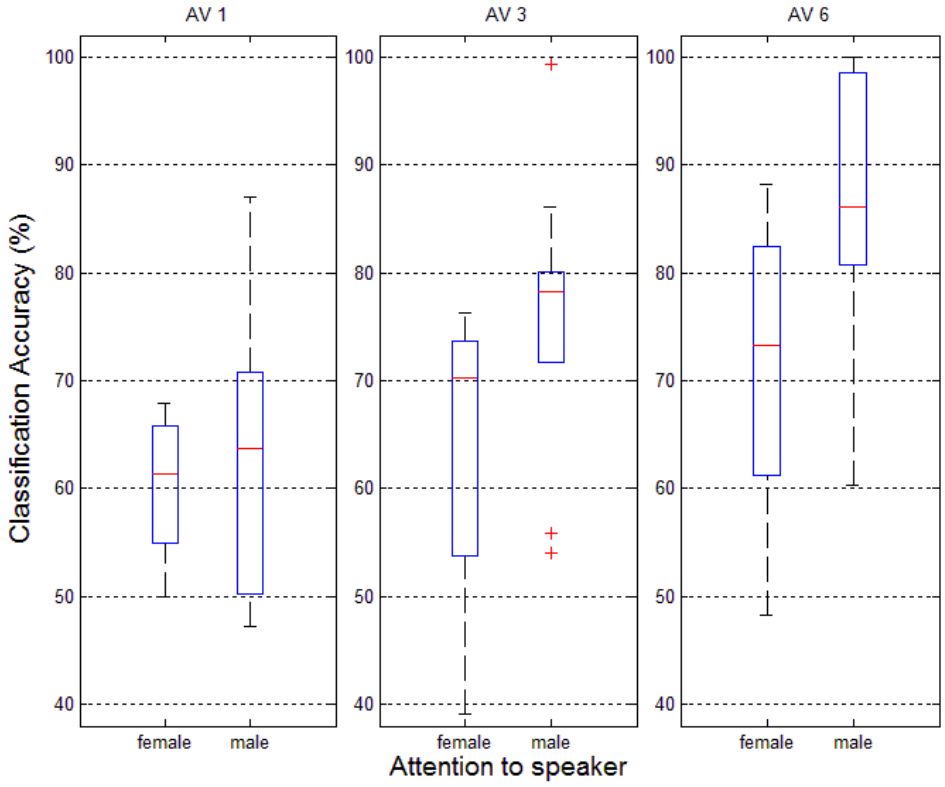

Figure 12: Density of classification accuracies of all participants (1-10) for all conditions (attention to female / male speaker) and analyses (AV 1: single trial; AV 3: average over 3 random epochs; AV 6: average over 6 random epochs).

3.5 Discussion

The results show that streams of spoken syllables evoke ASSR that are further affected by attention. Statistical analyses revealed that this effect of attention is nearly significant, while the gender of the speaker, respectively syllable rate in general, has no significant effect. Single trial classification of the recorded data showed on average an accuracy of 61 % correct with ITR of 0.2 bit/min (for the best subject

in the condition “attend to female speaker” 87 % correct with ITR of 1.3 bit/min). So it is possible to construct an auditory BCI that is based on focusing attention on streams of spoken syllables.

Classification of averaged data sets improved the average accuracy to 70 % (set of three random trials) and 79 % correct (set of six random trials) by keeping the ITR at about 0.2 bit/min. rANOVA of the six trial average results shows a significant dependence of classification accuracy and testing condition. So this shows that it might be easier to detect the direction of attention, when the listener is attending to the male speaker than to the female speaker. This effect is in line with some of the participants reporting that the syllables spoken by the male speaker are easier to follow than the ones spoken by the female speaker. This observation is corroborated by the behavioral data documented by the participants during the measurement. The amount of miscounted syllables /ti/ (female speaker) summed over all participants was 22, whereas in contrast the amount of miscounted syllables /ku/ (male speaker) has only been 10. In contrast to this hypothesis there was no significant dependence of the ASSR to speaker respectively syllable rate found in the averaged data, which also can be seen in Figure 11.

Kim et al. (2011) constructed an auditory BCI based on trains of pure tone bursts with presentation rates of 37 and 43 Hz. In one comparable condition, an analysis window of 20 s and two features were used for classification (p ca. 80 %, T=20 s). Comparing this to our averaged single trial results shows that the paradigm using streams of real spoken syllables presented in rates around 2-3 Hz yielded on average about 20 percent points lower classification accuracies. Increasing the analysis window of the syllable paradigm to 120 s (average over six trials) enhances the performance to comparable accuracy of about 79 %. This classification results can be explained by the properties of ASSR to signals with low modulation frequencies (like speech-rhythm/-envelope). Picton et al. (1987) reported, that ASSR evoked by amplitude modulated tones are most reliably recorded at modulation rates near 40 Hz. EEG recordings of ASSR to signals with low modulation frequencies contain a higher noise level and require, due to the slowness of the modulation signal, a longer time for the analysis.

Nakamura et al. (2013) recently presented an auditory BCI using artificial speech sentences generated by text-to-speech software with an amplitude modulation of about 40 Hz (modulation depth 50 %). They achieved an average classification accuracy of 78.6 ± 5.32 % with an ITR of 2.71 bits/min. Comparing this to the classification result of the syllable paradigm using an analysis window of 120 s, the performance in accuracy is similar, but the ITR of about 0.18 bits/min is much lower due to the long analysis time.

Hill et al. (2014) used a different approach for their auditory BCI. They classified event related potentials (like N1, N2 and P3) to streams of synthesized words coming from two different directions simultaneously (stream one (female voice): „yes“, „yep“; steam two (male voice): „no“, „nope“). They achieved a classification accuracy of 77 % on average using an analysis window of only 15 s length. Our syllable paradigm is on average less accurate by 15 % while using a five seconds longer analysis window (comparably accurate using a four times larger window). Hill et al. further tested their paradigm with

abstract stimuli using streams of beeps (512 Hz one direction, 768 Hz the other) and discovered, that the more natural and intuitive synthetic words are at least as effective as the beeps.

O'Sullivan et al. (2014) presented their participants two different stories from two directions simultaneously, one to the left and the other to the right ear. Participants were asked to attend to one story and to ignore the other. O'Sullivan et al. were able to estimate the envelope of the attended story from the EEG of the participants using the stimulus-reconstruction approach. A detector correlated the estimation of the envelope to the actual envelopes of the two stories and detected the direction of attention with an average accuracy of 82% (60 s analysis window). Our syllable paradigm performs 12 percent points less accurate using the same analysis window.

Visual BCI like for example the visual P300-Speller of Donchin et al. (2000; p=90 %, ITR=4.8 bit/min) are capable of high performance after short time windows. Further examples for high performing visual BCI systems are demonstrated by Hoffmann et al. (2008) and Sellers and Donchin (2006). They are using different pictures (television, telephone, lamp …; Hoffman et al.) or Words („yes", „no", „pass", „end"; Sellers and Donchin) in an visual oddball paradigm to classify P300 responses. Current visual BCI are outperforming auditory BCI with respect to classification accuracy and ITR (Furdea et al., 2009). One possible explanation for this effect could be a higher SNR in the measured EEG signal due to the bigger cortical surface of the visual system compared to the auditory cortex. Nevertheless BCI systems that are driven by auditory attention on acoustical stimuli are offering an alternative for physically disabled people.

Our results and the study of Nakamura et al. (2013) and Hill et al. (2014) show that it is possible to use synthesized speech or even spoken syllables as input for an auditory BCI. O'Sullivan et al. (2014) realized an auditory BCI based on natural speech using approaches like stimulus-reconstruction. This is a promising basis for further BCI studies to put the focus on natural speech as an input. Hill et al. reported that their subjects complained about the difficulty of understanding a paradigm based on artificial stimuli and further reported that the used beeps were harsh and mildly unpleasant. To make the use of such an interface as easy as possible for the patient, the focus should be on keeping the signal as natural as possible. This is a further argument against using amplitude modulation on speech signals to evoke ASSR. Stimulus-reconstruction methods can be used to open the opportunity for a speech driven BCI, but they are based on linear regression and least-square error methods for which it is necessary to have total information of the input signal. Using natural features of speech like relevance or speech-rhythm to evoke classifiable EEG potentials could be a big benefit for hearing impaired listeners using auditory BCI in the future.

3.6 Conclusion

Our paradigm showed that it is possible to classify ASSR evoked from streams of spoken syllables. For this reason speech-rhythm seems to be a feasible feature for a BCI to detect which of two speakers a listener is attending to. Nevertheless with an average ITR of 0.2 bit/min it would be necessary to wait 5 minutes to detect the direction of interest. Such a performance is not good enough for a BCI in a real life environment. In the future we are planning to optimize the classification methods for ASSR evoked by spoken sentences. The goal is to leave the speech signal unmodulated and natural as it is and to measure the ASSR to the speech-rhythm itself. In fact, in a real cocktail party situation we will face a lot of different problems, like for instance a in time leading competing speaker will automatically draw the attention of the subject away from the person he is trying listen to (Choi et al., 2014). Nevertheless, it should be possible to minimize the analysis window of the optimized classifier to enhance the performance of the BCI to a level that it could be used to control a hearing aid in a real life situation.

3.7 Acknowledgments

We would like to thank Stefan Debener and Maarten De Vos for fruitful discussions about BCI and signal processing. Furthermore we thank Christiane Thiel for her support and Tobias de Taillez for his contribution to the predecessor study. This study was funded by the Cluster of Excellence „Hearing4all” at the University of Oldenburg as well as the PhD program „Signals and Cognition” (Niedersächsisches Ministerium für Wissenschaft und Kultur). The contribution of Mathias Dietz was funded by the European Union under the Advancing Binaural Cochlear Implant Technology (ABCIT) grant agreement.

4 Auditory BCI based on a simple classification approach, using correlation between speech envelope and single-trial EEG, for sentence and speaker segregation.[2]

4.1 Abstract

Brain Computer Interfaces (BCI) are detecting and classifying neurophysiological responses in order to control electronic devices. Most BCI systems are based on visual stimulation or imagined muscle movement, only a few apply auditory input. Because auditory-based BCI are independent of visual or mobile skills, they could be an alternative for visually or physically disabled people.

This paper investigates the accuracy and speed of an auditory BCI approach using a simple classifier based on correlation between speech envelope and evoked single-trial EEG to segregate different speakers and sentences. Ten different classification conditions are investigated in total. In five conditions a segregation of two different speakers and six diotically presented sentences in quiet is tested. In five further conditions a speech-in-noise stimulus with a flat envelope and a non-speech stimulus, mimicking the envelope and the overall spectrum of one sentence, are added to segregate them from either a single sentence spoken from a male or female speaker in quiet.

Averaged over ten volunteers the single-trial classification of eight test conditions show significant above chance level accuracies using an analysis window of 1.6 s. For example, the segregation of six different test sentences shows a significant median recognition rate of 27.6 % (chance rate 16.7 %) with a corresponding information transfer-rate of 2.0 bits/min.

In conclusion it is possible to segregate sentences or speakers by classifying correlation between speech envelopes and evoked single-trial EEG using an analysis window of only 1.6 s duration, even under conditions with background noise. With this simple correlation based BCI approach it is further possible to achieve comparable performance to recent published speech driven BCI approaches based on adaptive filter methods.

---

[2] This Chapter is submitted as:

da Silva Souto C., Mauermann M., Kollmeier B.: Auditory BCI based on a simple classification approach, using correlation between speech envelope and single-trial EEG, for sentence and speaker segregation. *PLOS ONE submitted 11.04.2018*.

4.2 Introduction

A Brain Computer Interface (BCI) is defined as a device using brain activity to control electronic devices. Therefore, the BCI has to detect and classify neurophysiological changes due to specific tasks or stimulation and finally activate an external system respectively (Vidal, 1973; Brunner et al., 2011). In this way, BCI can be used to control external devices without manual interaction. The majority of BCI systems are based on neuronal responses to visual stimuli or use neuronal activation patterns that occur while a certain stimulus is perceived or while a specific action is imagined. Since 2006, several studies have investigated BCI controlled by auditory evoked potentials (e.g., Halder et al., 2010; Kim et al., 2011). Because auditory-based BCI are not dependent on visual skills, mobility or posture of the head and eyes, they could be an alternative for people that are visually disabled or completely paralyzed (as suggested by Sellers and Donchin, 2006; Halder et al., 2010; Kim et al., 2011). In the future BCIs might be used to control hearing devices, e.g. by selecting specific programs or filters in a hearing aid according to an acoustic object, like a specific speaker, that the listener is attending to and which is identified automatically by the BCI. For those BCI applications it would be necessary to focus on brain activity that is evoked by features of respective acoustic stimuli. The most important class of stimuli in this field is speech. Therefore, an auditory BCI approach is investigated here which is based on correlation between speech envelope and evoked EEG to segregate different speakers and sentences. Aiken and Picton (2008) discovered that the low frequency components (below 7 Hz) of speech envelopes are basically directly represented in the evoked electroencephalogram (EEG) of the auditory cortex. Further studies like O'Sullivan et al. (2014) and Ekin et al. (2016) used regression methods to estimate the envelope of a presented speech signal from the evoked EEG recordings (aka. stimulus-reconstruction method; Rieke et al., 1995; Mesgarani et al., 2009). O'Sullivan et al. (2014) simulated a competing speaker situation by presenting two spoken stories from two different talkers and directions simultaneously. Their BCI approach utilized the correlation of the estimated signal with the actual envelopes of both speech signals. With this method they were able to detect which story a listener was selectively attending to. The performance of a BCI depends on the accuracy and the speed of its classifier. Stimulus reconstruction methods that have been described by O'Sullivan et al. (2014) and Ekin et al. (2016) depend on rather long analysis windows of about 60 s. The speed and accuracy of an auditory BCI is critical to be feasible for controlling a hearing aid in a natural environment. The major goal of the current study is to investigate to what extent the ITR and latency of a speech-driven BCI as e.g., found in O'Sullivan et al. (2014) and Ekin (2016) can be preserved or even improved, using a simple correlation based classification approach depending on an short analysis window duration corresponding to the duration of a spoken sentence (here 1.6 s). Therefore, we segregated different spoken sentences or speakers, that are presented one at a time in quiet by directly classifying the correlation between the specific speech envelope and their corresponding single-trial EEG without the use of adaptive filter methods or high computational neural network approaches (Kottaimalai et al., 2013).

With respect to possible future applications of BCIs in hearing aids, realistic scenarios need to be taken into account, especially the influence of noise has to be considered. This includes external noise disturbing the acoustic stimulus and its envelope, as well as, an increasing internal noise related to increasing hearing loss, which may distract the measurable envelope characteristics in the EEG. Petersen et al. (2017) focused on this aspect by investigating the correlation functions between speech envelopes and the evoked EEG data for different signal-to-noise ratios (SNR; speech masked by another speaker) and different levels of hearing loss. They described that for listeners with varying degrees of hearing loss the neuronal tracking of the attended speech signal can be obtained more reliably than neuronal tracking of the ignored speech signal. Overall, correlations are reduced with decreasing SNR and increasing hearing loss. Since they did not implement a classifier in their study, it remains unclear to which extent correlation based classifiers are directly affected by these findings as well. Assuming the brain involves very specific processing strategies to enhance incoming speech against other stimuli, such strategies could enhance the EEG response to the speech envelope as well, while the contributions of other signal components (like noise) to the EEG are attenuated. To provide a first test of these assumptions, we added to our BCI approach a speech-in-noise condition that provides clearly intelligible speech. By adding a masking noise to one sentence and leaving the underlying speech envelope unchanged, a mixed signal with a flat envelope is created. If there is any kind of noise reduction mechanism in the brain which is reflected in the EEG recording and the classifier can use this information, it should be possible to segregate the underlying sentence of the noisy stimulus from another clean speech sentence, even when the classifier is trained with the clean envelopes of both sentences respectively. If this noise reduction mechanism is working effectively the resulting segregation accuracy should still be comparable to the condition when both sentences are presented as clean speech. Another related question is: (1) Does the BCI accuracy depend basically on the specific characteristics of speech envelopes, or (2) is it important for the structure of the resulting EEG signals that the respective stimulus is perceived as intelligible speech explicitly? To examine this question a non-speech stimulus mimicking the envelope and the overall spectrum of one sentence spoken by a male speaker was designed. It is tested whether the segregation from any of the sentences and this non-speech stimulus shows an accuracy that is comparable to the segregation from the original sentence spoken by the male speaker. If so, this would underline the importance of the speech like envelope for the classifier, while the intelligibility of the speech might have less effect.

4.3 Methods

4.3.1 Participants

Initially twelve healthy volunteers took part at this study (5 female, 7 male, mean age: 27, range: 19-37 years). All subjects had hearing thresholds of 20 dB HL or better at the audiometric frequencies between 125 Hz and 10 kHz (Auritec AT900), except for one volunteer who had a threshold in quiet of 25 dB HL at 8 kHz on the right ear. All subjects were right-handed according to the questionnaire by Oldfield (1971) except for one participant who obtained a score of 0 (ambidextrous). No reports of any present or previous psychiatric or neurological disorder were given. Participants were paid for participation. Written consent was obtained from each participant prior to the experiments. The experiments were approved by the local ethics committee of the University of Oldenburg.

4.3.2 Stimuli

The speech material used in this study was selected from a pool of the Oldenburg-sentence test (Wagener et al., 1999), in a version specially recorded as fluently spoken sentences for testing of automatic speech recognition systems (Meyer et al., 2015). All the sixteen selected sentences spoken in German are shown in Table 4 (see appendix). Eight sentences were spoken by a single female (indicated as F1 … F8) and eight by a single male speaker respectively (indicated as M1 … M8). The duration of the spoken sentences varies from about 1.6 s up to 2 s. Their recording was extended to a duration of 2 s by zero-padding at the end.

All sentences share the same syntactical structure, starting with a name followed by a verb, a numeral, an adjective and an object. This regular structure may provide an additional obstacle for classification purposes. For a partial compensation and in line to studies with less structured speech stimuli, the sentences of the test stimuli F1, F2, F3, M1 and M2 are kept as different as possible to make a classification easier. Therefore, no word of F1, F2, F3, M1 or M2 occurs in one of the other sentences. M3 provides the same sentence as F3, while being spoken by the male instead of the female speaker. If the investigated cortical EEG responses are affected not only by the characteristics of the physical speech stimulus (differences in pitch, speed rate, intonation etc.) but in addition by the content of the sentences, a difference in segregation performance between M3 and F3 compared to the segregation of other sentences might occur. The test stimuli F1, F2, M1, M2 and F3 (M3) have different plosive onsets. The varying words and speakers provide stimuli with rather different amplitude modulation spectra.

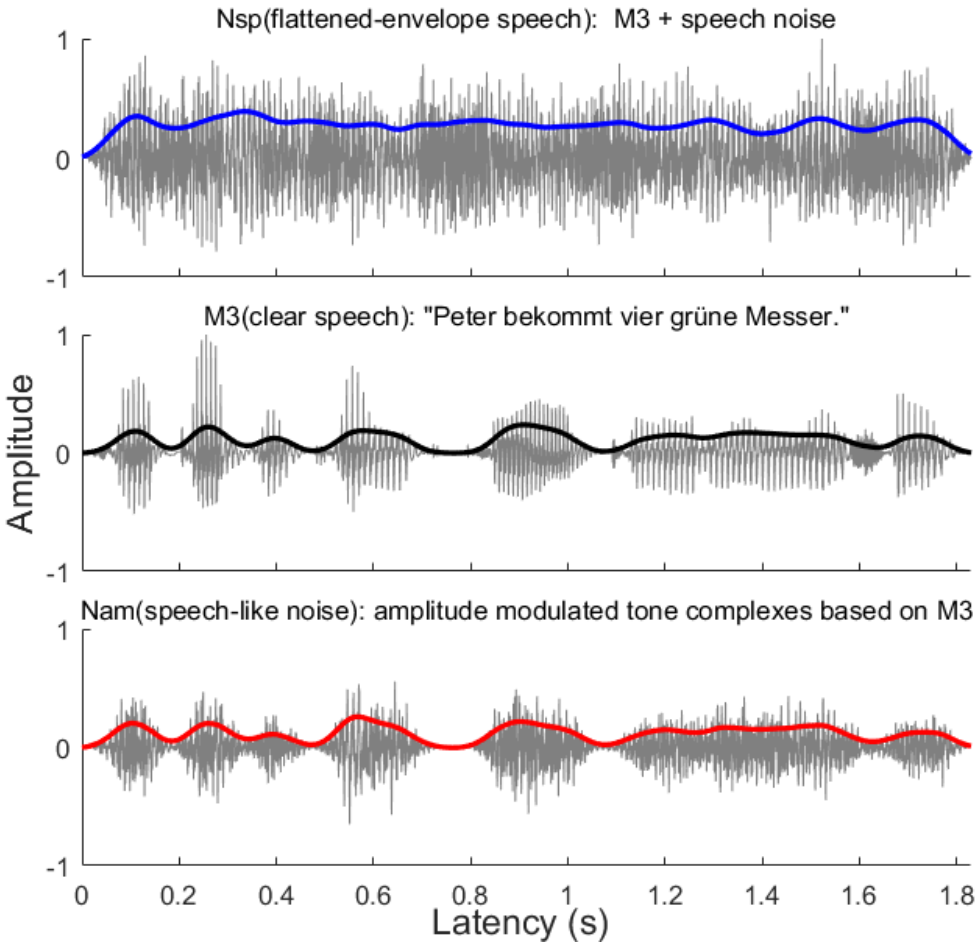

Figure 13: Middle: Time structure of M3 (clear speech). Top: Flattened-envelope speech (Nsp) generated by addition of speech noise to M3, while leaving the underlying speech envelope of M3 unchanged. Bottom: speech-like noise (Nam) generated by amplitude modulation of tone complexes. The tone complexes are based on the fundamental frequency of the male speaker and the modulation on the envelope of the same speaker. All three time signals are shown with their 8Hz low pass filtered envelope.

Two additional stimuli Nsp and Nam were generated to investigate the influence of noise and the relevance of speech characteristics for classification. Nsp is generated by adding a male speech noise to M3 to mask the underlying speech envelope, providing an SNR of 10 dB. The male speech noise is generated as a random phase tone complex. Therefore, the spectra of all sentences in the speech corpus spoken by the male speaker are calculated using a time window of 2s (according to the length of the longest sentence in this speech corpus). The amplitudes of the spectrum of the male speech noise are set in respect to the average level spectra of the speech corpus, their phases are randomized and the spectrum is transformed into the time domain. The noise added to the speech stimulus in the Nsp signal is cut out for the duration of the speech stimulus and ramped with 92 ms hanning shaped flanks on both sides. The stimulus Nam was designed as a noise stimulus approximating both the envelope and the spectrum of the M3 stimulus by combining the amplitude spectrum of M3 with a random phase spectrum. After transformation of this

complex spectrum into the time domain the resulting signal was multiplied with the envelope of the target signal M3. It should be noted that Nsp remains a clearly intelligible sentence with a flattened envelope (see Figure 13 top panel) whereas Nam provides no intelligible speech but has an envelope (compare Figure 13 middle and bottom panel) and spectrum which is highly comparable to the sentence M3.

The stimuli F1, F2, F3, M1, M2 and M3 are used as test stimuli for sentence and speaker segregation. To be in line with recent attention based BCI approaches like the one designed by O'Sullivan et al. (2014), the subjects had to actively attend to the presented sentences and answer related questions. Further benefits for the EEG evoked during active listening tasks in contrast to passive listening have been shown by Bennington et al. (1999) and O'Sullivan et al. (2015). The remaining ten sentences F4-F8 and M4-M8 were added to make this task for the subjects less easy and less repetitive. In addition, M3 is used as a basis to test possible effects of gender (F3), noise (Nsp) or relevance of speech characteristics (Nam) for classification.

4.3.3 Setup

Signal presentation was controlled digitally using custom made scripts in Matlab 2015a (Mathworks). The digital stereo signals were DA converted at a sampling rate of 44100 Hz using a Fireface UCX multichannel sound card (RME), then presented via a TDT HB7 (Tucker-Davis) headphone buffer using ER2 insert phones (Etymotic Research). All stimuli where presented diotically at 70 dB SPL.

EEG data were recorded with the Biosemi ActiveTwo system (7 channels, Electro-Cap International 10-20 system, Parker contact gel). All electrode offsets were set not higher than 20 mV, prior to data recording. Data was collected with the Biosemi software ActiView (6.03) at a sampling rate of 1024 Hz without any additional filtering. In order to allow exact temporal alignment of the EEG recordings with the arrival of the acoustical stimulus at the eardrum, a trigger pulse is played in parallel via the digital output channel (SPDIF) of the Fireface UCX with a delay of 1ms to compensate the acoustical delay of the ER-2 insert-phones. The usage of the ASIO sound drivers of the Fireface UCX is warranting a sample exact synchronization of all output channels of the Fireface UCX. The digital trigger pulses (specific amplitudes indicate the respective stimuli) are converted (using a custom-made converter) into a respective set of TTL trigger pulses to be captured by the Biosemi system synchronously with the EEG data.

The measurements took place in a double walled and shielded booth (IAC) while the participants were sitting in a comfortable chair focussing on a blank screen in front of them. The subjects were instructed to pay attention to the different diotically presented OLSA sentences and to answer related questions by pressing one of two buttons on a box. The questions were randomly generated and shown in-between presentations on a screen to help fix the subject's attention on the presented speech.

4.3.4 Task and Experimental design

The measurement paradigm starts by randomly presenting one of the 18 stimuli included in Table 4. After a break of 300 ms a question in German shows up in the center of the screen asking for a specific piece of information of one random part of the presented sentence. A question for every part of the presented sentence is possible. For example, possible questions subsequent to the presentation of F3 "Peter bekommt vier grüne Messer." (Peter gets four green knives.) would be: "Wie viele Messer?" (How many knives?) or "Wer bekommt Messer?" (Who is getting knives?) etc.. Underneath the question on the bottom left and right two possible answers show up in random order. One is correct and one incorrect, randomly picked out of the pool of sentences for the respective part of the OLSA sentence. The participant has a maximum of 10 s time to respond by pressing one of two buttons on a response box. Immediately after the response is given, the screen turns blank again. The whole process is repeated after a pause of 300 ms. This task was created to provide the subjects attention to the repetitive presented speech stimuli during the EEG recordings.

The measurement took place in a single session. Each session was split into 4 runs of about 15 min duration of EEG recording and breaks of five to ten minutes in between. On each run 196 stimuli were presented in random order (stim: F1-3/M1-3/Nsp/Nam with 22 repetitions each; stim: F4-8/M4-8 with 2 repetitions each). This leads to a total of 88 repetitions per test stimulus (F1-F3, M1-M3, Nsp and Nam).

4.3.5 Data processing

EEG data were referenced to the two averaged mastoid channels ((A1 + A2) / 2). The five electrodes around the vertex (Fcz, C1, Cz, C2, Cpz) are used as analysis channels. Comparable electrode-setups parallel to the auditory pathway are commonly used to measure auditory evoked EEG responses (Picton and Hillyard, 1974b). The raw data are digitally filtered using a 4th order bandpass from 2 to 8 Hz, to reduce electrode drift and high frequency noise. According to Pasley et al. (2012) this frequency band is optimal to observe envelope properties of the presented stimuli in EEG-Data. The EEG Data are separated into epochs according to corresponding trigger-information with epoch length of 1.6 s (according to the length of the shortest sentence in this speech corpus). A baseline correction is applied, estimating the DC off set using the average of the 100 ms before each epoch's onset and subtracting it. After applying an artifact threshold of 40 µV to the EEG data of each channel, to detect and exclude epochs with high amplitude artifacts, an average over the five channels is done. After down sampling the data to a sampling rate of 128 Hz it is sorted by trigger information corresponding to the different test sentences.

In order to provide comparable reference signals for the correlation with the EEG data, the envelope of each stimulus is extracted by calculating the absolute value of the respective analytic signal obtained with the help of a Hilbert transform of the broadband signals. Furthermore, each of the envelopes is low pass filtered at a cutoff frequency of 8 Hz (4th order, zero phase) and down sampled to a sampling rate of 128

Hz. The duration of the final reference envelopes is explicitly limited to the first 1.6 s, according to the length of the shortest test sentence. This limitation is done in conjunction with the limitation of the EEG epochs to 1.6 s to provide only correlation information between stimulus envelope and evoked EEG to the classifier and therefore prevent any unwanted classification strategies e.g., the detection of envelope length.

4.3.6 Classification and Statistical analysis

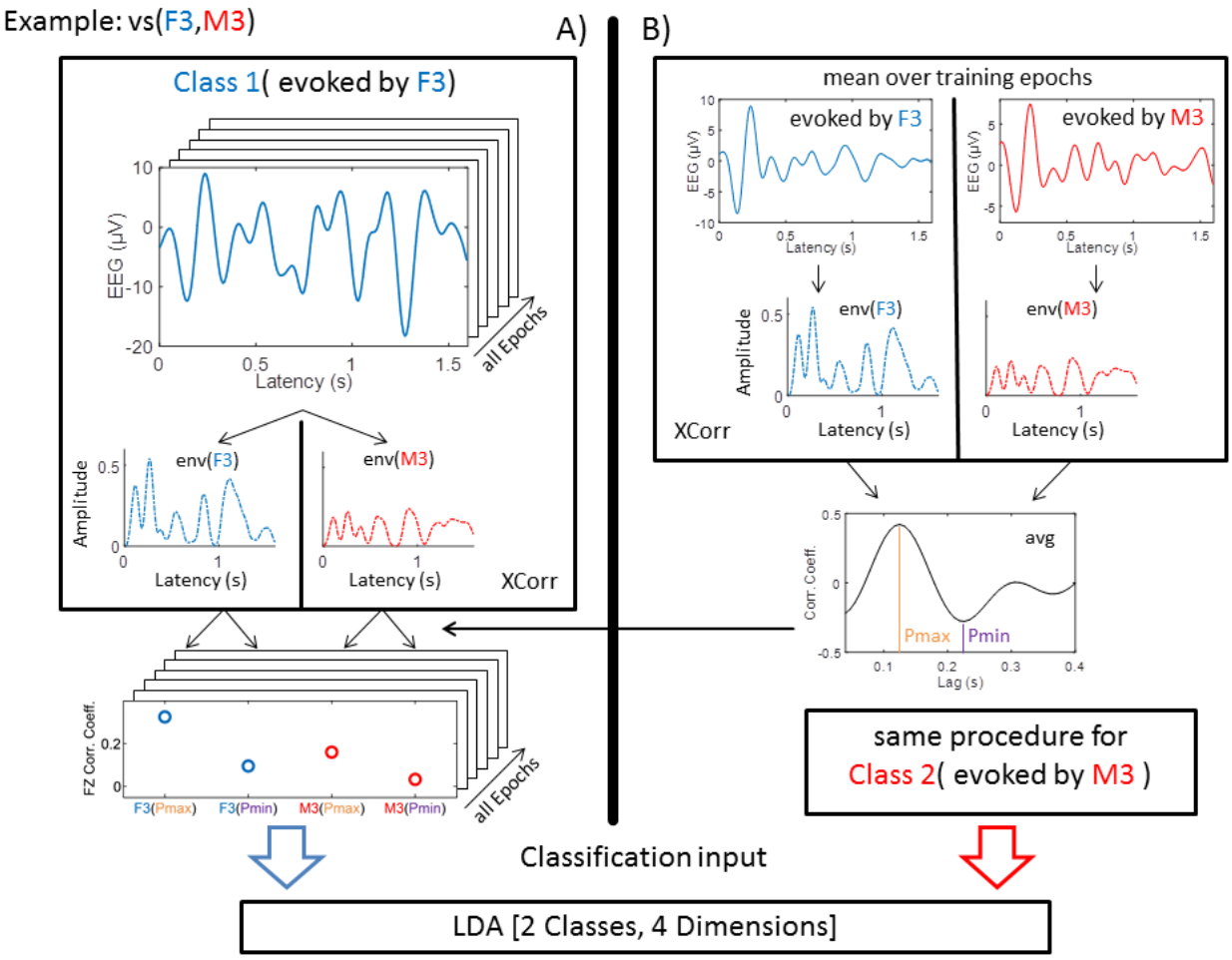

Figure 14: Preprocessing example for the classification input of the first class (EEG evoked by F3) of a linear discriminant analysis (LDA) to discriminate one of two attended speakers, speaking the same sentence. A)Top: Single-trial EEG epochs evoked by the third sentence of the female speaker (F3). A)Middle: The single-trial EEG epochs were cross correlated (mean-removed, normalized [autocorrelation equal 1 at zero lag] and Fisher z-transformed) to the envelopes of each of the two sentences (F3 and M3). B)Top: Mean over EEG epochs evoked by two sentences F3 and M3 that are used for training. B)Middle: Mean over EEG training epochs were cross correlated (mean-removed, normalized [autocorrelation equal 1 at zero lag]) to the envelope of the corresponding sentence. B)Bottom: The resulting Correlation coefficient sequences were averaged and the lags of the minimal and maximal peak (Pmax and Pmin) were detected within a limited lag range of 40 to 400

ms (corresponding to cortical auditory evoked potential; Picton et al., 1974a). A)Bottom: Correlation coefficients corresponding to the single-trial sequences at the specific lag positions Pmax and Pmin are used as Class1 input for the LDA. The LDA input of Class2 follow the same procedure respectively.

According to the limitations in measurement duration a maximum of 88 epochs are available for each condition. Therefore, a linear discriminant analysis (LDA) is used for the data classifications due to its robustness even for a low amount of available training samples per class (as shown e.g. by Hu and Yu, 2011). A 10-fold cross-validation was used to further maximize sample size of the test data. The cross-validation was done for each participant and discrimination condition individually. From each participant, we used the least number of available epochs (Emax; see appendix Table 5) evoked by one of the eight test stimuli (F1-3/M1-3/Nsp/Nam), to provide fair comparisons across the different test conditions. Emax epochs were divided in 10 folds. In each validation 9/10 of the data was used for training and 1/10 for testing (using the Matlab command cvpartition). All classification results across the 10 folds are taken to calculate the classification accuracy for the respective subject and test. Preprocessing of classification input is shown exemplarily in Figure 14 for one test condition (two speakers, same sentence; vs(F3,M3)). In each validation step, correlation coefficient sequences (CCS) are computed with mean-removed and normalized (autocorrelation equal 1 at zero lag) cross correlation between each single-trial EEG epoch and each of the reference envelopes (derived from the stimuli; see Figure 14 A)). To identify the optimal CCS lags for classification, correlation coefficient sequences were computed for each class – between the mean over EEG epochs used for training and the respective signal envelopes - and finally averaged (see Figure 14 B)). We refer to this correlation coefficient sequences of the averaged training data as CCS_{train_avg}. The lags of the minimal and maximal peak of the resulting CCS_{train_avg} are detected (limited Llagrange of 40 to 400 ms corresponding to cortical auditory evoked potential P1 to N2; Picton et al., 1974a) and used to select the LDA input from the CCS (see Figure 14 bottom). All CCS values were Fisher z-transformed before classification. Ten discrimination conditions were used for classification. In the list below the properties of the test conditions are described in the following order: subcategory, label of LDA classes, amount of LDA classes, degree of LDA dimension. For example, in test condition vs(F1,F2,F3,M1,M2,M3), the classifier has to identify one out of six possible sentences F1, F2, F3, M1, M2 and M3, by the correlation data of its respective evoked EEG response and the envelopes of those sentences. The sentences are presented one at a time, i.e., no detection of spatial attention is done.

1. Six sentences: vs(F1,F2,F3,M1,M2,M3); 6 Classes, 12 Dimensions
2. Three sentences: female speaker vs(F1,F2,F3); 3 Classes, 6 Dimensions
3. Three sentences: male speaker vs(M1,M2,M3); 3 Classes, 6 Dimensions
4. Two speakers: all sentences vs(F:[F1 F2 F3],M:[M1 M2 M3]); 2 Classes, 12 Dimensions
5. Two speakers: same sentence vs(F3,M3); 2 Classes, 4 Dimensions

6. Clear vs Noisy: clear male vs flattened-envelope speech vs(M3,N1); 2 Classes, 4 Dimensions
7. Clear vs Noisy: clear male vs speech-like noise vs(M3,N2); 2 Classes, 4 Dimensions
8. Clear vs Noisy: clear female vs flattened-envelope speech vs(F3,N1); 2 Classes, 4 Dimensions
9. Clear vs Noisy: clear female vs speech-like noise vs(F3,N2); 2 Classes, 4 Dimensions
10. Clear vs Noisy: clear female vs flattened-envelope speech (trained with clear speech envelope of male speaker) vs(F3,Nsp[M3]); 2 Classes, 4 Dimensions

To examine the distribution of the classification results of each test and subject, Monte Carlo permutation tests with 5000 iterations were done. At the beginning of each iteration, the order of the EEG data is randomized, then a 10-fold cross-validation is done like described top. The distributions were computed for classifications trained with correct Class-Labels (D_{CL}) as well as for classifications trained with randomized Class-Labels (D_{RL}). To test for statistically significant classification accuracies on single subject level the average value of the D_{CL} has to exceed the 95 % percentile (corresponding to a significance level of p = 0.05) of the respective D_{RL}. To estimate the statistical significance of the classification accuracy when averaged over subjects, a paired-sample t-test (one-tailed) is done between the mean values of the single subject D_{RL} in one vector and the mean values of the respective D_{CL}. To test for statistically significant differences in classification accuracy between test conditions, repeated analysis of variance (rANOVA) were done. A Mauchly-Test was applied to test for sphericity of the data. The assumption of sphericity was not violated in any of the conducted analyses.

Information transfer-rates (ITR) are calculated to indicate the bits per class that are generated by the BCI after the duration of one minute, according to the following equation (N: amount of classes, p: recognition rate, T: duration of epoch; Shannon and Weaver, 1964; Besserve et al., 2007; Speier et al., 2013):

$$\mathrm{ITR} = \frac{60}{\mathrm{T}}[\log_2 \mathrm{N} + \mathrm{p} \cdot \log_2 \mathrm{p} + (1-\mathrm{p}) \cdot \log_2 \left(\frac{1-\mathrm{p}}{\mathrm{N}-1}\right)],$$

$$\Delta\mathrm{ITR} = \frac{60}{\mathrm{T}} \cdot \log_2 \left(\frac{\mathrm{p} \cdot (\mathrm{N}-1)}{1-\mathrm{p}}\right) \cdot \Delta\mathrm{p}.$$

4.4 Results

Two noisy male participants were excluded due to low number of available epochs (Emax) (see Table 5 in appendix), therefore the Data from 10 volunteers entered the analysis. Figure 15 shows the resulting classification rates of the LDA for all participants and test conditions. The corresponding mean classification rates over subjects with standard errors of the mean and ITRs for each test condition are given in Table 3, together with the corresponding statistical results.

The classification of the test condition vs(F1,F2,F3,M1,M2,M3) shows a significant mean recognition rate of 27.64 ± 1.76 % (p = 0.00008) with corresponding ITR of 2.04 ± 0.62 bits/min. In other words, this BCI can

detect which of six sentences a listener attends to, at 100 % accuracy after less than 30 seconds. On single subject level nine of ten subjects show significant above chance level mean recognition rates. The best subject scored an accuracy of 37.20 ± 0.02 %, resulting in an ITR of 6.55 ± 0.05 bits/min. The test conditions related to three sentences and two speakers scored significantly above chance level mean recognition rates, while half or more of the single subjects show significant mean recognition rates.

In the 'Clear vs Noisy' test conditions only vs(M3,Nam) did not score at a mean classification accuracy significantly above chance. In test condition vs(M3,Nsp) no single subject scored higher than chance recognition rates. The rANOVA reveals no significant difference in classification accuracy between test condition vs(M3,Nsp) and vs(M3,Nam) ($F(1,9)=4.19$; $p=0.071$), but a trend is noticeable. Further no significant differences between test condition vs(F3,M3) and vs(M3,Nam) ($F(1,9)=2.83$; $p=0.127$), vs(F3,M3) and vs(M3,Nsp[M3]) ($F(1,9)=2.82$; $p=0.128$) were found.

The behavioral test results are shown in Table 6 (see appendix) for all subjects and test conditions. All subjects' answers to the questions, corresponding to the stimuli F1-3, M1-3 and Nsp, were nearly 100% correct. A paired-sample t-test (one-tailed) between behavioral results of M3 and Nam revealed that questions to Nam were answered significantly worse than questions to M3 ($p = 0.013$).

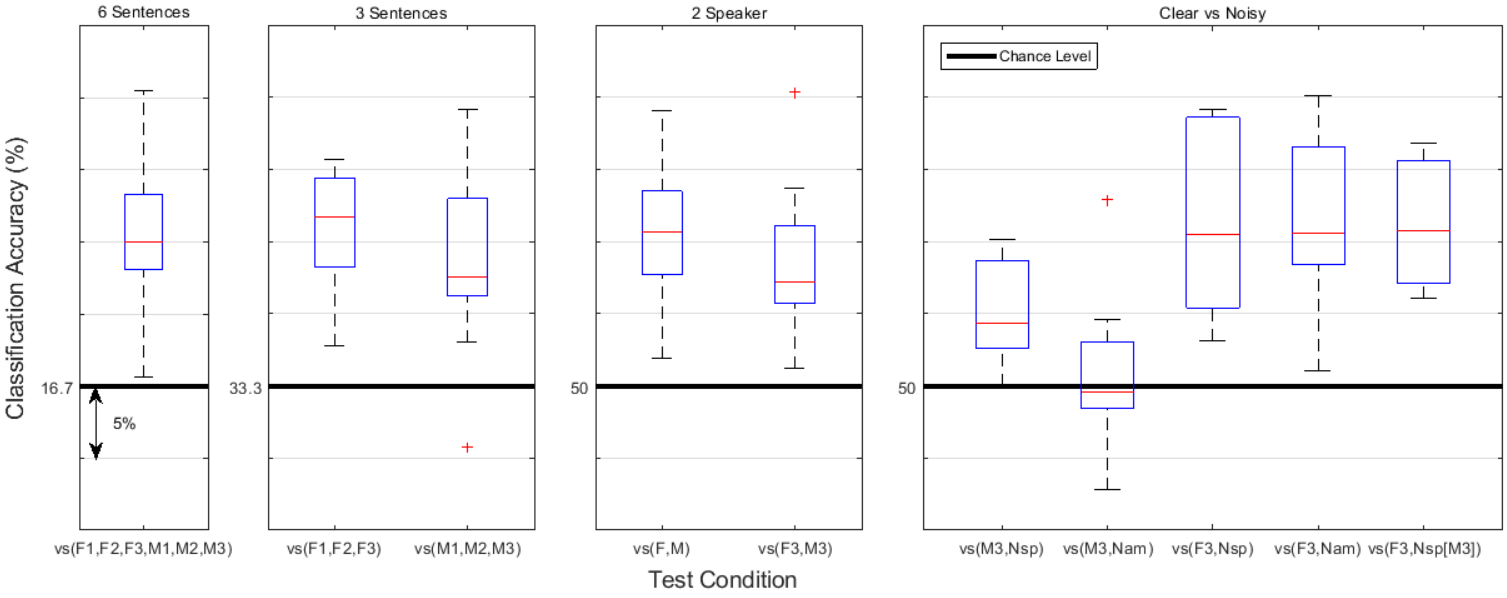

Figure 15: Distribution of achieved LDA classification rates for all participants and test conditions (bottom). Black lines indicate the chance level of each subcategory (top) which are shown on the corresponding y-axes. Dotted lines indicate a 5 % step of increase or decrease in classification accuracy.

Table 3: Averaged results of the Monte Carlo permutation tests are shown for each test condition. The mean classification accuracies of the distributions, obtained by training with randomizes Class-Labels (Chance) and correct Class-Labels (Acc), are averaged over subjects and presented with their standard errors of the mean. Further the corresponding statistical test results are presented. ITRs corresponding to the mean classification accuracies are shown for each test condition with standard errors of the mean.

Test Condition	Chance (%)	Acc (%)	t-test	Single	ITR (bits/min)
vs(F1,F2,F3,M1,M2,M3)	16.66 ± 0.01	27.64 ± 1.76	p = 0.00008	9/10	2.04 ± 0.62
vs(F1,F2,F3)	33.33 ± 0.01	43.86 ± 1.44	p = 0.0003	8/10	1.29 ± 0.35
vs(M1,M2,M3)	33.34 ± 0.02	42.02 ± 2.19	p = 0.007	5/10	0.89 ± 0.44
vs(F,M)	49.99 ± 0.01	60.33 ± 1.62	p = 0.00006	8/10	1.16 ± 0.37
vs(F3,M3)	49.99 ± 0.02	58.35 ± 1.79	p = 0.001	3/10	0.76 ± 0.33
vs(M3,Nsp)	49.98 ± 0.02	54.91 ± 1.16	p = 0.005	0/10	0.26 ± 0.12
vs(M3,Nam)	49.97 ± 0.02	50.49 ± 1.75	p = 0.47	1/10	0.00 ± 0.02
vs(F3,Nsp)	50.01 ± 0.03	61.25 ± 1.89	p = 0.0004	6/10	1.38 ± 0.47
vs(F3,Nam)	50.01 ± 0.02	61.30 ± 1.75	p = 0.0003	6/10	1.39 ± 0.44
vs(F3,Nsp[M3])	50.02 ± 0.02	61.03 ± 1.28	p = 0.00004	5/10	1.33 ± 0.31

4.5 Discussion

Overall, the results show that it is possible to segregate a speaker or a specific spoken sentence that a person actively listened to, from a known pool of speakers or sentences by using a simple classifier based on correlation coefficients between evoked single-trial EEG responses and the envelopes of all different speech stimuli considered in the respective task. However, due to the poor SNR of a single-trial EEG epoch and its short length of 1.6 s, an identification based only on the highest correlation coefficient leads to a classification at chance level. Therefore, the classifier has been extended to use input-vectors containing the correlation coefficients between a single-trial EEG epoch and the envelopes of all sentences of the specific test condition (e.g. segregation of six sentences: 6 Classes, 12 Dimensions). This approach provides a successful auditory BCI that can segregate one out of six sentences spoken by two different speakers, one

of three sentences spoken by a single speaker, two speakers each speaking three sentences and one sentence spoken by two different speakers. The median classification accuracies of these test conditions are statistical significantly above chance level ($p = 0.0008$). The presented BCI approach scores on average an ITR of 2.04 bits/min in the six sentence condition and ITRs in the range of 0.8 to 1.3 bits/min for the three sentences and two speaker conditions (see Table 3 for details).

O'Sullivan et al. (2014) simulated a naturalistic multi-speaker scenario by presenting two spoken stories dichotically to their participants. Their task was to attend to one story while ignoring the other one. Their approach allows the detection of selective attention on a single-trial basis (about 60 s) by classifying correlations between the envelope of the attended story and an estimated envelope signal. The estimate is generated with a linear regression model using information of the evoked EEG and the envelope of the attended story. They achieved on average a recognition rate of 89 % when using an analysis window of about 60 s, which corresponds to an ITR of 0.5 bits/min. The two speaker classification condition vs(F,M) of our BCI approach allows for the most fair comparison of both approaches. The two speaker condition scores on average a 0.66 bits/min higher ITR while relying on an over 58 s shorter analysis window.

The reconstruction filter used by O'Sullivan et al. (2014) is trained on the attended speaker alone, so their detector is only able to answer a very specific question: Is this the EEG evoked by the attended speaker or is it evoked by something else? Ekin et al. (2016) tried to overcome this by extending O'Sullivan's BCI approach using a similar auditory paradigm. In contrast to O'Sullivan et al. (2014), they built reconstruction filters for both, the attended and the unattended story based on a Capon minimum variance distortionless response beamforming method. With this method, they scored comparable classification accuracies (86.1 %) for an attended story decoder and could outperform O'Sullivan et al. (2014) in an unattended story decoder (80.6 %). Still relying on an analysis window of 60 s length, they achieved an ITR of about 0.42 bits/min for the attended story decoder and 0.29 bits/min for the unattended story decoder. In the two speaker condition vs(F,M) we score on average a 0.74 bits/min higher ITR than their attended story detector.

Biesmans et al. (2017) showed another speech driven BCI approach based on O'Sullivan et al. (2014). They enhanced the classification performance by solving a single least squares estimation over the entire training data set, instead of averaging over a set of least squares estimations for each single trial. To optimize the envelope extraction for the classifier, a comparison of various auditory modelling processes was done, using an analyzation Window of 30 seconds length. A simple model, based on a combination of power law relation (loudness model) and gammatone filter bank, showed the best classification accuracy of 81.5 % leading to an ITR of 0.62 bits/min. On average we score a 0.54 bits/min higher ITR in the two speaker conditions vs(F,M) while relying on an over 28 s shorter analysis window.

Recently Haghighi et al. (2016) presented yet another auditory BCI paradigm based on speech. They presented two competing audiobook stories in two different conditions to their four participants. In the

first condition, the two stories were presented diotically simultaneously and in the second they were presented dichotically. The task was to attend to one given story and ignore the other one. Haghighi et al. (2016) calculated cross correlation functions of the evoked EEG and the envelopes of the stories and classified them using a Regularized Discriminant Analyses. They tested varying analysis windows and achieved best results for a window length of 58 s. In the diotic condition they achieved an average classification accuracy of 95 % leading to an ITR of about 0.74 bits/min. Speaker discrimination in the dichotic condition scored an average accuracy of 86 %, resulting in an ITR of about 0.43 bits/min. We were able to segregate one of two speakers with a 0.73 bits/min higher ITR relying on a 56 s shorter analysis window. Haghighi et al. (2016) showed that a correlation based BCI approach that is not relying on adaptive filter methods is able to segregate competing speakers, but it still depends on long analysis windows.

Although using the ITR as a performance measure allows for a rather fair comparison across studies, it should be stressed that in the studies of O'Sullivan et al. (2014), Ekin et al. (2016), Biesmans et al. (2017) as well as Haghigi et al (2016), the spatial attention to one of two competing streams of speech was investigated, while the current study measures the performance of a BCI approach for speaker or sentence segregation, presenting only one sentence at a time. Future studies should investigate the performance of a basic correlation based classification approach in a competing speaker situation and further test its accuracy for short analysis windows.

The comparison of performance for several further conditions gives some insight into the relevance of the stimulus envelope for classification and the importance of using speech signals versus artificial stimuli with speech mimicking properties as an input for the BCI. To indicate the relevance of the stimulus envelope, it is necessary to test the performance of the classifier in a partially noisy scenario only having the information of the envelopes of the clear sentences. Therefore we classified in test condition vs(F3,Nsp[M3]) the correlation between the EEG evoked by a clear sentence from the female speaker (F3) and the respective stimulus envelope and the correlation between the EEG evoked by the same sentence spoken by a male speaker with speech noise (Nsp) and the envelope of the same sentence without speech noise (M3). The results show a median significantly above chance level classification accuracy of 61 % ($p = 0.00004$) with a corresponding ITR of 1.33 bits/min. This performance is comparable to the conditions vs(F3,M3) and vs(F3,Nsp), which indicate the capability of the classifier to generalize to speech in noise when trained with correlation data of evoked EEG and clean speech envelopes only. A possible interpretation is that the speech envelope is somehow preserved or even enhanced in the EEG signal, compared to the computed envelope of the presented speech in noise stimulus, which benefits the specifically trained classifier.

To see if the EEG simply replicates very characteristic envelopes of a stimulus independently if it is intelligible speech or not we classified a sentence spoken by a male speaker (M3) with an artificial speech like noise (Nam) mimicking the envelope and overall spectral content of M3. Assuming that the classifier

focuses predominantly on the envelope information of the stimuli, M3 and Nam should evoke very similar representations of their envelopes in the EEG, i.e. a classification should hardly be possible. Subsequently, the performance for a segregation of another clean sentence (e.g. F3) and M3 or Nam respectively should be comparable. Both assumptions are confirmed by the results and could be explained by the similar envelopes of M3 and Nam (the correlation of the envelopes of M3 and Nam is 0.955). The psychoacoustic data showed, that questions to Nam were answered significantly worse than questions to M3 ($p = 0.013$). However, although Nam is not intelligible speech, the responses in the listening task (for Nam presentation, valid M3 questions were asked) were clearly above chance level for ten out of twelve subjects. These subjects may have identified Nam as M3 due to the very similar signal properties or they guessed the right answers due to the repetitive presentations of sentences with identical content (F3, M3 and Nsp) during the experiment.

4.6 Summary and Conclusions

In this study, we show that it is possible to segregate sentences or speakers clearly above chance level by classifying correlations between envelope of presented speech signals and evoked EEG on a single-trial basis using an analysis window of only 1.6 s duration. The performance of the presented BCI approach is comparable to recent published speech driven BCIs using longer analysis windows and adaptive filter methods. Furthermore, it appears to be able to generalize to speech in noise when trained with correlation data obtained by clean speech envelopes end evoked EEG only. Comparing the classification results of clean speech versus speech in noise on the one hand and clean speech versus a stimulus with speech mimicking properties on the other, provides some indication that the envelope of the acoustically presented stimulus is the most relevant factor for correlation based classification of single-trial EEG data. For future applications, it is still necessary to further improve the performance of current auditory BCI systems in terms of speed and accuracy to make them more reliable in natural acoustic situations with competing speakers, disrupting noises and hearing-impaired users.

4.7 Acknowledgments

We would like to thank Florian Schmidt and Oliver Behler for fruitful discussions about signal processing and statistical analysis of EEG data.

4.8 Appendix

Table 4: Gender, synonyms and German content of the 18 used stimuli.

Speaker	Stimulus	
female	F1:	"Doris hat zwölf große Blumen."
	F2:	"Britta gibt sieben rote Sessel."
	F3:	"Peter bekommt vier grüne Messer."
	F4:	"Tanja mahlt acht weiße Dosen."
	F5:	"Ulrich gewann zwölf rote Messer."
	F6:	"Britta schenkt zwei weiße Schuhe."
	F7:	"Wolfgang sieht vier nasse Schuhe."
	F8:	"Stephan gibt fünf grüne Steine."
male	M1:	"Tomas kauft neun schwere Tassen."
	M2:	"Kerstin nahm acht kleine Dosen."
	M3:	"Peter bekommt vier grüne Messer."
	M4:	"Kerstin bekommt zwei rote Tassen."
	M5:	"Tanja kauft acht nasse Messer."
	M6:	"Ulrich verleiht sieben große Ringe."
	M7:	"Nina bekommt achtzehn weiße Blumen."
	M8:	"Doris gewann sieben teure Bilder."
noise	Nsp:	M3 + speechnoise
	Nam:	amplitude modulated tone complex

Table 5: Least number of available epochs (Emax) evoked by one of the eight test stimuli (F1-3/M1-3/Nsp/Nam) of each subject.

Subjects	SJ 1	SJ 2	SJ 3	SJ 4	SJ 5	SJ 6	SJ 7	SJ 8	SJ 9	SJ 10	SJ 11	SJ 12
Emax	73	75	74	35	64	63	85	64	16	69	77	71

Table 6: Behavioral results of all participants. Amount of correct responses to questions corresponding to the eight stimuli (F1-3/M1-3/Nsp/Nam).

Subjects	F1	F2	F3	M1	M2	M3	Nsp	Nam
SJ 1	88	88	88	88	88	88	88	87
SJ 2	88	88	88	88	88	88	88	84
SJ 3	88	88	88	84	88	86	85	84
SJ 4	86	87	88	88	87	86	88	86
SJ 5	83	86	88	87	88	87	88	82
SJ 6	87	88	86	88	86	87	88	68
SJ 7	88	88	88	88	88	88	88	83
SJ 8	87	87	87	85	87	87	87	80
SJ 9	88	88	88	88	88	88	88	83
SJ 10	87	87	88	87	88	87	87	43
SJ 11	88	87	87	87	88	88	88	51
SJ 12	87	88	88	88	88	88	88	86

5 Influence of speech-simulating noise on a simple correlation based speech driven auditory BCI approach and the relation of classification accuracy to perceived speech.

5.1 Abstract

Brain Computer Interfaces (BCI) based on auditory stimulation can focus on brain activity that is evoked by features of speech. This may provide interesting applications in the future, like the selection of specific background noise filters on a hearing device, according to the ability of the listener to understand speech. This paper investigates the performance of an auditory BCI approach in a speech-in-babble-noise scenario classifying two different sentences at three different signal-to-noise ratios (SNR), while the subjects performed a concurrent psychoacoustic attentive task. Therefore, a simple classification approach with an analysis window of 1.2 s is used based on the correlation of clean speech envelopes and evoked single-trial EEG to test the BCIs ability to generalize to speech in noise and to observe a possible relation between its performance and the human speech intelligibility.

Averaged over twelve listeners, the single-trial classification of all test conditions show significant above chance level accuracies. For example, the classification of two different sentences under the best SNR condition (-11.6 dB) shows a median recognition rate of 62 % with a corresponding information transfer rate (ITR) of 2.2 bits/min (ITR drops by about 3 % for a SNR decrease of 1.5 dB). The classification results further show a significant dependence on the rate of perceived speech.

In conclusion it is possible to segregate sentences under varying levels of speech-simulating noise by simple classification of the correlation between clean speech envelopes and evoked single-trial EEG using an analysis window of only 1.2 s duration. Hence, a performance comparable to recent BCI approaches is achieved. Since a relation between the BCIs performance and the human speech intelligibility exists, a possible future application could be an automatic noise filter selection on a BCI-controlled hearing aid.

5.2 Introduction

A Brain Computer Interface (BCI) can detect and classify neurophysiological changes due to specific tasks or stimulation, to further control electronic devices without manual interaction (Vidal, 1973; Brunner et al., 2011). The majority of BCI systems are based on visually evoked neuronal responses or use neuronal activation patterns that occur while a specific action is imagined. Since 2010, several studies have investigated BCI controlled by auditory evoked potentials (e.g., Halder et al., 2010; Kim et al., 2011). As auditory-based BCI do not rely on visual skills, motor skills or posture of the head and eyes, they could be a further alternative for people who are visually disabled or completely paralyzed (Sellers and Donchin, 2006; Halder et al., 2010; Kim et al., 2011). In the future auditory-based BCIs might be used to control hearing devices, e.g., by pointing specific spatial filters in a hearing aid to an acoustic object, like a specific attended speaker, or selecting a specific background noise filter according to the ability of the listener to understand the acoustic signal, like speech. For such a BCI applications it would be necessary to focus on brain activity that is evoked by features of speech. The current study therefore considers the objective detection of attended speech in a sentence discrimination task at different speech-to-noise ratios using speech-simulating noise as a background.

Aiken and Picton (2008) discovered, that low-frequency-components of speech envelopes (below 7 Hz) can be observed in the evoked EEG of the auditory cortex. O'Sullivan et al. (2014) and Ekin et al. (2016) exploit this frequency band to estimate the presented speech envelope from the measured EEG using linear regression models (aka. stimulus-reconstruction method; Rieke et al., 1995; Mesgarani et al., 2009). O'Sullivan et al. (2014) presented two spoken stories from two different speakers and directions simultaneously, to simulate a competing speaker scenario. They were able to detect which story a listener was selectively attending to, by correlating the estimated signal with the actual envelopes of both speech signals. This has been used as the basis for speech driven BCI in a series of successive studies (see Ekin et al., 2016; Biesmans et al., 2017). The performance of BCI approaches depend on the accuracy and the speed of their underlying classifier and can be compared by the respective information transfer-rates (ITR). The performance and processing delay (or latency) of auditory BCIs is a critical limitation for their potential application to the control of hearing aids in natural environments.

Therefore, in a preceding study of Chapter 4 tested the accuracy and speed of a simple auditory BCI approach based on correlations between speech envelope and evoked single-trial EEG using an analysis-window of 1.6 s duration (which is fairly corresponding to the duration of a short spoken sentence) for the segregation different speakers and sentences. With this BCI it is possible to segregate sentences or speakers clearly above chance level under five different classification conditions, six different diotically presented sentences and two different speakers in quiet with comparable performance to recently published speech-driven BCIs using longer analysis windows and adaptive filter methods. The comparison of classification performance of the conditions (1) clean speech versus speech in noise (trained with

correlation data obtained by the speech envelopes and evoked EEG from clean speech only) and (2) clean speech versus a stimulus with speech mimicking properties (mimicking the envelope as well as the overall spectrum of one sentence) led to two assumptions: (1) The BCI appears to generalize to speech in noise. (2) The envelope of the presented signal appears to be the most relevant factor for a correlation-based BCI approach even if the signal is not perceived as speech.

In the current study speech material from the Oldenburg children sentence test (OLKISA) is used. The OLKISA (Wagener and Kollmeier, 2005) is a psychoacoustic sentence matrix test for children in German. It is a shortened version of the OLSA (which was the pool of speech material used in Chapter 4) using only three words instead of five per sentence. Both sentence tests are generated to be sustainable in disturbing noise situations, and to have a steep discrimination function (Kollmeier et al., 2015; Wagener et al., 1999; Wagener and Brand, 2006). The tests are used for a variety of applications, e.g. clinical audiology and hearing aid fitting and provide a realistic test scenario for our auditory BCI approach to further investigate speech-in-noise performance under varying signal-to-noise ratios (SNR). A classifier needs multiple repeated EEG recordings of the same sentence to obtain sufficient amount of data for training and testing. Therefore, the short version (OLKISA) is used in the current study to keep measurement duration for each subject in an acceptable range within a single measurement session.

The major question of the current study is, to which extent the BCI generalizes to speech in noise during an attentive task when trained on clean speech only. A second question is, if the BCI still provides a performance above chance level at SNRs at which human subjects already show clearly reduced speech intelligibility. In other words: Is there a relation between decrease in BCI performance and human speech recognition? If so, the classification accuracy of a BCI-d controlled hearing aid could be used, e.g., to select a noise reduction algorithms suitable for the respective acoustical situation or even to estimate the instantaneous speech recognition performance of the listener. Addressing the first question the performance of the correlation-based auditory BCI approach is tested in a speech in speech-simulating noise scenario classifying two different sentences at three different SNR conditions within a short time window of just 1.2 s duration (which is the duration of the shortest sentence used here). According to the shorter sentences of the OLKISA in comparison to the OLSA the analysis time window used here is even shorter than in Chapter 4. The masking noise is the specific stationary speech noise of the OLKISA and OLSA tests (olnoise) which provides the same long term spectrum as the speech material to simulate a realistic scenario with a speaker masked by different levels of speech-shaped noise. Furthermore, the psychoacoustic measurement of individual speech intelligibility while simultaneously recording the EEG may provide some insights into the second question about BCI performance in relation to human speech intelligibility.

5.3 Methods

5.3.1 Participants

Twelve subjects (6 female, 6 male) in the age of 20-38 years (mean: 27 years) participated in the study. Ten subjects had hearing thresholds of 20 dB HL or better at all audiometric frequencies between 125 Hz and 10 kHz measured with a clinical audiometer (Auritec AT900). Two subjects had a threshold in quiet of 25 dB HL on the right ear at a single frequency, Subject 4 at 8 kHz and Subject 5 at 10 kHz respectively. All subjects were right-handed according to the questionnaire by Oldfield (1971) except for two participants who obtained a score of 0 and -10 (bi-manual). No participant reported any present or previous psychiatric or neurological disorder. The participants got an expense allowance of 10 € per hour. Written consent was obtained from each participant prior to the experiments. The experiments were approved by the local ethics committee of the University of

5.3.2 Setup

Signal presentation was controlled digitally using a custom script under Matlab 2015a (Mathworks). The digital signals were DA converted at a sampling rate of 44100 Hz using a FireFace UCX sound card (RME), then presented via a TDT HB7 (Tucker-Davis) headphone buffer using ER2 insert phones (Etymotic Research).

EEG data were recorded using a Biosemi ActiveTwo system (7 channels, Electro-Cap International 10-20 system, Parker contact gel) using the Biosemi software ActiView (6.03) at a sampling rate of 1024 Hz without any additional filtering. Electrode offsets were set not higher than 20 mV, prior to data recording. Exact temporal alignment of the EEG recordings with the arrival of the acoustical stimulus at the eardrum is ensured by transmitting trigger pulses via the digital output channel (SPDIF) of the Fireface with a delay of 1 ms to the presented stimulus to compensate the acoustical delay of the ER-2 insert-phones. The digital trigger pulses were converted to TTL trigger pulses (using a custom converter) and captured by the Biosemi system synchronously with the EEG data.

The measurements took place in a shielded booth (IAC) while the participants were sitting in a comfortable chair looking at a fixation cross in front of them. Both, speech stimuli and masking noise were presented diotically. The subjects were instructed to attend the presented sentences and repeat the words they understood. The instructor of the measurement was able to hear the participants via a hands-free speaking system.

5.3.3 Stimuli and Experimental design

The speech material used in this study was the german Oldenburg-children-sentence test (HörTech, Oldenburg) produced by the standard male speaker employed. From the speech material, the two

sentences S1 ("Drei kleine Bilder.", in English “Three little pictures.”) and S2 ("Sieben große Tassen.", in English “Seven big cups.”) are chosen as test stimuli for classification, due to their different onset syllables (S1: plosive, S2: fricative) and envelope spectrums (see Figure 16, middle). The durations of the two spoken sentences are 1.2 s (S1) and 1.3 s (S2). The respective speech noise (olnoise) with the same long-term spectrum as the average speech is added diotically to the stimuli as a quasi-running noise (from a noise signal of 20 s duration, a 2 s segment is cut out with randomly varying starting points for each sentence presentation).

All measurements for each subject took place in a single session of about five hours duration (including several long breaks of 20 minutes and longer). At the beginning of the session an audiogram was measured, followed by a ca. 10 min OLKISA measurement (without EEG recording) to obtain the individual speech recognition threshold (SRT, i.e., SNR level at which speech is perceived 50% correct). The individual SRT was estimated from the measurement of a single test list after measuring five trainings lists (14 sentences per list) beforehand to reduce possible training effects of the OLKISA (Wagener und Kollmeier, 2005). In the third part of the session, EEG measurements were performed presenting three SNR conditions. In condition C50, the noise level was adjusted according to the individually measured SRT (mean over subjects in current study: -13.1 dB). In two further conditions, fixed variations of these individual SNRs are used to change the expected speech recognition level to about 30 % (C30: SRT -1.5 dB) and 70 % (C70: SRT +1.5 dB), according to the properties of the average OLKISA discrimination function (Wagener und Kollmeier, 2005). The total EEG session contained 756 stimuli, which consist of 84 repetitions per test stimulus and SNR condition (C30(S1,S2), C50(S1,S2), C70(S1,S2)), 12 OLKISA lists at C70 and 6 lists at C50 (14 sentences per list; included to distract the subjects). The measurement was split into 5 runs with about 150 stimuli and breaks of five to fifteen minutes in between.

During the measurement the OLKISA sentences included in the selected test lists were presented in random order. All sentences were presented in noise (olnoise). In each trial the masking noise starts 500 ms before, and ends about 500 ms after the presentation of the sentence. All sentences were presented at a fixed level of 65 dB SPL, while the noise levels were adjusted in each trial according to the individually measured SRT and the condition-specific SNR level (C30, C50 or C70). The participants were instructed to repeat the understood words of the sentence or to tell that they understood none of the words. The instructor was situated outside the EEG recording booth and acoustically monitored the participant via a free-field intercom and manually marks the words that have been understood in the measurement software. After completing the data input by the instructor, the next stimulus started automatically after an additional pause of 500 ms. The EEG was measured during the whole run, but only the recorded data during stimulus presentation are used for classification.

5.3.4 Data processing

Five electrodes around the vertex (Fcz, C1, Cz, C2, Cpz) were chosen as analysis channels and referenced to the average over both mastoid channels ((A1 + A2) / 2). To reduce electrode drift and high frequency noise, a digital bandpass filter from 2 to 8 Hz (4th order, zero phase) was applied to the raw data. Pasley et al. (2012) reported for this frequency band an optimal observation of stimuli envelope properties in the EEG-Data. The EEG Data is separated into epochs according to corresponding trigger-information with epoch length of 1.2 s (according to the length of the shortest sentence in the speech corpus). The DC offset was estimated and corrected by averaging the 100 ms before each epoch and subtracting it from each epoch respectively. After applying an artifact threshold of 40 µV to the EEG data of each channel, to detect and exclude epochs with high amplitude artifacts, an average over the five channels is done. After down sampling the data to a sampling rate of 128 Hz it is sorted by trigger information corresponding to the different test sentences.

The envelope of each stimulus is extracted by calculating the absolute value of the respective analytic signal obtained with a Hilbert transform of the broadband signals to provide comparable reference signals for the correlation with the EEG data. Each of the envelopes is further low pass filtered at a cutoff frequency of 8 Hz (4th order, zero phase) and down sampled to 128 Hz. The duration of the envelopes is limited to the first 1.2 s due to the length of the shortest presented sentence.

5.3.5 Classification and Statistical analysis

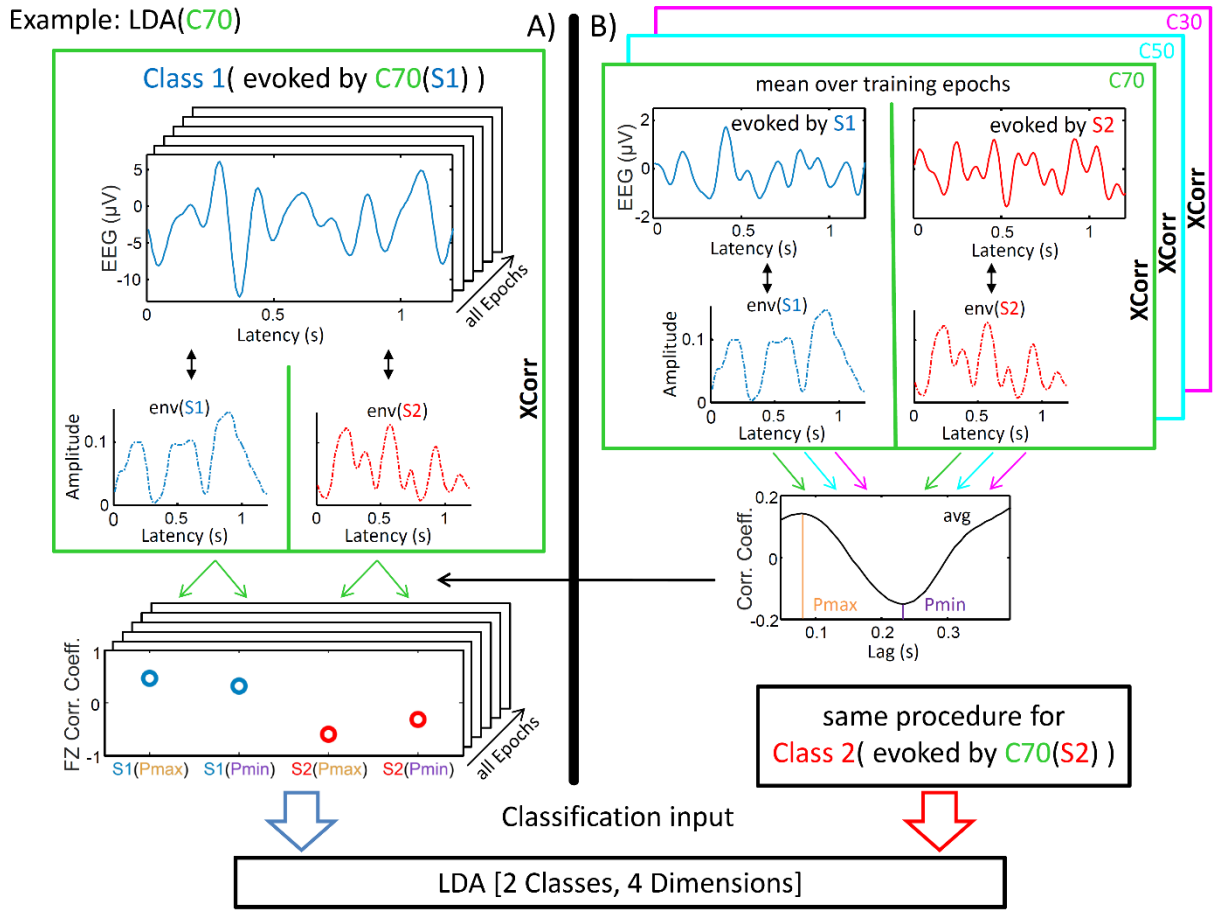

Figure 16: Preprocessing example for the classification input of the linear discriminant analysis (LDA) of test condition C70. A) Top: Single-trial EEG epochs evoked by the corresponding OLKISA sentence C70(S1). A) Middle: The single-trial EEG epochs were cross correlated (mean-removed, normalized [autocorrelation equal 1 at zero lag] and Fisher z-transformed) to the clear envelopes of each of the two sentences (S1, S2). B) Top: Mean over EEG epochs evoked by two sentences C70(S1) and C70(S2) that are used for training. B) Middle: Mean over EEG training epochs were cross correlated (mean-removed; normalized, i.e., the autocorrelation is equal to 1 at zero lag) to the envelope of the corresponding sentence. B) Bottom: The resulting Correlation coefficient sequences of each test condition and class, were averaged and the lags of the maximum and minimum (Pmax and Pmin) were detected within a limited lag range of 40 to 400 ms (corresponding to cortical auditory evoked potential; Picton et al., 1974). A) Bottom: Correlation coefficients corresponding to the single-trial sequences at the specific lag positions Pmax and Pmin are used as Class1 input for the LDA. The LDA input of Class2 follow the same procedure respectively (Chapter 4).

The classification procedure is almost identical to the preceding study of Chapter 4. The EEG-Data was classified via a linear discriminant analysis (LDA) using the function “classify” under MATLAB 2015a (Mathworks), to provide a rather robust classification for a low amount of available training samples per

class (here 84 epochs in total per condition and class; as shown e.g. by Hu and Yu, 2011). To further maximize the size of available test samples a 10-fold cross-validation was done for each participant and discrimination condition individually. To provide a fair comparisons across the different test conditions we used the least number of available epochs (Emax; see Table 7) evoked by one of the six test stimuli (C30(S1,S2), C50(S1,S2), C70(S1,S2)). Emax epochs were divided in 10 folds using the Matlab command cvpartition. In each validation 9/10 of the data was used for training and 1/10 for testing. The classification results across all 10 folds are taken to calculate the classification accuracy for the respective subject and test.

Preprocessing of the LDA input is exemplary shown in Figure 16 for one test condition (C70). In each validation step, correlation coefficient sequences (CCS) are computed with mean-removed and normalized (autocorrelation equal 1 at zero lag) cross correlation between each single-trial EEG epoch and each of the reference-stimulus envelopes (derived from the clear stimuli S1 and S2; see Figure 16 A)). The identification of the CCS lags to be used for the classification is obtained in four steps (see Figure 16 B)). (1) All EEG epochs used for training within one class and test condition are averaged. (2) For each class and test condition correlation coefficient sequences were computed between the results from (1) and the respective clear signal envelopes. (3) The results from (2) are averaged over classes and conditions and are referred as CCS_{train_avg}. (4) The lags of the maximum and minimum of the resulting CCS_{train_avg} are detected (within a limited lag range of 40 to 400 ms corresponding to cortical auditory evoked potential P1 to N2; Picton et al., 1974) and used to select the LDA input from the CCS (see Figure 16 bottom). The correlation coefficients were Fisher z-transformed before classification. Three discrimination conditions were used for classification testing the three SNR presentation levels ((1) C30, (2) C50 and (3) C70). In each condition the classifier had to discriminate whether a certain single EEG trial was evoked by S1 or S2.

To examine the distribution of the classification results of each test and subject, Monte Carlo permutation tests with 10000 iterations were done. At the beginning of each iteration, the order of the EEG data is randomized, then a 10-fold cross-validation is done like described top. The distributions were computed for classifications trained with correct Class-Labels (D_{CL}) as well as for classifications trained with randomized Class-Labels (D_{RL}). To test for statistically significant classification accuracies on single subject level the average value of the D_{CL} must exceed the 95 % percentile (corresponding to a significance level of $p = 0.05$) of the respective D_{RL}. To estimate the statistical significance of the classification accuracy when averaged over subjects, a paired-sample t-test (one-tailed) is done between the mean values of the single subject D_{RL} in one vector and the mean values of the respective D_{CL}. Statistical significance of classification accuracy values of three different SNR conditions was tested using a (3x1) repeated analysis of variance (rANOVA), while the within subject factors are SRT (C30, C50, C70). A Mauchly-Test was applied to test for sphericity of the data and in case of significance, Greenhouse-Geisser correction was applied.

For a better comparison with other studies the bits per class that are generated by the BCI after the duration of one minute (Information transfer-rates, ITR), are calculated due to following equation (N: amount of classes, p: recognition rate, T: duration of epoch in seconds; Shannon and Weaver, 1964; Besserve et al., 2007; Speier et al., 2013):

$$\mathrm{ITR} = \frac{60}{\mathrm{T}}[\log_2 \mathrm{N} + \mathrm{p} \cdot \log_2 \mathrm{p} + (1-\mathrm{p}) \cdot \log_2 \left(\frac{1-\mathrm{p}}{\mathrm{N}-1}\right)],$$

$$\Delta\mathrm{ITR} = \frac{60}{\mathrm{T}} \cdot \log_2 \left(\frac{\mathrm{p} \cdot (\mathrm{N}-1)}{1-\mathrm{p}}\right) \cdot \Delta\mathrm{p}.$$

Table 7: Least number of available epochs (Emax) evoked by one of the two test stimuli presented at three SNR levels of each subject.

Subject ID	1	2	3	4	5	6	7	8	9	10	11	12
Emax	75	83	62	83	80	79	81	56	74	77	83	83

5.4 Results

Figure 17 summarizes the LDA results, e.g., the mean of the D_{CL} for all participants and test conditions. The corresponding median classification rates over subjects with standard errors of the mean and ITRs for each test condition are given in Table 8. In Test condition C70, the mean D_{CL} classification accuracies over subjects are significant above chance level (p = 0.00003) with a median recognition rate of 62.2 % ± 1.8 % and corresponding ITR of 2.2 ± 0.7 bits/min. On single subject level eight of twelve subjects score a mean D_{CL} recognition rate significant above chance level. The best BCI performance for condition C70 was obtained for Subject 8 with an accuracy of 72.54 ± 0.02 %, resulting in an ITR of 7.60 ± 0.01 bits/min. The test conditions C50 and C30 scored significantly above chance level, with median recognition rates of 58.1 ± 2.0 % for C50 (p = 0.003; eight out of twelve significant on single subject level), and 55.7 ± 1.1 % for C30 (p = 0.0004; three out of twelve significant on single subject level). For subject 2 the BCI was even able to score an average classification accuracy of 57.04 ± 0.02 % (95 % percentile of DRL is 56.63 %) with an ITR of 0.718 ± 0.004 bits/min in test condition C30 with a presented SNR of -15.1 dB, where the subject was able to understand only 22.4 ± 2.7 % of the presented speech. In the investigated SNR range, the average classification accuracies, as shown in Table 8, drop at a rate of about 3 % for a SNR decrease of 1.5 dB. The rANOVA of the results reveals a significant dependence of classification accuracy and SNR ($F(1.3,13.8)=11.09$; $p=0.003$; Greenhouse-Geisser corrected).

The correct responses of the subjects corresponding to all presented stimuli can be seen in Table 9. At the adaptively measured SRT obtained by the pre-test a speech recognition of 50 % is expected in the main test as well (C50). For a 1.5 dB lower (C30) or higher (C70) SNR we would expect speech intelligibilities of about 30 % or 70 %, respectively (Wagener and Kollmeier, 2005). However, the average speech recognition for the often presented sentences S1 and S2 is clearly higher than expected (C30: 36.8 %, C50: 65.6 % and C70: 84 % correct understood words). However, the averaged speech recognition of the less repetitive presented standard OLKISA test list (which was interleaved with the test for the two standard sentences in the joint EEG- and speech recognition measurement paradigm) is close to the respective expected value (C50: 51.4 % and C70: 73.7 % correct understood words). For comparison in Table 10. the SRT that were individually obtained prior to the EEG measurements with an adaptive procedure are given for each subject ("Pre-Test") and in addition to SRT values ("Fit"), that were calculated by fitting a psychometric function (according to Equation 1, RS: recognized speech, SNR: signal-to-noise ratio, SRT: speech recognition threshold, m: rise of function at SRT; Wagener and Kollmeier, 2005) to the individual SNR levels for all three EEG conditions (see chapter 5.3.3) and the corresponding individual speech recognition levels shown in Table 9:

$$\mathrm{RS(SNR)} = \frac{1}{1+\mathrm{e}^{(4\,m\,(SRT-SNR))}}. \quad \text{Equation 1}$$

Except for subject 4, 6 and 8 all prior measured SRT ("Pre-Test") are within a deviation of 0.1 to 1.1 dB to the calculated SRT-results ("Fit").

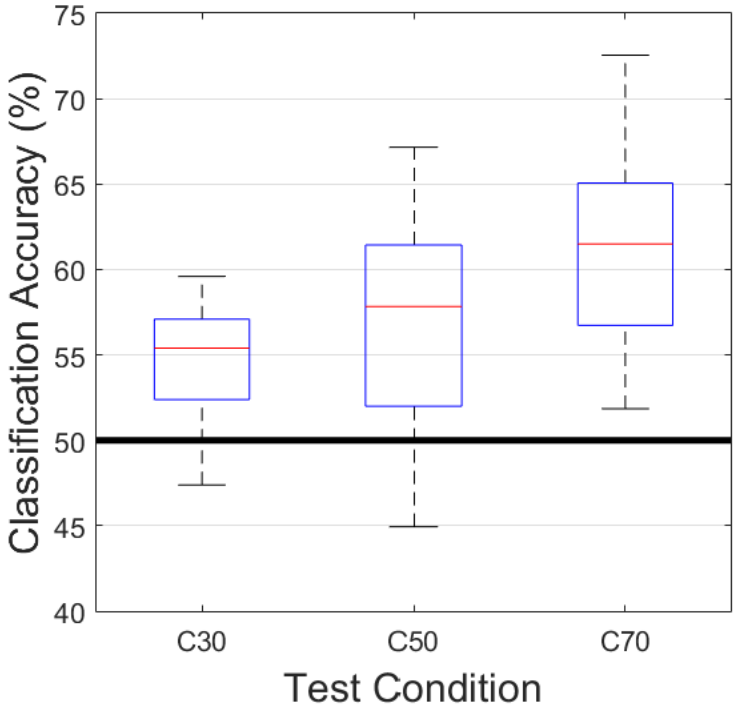

Figure 17: Distribution of achieved LDA classification rates for all participants and test conditions (C50: the noise level adjusted according to the individually measured SRT [mean over subjects: -13.1 dB]; C30: SRT -1.5 dB; C70: SRT +1.5 dB).

Table 8: Averaged results of the Monte Carlo permutation tests are shown for each test condition. The median over subjects' classification accuracies and corresponding standard errors of the mean are calculated with the mean classification accuracies of the distributions, obtained by training with randomizes Class-Labels (Chance) and correct Class-Labels (Acc). Further the corresponding statistical test results are presented. ITRs corresponding to the median classification accuracies are shown for each test condition with standard errors of the mean.

Test Condition	Chance (%)	Acc (%)	t-test	Single	ITR (bits/min)
C30	49.96 ± 0.02	55.71 ± 1.05	p = 0.0004	3/12	0.47 ± 0.17
C50	49.99 ± 0.01	58.14 ± 1.96	p = 0.003	8/12	0.96 ± 0.46
C70	49.99 ± 0.02	62.16 ± 1.83	p = 0.00003	8/12	2.16 ± 0.66

Table 9: Behavioral data of the subjects showing the average percentage of correctly repeated words (with standard error of the mean) of all sentences that were presented during the three different test conditions (C30, C50 and C70) and of all sentences contained in the random OLKISA-lists that were presented at two different noise levels [C50(OLKISA) and C70(OLKISA)]. Values in red indicate the subjects (4, 6 and 8) that might have used additional non-task-associated strategies (as discussed in chapter 5.5).

Subject	C30	C50	C70	C50(OLKISA)	C70(OLKISA)
1	25.4 ± 2.4	56.3 ± 2.9	76.2 ± 2.1	40.8 ± 4.2	72.2 ± 2.5
2	22.4 ± 2.7	58.9 ± 3.0	89.7 ± 1.5	52.4 ± 4.2	77.6 ± 2.5
3	39.7 ± 3.2	73.0 ± 2.5	87.9 ± 1.7	57.5 ± 4.3	77.8 ± 2.1
4	74.8 ± 2.8	88.3 ± 1.9	95.0 ± 1.1	72.2 ± 3.6	83.7 ± 1.9
5	27.4 ± 3.2	61.9 ± 3.3	81.7 ± 2.3	50.8 ± 4.7	71.4 ± 2.9
6	57.7 ± 3.1	78.4 ± 1.8	83.3 ± 1.4	57.9 ± 4.5	82.1 ± 2.2
7	26.0 ± 2.4	51.0 ± 2.8	77.4 ± 2.0	46.0 ± 4.1	66.5 ± 2.6
8	61.1 ± 3.2	82.1 ± 2.3	93.1 ± 1.2	58.7 ± 4.0	80.2 ± 2.1
9	22.8 ± 2.7	61.7 ± 3.1	83.7 ± 2.0	40.9 ± 4.2	65.3 ± 2.7
10	39.3 ± 3.2	70.8 ± 2.7	89.3 ± 1.8	49.6 ± 4.0	71.6 ± 2.5
11	27.2 ± 2.6	66.5 ± 2.3	82.9 ± 1.6	46.8 ± 3.7	71.0 ± 2.4
12	17.5 ± 2.0	38.3 ± 2.6	68.1 ± 2.5	43.3 ± 3.8	65.1 ± 2.7
Avg.	36.8 ± 5.3	65.6 ± 4.0	84.0 ± 2.2	51.4 ± 2.6	73.7 ± 1.9

Table 10: Individualy measured speech recognition thresholds (signal-to-noise ratio level at 50% speech intelligibility) for each subject from the preceding OLKISA measurement ("Pre-Test") and estimated SRT from the discrimination function (Equation 1) fitted to the individual behavioral data of the sentences S1 and S2 during the EEG measurement ("Pre-Test"). The red markings indicate the subjects that might have used non-task associated strategies (as discussed in chapter 5.5).

SRT	SJ 1	SJ 2	SJ 3	SJ 4	SJ 5	SJ 6	SJ 7	SJ 8	SJ 9	SJ 10	SJ 11	SJ 12
Pre-Test	-13.7	-12.6	-12.9	-12.9	-12.3	-12.1	-13.5	-13.3	-13.0	-12.8	-13.6	-14.6
Fit	-13.9	-13.0	-14.0	-16.2	-12.7	-14.4	-13.6	-15.4	-13.4	-13.8	-14.2	-14.0

5.5 Discussion

The results show that even under the interfering of speech-simulating noise presented at varying levels, it is still possible to segregate two sentences using the simple correlation based classification approach of Chapter 4 although relying here on a shorter analysis windows of only 1.2 s instead of 1.6 s duration. The results further indicate the ability of the BCI to generalize to speech in noise when trained on clean speech envelopes. Classification accuracies over subjects were significant above chance level for all presented SNRs ($p < 0.003$) and therefore actually under conditions with reduced speech intelligibility in humans. The BCI approach presented here scores an ITR of 2.2 bits/min under the best SNR condition (C70, mean SNR: -11.6 ± 0.2 dB, mean subjects speech intelligibility: 84.0 ± 2.2 %). In other words, within less than 28 seconds this BCI can detect which one of two sentences a listener had heard under the disturbance of speech-simulating noise with an accuracy of 100 %. For Subject 2 at a SNR of -15.1 dB when the subject was able to understand only 22.4 % of the presented speech, the BCI was still able to achieve a classification accuracy of 57%, significantly above chance level ($p < 0.05$) with an ITR of 0.72 bits/min. The classification results show a significant dependence on the human speech intelligibility ($p = 0.003$). This indicates the feasibility of BCI applications that select noise reduction algorithms which are appropriate for respective acoustical situations on a hearing aid.

The most comparable condition of the previous study (Chapter 4) to the current investigation in respect to speech in speech-simulating noise is the discrimination of the EEG evoked by a clear sentence from the female speaker and the respective stimulus envelope against the correlation between the EEG evoked by the same sentence spoken by a male speaker with speech background noise and the clear envelope of the same sentence. The respective results in the previous study showed a mean classification accuracy over subjects of 61.03 ± 1.28 % with a corresponding ITR of 1.33 ± 0.31 bits/min. The C70 condition of the present BCI performs with a nearly identical classification accuracy, while both spoken sentences are disturbed with a higher level of background speech-simulating noise [Chapter 4 employed masked stimuli at a SNR of 10 dB, here the mean over subject SNRs are -14.6 dB (C30), -13.1 dB (C50) and -11.6 dB (C70)]. Decreasing the SNR by 1.5 dB by raising the noise leads on average to a 20 % decreased speech intelligibility in the human subjects and to a performance drop of about 3 % of BCI accuracy. The ITR corresponding to the median classification accuracy of condition C70 is about 0.8 bits/min higher than the ITR obtained in the previous study, explained by the slightly higher classification accuracy and the 0.4 s shorter analysis window. Even in the worst SNR condition tested here, the classification accuracies over subjects are still significantly above chance level. These results support the assumption of the previous study that the BCIs are able to generalize to speech in noise when trained with correlation data obtained by clean speech envelopes and evoked EEG only.

The overall performance of this intentionally simple BCI approach is comparable to other studies in the field. O'Sullivan et al. (2014) and Ekin et al. (2016) simulated a competing speaker scenario by presenting

two spoken stories dichotically to their participants. Their task was to attend to one story while ignoring the other one. Both BCI approaches detect selective attention on a single-trial basis of about 60 s by classifying correlations between the envelope of the attended story and an estimated envelope signal based on a linear regression model using information of the evoked EEG and the envelope of the attended story. O'Sullivan et al. (2014) achieved on average a recognition rate of 89 % with a corresponding ITR of 0.5 bits/min. Our BCI approach scores on average a 1.66 bits/min higher ITR in the C70 condition. Ekin et al. (2016) extended the BCI approach of O'Sullivan et al. (2014) by the addition of a reconstruction filter for the unattended story, both filters basing on a Capon minimum variance distortionless response beamforming method. They scored classification accuracies of 86.1 % for an attended story decoder (ITR of 0.42 bits/min) and 80.6 % in an unattended story decoder (ITR of 0.29 bits/min). The ITR of our BCI approach is on average 1.74 to 1.87 bits/min higher ITR for condition C70. The ITR of condition C50 is 0.54 bits/min higher than the attended story decoder. The BCIs ITR for the worst SNR condition (C30) is comparable to the attended story decoder. Biesmans et al. (2017) enhanced the classification performance of O'Sullivan et al. (2014) BCI approach by applying the following three changes. (1) The analyzation window was decreased to 30 seconds length. (2) A single least squares estimation over the entire training data set was solved, instead of averaging over a set of least squares estimations for each single trial. (3) An optimized envelope extraction was tested by comparing various of auditory modelling processes. A simple model, based on a combination of power law relation (loudness model) and gammatone filter bank, showed the best classification accuracy of 81.5 % leading to an ITR of 0.62 bits/min. Our BCI approach scores on average a 1.54 bits/min higher ITR in the C70 condition. In contrast to O'Sullivan et al. (2014), Ekin et al. (2016) and Biesmans et al. (2017), we do not detect spatial attention to one trained speaker in a competing speaker scenario, but we test our simple correlation-based BCI approach using rather high levels of background babble noise with an analysis window of only 1.2 s.

The average speech intelligibilities for the conditions C50 and C70 of the standard OLKISA test lists, measured in parallel with the EEG, are close to the expected 50 % (i.e., 51.4 ± 2.6 %) and 70 %, respectively (i.e.,73.7 ± 1.9 %). This indicates that the individual SRTs determined in the OLKISA pretests are valid. However, an unusual high repetition was required for the test sentences S1 and S2 to obtain a sufficient amount of EEG epochs for classification. Consequently, the average speech intelligibility of the two often-repeated test sentences S1 and S2 is higher than expected (about 7 % higher for test condition C30, and about 15 % for the test conditions C50 and C70). A further OLKISA specific training effect in this order of magnitude appears unlikely because several training lists were applied to the subjects before and, in addition, an unspecific training effect was not observed for the complete test according to the trends in the training lists. One explanation could be, that the high number of repetitions of the test sentences S1 and S2 allowed the subjects to develop additional non-task associated strategies, like guessing which sentence they had heard due to any specific auditory cue (such as, e.g., the duration or perceived rhythmic pattern of one sentence as opposed to the other) and just repeating the words they recall from a previous

presentation of the same sentence at a better SNR condition. The individual comparison of SRT estimated from the discrimination function fit (see Equation 1) for the prior measured SRT levels provides some evidence that if so, only some of the subjects were using those non-tasks associated strategies. In seven subjects, the deviation of estimated and measured SRT is lower or equal to 0.6 dB (see Table 10). In contrast, Subject 4, 6 and 8 show highly increased estimated SRTs (see Table 10), which suggests in addition to their high speech intelligibility values shown in Table 9, that they most probably took advantage of additional cues. For future studies combining EEG recordings with simultaneous speech recognition measurements, these findings suggest that the number additional OLKISA lists beside the required repeated test stimuli should be increased - despite an increased measurement effort (which in the current study took about 5 hours per subject in one session). It should be noted that such an interleaved speech recognition and EEG recording experiment is advantageous because it provides better distraction of the subjects and it should compensate the bias of the speech intelligibility data caused by the high amount of repeated sentences.

5.6 Summary and Conclusions

In this study we tested an auditory BCI approach classifying sentences in speech-simulating babble-noise, at SNRs where the users already showed clearly reduced speech intelligibility. As a concurrent psychoacoustic task we measured in parallel to the EEG recordings the speech recognition rate to a more extended speech test. This BCI approach can segregate two sentences presented at three different SNR conditions clearly above chance level by classifying correlations between clean speech envelopes of presented signals and the respective evoked single-trial EEG using an analysis window of only 1.2 s duration. The performance of this approach is comparable to predecessor previous study and recently published speech-driven BCIs. As assumed in the previous study, these results show the ability of the BCI to generalize to speech in noise when trained with correlation data obtained from clean speech envelopes and the respective evoked EEG only. The classification results further show a significant dependence on the individual speech recognition rate, which indicates a possible use for speech-recognition-driven BCI applications like e.g. noise reduction strategy selection on a hearing aid.

5.7 Acknowledgments

We would like to thank Jana Mueller, Florian Schmidt and Oliver Behler for fruitful discussions about signal processing and statistical analysis of EEG data and Matthias Vormann (Hörzentrum Oldenburg GmbH) for providing us with a MATLAB version of the OLSA/OLKISA.

6 General Summary and Discussion

The current thesis describes important steps on the way towards an auditory BCI paradigm based on speech that could be used to control hearing devices in the future. When developing the paradigm of Chapter 3, only few BCI auditory paradigms existed so far (e.g., Halder et al., 2010; Kim et al., 2011). They mainly utilized artificial stimulation or artificially modified speech. As one step towards a more realistic scenario, the study in Chapter 3 utilized a paradigm presenting streams of real speech syllables at speech-rhythm rates. A competing speaker situation was simulated to investigate if long term components of the evoked EEGs spectrum are feasible for selective attention detection. The results of Chapter 3 show that it is possible to measure ASSR to two competing streams of real speech syllables and to detect the selective attention of the listener by classifying the attention-generated amplitude differences of the respective long-term frequency bins and their first harmonics during a single-trial EEG of 20 s length. Therefore, speech-rhythm seems to be a feasible feature of speech to drive an auditory BCI detecting the selective attention of a listener in a competing speaker scenario. However, an average classification accuracy of 61 % leading to an ITR of 0.2 bits/min is not practical for usage in a real-life environment, since the BCI would need 5 min to detect the attended speaker with an accuracy of 100 %.

Visual-driven BCI paradigms like the visual P300-Speller of Donchin et al. (2000), who scored a classification accuracy of 90 % and an ITR=4.8 bits/min, still outperform current auditory BCI approaches (as described by Furdea et al., 2009; Chang et al., 2013). One explanation is the better SNR of the visual evoked EEG due to the higher amplitude of visual evoked potentials due to the bigger cortical structure of the visual system in contrast to the auditory cortex. However, for hearing devices in several scenarios, visual driven BCI are only of limited benefit and auditory BCI provide a useful alternative for physically and visually disabled people. Therefore, further efforts should focus on enhancing the speed and accuracy of auditory BCI. In Hill et al. (2014) some subjects reported about problems with the understanding of auditory paradigms- when using artificial stimuli or they describe the perception of beeps to be harsh and mildly unpleasant. Therefore, future auditory approaches should be rather focused on natural speech as stimulation than amplitude modulated speech or artificial signals.

The next step towards a speech driven BCI was to create a paradigm segregating spoken sentences. Therefore, in several pilot studies the correlation was investigated between the envelope of a sentence and the respective evoked EEG as well as the correlation between the spectrum of the stimulus envelope and the spectrum of the evoked EEG data (see General appendix 8 for an overview). Neither a segregation of spoken sentences based on classification of ASSR to speech-rhythm nor a classification due to correlation between the respective long term spectra was possible. One major reason is probably that the variation of speech-rhythm during a sentence is too large to allow stable ASSR measurements. Further testing lead to the promising paradigm presented in Chapter 4. The suggested auditory BCI paradigm is capable of segregating two speakers and six sentences, based on rather simple correlation evaluation between the

envelopes of all presented spoken OLSA sentences and a single-trial EEG epoch of 1.6 s duration evoked by one of the sentences. A segregation with significantly above chance level ($p < 0.01$) mean classification accuracies was possible in the test conditions segregating one out of six sentences spoken by two different gendered speakers, one out of three sentences spoken by either a single male or female speaker, one out of two differently gendered speakers each speaking three sentences and the same sentence spoken by two differently gendered speakers. It should be noted, that the segregations were not possible by simple detection of the highest correlation coefficient, probably due to the poor SNR of the recorded single-trial EEG epochs. Therefore, the LDA input-vectors where extended with correlation coefficients between a single-trial EEG epoch and the envelopes of all sentences of the specific test condition. With this approach it was possible to score an average ITR of 2.04 bits/min for the six-sentence condition (ITRs of 0.8 to 1.3 bits/min for segregations between three sentences or two speakers) which is an improvement of speed in comparison to the results of Chapter 3. These results are further comparable, or even higher than the performances of the stimulus-reconstruction approaches presented by O'Sullivan et al. (2014) (average ITR of 0.5 bits/min), Ekin et al. (2016) (average ITR for attended story detection of 0.42 and 0.29 bits/min for the unattended story) and Biesmans et al. (2017) (average ITR of 0.62 bits/min, using a combination of power law relation (loudness model) and gammatone filter bank for stimulus envelope extraction), who use adaptive filter methods . However, in contrast to O'Sullivan et al. (2014), during further pilot tests (see Chapter 8.4) it was not possible to segregate competing sentences or speakers with the BCI approach of Chapter 4 on a single-trial basis.

The approach presented here further appears to be able to generalize to speech in noise when trained with correlation data between the evoked EEG and respective clean speech envelopes of the presented stimuli. This assumption is raised by the comparable performance of the three test conditions (1) segregating female and a male speaker speaking the same sentence (F3 vs M3), (2) same scenario while the male speaker is masked with noise (F3 vs Nsp) and (3) same scenario as (2), but the classifier possesses only the correlation coefficients between the respectively evoked single-trial EEG epochs and the clean speech envelopes of the presented sentences (F3 vs Nsp[M3]). Furthermore, the results of Chapter 4 provides evidence that signal envelopes of the presented stimulus seem to be replicated in the EEG regardless if the stimulus was perceived as intelligible speech or not and that the classifier seem to focus predominantly on this envelope information. This assumption is supported by following two findings: (1) a segregation of a test sentence M3 spoken by a male speaker against an artificial noise-stimulus Nam, which is mimicking the envelope and overall spectral content of M3 was not possible. (2) On the other hand, the segregation of other clean sentences against the test sentence M3 (e.g., F3 vs M3) showed comparable results to respective segregations against Nam (e.g., F3 vs Nam). This can be explained by the similarity of the signal envelopes of M3 and N2 (correlation of 0.955). The psychoacoustic data of the listening task showed, that questions to Nam were answered significantly worse than questions to M3 ($p = 0.013$). Though, ten out of twelve subjects showed responses clearly above chance level when answering questions to M3 while

listening to Nam. These subjects might have identified Nam as M3 due to the similar signal-properties or they guessed the right answers due to the high number of repeated test-sentences containing the same words (F3, M3 and N1 containing: "Peter bekommt vier grüne Messer.").

In Chapter 5 an auditory BCI using a simple correlation-based classification approach as suggested in Chapter 4 was tested under the disturbance of three different levels of speech-simulating noise. The results show that this BCI approach is still able to segregate two spoken sentences clearly above chance level even when using an even shorter analysis window of 1.2 s. At these SNR levels the users showed clearly reduced speech intelligibility, which was measured in a psychoacoustic task parallel to the EEG recordings. The data reveal a significant dependence of the individual classification accuracy on the (decreasing) SNR levels ($p = 0.003$), in line with the human speech intelligibility. A SNR decrease of 1.5 dB leads to an average reduction of about 20 % in speech intelligibility and about 3 % reduction in the classification accuracy of the BCI. A significant above-chance level classification (accuracy of 57.04 %, ITR of 0.72 bits/min, $p < 0.05$) was possible at an individual speech recognition level as low as 22.4 % (subject 2, SNR of -14.1 dB). The results provide some necessary properties of the BCI (like the stability to noise and a relation of resulting classification accuracy to the human speech intelligibility) for a future control on a hearing aid in a natural hearing environment. Exemplary an aided patient is attending a speaker on a party with several disturbing speakers in the background that are getting louder over time. The BCI could detect a decrease in speech intelligibility of the patient by monitoring a decrease of classification accuracy to the attended speaker, thereupon activating an appropriate noise reduction algorithm to mask the disturbing noises of the background.

Since it was possible to segregate sentences at all three different SNR levels significantly above chance level ($p < 0.01$), the results show furthermore the capability of the current single-trial BCI approach to generalize to speech in noise, when trained with correlation data of clean speech envelopes end evoked single-trail EEG. This clearly supports the respective assumption made in Chapter 4. The performance of the BCI is overall comparable with the results of Chapter 4 and therefore as well comparable to the ones of O'Sullivan et al. (2014), Ekin et al. (2016) and Biesmans et al. (2017). For example, in the best SNR condition (C70, mean SNR: -11.6 ± 0.2 dB, mean subjects speech intelligibility: 84.0 ± 2.2%) the BCI of Chapter 5 scored on average an ITR of 2.2 bits/min, while the approach of Chapter 4 was able to segregate two speaker in a speech in speech-simulating noise situation with an accuracy of 61.3 ± 1.9 % and an ITR of 1.4 ± 0.5 bits/min, when trained with correlation data obtained by clean speech envelopes (O'Sullivan et al. (2014): ITR of 0.5 bits/min; Ekin et al. (2016): ITR of 0.42 bits/min [attended story], ITR of 0.29 bits/min [unattended story]; Biesmans et al. (2017): ITR of 0.62 bits/min).

It was further noticeable, that the average speech intelligibility levels obtained from the often repeated test sentences (as needed for the BCI training and analysis) was better than expected in normal hearing subjects and better than obtained for the standard OLKISA test lists presented to distract the subjects. The often

repeated presentation of the test sentences could have allowed the subjects to develop non-task associated strategies to perform better at the speech recognition task as the respectively expected thresholds. Future studies combining speech-driven BCI approaches with concurrent psychoacoustic speech recognition measurements might have to increase the amount of distraction from OLKISA lists to compensate the bias, although it would exceed the bearable length for one measurement session (The measurement sessions here took about 5 hours per subject).

In summary, the auditory BCI approach presented in this thesis shows some promising evidence that a future BCI-controlled hearing aid based on speech might be possible. To accomplish this goal, it is still necessary to further enhance the performance of the BCI as well as to investigate the representation of speech in the EEG to obtain additional founded knowledge of classifiable speech evoked EEG potentials.

7 Conclusion and Outlook

The following conclusion can be drawn from the findings in the current thesis:

- ASSR to competing streams of spoken syllables presented at fixed rates are a feasible EEG-feature to detect spatial auditory attention on a single-trial basis (see Chapter 3). If this approach would be feasible for natural speech stimuli as well, its performance would be fast and accurate enough to be considered for practical use in a realistic scenario.
- Neither ASSR nor the long-term spectrum of the EEG appears to be a promising feature for auditory BCI that are based on real speech. Several pilot studies (see Appendix) indicate that ASSR are neither clearly detectable to the speech-rhythm of a spoken sentence in quiet, nor to the fundamental frequency of a vowel of a single word or syllable. One important reason is probably that ASSR measurements are too sensitive to the natural variation of speech rhythm and the respective variability of the modulation frequency.
- Reasonable results can be obtained by using a BCI approach using the direct correlation between evoked single-trial EEG and the respective speech envelope (see Chapters 4 and 5) even when a simple classifier is employed that is computational more efficient than adaptive filter methods typically used in comparable BCI approaches with similar performance. The correlation-based BCI approach is capable of segregating different speakers as well as several speech-test sentences in quiet.
- The correlation-based BCI approach generalizes to speech in noise, the BCI still performs above chance level even at SNR levels at which the human listeners already show clearly reduced speech intelligibility. A relation was found (Chapter 5) between classification accuracy and the speech intelligibility of the user.

The findings that a computational efficient simple correlation-based classifier is capable to segregate speakers and sentences from EEG recordings, while the respective BCI generalizes to speech in noise, as well as the finding that the drop in BCI performance with decreasing signal-to-noise ratio reflects to some extent the human performance in speech intelligibility provide promising directions for further steps on the way towards a BCI control for hearing aids. In future tests natural speech should still be used for stimulation, since speech appears to be a feasible stimulus for the control of BCI and it is the predominant signal in the acoustical scene (i.e., multiple competing speaker masked by background noise) that an auditory BCI aims to analyse.

When looking for efficient BCI approaches for applications like hearing aids, several constrains have to be considered like the need of real time processing, the need of limited energy consumption as well as limited computational power. Therefore, it appears worthwhile to investigate further ways to combine the given information from stimulus and EEG response in an optimized way (like the combination of multiple

correlation coefficients here), to provide robust features for a simple, i.e., computational efficient BCI classifier that cover a different application range than more expensive (in terms of energy consumption and computational resources) and potentially more powerful approaches like adaptive filtering or deep neural networks (Kottaimalai et al., 2013). The present BCI approach showed a dependence between its classification accuracy and the SNR of the acoustical situation (which is in line with the human speech intelligibility) under the disturbance of noise (see Chapter 5). This relation might be used to activate an appropriate noise filter to the given acoustical situation on a listening device like a hearing aid. For example, a detected change of the classification accuracy of the BCI (detecting an attended speaker) and the SNR of the given acoustical scene (e.g., using the microphone array of the device) over time, could be used to activate an appropriate noise filter. Overall, the performance (accuracy and speed) of the present approach still needs to be enhanced as well as spatial attention detection to competing speakers needs to be implemented to be feasible for future real-life applications. In the following two passages we will discuss possible solutions to accomplish this goal and further problems that could accrue on the way.

One major problem shared by the commonly used correlation-based BCI approaches is their limitation to a certain pool of stimuli which clearly limits the generalization to arbitrary stimuli and acoustic scenes. To overcome this limitation, a more detailed knowledge about the representation of speech stimuli in the EEG would be very useful. Therefore, future experiments should take one step back from single trial detection paradigms with disturbance of noise and rather focus on detailed investigations on speech evoked EEG measurements with a good SNR, obtained by averaging data from repeated measurement trials. The knowledge about stimulus-related characteristics or features in the EEG response would in turn help to search more targeted for specific EEG features in noisy single trial EEG. For example, the use of a respectively adapted "stimulus envelope" for the correlation with EEG data might be beneficial, like e.g. Petersen et al. (2017) presented, who found that the speech-onset envelopes (computed by taking the first derivative of the speech envelope before it is further half-wave rectified) are represented better in the evoked EEG then the actual speech envelopes, or Biesmans et al. (2017), who enhanced the classification performance of the BCI approach of O'Sullivan et al. (2014) by extending the stimulus envelope extraction with a combination of power law relation (loudness model) and gammatone filter bank. In contrast to the findings of Petersen et al. (2017), first tests introducing the speech-onset envelopes into the BCI paradigms presented in Chapter 4 and 5, only showed marginal higher classification accuracies in some test conditions. Some further tests were made in parallel to the study presented in Chapter 4 that correlated the EEG data with the output of several different loudness models (e.g. Glasberg and Moore, 2002) instead of stimulus envelope. This, however, did not provide any benefit. This may indicate that loudness is not a good representation of speech stimuli in cortical EEG responses. Searching for other features of speech stimuli related to highly characteristic pattern in the EEG is therefore highly recommended for future studies. It might be beneficial to investigate, if speech features extraction methods that attempt to simulate the processing stages across the auditory system (converting the speech stimuli into some kind of

“internal representation” of speech within the auditory system), are beneficial for a correlation based classifier. Three examples for such models, which are already successfully used by other speech related systems like, e.g. (noise-robust) automatic speech recognition (ASR) systems, are the perception model (PEMO; Dau et al., 1997), Amplitude Modulation Spectrograms (AMS; Moritz et al. 2011) and the Gabor filterbank (Schädler et al., 2016).

To address future ways to overcome the problem of a limited pool of stimuli we assume now that the knowledge about the EEG representation of a set of specific speech characteristics (e.g. the stimulus envelope or its respective onsets) is given. Then for every given speech stimulus the respective EEG pattern should occur in the same order as the corresponding specific speech characteristics in the given stimulus. This assumption is very similar to the underlying computational prerequisites for ASR where a given output stream (for ASR: speech waveform, for BCI: EEG waveform) is produced by a specific input stream (for ASR: talker and string of phonemes, for BCI: spoken speech features). An approach which would use information like these may provide clear benefits for the identification of an arbitrary but given stimulus (like the suprasegmental properties of ASR are beneficial for the detection of whole sentences due to context). This does not mean, that in near future an improved BCI is able to predict or reconstruct the content of sentences that a subject had heard. Even with clearly improved classification strategies, this still appears rather unrealistic for a long time. However, the prediction whether an arbitrary speech stimulus was heard or not appears to be a reasonable goal. This can be seen as an extension of the sentence segregation task described in chapter 4 to an open set of stimuli. Such a BCI could be used, e.g., to detect specific command words or sentences while ignoring unknown speech signals. Furthermore, it probably could be trained with a rather small set of optimized words or sentences that focuses on the most essential characteristic speech features.

The BCI shown in this thesis presents some promising glimpses into the direction of a future BCI controlled hearing aid. For this purpose, it is still necessary to further investigate auditory BCI paradigms based on speech in natural acoustic environments. Nevertheless, this BCI approach is capable of segregating sentences and speakers even under the disturbance of noise, when a limited pool of sentences is used. In a situation where a limited set of sentences would still be sufficient, and the user is not able to see or move, this approach could provide a more natural alternative to an auditory P300 speller based on artificial sounds like the one shown by Klobassaa et al. (2009). For example, a locked-in-state patient living in a caretaking facility would only need a certain amount of specific sentences to communicate to his caretaker and his quality of life could be enhanced by the use of present BCI.

8 General appendix

This General appendix provides a brief overview of the paradigms and results of pilot studies that were done before and parallel to the work presented in Chapters 3-5. The first three pilot studies indicate the transition process from the BCI approach investigated in Chapter 3 to the BCI approach used in Chapter 4 and 5. During this process the focus of the approach shifted from the classification of long term spectrum information of the EEG related to speech-rhythm to classification of correlation-coefficients of evoked speech and speech envelopes. The last pilot study tested the BCI approach of Chapter 4 in a competing speaker scenario.

8.1 Fundamental frequency of vowels in the EEG

Rationale: This approach aimed to use information of the long term spectrum of the evoked EEG to words like e.g. the fundamental frequencies of the presented vowels for classification.

Paradigm: The two names "Peter" and "Ulrich" from the OLSA (male speaker) were presented 1000 times one at a time diotically to three subjects at 58 dB RMS while they watched a movie. The measuring setup and signal processing was comparable to Chapter 3.

Results: The EEG recordings showed no evidence of the fundamental frequencies of the presented vowels in the long term spectrum.

8.2 Segregation of sentences

Rationale: After the failure of the pilot shown in Chapter 8.1 this approach aimed to further investigate the spectral domain in addition of the time domain of the EEG evoked by speech to find classifiable potentials.

Paradigm: The two OLSA sentences "Peter kauft achtzehn nasse Schuhe." and "Kerstin kauft zwölf alte Blumen." (male speaker) were presented 500 times one at a time diotically to three subjects at 58 dB RMS while they watched a movie. The measuring setup and signal processing was comparable to Chapter 3. This pilot was repeated for one subject with two different OLSA sentences ("Stephan hat fünf grüne Bilder." and "Nina sieht vier grüne Tassen."; male speaker) to obtain additional comparison data.

Results: The long term spectrum and the time signal of the recorded EEG responses were analyzed. A segregation of the two sentences was not possible, neither using a correlation of the long term spectrum of single trial EEG and stimuli nor using a correlation of the single trial EEG time signal and the stimuli envelopes. The simple classification of correlation coefficients at single trial basis leads to rates at chance level, but the concept seemed promising, when averaged over epochs.

8.3 Segregation of sentences and speakers

Rationale: After the promising correlation concept shown in Chapter 8.2, this approach aimed to further test and enhance this correlation approach to make it usable for single trial classification.

Paradigm: Eight OLSA sentences from four different speakers (two females and two males, each speaking two different sentences) were presented diotically with 160 repetitions at 58 dB RMS while the subjects watched a movie. Five subjects participated. The measuring setup and signal processing was comparable to Chapter 8.2. Here a list of the presented stimuli is given (Stimuli nomenclature (ABC): A = 1 female / 2 male; B = speaker 1, 2, 3, 4; C = number of sentence 1, 2):

- 111: "Peter gibt sieben teure Autos."
- 112: "Britta verleiht elf alte Bilder."
- 121: "Ulrich gewann zwölf rote Messer."
- 122: "Nina sieht vier grüne Tassen."
- 211: "Tomas kauft neun schwere Tassen."
- 212: "Kerstin nahm acht kleine Dosen."
- 221: "Wolfgang bekommt elf schöne Messer."
- 222: "Stephan mahlt vier schwere Dosen."

Results: The results of this pilot were in line with Chapter 8.2. Several classification scenarios were tested, but no segregation based on ASSR or correlation of spectral information was possible. Therefore, spectral information of the recorded EEG does not seem to be feasible for a speech driven BCI. Further attempts of a simple classification of correlation data in time domain bases on the single trial data were without success. However, a successful proof of concept was possible using averaged data for a specific cross-correlation window. Different attempts of combining correlation data of single trial EEG and all given speech envelopes for classification input lead to the successful BCI approach presented in Chapter 4.

8.4 Competing sentences

Rationale: This pilot aimed to test the successful BCI approach shown in Chapter 4 in a competing speaker situation.

Paradigm: Four OLSA sentences from two different speakers (two females and two males) were used in a competing speaker paradigm. The two participating subjects were asked to attend to one speaker, ignore the other one and count the amount of presented target words spoken by the speaker of interest, while the German word "Tassen" was the target. The measuring setup and signal processing was comparable to Chapter 4. Following stimuli were presented with 60 repetitions at 58dB RMS during this experiment

(Stimuli nomenclature (ABC) = A[1: W left + M right, 2: M left + w right] B[1: attention do W , 2: attention to M] C[0: 0 targets, 1: 1 target, 2: 2 targets]:

- 110 : Wn + Mn; attention to women
- 120 : Wn + Mn; attention to man
- 210 : Mn + Wn; attention to women
- 220 : Mn + Wn; attention to man
- 111 : Wt + Mn; attention to women
- 121 : Wn + Mt; attention to man
- 211 : Mn + Wt; attention to women
- 221 : Mt + Wn; attention to man
- 112 : Wt + Mt; attention to women
- 122 : Wt + Mt; attention to man
- 212 : Mt + Wt; attention to women
- 222 : Mt + Wt; attention to man
- Wn: women, no target, "Ulruch gewann zwölf rote Messer."
- Wt: women, target, "Nina sieht vier grüne Tassen."
- Mt: man, target, "Tomas kauft neun schwere Tassen."
- Mn: man, no target, "Kerstin nahm acht kleine Dosen."

Results: Different variations of the current BCI approach (see Chapter 4) in respect of combining correlation information for classification input to segregate competing sentences were not successful, i.e., provide only classification accuracies at chance level.

9 References

Aiken S. J. and Picton T. W.: *Human cortical responses to the speech envelope*. Ear and Hearing 29: 139-157 (2008).

Bennington J. Y., Polich J.: Comparison of P300 from passive and active tasks for auditory and visual stimuli. Journal of Psychophysiology 34: 171 - 177 (1999).

Besserve M., Jerbi K., Laurent F., Baillet S., Martinerie J., Garnero L.: *Classification methods for ongoing EEG and MEG signals.* Biological Research 40: 415-437 (2007).

Biesmans W., Das N., Francart T., Bertrand A.: *Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario.* TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING 25(5): 402-412 (2017).

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

Birbaumer N.: Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. Psychophysiology 43: 517-532 (2006).

Blankertz B., Sannelli C., Halder S., Hammer E. M., Kübler A., Müller K. R., Curio G., Dickhaus T.: *Neurophysiological predictor of SMR-based BCI performance.* NeuroImage 51(4): 1303-1309 (2010).

Brunner P., Bianchi L., Guger C., Cincotti F., Schalk G.: *Current trends in hardware and software for brain-computer interfaces (BCIs).* J. Neural Eng. 8: 025001 (2011).

Chang M., Nishikawa N., Struzik Z. R., Mori K., Makino S., Mandic D., Rutkowski T. M.: *Comparison of P300 Responses in Auditory, Visual and Audiovisual Spatial Speller BCI Paradigms.* Proceedings of the Fifth International Brain-Computer Interface Meeting (2013). arXiv:1301.6360

Choi I., Wang L., Bharadwaj H., Shinn-Cunningham B.: Individual differences in attentional modulation of cortical responsescorrelate with selective attention performance. Hearing Research 314: 10-19 (2014).

Comerchero M. D., Polich, J.: *P3a and P3b from typical auditory and visual stimuli.* Clinical neurophysiology, 110(1): 24-30 (1999).

Cristianini N., Shawe-Taylor J.: *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, (2000).

Dau T., Kollmeier B., Kohlrausch A.: *Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers.* The Journal of the Acoustical Society of America 102(5): 2892-2905, (1997).

Dau T., Wegner O., Mellert V., Kollmeier B.: *Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion.* The Journal of the Acoustical Society of America, 107(3): 1530-1540 (2000).

Donchin E., Spencer K. M., Wijesinghe R.: The Mental Prosthesis: *Assessing the Speed of a P300-Based Brain-Computer Interface.* IEEE TRANSACTIONS ON REHABILITATION ENGINEERING 8(2): JUNE (2000).

Durban P.: *Artefakte bei der Selbstkontrolle langsamer Hirnpotentiale.* Eberhard-Karls-Universität zu Tübingen, Dissertation (2006).

Ekin B., Atlas L., Mirbagheri M., Lee A. K. C.: An alternative approach for auditory attention tracking using single-trial EEG. ICASSP (2016).

Farwell L. A., Donchin E.: *Talking off the top of your head: toward a mental prosthesis ultilizing event-related brain potentials.* Dept. of Psychology and Cognitive Psychophysiology Laboratory, University of Illionis at Urbana-Champaign, Champaign, IL 61820 (1988).

Fisher R. A.: *The use of multiple measurements in taxonomic problems.* Annals of human genetics 7(2): 179-188 (1936).

Friedman D., Cycowicz Y. M., Gaeta H.: *The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty.* Neuroscience & Biobehavioral Reviews, 25(4): 355-373 (2001).

Furdea A., Halder S., Krusienski D. J., Bross D., Nijboer F., Birbaumer N., Kübler A.: *An auditory oddball (P300) spelling system for brain-computer interfaces.* Psychophysiology 46: 617-625 (2009).

Glasberg B. and Moore B.: A Model of Loudness Applicable to Time-Varying Sounds. JAES 50: 331-342 (2002).

Grabe E., Low E.L.: *Acoustic correlates of rhythm class.* Laboratory Phonology. Vol. 7 - Mouton de Gruyter Berlin: 515–546 (2002).

Haghighi M., Moghadamfalahi M., Nezamfar H., Akcakaya M., Erdogmus D.: *Toward a brain interface for tracking attendet auditory sources.* International workshop on machine learning for signal processing, Italy, Salerno sept. 13-16 (2016).

Halder S., Rea M., Andreoni R., Nijboer F., Hammer E. M., Kleih S. C., Birbaumer N., Kübler A.: *An auditory oddball brain-computer interface for binary choices.* Clinical Neurophysiology 121: 516-523 (2010).

Herdman A. T., Lins O., Van Roon P., Stapells D. R., Scherg M., Picton T. W.: *Intracerebral sources of human auditory steady-state responses.* Brain topography 15(2): 69-86 (2002).

Hill N. J., Ricci E., Haider S., McCane L. M., Heckman S., Wolpaw J. R., Vaughan T. M.: *A practical, intuitive brain–computerinterface for communicating 'yes' or 'no' by listening.* Journal of Neural Engineering 11: 035003 (2014).

Hoffmann U., Vesin J. M., Ebrahimi T., Diserens K.: *An efficient P300-based brain-computer interface for disabled subjects.* Journal of Neuroscience Methods 167(1): 115-125 (2008).

Hoke M., Ross B., Wickesberg R., Lütkenhöner B.: *Weighted averaging theory and application to electric response audiometry.* Electroencephalography and Clinical Neurophysiology 57(5): 484-489 (1984).

Hu Y., Yu D.: *The Comparison of Five Discriminant Methods.* International conference on management and service science, China, Wuhan aug. 12-14 (2011).

Kim D. W., Hwang H. J., Lim J. H., Lee Y. H., Jung K. Y., Im C. H.: *Classification of selective attention to auditory stimuli: Toward vision-free brain-computer interfacing.* Journal of Neuroscience Methods 197: 180-185 (2011).

Klobassaa D. S., Vaughana T. M., Brunnera P., Schwartzd N. E., Wolpawa J. R., Neuperb C., Sellersa E. W.: *Toward a high-throughput auditory p300-based brain-computer interface.* Clinical Neurophysiology 120(7): 1252-1261 (2009).

Kollmeier B., Warzybok A., Hochmuth S., Zokoll M. A., Uslar V., Brand T., Wagener K. C.: *The multilingual matrix test: Principles, applications, and comparison across languages: A review.* International Journal of Audiology 54(2): 3-16 (2015).

Kottaimalai R., Pallikonda Rajasekaran M., Selvam V., Kannapiran B.: *EEG Signal Classification using Principal Component Analysis with Neural Network in Brain Computer Interface Applications.* International Conference on Emerging Trends in Computing, Communication and Nanotechnology (2013).

Kübler A., Kotchoubey B., Kaiser J., Wolpaw J. R., Birbaumer N.: *Brain–computer communication: Unlocking the locked in.* Psychological Bulletin 127(3): 358-375 (2001).

Lopez M. A., Pomares H., Pelayo F., Urquiza J., Perez J.: *Evidences of cognitive effects over auditory steady-state responses by means of artificial neural networks and its use in brain-computer interfaces.* Neurocomputing 72(16-18): 3617-3623 (2009).

Martinez, Mario Castro A., Moritz N., Meyer B. T.: *Should deep neural nets have ears? The role of auditory features in deep learning approaches.* Interspeech, September: 2435-2439 (2014).

Mesgarani N., David S. V., Fritz J. B., Shamma S. A.: *Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex*. Journal of Neurophysiology 102: 3329-3339 (2009).

Meyer B. T., Kollmeier B., Ooster J.: Autonomous measurement of speech intelligibility utilizing automatic speech recognition. In Proc. Interspeech (2015).

Mika S., Ratsch G., Weston J., Scholkopf B.,Mullers K. R.: *Fisher discriminant analysis with kernels.* Neural networks for signal processing IX, (1999).

Mika S.: *Kernel fisher discriminants.* Dissertation, Berlin (2003).

Moore B. C. J., Thwaites A., Glasberg B. R., Nimmo-Smith I., Marslen-Wilson W. D.: *Representation of Instantaneous and Short-Term Loudness in the Human Cortex.* Frontiers in Neuroscience 10: 183 (2016).

Moritz N., Anemuller J., Kollmeier B.: *Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments.* In Proceedings of ICASSP: 5492-5495 (2011).

Näätänen R., Paavilainen P., Rinne T., Alho K.: *The mismatch negativity (MMN) in basic research of central auditory processing: a review.* Clinical neurophysiology 118(12): 2544-2590 (2007).

Nakamura T., Namba H., Matsumoto T.: *Classification of Auditory Steady-State Responses to Speech Data.* 6th Annual International IEEE EMBS Conference on Neural Engineering, San Diego -California (6 - 8 November, 2013).

O'Sullivan J. A., Power A. J., Mesgarani N., Rajaram S., Foxe J. J., Shinn-Cunningham B. G., Slaney M., Shamma S. A., Lalor E. C.: *Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG*. Cerebral Cortex Advance Access (2014).

O'Sullivan J. A., Shamma S. A., Lalor E. C.: Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. The Journal of Neuroscience 35(18): 7256 - 7263 (2015).

Pasley B. N., David S. V., Mesgarani N., Flinker A., Shamma S. A., Crone N. E., Knight R. T., Chang E. F.: *Reconstructing speech from human auditory cortex.* PLoS Biology 10:1 (2012).

Petersen E. B., Wöstmann M, Obleser J., Lunner T.: Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J. Neurophysiol 117: 18-27 (2017).

Pritchard W. S.: *Psychophysiology of P300.* Psychological Bulletin, 89(3): 506-540 (1981).

Picton T. W., Alain C., Otten L., Ritter W., Achim A.: *Mismatch negativity: different water in the same river.* Audiology and Neurotology 5(3-4): 111-139 (2000).

Picton T. W., Hillyard S. A.: *Human Auditory Evoked Potentials. II: Effects of Attention.* Electroencephalography and Clinical Neurophysiology 36: 191-199 (1974).

Picton T. W., John M. S., Dimitrijevic A., Purcell D.: *Human auditory steady-state response.* International Journal of Audiology 42(4): 177-219 (2003).

Picton T. W., Skinner C. R., Champagne S. C., Kellett A. J. C., Maiste A. C.: *Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone.* Journal of Acoustical Society of America 82: 165–78 (1987).

Riedel H., Granzow M., Kollmeier B.: Single-sweep-based methods to improve the quality of auditory brain stem responses. Part II: averaging methods. Journal of Audiological Acoustics 40(2): 62-85 (2002).

Rieke F., Bodnar D. A., Bialek W.: Naturalistic Stimuli Increase the Rate and Efficiency of Information Transmission by Primary Auditory Afferents. Biological Sciences 262: 259-265 (1995).

Schädler M. R., Warzybok A., Ewert S. D., Kollmeier B.: *A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception.* Journal of Acoustical Society of America 139 (2016).

Scherg M.: *Akustisch evozierte Potentiale: Grundlagen-Entstehungsmechanismen-Quellenmodell.* (1991).

Sellers E. W., Donchin E.: *A P300-based brain-computer interface: Initial tests by ALS patients.* Clinical Neurophysiology 117: 538-548 (2006).

Shannon C. E., Weaver W.: *Mathematical Theory of Communication.* The University Illinois Press, Urbana (1964).

Schmidhuber J.: *Deep learning in neural networks: An overview.* Neural Networks 61: 85-117 (2015).

Speier W., Arnold C., Pouratian N.: Evaluating True BCI Communication Rate through Mutual Information and Language Models. PLOS ONE 8:10 (2013).

Vidal J. J.: *Toward direct Brain-Computer Communication.* Annu. Rev. Biophys. Bioeng. 2: 157-180 (1973).

Wagener K. and Brand T.: Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. International Journal of Audiology 44: 144-156 (2005).

Wagener K. and Kollmeier B.: Evaluation des Oldenburger Satztests mit Kindern und Oldenburger Kinder-Satztest (Evaluation of the Oldenburg sentence test with children and the Oldenburg children's sentence test). Z Audiol 44(3): 134-143 (2005).

Wagener K., Kühnel V., Kollmeier B.: Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test. Zeitschrift Fur Audiologie 38: 4-15 (1999).

Wagener K., Kühnel V., Kollmeier B.: Entwicklung und Evaluation eines Satztests in deutscher Sprache III: Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test). Z Audiol 38(3): 86-95 (1999c).

Zschocke S., Hansen H. C.: *Klinische Elektroenzephalographie. Entstehungsmechanismen des EEG.* Springer-Verlag Berlin Heidelberg (2012).

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Außerdem versichere ich, dass ich die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichung, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe.

_____

Oldenburg, den 26. März 2019