Machine learning approaches for detection of robot errors in human brain signals

Joos Behncke

Technische Fakultät
Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

Betreuer: Prof. Dr. Wolfram Burgard

Machine learning approaches for detection of robot errors in human brain signals

Joos Behncke

Dissertation zur Erlangung des akademischen Grades Doktor der Naturwissenschaften
Technische Fakultät, Albert-Ludwigs-Universität Freiburg

Dekan	Prof. Dr. Rolf Backofen
Erstgutachter	Prof. Dr. Wolfram Burgard Albert-Ludwigs-Universität Freiburg
Zweitgutachter	Prof. Dr. Tonio Ball Albert-Ludwigs-Universität Freiburg
Tag der Disputation	30. April 2020

Zusammenfassung

Fehler zu begehen ist menschlich. Jedoch eigene Fehler zu reflektieren und daraus zu lernen, ist eine unschätzbar wertvolle Fähigkeit. Dabei ist das Erkennen von Fehlern oder Irrtümern ebenso grundlegend wichtig wie die Fähigkeit Anpassungen vorzunehmen, was in ähnlichen oder gleichen Szenarien ein Wiederholen von Fehlern verhindern könnte. Diese Prozesse ermöglichen eine effiziente Gestaltung für zukünftige Aufgaben. Gleiches gilt für intelligente Robotersysteme. Auch hier können fehlerhafte Ausführungen auftreten, was bei der Zusammenarbeit mit Menschen zu einem kritischen Sicherheitsproblem führen kann. Aus diesem Grund ist es notwendig, auftretende Fehler rechtzeitig zu erkennen, nicht nur um das System zu stoppen, sondern auch, um aus diesen Fehlern zu lernen und in Zukunftsszenarien nicht wieder dieselben Fehler zu begehen. Adaptive Systeme können Fehler in Echtzeit erkennen und Anpassungen vornehmen, um ein schnelles Lernen zu ermöglichen.

In der industriellen Fertigung wird bereits eine Vielzahl intelligenter und autonomer Robotersysteme eingesetzt, wobei die Stichworte *Künstliche Intelligenz* (KI) und *Industrie 4.0* immer häufiger fallen. Ein Teilaspekt dieser hochkomplexen Vernetzungsprozesse wird auch von kollaborativen Systemen zwischen Mensch und Maschine gespielt und Konzepte, die auf physiologischen Daten basieren, rücken in den Fokus. Hirnsignale sind wahrscheinlich die komplexeste, aber auch die vielversprechendste Form von Kontrollsignalen. Echtzeitanalysen können die Effizienz deutlich steigern und Sicherheitssysteme in Szenarien der direkten Zusammenarbeit unterstützen. Auch in Bereichen wie dem Gesundheitswesen, in denen intelligente Roboter-Assistenten zahlreiche Aufgaben übernehmen könnten, ist eine intuitive Steuerung unerlässlich.

Lösungen auf Basis von Robotersystemen werden in der Regel nicht ausschließlich vom Anwender gesteuert. In der Regel sind es vor allem autonome intelligente Subsysteme, die entscheidende Schritte in einem Prozess übernehmen und nur grob durch menschliche Hilfe gesteuert werden. Gerade in der industriellen Produktion wird dieses Kooperationsmodell häufig eingesetzt. In anderen Bereichen ist der Stand der Technik noch nicht so weit fortgeschritten, aber zahlreiche Forschungsprojekte beschäftigen sich mit diesem Thema. Beispielsweise verfolgen Hilfssysteme für Menschen mit motorischen Defiziten das Ziel, den Nutzern mehr Autonomie zurückzugeben und autonome Roboterassistenten ermöglichen lüssigkeitsaufnahme ohne weitere menschliche Hilfe. Um weiter den Prozess des autonomen Trinkens zu optimieren, gibt es zum Beispiel Studien, die versuchen, den Füllstand in einem Becher zu erfassen oder den Gießprozess, basierend auf maschinellen Lerntechniken zu, erfassen. Ebenso werden ganzheitlichere Systeme entwickelt, die es unter anderem ermöglichen, über bewusste Hirnsignale mittels eines übergeordneten Frameworks mit einem intelligenten Roboter-Service-Assistenten zu kommunizieren.

Auch wenn diese Subsysteme an sich funktionieren, braucht es einen Benutzer, der

diese Systeme in Gang setzt und bewusst steuert. Erfolgt die Kommunikation über Hirnsignale des Benutzers, spricht man von einem *Brain Computer Interface*, kurz BCI. Diese Schnittstellen wurden zunächst vor allem für schwerstgelähmte Patienten entwickelt, aber auch häufig in ganz anderen Bereichen wie der Unterhaltungsindustrie eingesetzt. Die Schwierigkeit bei der Arbeit mit Hirnsignalen besteht jedoch darin, die gewünschten Mustertypen zu erkennen und von anderen Typen zu unterscheiden. Für die praktische Anwendung ist eine hohe Zuverlässigkeit der Detektionssysteme erforderlich. Einige maschinelle Lerntechniken wie zum Beispiel *linear discriminant analysis* (LDA), *support vector machines* (SVM) oder auch *common spatial patterns* (CSP) haben sich bei der Klassifizierung von Hirndaten etabliert. Die Leistungen liegen jedoch nicht unbedingt in den gewünschten und für die Praxis erforderlichen Bereichen, wie zum Beispiel bei der Dekodierung fehlerbezogener Signale. Solche Fehlersignale können in vielerlei Hinsicht zur Verbesserung eines BCI-Systems beitragen. Ein wichtiger Beitrag dieser Arbeit ist die Untersuchung dieser Signaltypen und die Optimierung ihrer Klassifizierung mit verschiedenen maschinellen Lerntechniken. Dabei werden unter anderem Methoden aus dem Bereich des Deep Learning eingesetzt, die in jüngster Zeit enorme Erfolge im Bereich der maschinellen Bildverarbeitung erzielt haben - eine Aufgabe, bei der der Mensch bisher jedem Maschinensystem überlegen war.

Eine weitere Schwierigkeit bei der Verwendung von BCIs besteht darin, dass ein für einen einzelnen Benutzer eingerichtetes System nicht allgemein verwendbar ist, sondern individuell an die zerebralen Antworten jedes weiteren Benutzers angepasst wird. Signale verschiedener Nutzer können völlig andersartig erscheinen und eine unterschiedliche räumliche Verteilung aufweisen, weshalb es notwendig ist, das System an den jeweiligen Nutzer anzupassen. Auch für den Fall, dass es nur von einem Benutzer verwendet wird, kann es sein, dass dieses funktionierende System aufgrund von Nicht-Stationaritäten nach einer gewissen Zeit neu kalibriert werden muss. Besonders hilfreich könnten hier Systeme sein, die Informationen für eine bestimmte Aufgabe speichern und die generalisierten Informationen für eine breite Anwendung zur Verfügung stellen können. Dies ist gerade dann von Vorteil, wenn nur wenige Daten zur Verfügung stehen, eine zuverlässige Dekodierung aber dennoch erforderlich ist.

Verschiedene motorische Muster können als Steuersignale für BCIs verwendet werden, aber auch kognitive Prozesse können diese Aufgabe erfüllen. Betrachtet man eine direkte Zusammenarbeit zwischen Mensch und Roboter, so kann, wie bereits erwähnt, eine schnelle und zuverlässige Fehlererkennung zu erhöhten Sicherheitsstandards beitragen, aber auch eine Verfeinerung der Steuerung ermöglichen. Eine komplizierte Integration eines Befehls *FEHLER* über ein alternatives Steuersignal im menschlichen EEG und dessen Detektion sind dabei unerwünscht. Vielmehr ist ein natürliches intrinsisches Muster erforderlich, das direkt bei der Wahrnehmung von Fehlern in den Signalen des Gehirns erzeugt wird, ob bewusst oder unbewusst, und sofort genutzt werden kann. Eines der am meisten erforschten Phänomene in diesem Zusammenhang ist eine negativer Abfall im Potential des EEGs, der bereits in den ersten 300 ms mit dem Auftreten eines Fehlers sichtbar ist, die *error-related negity* (ERN) oder *error negativity* (Ne). Dieses ereignisbezogene Potential tritt auf, wenn ein Fehler beobachtet oder begangen wird, wobei die Amplitude des Signals durch die subjektive Bedeutung sowie die subjektive

Wahrnehmung des Fehlers moduliert wird. Darüber hinaus gibt es im Spektralbereich nur wenige Untersuchungen, die die Komplexität der Fehlerverarbeitung über einzelne Komponenten hinweg beschreiben können. Neben der Relevanz für eine optimale Erkennung von Fehlersignalen hat die Untersuchung von Fehlerprozessen einen hohen wissenschaftlichen Wert. Neue Erkenntnisse in diesem Bereich können zum Verständnis der zeitlichen und räumlichen Fehlerverarbeitung beitragen und einen tieferen Einblick in das Zusammenspiel der verschiedenen Hirnareale geben. Ebenso ist die funktionelle Aufteilung des Gehirns bis heute nicht vollständig erklärt und verstanden.

Nicht-invasive Verfahren wie das Oberflächen-EEG, die die Spannung auf der Kopfhaut messen, können nur Hirnsignale aufnehmen, die oberflächennah erzeugt werden und sind daher weitgehend auf oberflächennahe kortikale Hirnareale beschränkt. Darüber hinaus führt der Abstand zum signalgebenden Gewebe zur Überlagerung vieler verschiedener Signale. Die Aufnahmen stellen somit vielmehr die Gesamtheit der Informationen mehrerer Hirnareale dar. Die Filterwirkung des Schädels und Artefakte durch muskuläre Aktivität erschweren Messungen dabei zusätzlich. Im Gegensatz dazu bieten Methoden, die Signale intrakraniell aufnehmen, das heißt innerhalb des Schädels, mehrere Vorteile. Um jedoch zuverlässige Aussagen über neurophysiologische Aktivitäten und Zusammenhänge von Hirnarealen treffen zu können, ist eine genaue Zuordnung zum darunter liegendem Gewebe unerlässlich. Dies kann durch Abbildung auf den Atlas eines Standardgehirns erreicht werden, der die Bereiche des Gehirns zytoarchitektonisch, das heißt basierend auf der zellulären Zusammensetzung des Gewebes, mit post mortem Gehirnen abbildet. Allerdings wird das weiche Hirngewebe während der Implantation verformt, was den Vergleich einer individuellen Struktur mit dem allgemeinen Fall erschwert. Darüber hinaus kann ein *Standard*-Gehirn nur grob allgemein formuliert werden, hat aber nicht unbedingt Gültigkeit für jeden Punkt, der für die einzelnen Gehirne in Betracht gezogen wird und berücksichtigt keine individuellen Eigenschaften. Gerade in Regionen, in denen verschiedene Bereiche aneinandergrenzen, kann eine zuverlässige Zuordnung kritisch werden.

Aus den eben aufgezeigten Problemen und Herausforderungen ergeben sich im Hinblick auf die Fehlererkennung anhand menschlicher Hirnsignale folgenden Fragen, die in dieser Arbeit behandelt werden:

- Kann das nicht-invasive EEG menschlicher Beobachter verwendet werden, um fehlerhafte Ausführungen von Robotersystemen zu entschlüsseln?
- Kann die Klassifizierung von fehlerbezogenen Signalen durch den Einsatz von convolutional neural networks (CNNs) verbessert werden? Hilft die Visualisierung dabei, die Ergebnisse neurophysiologisch zu interpretieren?
- Ist es prinzipiell möglich, Robotertypen basierend auf dem EEG zu unterscheiden und wie beeinflusst der Robotertyp die Fehlererkennung? Haben Kriterien wie die Anzahl menschlicher Ähnlichkeitsmerkmale einen Einfluss auf die Ergebnisse?
- Kann die Fehlerdekodierung mit CNNs basierend auf intrakraniellen Aufzeichnungen erfolgen, wenn der Benutzer selbst Fehler begeht? Hängt das Ergebnis davon ab, wie stark der Fehler subjektiv wahrgenommen wird oder wie real er erscheint?

- Sind CNNs in der Lage in einer Aufgabe generalisierende Informationen zu lernen, um sie in einer anderen Aufgabe gewinnbringend einzusetzen?
- Kann das intrakranielle EEG zum Verständnis der zeitlichen und räumlichen Verarbeitung von Fehlerprozessen beitragen? Gibt es Hinweise auf entscheidende Merkmale in diesen Prozessen?
- Wie können die Schwierigkeiten bei der Zuordnung von intrakraniellen Elektrodenkontakten zu den darunter liegenden Hirnarealen gelöst werden, wenn zum Beispiel Verformungen durch Implantation, aber auch individuelle Eigenschaften das Prozedere erschweren?

Zunächst kann grundsätzlich gezeigt werden, dass eine fehlerhafte Roboterausführung einer instruierten Aufgabe basierend auf dem nicht-invasiven EEG menschlicher Beobachter unter Verwendung des verbreiteten Common Spatial Pattern (CSP) Algorithmus dekodiert werden kann. Es ist ebenfalls möglich, den Typus des Roboters anhand des EEG zu unterscheiden. Zu diesem Zweck wurden zwei Experimente konzipiert und durchgeführt. Der visuelle Reiz entsteht dabei durch eine richtig oder falsch ausgeführte Einschenk- oder Hebeaufgabe des Roboters, zwei grundlegende und einfache Prozesse einer Mensch-Roboter-Zusammenarbeit. Insgesamt 18 Probanden nahmen an der passiven Beobachtungsaufgabe teil und prüften, ob sie erkennbare Fehlermuster erzeugen, während die Roboter die angewiesenen Aufgaben ausführten. Der daraus resultierende Datensatz stellt somit eine geeignete Grundlage für die Untersuchung von fehlerbezogenen EEG-Phänomenen dar.

Unter Verwendung eines deep convolutional neural networks kann die Qualität der Fehlerdetektion deutlich und signifikant verbessert werden. Das bisher untersuchte Problem der Fehlerdekodierung wird auf Seiten des Dekodierers optimiert. Das CNN schneidet deutlich besser ab, sowohl als die vorherige CSP-Implementierung als auch als die konventionelle lineare Diskriminanzanalyse (LDA), die hier zum Vergleich verwendet wird. Dies bestätigt den allgemeinen Trend der EEG-Analyse, auf neue Methoden wie CNNs zurückzugreifen und erweitert ihn um den Bereich der Fehlererkennung. Die Visualisierung der Ergebnisse zeigt wichtige Eigenschaften der Klassifikationsmerkmale, die für dieses Problem in der Zeitdomäne zu liegen scheinen. Geht es um die Unterscheidung zwischen Robotertypen, so übertreffen die CNNs ebenfalls die Analysen mit früheren Standards. Die Ergebnisse geben erste Hinweise auf Unterschiede in der Wirkung von Robotern auf den Menschen bezüglich der menschlichen Ähnlichkeitsskalen des Roboters. Die Visualisierung der erlernten Klassifikationsmerkmale liefert Informationen über die zugrundeliegenden Prozesse. Darüber hinaus wird untersucht, inwiefern die Klassifikation von Fehlern von der Anzahl der menschenähnlichen Merkmale eines Roboters abhängt.

Des Weiteren wird eine neue Methode zur Zuordnung von intrakraniellen Elektrodenkontakten zu Hirnarealen anhand von Magnetic Resonance Imaging (MRI) vorgestellt. Der Algorithmus ist besonders nützlich, wenn Implantationen im subduralen Raum zu Verformungen des Gehirns führen und eine genaue Zuordnung von Elektrodenkontakten zu Hirnarealen schwierig wird. Das Verfahren zielt darauf ab, die Bereiche zu bestimmen, die zu dem von einem Elektrodenkontakt aufgezeichneten Signal beitragen

und basiert auf der Berücksichtigung individueller Eigenschaften des Gehirns und einer Rücktransformation in die kortikalen Bereiche des Gehirns. Als Ergebnis werden Wahrscheinlichkeiten für den möglichen Einfluss eines Hirnareals auf den entsprechenden Elektrodenkontakt ausgegeben. Das Verfahren wurde in eine Softwareumgebung eingebettet, die eine benutzerfreundliche Anwendung ermöglicht. Das Softwarepaket beinhaltet die visuell unterstützte Identifikation der Elektrodenkontakte, die automatische Zuordnung und eine 3D-Visualisierung inklusive Virtual-Reality-Export. Insbesondere im Bereich der klinischen Forschung ist es möglich, selbst ohne Programmierkenntnisse von den Methoden zu profitieren und Erkenntnisse über die genaue Position der Elektrodenkontakte zur besseren Interpretation auftretender Phänomene zu erlangen. Die Software ist frei verfügbar und in der von Wissenschaft und Industrie weit verbeiteten Software MATLAB eingebettet.

Neben fehlerhafter Ausführung jeglicher Effektoren können Fehler auch vom Anwender selbst begangen werden. Es wird ein Paradigma vorgestellt, das sowohl auf motorische als auch auf kognitive Antworten untersucht werden kann und zu einem immensen Datensatz von intrakraniellen Aufnahmen beigetragen hat. Dazu gehören Aufnahmen von 47 Epilepsiepatienten, die alle unterschiedliche Implantationen aufwiesen und damit zahlreiche verschiedene Hirnareale abdeckten. Darüber hinaus wurde diesem Datensatz ein weiterer Vergleichssatz hinzugefügt, der Aufzeichnungen eines weiteren Paradigmas für 15 der 47 Patienten enthält, das auch die Analyse von Fehlersignalen ermöglicht. Es werden Analysen vorgestellt, die grundlegende Erkenntnisse über die Verarbeitung von Fehlern im menschlichen Gehirn geben, sowohl auf temporaler als auch auf spektraler Ebene. Darüber hinaus werden gemeinsame und damit möglicherweise allgemeine Merkmale in der zerebralen Verarbeitung der hier untersuchten Fehler in den beiden unterschiedlichen Paradigmen aufgedeckt. Die Ergebnisse können zur Untersuchung der Verallgemeinerung von fehlerbezogenen Mustern für den Transfer über Paradigmen hinweg beitragen.

Basierend auf den eben vorgestellten Daten wird das Potential von CNNs zur Fehlererkennung durch intrakranielles EEG nachgewiesen. Der Versuch, allgemeine fehlerbezogene Informationen auf weitere Paradigmen zu übertragen, zeigt, dass ein Vortrainieren der CNNs zu signifikanten Verbesserungen führen kann, insbesondere bei wenig verfügbaren Daten. Dies ebnet den Weg für das Antrainieren eines allgemeinen Fehlermusters, das für viele verschiedene Fehlertypen geeignet ist und somit schneller und zuverlässiger erkannt werden kann. Insbesondere in Verbindung mit BCIs kann diese Generalisierungsmethode helfen, die große Anzahl potenziell auftretender Fehler, aber auch andere Arten von Kontrollsignalen, zu erkennen.

Die dieser Arbeit zugrundeliegenden Experimente wurden allesamt systematisch geplant, durchgeführt und sorgfältig ausgewertet. Alle vorgestellten Methoden wurden ebenfalls mit größter Sorgfalt implementiert und Ergebnisse auf statistische Relevanz überprüft sowie anhand der Literatur gegengeprüft. Zusammenfassend stellt diese Arbeit einen wertvollen Beitrag dar, Fehler in einer Mensch-Roboter-Kooperation anhand menschlicher Hirnsignale zu erkennen, wobei die entwickelten Methoden nicht zwangsläufig auf diesen Typus von Signal beschränkt sein müssen. Neben hervorgebrachten fundamentalen Erkenntnissen zur fortlaufenden neurophysiologischen Erforschung von Fehlersignalen im Speziellen, trägt die entwickelte Methode ELAS allgemein dazu bei, verlässliche Zuord-

nung von Elektroden zu Hirnarealen und somit korrekte Interpretationen von zerebralen Phänomenen zu garantieren.

Abstract

An industry in which processes are becoming increasingly complex and terms such as artificial intelligence, networking and increasing digitization are becoming more an more important, reliable automated processes and security are fundamental. In cooperation with humans, a reliable error detection subsystem in an automated intelligent robotic system can provide increased security and adaptivity. If this system is based on human brain signals, it is called Brain Computer Interface (BCI). For practical applicability, however, high decoding accuracies are required for the detection of errors, which all implementations currently have to struggle with.

First, this thesis classifies errors committed by robots using conventional methods by means of electroencephalography (EEG) of a human observer. Using deep convolutional neural networks (CNNs) the performance can be significantly improved. This also applies to the differentiation of robot types and conclusions can be drawn about the appearance of the robot. In a second approach, the potential of the CNN architecture for error detection is confirmed on intracranial EEG (iEEG), where errors are generated by incorrect execution of the user himself. In order to imitate everyday situations in which little data is available for training, information is transferred across paradigms, finetuned with successively increasing available data and then classified. This leads to a significant improvement of performance in the case of little data in fine tuning.

Surface EEG cannot pick up signals directly at the tissue and recordings originate largely from the cortical areas near the surface, whereas iEEG is not limited by these circumstances. In order to gain a deeper understanding of error processing in the brain, the data obtained from the iEEG are examined for error-related information and the comprehensive involvement of the different areas is revealed. The power increase turns out to be a dominant feature of error processing.

A newly developed algorithm for the assignment of electrode contacts to brain areas is presented to compensate for deformations during implantations and individual differences in human brains. This algorithm is based on cortical retransformation and individual landmarks, and uses probabilistic, cytoarchitectonically-defined maps, and improves the assignment in the evaluation. The algorithm is also embedded in a user-friendly interface that can be used without any programming experience.

All methods presented were implemented with great care and results were checked for statistical significance and verified by the literature. In summary, this work represents a valuable contribution to detection of errors in human-robot cooperations by means of human brain signals, although the methods developed do not necessarily have to be limited to this type of signal. In addition to the fundamental findings gained in the ongoing neurophysiological research of error signals in particular, the developed ELAS method contributes in general to correct interpretations of cerebral phenomena.

Acknowledgments

In addition to the time required, a lot of hard work and stamina is necessary to finish a doctoral thesis, which requires a lot of support. At this point I would like to pay tribute to all those who contributed to the completion of the thesis with their ideas and encouragements.

First of all, I would like to thank Wolfram Burgard and Tonio Ball, my supervisors and mentors, who integrated me into their team of excellent colleagues and provided a stimulating working environment. During my work in the Autonomous Intelligent Systems group and the Medical AI Lab, I was able to learn from their experience in science, but also from their publishing and presentation skills. Within the BrainLinks-BrainTools Cluster of Excellence I also had the unique opportunity to gain insights into different disciplines and to work on interdisciplinary projects.

I would also like to thank Andreas Schulze-Bonhage and Petr Marusič who made the work with the invaluable intracranial data possible. In this context, I would like to appreciate the help of the physicians and medical technical assistants of the epilepsy centres in Freiburg and Prague, especially André Haak.

The creation of the huge intracranial data set would not have been possible alone. At this point, a big thankyou goes to Jiří Hammer and Martin Völker, also for technical discussions. Likewise I would like to mention Dominik Welke and Marina Hader, with whose help the EEG recordings of the robot observations could be realized. Furthermore, I would like to thank Robin Schirrmeister for bringing me closer to machine learning based on deep convolutional neural networks.

The pleasant contact with the colleagues of the Medical AI Lab and the fruitful but also entertaining conversations contributed to not losing the focus. That is why I would like to take this opportunity to thank my colleagues at the lab. In detail, I would like to mention Martin Glasstetter, Katrin Usai, Markus Kern, Alexis Gkogkidis, Stephan Hertweck, Sofie Berberich, Pia Hagen-Wiest, Bella Diekmann, Lukas Gemein and Roland Berkemeier.

Furthermore, I would like to thank the colleagues of the Autonomous Intelligent Systems group, but especially Chau Do for the inspiring discussions during the joint appearances at the Baden-Württemberg Stiftung. Thank you to all co-authors of the journal and conference publications.

I would like to appreciate the support of Carmen Schneider and Susanne Bourjaillat in technical and administrative questions.

I also would like to gratefully acknowledge that the German Research Foundation (DFG) generously supported the work on this thesis within the BrainLinks-BrainTools Cluster of Excellence (grant number EXC 1086).

Finally, it remains to be said that I am lucky to be surrounded by wonderful people. A huge thankyou to my family, my friends and especially to Adriana.

Contents

Chapter 1

Introduction

Making mistakes is human. However, the ability to reflect our own mistakes and to learn from them is one of the most valuable qualities we possess. To this, recognizing mistakes or errors is fundamental, likewise the ability to make adjustments, that might prevent once more a fail in recurring scenarios. These processes enable an efficient design of approaching tasks. The same applies to intelligent robotic systems. Their application can also lead to faulty execution, which can be a critical safety problem when collaborating with humans. For this purpose it is necessary to recognize occurring errors in time to stop the system, but also to learn from these errors in order not to commit the same errors again in future scenarios. Adaptive systems can detect errors in real time and make adjustments to enable rapid learning.

A wide range of intelligent and autonomous robot systems is already being used in industrial manufacturing, where the keywords *Artificial Intelligence* (AI) and *Industry 4.0* are currently attracting more and more attention. A partial aspect of this highly complicated networking processes is also played by collaborative systems between man and machine, and concepts based on physiological data come into focus. Brain signals are probably the most complex but also the most promising form of control signals. Real-time analyses can significantly increase efficiency and support safety systems in scenarios of direct cooperation. Also for instance in fields like healthcare, where intelligent robotic service assistants could take on numerous tasks, an intuitive control is fundamental.

Solutions based on robotic systems are usually not exclusively controlled by the user. As a rule, it is primarily autonomous intelligent subsystems that take over decisive steps in a process and are only roughly controlled by the addition of human help. Especially in industrial production, this model of cooperation is often used. In other areas, the state of the art is not yet so advanced, but numerous research projects are dealing with this topic. Auxiliary systems for people with any motor deficits, for example, pursue the goal of giving more autonomy back to users and, e.g., autonomous robotic assistants can enable intake of fluids without further human care. To give further examples, in order to optimize the process of autonomous drinking, there are also efforts to detect the fluid level in a cup or to learn the pouring process based on machine learning techniques. More holistic systems are developed likewise, which for example enable users to communicate with an intelligent robotic service assistant via conscious brain signals by means of a high-level framework.

Even if these subsystems function per se, it requires a user who deliberately controls

and sets in motion these systems. If the communication takes place via brain signals of the user, one speaks of a *Brain Computer Interface*, or BCI for short. These interfaces were initially developed primarily for severely paralyzed patients, but are also frequently used in completely different areas such as the entertainment industry. However, the difficulty of working with brain signals is the recognition of desired pattern types and their distinction from other types. For practical application a high reliability of the detection systems is necessary. Some machine learning techniques as for example *linear discriminant analysis* (LDA), *support vector machines* (SVM) but also *common spatial patterns* (CSP) have become established methods when it comes to classifying brain data. However, the performances are not necessarily in the desired and for practical purposes required ranges, for instance when it comes to decoding error-related responses. Though, such error signals can conduce to the improvement of a BCI system in many ways. An important contribution of this thesis is the investigation of these signal types and the optimization of the their classification using different machine learning techniques. Among other things, methods from the field of deep learning are used, which have recently achieved enormous success in the field of machine vision, a task in which until now humans have been superior to any machine system. Fig. 1.1 exemplarily shows possible faulty robotic executions in an autonomous drinking scenario.

A further difficulty with the use of BCIs is that a system established for a user is not generally defined, but individually adapted to the responses of this user. Signals of different users can appear completely diverse and exhibit distinct spatial distribution, which is why it is necessary to adapt the system to the individual user. And even when used by only one user, a functioning system may need to be recalibrated after a certain period due to non-stationarities. Systems that can store information for a specific task and make the generalizing part available for a broad application could be particularly helpful here. This is also the case if only little data is available, but reliable decoding is still needed.

Various motor patterns can be used as control signals for BCIs, but also cognitive processes can accomplish this task. As already mentioned, if one considers a direct cooperation between humans and robots, fast and reliable error recognition can contribute to increased safety standards, but also allows the refinement of control. A complicated integration of a command ”error” via another substitute control signal in the human EEG and its detection would be undesirable. Rather, a natural intrinsic pattern is required that is generated directly upon, conscious or unconscious, perception of errors in the signals of the brain and can be tapped immediately. One of the most researched phenomena in this context is a negative deflection in the EEG, already visible in the first 300 ms with the occurrence of an error, the *error-related negativity* (ERN) or *error negativity* (Ne). This *event-related potential* (ERP) occurs when observing or committing an error, where the amplitude of the deflection is modulated by the subjective importance and perception of the error. Beyond that, for example in the spectral domain, there are few investigations which can describe the complexity of the processing of errors beyond single components. Besides the relevance for an optimal detection of error signals, the investigation of error processes has a high scientific value. New findings in this field can contribute to an understanding of the temporal and spatial processing of error processing and give a deeper

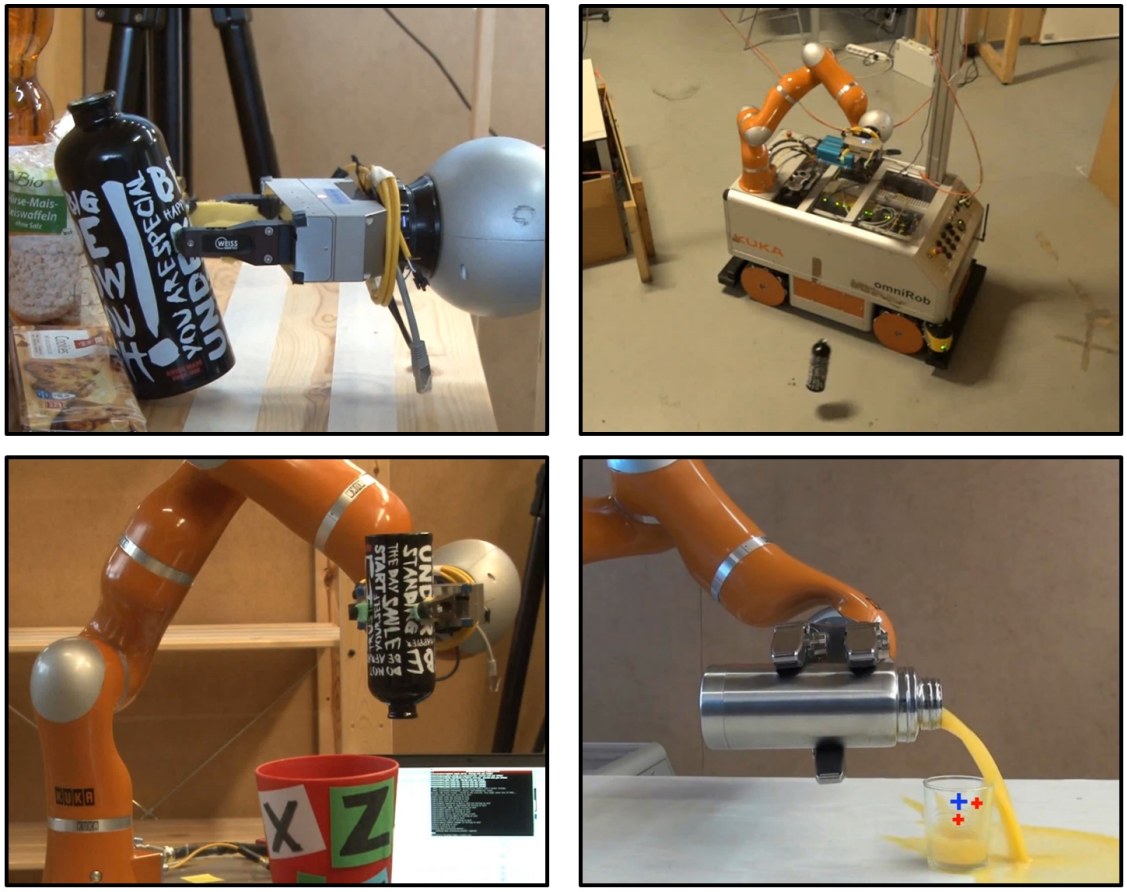

Figure 1.1: Erroneous robotic execution. Examples of situations where a human robot interaction might go wrong: robotic arm LBR iiwa (LBR iiwa, KUKA Roboter GmbH, Augsburg, Germany) is unsuccessful in executing instructed task in an autonomous drinking scenario.

insight into the interaction of the different brain areas. Similarly, the functional allocation of the brain cannot be fully explained and understood to this day.

Non-invasive methods such as the surface EEG, which measure the voltage at the scalp, can only pick up brain signals that are generated close to the surface and are thus largely restricted to cortical brain areas near the surface. In addition, the distance to the signal-generating tissue leads to the superposition of many different signals and the recordings rather represent the entirety of the information of several areas. The filtering effect of the intermediate skull and the artifacts from muscles also hardly simplify the task. By contrast, methods that pick up signals intracranial, i.e. within the skull, offer several advantages. Though, if reliable statements about neurophysiological activities and connections of brain areas are to be made, an exact assignment to the underlying tissue is indispensable. This can be achieved by mapping to the atlas of a standard brain, which generally maps the areas of the brain cytoarchitectonically, i.e. based on the cellular composition of the tissue, using post mortem brains. However, the soft cerebral tissue is deformed during implantation, which makes the comparison of an individual structure to the general case

more difficult. In addition, a ”standard” brain can only be roughly formulated in general terms, but does not necessarily have validity for every point considered for the individual brains and does not take individual characteristics into account. Especially in regions where different areas border each other, a reliable assignment can become critical.

With regard to error detection based on human brain signals, the following questions arise from the problems and challenges presented, and are dealt with in this thesis:

- Can the non-invasive EEG of human observers be used to decode faulty execution of robotic systems?
- Can the classification of error-related signals be improved by using deep learning? Does visualization help to interpret the results neurophysiologically?
- Is it in principle possible to distinguish robot types based on the EEG and how does the type of robot affect the detection of errors? Do criteria such as the number of human similarity characteristics have an influence on the results?
- Can error decoding take place deep learning based on intracranial recordings when users make errors themselves? Does the performance depend on how strong the error is subjectively perceived or how realistic it seems to be?
- Are convolutional neural networks (CNNs) able to learn generalizing information in one task in order to use it profitably in another task?
- Can the intracranial EEG contribute to an understanding of the temporal and spatial processing of error processes? Is there any indication of decisive features in these processes?
- How can the difficulties in assigning intracranial electrode contacts to underlying brain areas be solved, if for example deformations due to implantation but also individual characteristics complicate the procedure?

1.1 Outline

The thesis is structured in the following way. First the basics of the thesis are explained in Chap. 2. Since the work takes up numerous concepts and analysis methods of neuroscience in addition to the main focus on computer science, it is difficult to survey both topics in their entirety. Therefore this chapter raises a complete theoretical basis for the ideas and applications in this work to be able to understand them. It also lists methods and algorithms that rather would belong to the results section. However, these connections are already dealt with in this chapter in order to guarantee the completeness of the method part from A to Z and thus to make the theory clearer. In addition, the implemented methods are used several times and thus only have to be derived and explained once at a central point. However, in the following chapters it is made clear which of the methods have been developed in the course of this thesis. Beginning with neurophysiology, the methods

also explain the essentials of the used recording techniques, the application of BCIs and spectral decomposition, and extensively discuss applied machine learning techniques, concluding with the introduction of statistical testing.

The results are divided into six chapters, roughly representing each a separate study. An overview is given in Fig. 1.2. In Chap. 3 it can basically be shown that incorrect robotic execution of an instructed task can be decoded based on the non-invasive EEG of human observers, making use of the conventional common spatial patterns (CSP) algorithm. It is also possible to distinguish the type of robot based on the EEG. For this purpose two experiments were designed and carried out. The visual stimulus is created by a correctly or incorrectly performed pouring or lifting task by a robot, two basic and simple processes of a human-robot collaboration. Altogether 18 subjects participated in the passive observation task, testing whether they generate recognizable error patterns while watching the robots perform the instructed tasks. The resulting data set thus represents a suitable basis for the investigation of error-related EEG phenomena.

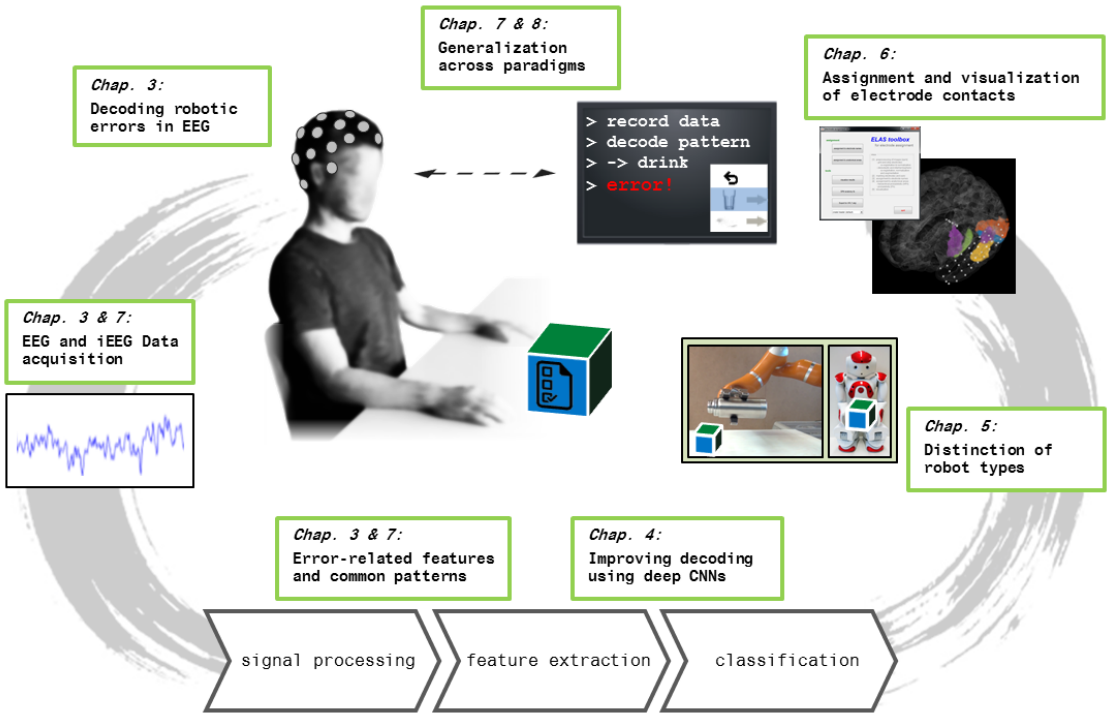

Figure 1.2: Outline of the chapter's content. Scheme of a collaborative human-robot-interaction based on (intracranial) brain recordings. The scene illustrates a situation with appearance of erroneous control or faulty execution of a robotic effector.

Chap. 4 describes the obvious and significant improvement of error-decoding performance using a deep convolutional neural network. The chapter takes up the problem investigated so far and starts with the optimization on the part of the decoder. The CNN performs significantly better, both as the previous CSP implementation and as the conventional linear discriminant analysis (LDA) used for comparison. This confirms the general trend of EEG analysis reverting to new methods like CNNs and extends it to the

field of error detection. The visualization of the results reveals important properties of the features that seem to be in the time domain in this problem. The CNNs also beat previous standards at the level of differentiation between robot types, which is dealt with in Chap. 5. The results give possible first indications of differences in the effect of robots on humans, with respect to human similarity scales. Visualization of learned features provides information about the underlying processes. In addition it is investigated to what extent the decoding of errors depends on the number of human-like features of a robot.

In Chap. 6, a novel method for assignment of intracranial electrode contacts to brain areas based on magnetic resonance imaging (MRI) is presented. The algorithm is especially useful when implantations in subdural space lead to deformations of the brain and an exact assignment of electrode contacts to brain areas becomes difficult. The method aims to determine the areas that contribute to the signal that is recorded by an electrode contact and is based on the consideration of individual characteristics of the brain and a retransformation to the cortical areas of the brain. As a result, probabilities for the possible influence of an area on these electrode contacts are given. The method has been embedded in a software environment that allows a user-friendly application. The software package includes the visual-supported identification of the electrode contacts, the automatic assignment and the 3D visualization, including virtual reality export. Especially in the field of clinical research it is possible to benefit from the methods without programming knowledge and to gain knowledge about the exact position of electrode contacts for better interpretation of occurring phenomena. The software is freely available and implemented in MATLAB, which is widely used in science and industry.

Apart from the observation of faulty execution, errors can also be committed by the user himself. Chap. 7 presents a paradigm that can be investigated for both motor and cognitive brain signals, and has contributed to an immense data set of intracranial recordings. This includes recordings of 47 epilepsy patients, who had different implantations and thus covered numerous different brain areas. In addition, a further comparative set was added to this data set, which contains recordings of a further paradigm for 15 of the 47 patients, which also permits analyses of error signals. In Chap. 7 analyses are presented which give basic insights into the processing of errors in the human brain, both on temporal and spatial scales. In addition, the different paradigms reveal the common and thus possibly general features in the cerebral processing of errors examined here. The results may contribute to the investigation of generalizing error-related patterns for the transfer across paradigms.

In Chap. 8, based on the recordings of the paradigms presented in Chap. 7, the potential of CNNs to detect errors by intracranial EEG is shown. The attempt to transfer general error-related information to further paradigms shows that pretraining of the CNNs can lead to significant improvements, especially in the case of little available data. This paves the way for the training of a general error pattern that fits many different error types and can therefore be recognized faster and more reliably. Especially in connection with BCIs, this generalizing method can help to detect the large number of potentially occurring errors, but also other types of control signals.

Chap. 9 finally concludes the results and insights of this thesis, and works out further scientific questions for future research.

1.2 Key Contributions

This thesis provides a scientific contribution to the optimization of error detection by means of human (intracranial) brain recordings, both in case the signals are triggered by faulty execution of the robot as well as by errors committed by a user himself. It also contributes to the neurophysiological understanding of error-related patterns. In detail the key contributions are as follows:

- For the analysis of error-related responses in human EEG, while observing robot errors, a huge data set comprising recordings of altogether 18 participants was generated. Initial analyses with standard methods showed that these error patterns as well as the type of robot can be decoded (Chap. 3). The resulting data set thus represents a suitable basis for the investigation of error-related EEG phenomena.

- The novel application of convolutional neural networks in the field of error-decoding based on human EEG shows that in this context deep CNNs perform significantly better than previous standards (Chap. 4).

- The huge potential of CNNs compared to conventional methods can also be shown for the differentiation of robot types by means of human EEG. The results give possible first indications of differences in the effect of robots on human similarity scales (Chap. 5).

- An algorithm for the assignment of intracranial electrode contacts to underlying brain areas was co-developed and implemented. A self-designed software environment has been designed, embedding the novel algorithm, which allows a user-friendly application. The tool includes the visual-supported identification of the electrode contacts, an automated assignment, a 3D visualization and an export for virtual reality application (Chap. 6).

- Based on a paradigm that can be investigated for both motor and cognitive brain signals, an immense data set of intracranial recordings could be created, including recordings of 47 epilepsy patients. Each participant exhibited different implantations and thus covering different brain areas. In addition, a further comparative set was added to this data set, which contains recordings of a further paradigm for 15 of the 47 patients, permitting analyses of error signals. Based on this data, analyses could be performed that revealed insights of temporal and spatial features during error-processing and that disclosed similarities in error patterns for different paradigms but same electrode contacts (same patients) (Chap. 7).

- Studies on a transfer of error-related patterns across paradigms resulted in a significant improvement of decoding accuracies in the case of small amounts of data. Furthermore, the CNN shows again high accuracies for the decoding of the error-related patterns in both cases (Chap. 8).

1.3 Publications

This thesis is based on our previous work presented in the following peer-reviewed journal papers and conference proceedings.

- J. Behncke, J. Hammer, A. Kalina, P. Marusič, A. Schulze-Bonhage, W. Burgard, and T. Ball. A core system for error processing delineated by intracranial eeg. *NeuroImage*, 2020. *submitted*.
- J. Behncke*, M. Kern*, J. Rüscher*, A. Schulze-Bonhage, and T. Ball. Probabilistic neuroanatomical assignment of intracranial electrodes using the elas toolbox. *Journal of Neuroscience Methods*, 2019. *These authors contributed equally.
- J. Behncke, R. T. Schirrmeister, M. Völker, J. Hammer, P. Marusič, A. Schulze-Bonhage, W. Burgard, and T. Ball. Cross-paradigm pretraining of convolutional networks improves intracranial eeg decoding. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.
- J. Behncke, R. T. Schirrmeister, W. Burgard, and T. Ball. The role of robot design in decoding error-related information from eeg signals of a human observer. *6th International Congress on Neurotechnology, Electronics and Informatics (NEUROTECHNIX)*, 2018.
- J. Behncke, R. T. Schirrmeister, W. Burgard, and T. Ball. The signature of robot action success in eeg signals of a human observer: Decoding and visualization using deep convolutional neural networks. In *6th International Winter Conference on Brain-Computer Interface*, pages 1–6. IEEE, 2018.
- D. Welke*, J. Behncke*, M. Hader, R. T. Schirrmeister, A. Schönau, B. Eßmann, O. Müller, W. Burgard, and T. Ball. Brain responses during robot-error observation. *Kognitive Systeme*, 2017. *These authors contributed equally.

The following publications that are not included in this thesis but also originate from the author's work at the research group.

- M. Völker, J. Hammer, R. T. Schirrmeister, J. Behncke, L. D. Fiederer, A. Schulze-Bonhage, P. Marusič, W. Burgard, and T. Ball. Intracranial error detection via deep learning. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.
- F. A. Heilmeyer, R. T. Schirrmeister, L. D. Fiederer, M. Völker, J. Behncke, and T. Ball. A framework for large-scale evaluation of deep learning for eeg. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.
- M. Völker, L. D. Fiederer, S. Berberich, J. Hammer, J. Behncke, P. Kršek, M. Tomášek, P. Marusič, P. C. Reinacher, V. A. Coenen, et al. The dynamics of error processing in the human brain as reflected by high-gamma activity in noninvasive and intracranial eeg. *NeuroImage*, 173:564–579, 2018.

1.4 Collaborations

This thesis was realized within the projects *BMI-Bot* (grant by the *Baden-Württemberg Stiftung*) and *Micro-Rec* of the interdisciplinary cluster of excellence *BrainLinks-BrainTools* and benefits from the cooperation with several other researchers. Wolfram Burgard and Tonio Ball acted as supervisors and mentors for this thesis and contributed with valuable ideas and feedback to the development of this thesis and the underlying publications. Thanks to Andreas Schulze-Bonhage and Petr Marusič, there was access to human intracranial brain data, which made parts of the subsequent analyses possible in the first place. Further collaborations are listed below for the individual chapters:

- Chap. 3: The underlying data set, based on non-invasive EEG, was generated cooperatively with the assistance of Marina Hader in the course of the master thesis of Dominik Welke. The analyses originated from the fruitful discussion with Robin Tibor Schirrmeister, whereby Andreas Schönau, Boris Eßmann and Oliver Müller were also involved in the completion of the publication. The related publication is [360].

- Chap. 4: For this chapter, the data set described in Chap. 3 was used. The analyses were made possible by the support of Robin Tibor Schirrmeister, who developed the open-source toolbox BRAINDECODE to create deep learning architectures. The related publication is [31].

- Chap. 5: Here, too, the analyses benefited from the aforementioned data set and the analysis methods. The related publication is [30].

- Chap. 6: The essential idea for the assignment algorithm originated from the mental work of Tonio Ball and was developed together with Markus Kern and Johanna Rüscher. The technique for estimating errors during co-registration was developed together with Markus Kern. Markus Kern and Johanna Rüscher realized the evaluation of the method. The related publication is [33].

- Chap. 7: The data for the investigation of intracranial signals were generated together with Martin Völker but especially Jiří Hammer, based on experimental design of the aforementioned. The related publication is still in preparation [34].

- Chap. 8: In this chapter, analyses are performed upon the data set based on intracranial recordings, which is described in Chap. 7. Analyses partly were made possible due to the toolbox of Robin Tibor Schirrmeister. The related publication is [32].

The figures in the particular chapters are taken from the cited publications or are modified versions of them. For all other pictures and figures in this thesis, either a source is quoted or the graphic originates from the creative work of the author.

1.5 Acronyms and Notations

Acronym	Phrase
ACC	anterior cingulate cortex (brain area)
ANN	artifical neural networks
BCI	brain computer interface
CDT	car driving task (experimental paradigm)
CNN	convolutional neural network
CS	central sulcus
CSF	cerebrospinal fluid
CSP	common spatial pattern
ECoG	electrocorticogram
EEG	electroencephalogram
EFT	ERIKSEN flanker task (experimental paradigm)
ELU	exponential linear unit
ERN/Ne	error-related negativity, error negativity
ERP	event-related potential
ErrP	error potential
FBCSP	filter bank common spatial pattern
FFT	fast FOURIER transform
HPA	hierarchical probabilistic assignment
iEEG	intracranial electroencephalogram
LDA	linear dicriminant analysis
LFP	local field potential
LOT	lifting observation task (experimental paradigm)
LS	lateral sulcus
MFC	medial frontal cortex (brain area)
ML	machine learning
MLP	multilayer perceptron
MNS	mirror neuron system
MRI	magnetic resonance imaging
Pe	error positivity
PFC	prefrontal cortex (brain area)
POT	pouring observation task (experimental paradigm)
rLDA	regularized linear dicriminant analysis
SMA	supplementary motor area (brain area)

Symbol	Description
x	scalar value
$\mathbf{x}$	row vector
$\mathbf{x} = (x, y, z)$	vector with scalar values
$\mathbf{X}$	$m \times n$ matrix
$\|\|\mathbf{x}\|\|$, $\|\|\mathbf{X}\|\|$	EUCLIDEAN norm of vector, matrix
$\|x\| = \|\|x\|\|$	absolute value of scalar
$\mathbf{x}^\top$	column vector
$\mathbf{X}^\top$	transpose of $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of $\mathbf{X}$
$\mathbf{X}^H$	adjunct of $\mathbf{X}$
$\mathbf{x}^+$	vector, only even coefficients of $\mathbf{x}$
$\mathbf{x}^-$	vector, only odd coefficients of $\mathbf{x}$
$\mathbf{\Sigma}$	covariance matrix
x_i	entry i of vector $\mathbf{x}$
$\mathbf{x}_i$	vector i of a set of vectors
$\mathbf{X}_{i,:}$	row vector i of matrix $\mathbf{X}$
act	activation function
$\arg\min_\gamma$	minimum under argument γ
corr	correlation
cov	covariance
log	natural logarithm
net_j	input for neuron j
$pred(j)$	predecessors of neuron j
var	variance
$\tilde{x}$	FOURIER transform of x
$g * h$	convolution
dx/dy	derivative
$\partial x/\partial y$	partial derivative
δ_{ij}	KRONECKER delta
$\vec{\nabla}$	nabla operator
$\Delta = \vec{\nabla}^2$	LAPLACE operator
$P(X)$	probability distribution of random variable X
$\mathcal{L}$	loss function
$\mathbb{C}$	space of complex numbers
$\mathbb{R}$	space of real numbers
$\mathcal{C}$	space of classification targets/classes

Chapter 2

Background and Methods

This chapter addresses the theoretical background of the thesis, considering that the applied methods primarily originate from two main disciplines, computer science and neuroscience. Hence, it is important to understand the underlying principles of both sides to obtain a comprehensive overview for an assessment of the results. This chapter is not only entitled to explain the underlying concepts and to motivate the issue of this thesis. Instead, self-implemented methods and algorithms as well as design choices are likewise already presented. The reason behind this structure is to introduce the multiply used machine learning algorithms generally and only once, and moreover to present the entire theory coherently from first to last. In the first section, Sec. 2.1, a brief overview of the brain is given and the theory of signal generation and transport is explained. Sec. 2.2 addresses the techniques to record those signals. The following Sec. 2.3 introduces the concepts for brain machine interfaces and the neural signals that drive their control, particularly regarding detection of errors. Sec. 2.4 presents the utilized methods for spectral decomposition, providing the basis for analyses in frequency domain. The most comprehensive part of this chapter is used by the conception of machine learning techniques, Sec. 2.5. Here, the three main decoding approaches of this thesis are discussed and connected tools are introduced. A last section, Sec. 2.6, motivates the use of statistics and explains the applied techniques.

2.1 Neurophysiology

Generally speaking, neurophysiology outlines the science of the nervous system. In the following, the underlying principles will be exposed in a top down manner. The most fundamental description divides the nervous system into the central and the peripheral nervous system. At this, the peripheral nervous system comprises all nerves that are not covered by the central nervous system (brain and spinal cord), and spins a dense net throughout the entire body. The central nervous system embodies the central human processing unit. It is responsible for the retention and processing of information that is incorporated from the exterior by the sensory organs. It also subserves the control and adjustment of internal organs and the coordination of the entire motoric capacities. After a short introduction of the brains structure, the nerve cells or neurons are characterized and the signal transport is described.

2.1.1 The Brain

Roughly, the brain can be subdivided into four main areas, see Fig. 2.1. The *cerebellum* is responsible for motor coordination, taking care of balance and fine tuning movements. Due to its vicinity to the spinal cord, it processes direct sensory, acoustic and visual information. Likewise, parts of learning processes are ascribed to the cerebellum. Among others, the *diencephalon* comprises the thalamus and the hypothalamus, and is preceding the telencephalon according to sensory and motor signals. Furthermore, it makes decisions about the prioritization of certain signals, while the hypothalamus controls a variety of body internal processes. In contrast, the *brain stem* (truncus cerebri) controls fundamental functions like respiration, blood pressure or reflexes. Abstract brain-teasers are processed by the *cerebrum* or *telencephalon*, comprising among others the cerebral cortex. The cerebral cortex constitutes approximately a fifth of the cells of the entire brain and has developed during evolution the most. It plays an essential role in the derivation of electrical signals on the scalp, generating signals from important functional centres. For example, projections of the visual pathways lead to the visual cortex in the occipital lobe. The auditory cortex in the upper regions of the temporal lobe serves to process acoustic signals and provides the necessities for hearing and speech. Associative areas in the anterior part of the frontal lobe are assigned to memory, higher thinking and cognitive processes. Control of movement is accomplished by the motor cortex, while the parietal lobe handles the somatosensory processes and e.g. ensures for the capability to calculate. Several functions, especially high-level cognitive tasks, are not always easy to locate and can vary among individuals. The method of mapping motor-sensory cortex functions was discovered when electrical stimulation led to the illusion of touch or movement [263]. Recapitulating, the general task of all areas is to process the huge mass of information given by the cells.

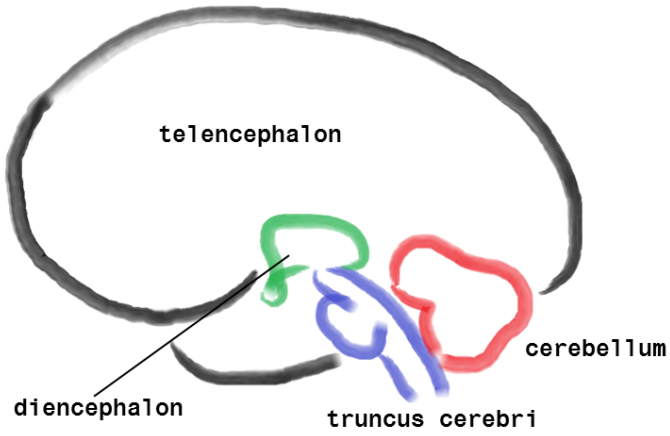

Figure 2.1: Sketch of the brain. The four main parts of the brain are formed by the telencephalon, the diencephalon, the cerebellum and the truncus cerebri.

The cells of the nervous system can be subdivided into two different types: *glia cells* and *nerve cells* or *neurons*, respectively. Glia cells build the supporting structure for the nerve cells and care for electrical insulation by enveloping them. In addition, they support by providing transport of substances, care for fluid exchange and accomplish homeostasis. The nerve cell is the structural and functional basic unit of the nervous system and will be regarded more closely in the following, supported by the theory given in [183].

2.1.2 Signal Generation and Transport

The neurons structure is depicted in Fig. 2.2, where the fundamental information processing takes place. Incoming signals, such as from preceding neurons, are received by specific transition points, called *synapses*. Besides, the particular informative parts are weighted at this step, before the *dendrites* collect all the information and transfer it to the *cell body* or *soma*. Inside the soma the weighted information parts are added up. As soon as the aggregated signal exceeds a certain threshold, the neuron generates an electrical signal which is passed to succeeding neurons. The output signal, formally known as *action potential*, arises from the origin hill of the *axon*, the so-called *axon hillock*, where the activation takes place. The electrical insulated axon fulfills the function of transferring the action potential to adjacent units.

In Fig. 2.2C the underlying biological model of a neuron is described by a simple mathematical approach [185]. Here, the single x_i stand for the input signal at the synapses. At these neuronal connections the inputs are weighted by the w_i and transferred by the dendrite. The soma finally adds up the weighted inputs:

$$\sum_i w_i x_i. \tag{2.1}$$

The weighted sum is compared with a certain threshold, here represented by the *heaviside step function*, and passed as output y to succeeding neurons.

$$y = \Theta\Big(\sum_i w_i x_i\Big). \tag{2.2}$$

Hereafter, the transport of the electrical signal will be examined according to the cellular ion balance. The information processing is hereby supported by the specific properties of the passive semipermeable membrane of the neuron. If the neuron is not excited, the membrane takes over a state of electrostatic balance between intracellular and extracellular space. The interior of the cell contains a high concentration of potassium ions (K^+), whereas exterior the sodium ions (Na^+) predominate. During resting state the different charges make for a basic voltage across the membrane of approximately $V_r = -70\,mV$, called *resting potential*, which is always considered according to the exterior. The equilibrium of $\neq 0\,mV$ is explained in the following.

The magic of these circumstances rests in the membrane itself. By reason of *ion channels* in the membrane, the cell membrane is selectively permeable. The ion channels that are active in resting state are highly permeable to K^+ ions, but significantly less

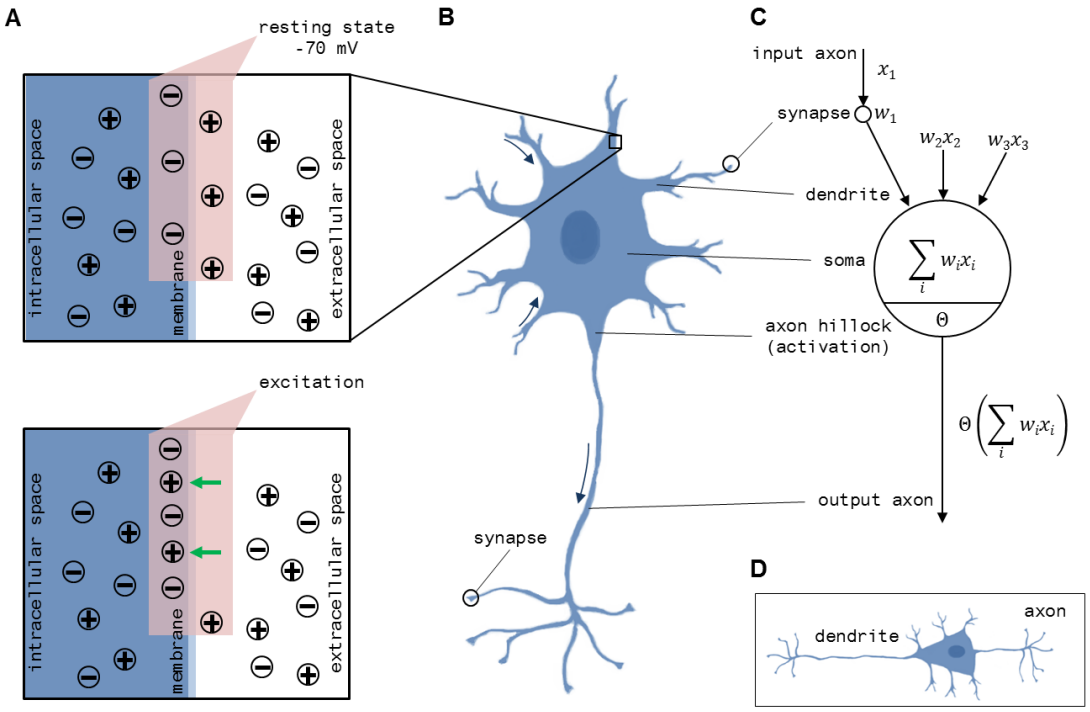

Figure 2.2: Schematic description of a neuron and its underlying processes. A Top: charges at the membrane in case of a resting state. During resting state the charges maintain in an equilibrium of $-70\,mV$. Bottom: when stimulated by an excitatory postsynaptic potential (EPSP), the ion channels in the membrane change their permeability resulting in an Na^+ flux into the interior of the cell, what increases the potential across the membrane. *Inspired by [329]* **B** Sketch of a neuron. The arrows indicate the direction of the signal propagation. **C** Mathematical model inspired by the biological neuron [185]. The input of a preceding neuron is weighted at the synapse. The dendrites translate the weighted signals to the soma, where it is added up. At the axon hillock, the output signal is initiated by the activation Θ as soon as a certain threshold is exceeded. The axon propagates the output to succeeding neurons. **D** Sketch of a pyramidal cell, which only appears in the cerebral cortex. It can be characterized by its pyramid-like body and the long dendrites. The pyramidal cells provide the largest contribution to the potential on the scalp.

permeable to Na^+ ions. Thus, the concentration gradient leads to an outward K^+ ion flux, while the electrical gradient that emerges from the remaining negative charges counteracts. By contrast, due to the low interior Na^+ concentration *and* the electrical gradient, the exterior Na^+ ions tend to diffuse into the interior, but inhibited by the ion channels. Furthermore, the active membrane pump (the protein adenosine triphosphate, ATP) acts against the flux of ions. This $Na^+ - K^+$ pump holds down the Na^+ concentration inside the cell, while keeping up the K^+ concentration. An equilibrium is established at the just mentioned resting potential. If the neuron is excited by preceding neurons, the potential at the membrane increases significantly. As soon as the potential exceeds a value of roughly $V = -55\,mV$, the membrane becomes considerably more permeable for Na^+ than for K^+ ions. The explosive change, also referred to as *depolarization*, generates a sharp increase of the membrane potential leading to a voltage peak of about $V = 40\,mV$, called *action potential*. Incidentally, the state reverts the relations and the *repolarisation* begins. Because of the inertia of the channels the system subsequently occupies the state

of *hyperpolarization*, being a little less negative than the potential during resting state. Ultimately, after a total refractory period of less than $2\,ms$, resting state takes over. In this way, the stream of information takes place across the cell. By means of the axon, the neuron can stimulate or inhibit adjacent neurons by generating either an exciting or an inhibiting postsynaptic potential (EPSP or IPSP) at the according synapses. This potential can persist for more than $10\,ms$ and consequently admit a large temporal and spatial summation.

According to the NERNST equation, which is derived from thermodynamic principles, the *equilibrium potential* E_χ can be calculated for any ion χ:

$$E_\chi = \frac{RT}{z_\chi F} \log \frac{[\chi]_o}{[\chi]_i}, \tag{2.3}$$

where R is the gas constant, T the temperature given in Kelvin, z_χ the valence of the ion χ and F the FARADAY constant. $[\chi]_o$ and $[\chi]_i$ define the concentrations of the ion inside and outside the cell. The NERNST equation can be regarded as a first approximation of the description of the resting potential over the cell membrane. Until now, the contribution of chloride (Cl^-) has not been mentioned for convenience. However, the equilibrium potential of chloride E_{Cl} also biases the *membrane potential* V_m, defined as the difference between the potential V_i inside the cell and outside the cell V_o, and is therefore considered in the GOLDMAN equation [144]. This equation describes the membrane potential in dependence on ionic permeability and concentration:

$$V_m = V_i - V_o \tag{2.4}$$

$$= \frac{RT}{F} \log \left(\frac{P_K[K^+]_o + P_{Na}[Na^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Na}[Na^+]_i + P_{Cl}[Cl^-]_o} \right). \tag{2.5}$$

Here, P_χ denotes the permeability of ion χ. The GOLDMAN equation is defined in a stationary state, what means that the sum of all ion fluxes $\sum_\chi I_\chi$ equals zero. When the state changes, for example when the permeability for K^+ ions is exceptionally high ($P_K \gg P_{Na}$ and $P_K \gg P_{Cl}$), the GOLDMAN equation reduces to the NERNST equation:

$$V_m \cong \frac{RT}{F} \log \frac{[K^+]_o}{[K^+]_i}. \tag{2.6}$$

Consider that $z_K = +1$. An equivalent circuit model of the resting membrane can be used to calculate the ion flux I_χ [183]. Here, the electrogenic influence of the Na^+-K^+ pump can be neglected. According to the model, going from inside to the outside for example across the K^+ branch, the total potential difference can be determined as:

$$V_m = E_K + \frac{I_K}{G_K}. \tag{2.7}$$

Here, the K^+ cell conductance of the membrane is defined as the product between number of resting state K^+ channels and conductance of an individual K^+ channel g_K, $G_K = N_K g_K$. Hence, the ion flux for K^+ can be described as:

$$I_K = G_K(V_m - E_K). \tag{2.8}$$

The *point current source model* or *standard model* is a technique to derive the *local field potential* (LFP, see section 2.2) from networks of neurons and is widely applied to model extracellular potentials [96, 267, 278, 284]. It is a rather simple approach which assumes that the LFP is generated by transmembrane currents and the neurons are embedded in an ohmic or perfectly resisting medium. Furthermore, the electric potential can be considered to be generated by point current sources and exhibiting a spherical symmetry. Then, the potential $V(\mathbf{r})$ at any position $\mathbf{r}$ in space is calculated by

$$V(\mathbf{r}) = \frac{1}{4\pi\sigma} \frac{I_b}{||\mathbf{r} - \mathbf{r}_b||}, \tag{2.9}$$

where $\mathbf{r}_b$ defines the position of the current source b and σ the electrical conductivity of the extracellular medium. As a matter of fact, $V(\mathbf{r})$ is solution of the LAPLACE equation, $\Delta V = 0$. In electrostatics, electrical potentials in an uncharged space suffice the LAPLACE equation. Assuming several point current sources, the potential at $\mathbf{r}$ can be computed as the linear superposition of all point current sources,

$$V(\mathbf{r}) = \frac{1}{4\pi\sigma} \sum_b \frac{I_b}{||\mathbf{r} - \mathbf{r}_b||}. \tag{2.10}$$

Finally, when the potential at a certain position is described as a time-dependent, discrete and finite signal, it assumes the form of a matrix:

$$\mathbf{V} \in \mathbb{R}^{e \times t}, \tag{2.11}$$

with t discrete time steps and the activity at a number of e different positions, which will be later considered as electrodes (electrode contacts) or channels. At this, electrode describes the physical sensor whereas channel refers rather to the signal recorded by an electrode.

2.2 Recording Techniques

Due to the electrical activities of the neurons, an electrical field is generated, inducing local compensation currents inside the dendrites and the extracellular areas [19]. The excitation of neurons is similarly accompanied by a magnetic field that is perpendicular to the electrical field. Albeit, the fields of a single neuron are extremely weak and measurements only succeed if neural tissue is hit directly by a sensor. The commonly called *single unit activity* (SUA) can be measured with the help of microelectrodes [54]. On larger scales such a measurement is not possible any more, but synchronization of electrical activity of adjacent neurons can be summarized as the *local field potential* (LFP) [67, 247]. Both with non-invasive (*electroencephalography*) and invasive (*intracranial electroencephalography*) measurement methods these effects can now be recorded. Here, the technologies can differ in their temporal and spatial resolution. Hereafter, the methods are described which were used to gather the extensive set of data.

2.2.1 Electroencephalography

Electroencephalography (EEG) is a non-invasive methods, recording changes of electrical potentials between electrodes on the scalp. In 1929 Hans Berger was the first to apply this method to acquire access to human brain activity [35]. Rapidly, it showed that physiological states could be reflected by reference to this method.

When a brain region is engaged in a certain task, a cluster of organized neurons, showing similar or synchronized activity, takes action. The large number of simultaneously working neurons generates an electrical dipole in the substructures of the brain. This dipole can be measured on the scalp as a scalp potential, depending on size, location and orientation of the dipole. There, due to conducting properties of the different transmigrated layers (see Fig. 2.3), the measurable signal is strongly attenuated by a factor of about 1000. Thereby, the layers, for instance the brain, the cerebrospinal fluid (CSF), the bone or the skin, appear like a low-pass filter. While intracranial measurement methods tap voltages of approximately $\pm 200\, mV$, the voltages at the surface merely amount to about $\pm 30\, \mu V$. Besides, it is difficult to make an assumption about the exact position of the origin of the signal, what makes the measurement considerably more difficult that intracranial [22]. Moreover, the projection of the neuronal dipole onto the scalp describes a highly complex mathematical problem, also known as the *EEG forward problem* [19]. A huge challenge in creating EEG recordings is the prevention or reduction of artifacts that are not caused by neural activity. Typical artifacts in EEG include muscle activity, eye movements, eye blinks or electrical stray signals from exterior sources. Contrary to their disadvantages, EEG recordings exhibit remarkable temporal resolution.

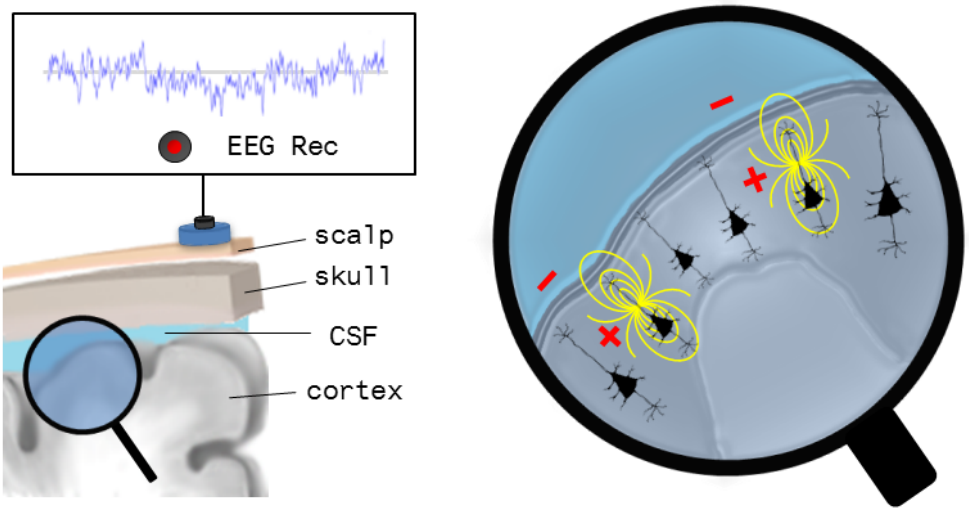

Figure 2.3: Generation of EEG signals. Left: the neurons propagate the electrical signals to the cortex, where the cerebrospinal fluid (CSF), the skull and the scalp have to be transmitted. At the scalp, the voltage can be measured by the EEG electrode. Right: schematical description of the pyramidal cells at the cortical surface, generating dipoles by the propagation of the electrical signals. Synchronous activity of several neighbouring neurons generate the local field potential (LFP). *Inspired by [1].*

Basically, the commonly called pyramidal cells contribute to the potential at the scalp, see Fig. 2.2D and Fig. 2.3. These cortical neurons are orthogonally oriented towards the cortical surface and are apparent most notably because of their pyramid-shaped cell body together with the long dendrites. The pyramidal cells can be found anywhere located in the cortex, belonging to the gray matter, but showing a predominant consistent orientation: while the axon lies in the direction of deeper layers, the long dendrites are oriented towards the surface. Because of their distance to the cortical surface, electrical fields that are generated by neurons laying in vicinity to the sulci are more difficult to detect, hence, the pyramidal cells close to the gyri constitute the substantial part of the signals, recorded by the EEG.

EEG usually is recorded unipolar, whereby the voltage of the channels is tapped against the voltage of one or several reference electrodes. To obtain reproducible results the electrodes are placed according to a certain scheme, such as the standard 64-channel montage in Fig. 2.4. For reasons of clarity this montage was used in Fig. 2.4 to convey a rough impression of the arrangement of electrodes. The measurements that underlie this thesis made use of a wet-gel 128-channel WAVEGUARD EEG cap (ANT NEURO, Enschede, Netherlands), persisting of sintered Ag/AgCl electrode elements. The data was recorded using a sampling rate of $5\,kHz$ (AC, 1250-Hz antialiasing low-pass filter). As a reference, the electrode Cz was selected, whereas the ground was located between AFz and Fz. Here, he naming of the electrodes vaguely reflects the electrodes distribution over the cortex, where they can roughly be assigned to the different lobes, see Fig. 2.4. The non-invasive measurement setup was optimized for high-frequency responses. To reduce disturbing electromagnetic signals, the recordings were located inside a cabin, serving as a FARADAY cage, which was equipped with an active electromagnetic shielding ("MRSHIELD" - CFW Trading Ltd, Heiden, Switzerland). The information exchange with exterior components and processing devices was done via fiber optic cables to sustain the shielding. Moreover, all electrical devices inside the cabin were supplied by DC batteries. Beside the electromagnetic shielding, the cabin also provided a good muting against sounds and vibration that might disturbed the participant during the measurements.

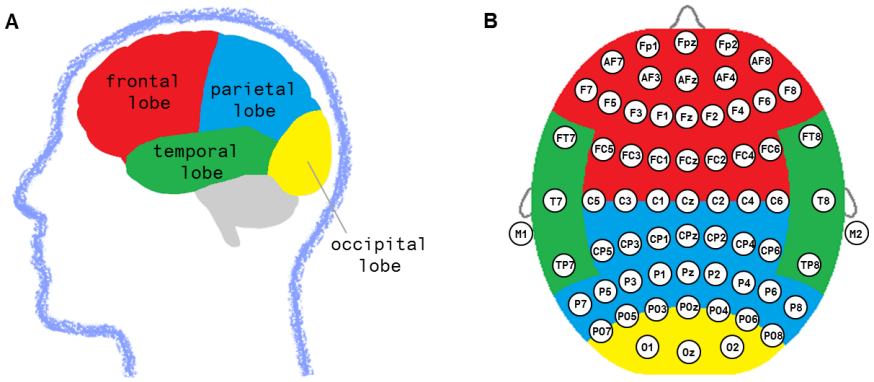

Figure 2.4: Illustration of the cortical areas. A Subdivision of the cortex into frontal, parietal, temporal and occipital lobe. **B** Schematic standard 64-channels montage over underlying lobar allocation, with colors matching the lobar color code from **A**. *Inspired by [1].*

2.2.2 Intracranial Electroencephalography

The method of intracranial electroencephalography (iEEG), including the electrocorticogram (ECoG) [177, 335] and the stereoelectroencephalography (stereo EEG or SEEG) [82, 326], has a notable latter history than the EEG. But, recordings gathered by intracranial methods exhibit a substantially higher signal quality. Here, the voltage is tapped directly at the tissue, in case of ECoG directly on the cortex, and does not have to transmit CSF, skull and skin. Moreover, there is no occurrence of noise sources like muscular activity, that have to be considered for EEG. Intracranial recordings have a higher spatial resolution and exhibit a higher signal-to-noise ratio than non-invasive recordings [181, 205, 283]. Another advantage of the invasive method is the more accurate assignment to underlying brain areas due to the direct contact. Various studies not only address the mapping of the brain, but also investigate possibilities to assign electrode contacts to the underlying brain areas and push the methods into the direction of more and more accurate assignments. Here, electrode contact refers to an individual sensor that records the signals of one channel. The assignment technique described in Chap. 6 as well as the corresponding *ELAS* [1] interface for assignment of intracranial electrodes, which has been developed as part of this thesis, make an important contribution to this aspire.

The advantages of invasive methods are not in the least to be counterbalanced to the circumstances, risks and complications of their application, since the electrodes operatively have to be positioned. This thesis benefits of the extraordinary possibility to work with patients from the epilepsy center in Freiburg, Germany, as well as the epilepsy center of the Motol University Hospital in Prague, Czech Republic, who gave their voluntary and informed consent for contribution. Patients that are received in such centers, generally have many seizures but do not respond to medication with anti-epileptica. In such cases, a resection of tissue that trigger the seizures can be considered, resulting in a lower intensity and quantity of seizures. And yet, the positioning of the implantation exclusively complies with the medical outcome of the epilepsy and does not necessarily cover areas, that would be optimal for a special application controlled via brain signals. Nonetheless, the information gathered in the course of a paradigm are versatile and of inestimable scientific value. Within the scope of the implantation period the patients kindly participated in the experiments, that indeed did not benefit the group of epilepsy patients, but rather a broad spectrum of people with other disorders or deficits.

The electrodes which were used in this thesis can distinguished as follows. Electrodes used for ECoG exhibit a two-dimensional arrangement which is directly placed on the cortex. This can be in form of a grid, describing a $n \times m$ array of electrode, or a strip, whereby the electrodes are placed in row. Equally, depth electrodes have been applied, showing a considerably less complicated implantation by being minimally invasive embedded. They are directly introduced into the subsurface tissue and can investigate brain areas that lie in deeper structures. The acquired intracranial EEG signals at the epilepsy center in Freiburg were recorded by means of a COMPUMEDICS amplifier (Singen, Germany) at a sampling frequency of $2\,kHz$, meanwhile the epilepsy center of the Motol University Hospital in Prague made use of the SCHWARZER EPAS amplifier (Munich, Germany) and

[1] https://github.com/joosbehncke/elas

the NICOLET EEG C-series amplifier (Pleasanton, USA), recording at a sampling rate of $512\,Hz$. The used depth electrodes were composed of platinum-iridium contacts (DIXI MEDICAL, Lyon, France and AD-TECH, Racine, WI, USA).

2.3 Brain Computer Interface

Generally, a *brain computer interface* (BCI) can be defined as an interface, that enables a human to control a machine without using the peripheral nervous system, and therefore establishes a direct communication path between the brain and an external effector. First reference in a scientific context dates from the work of *Vidal* in the early 1970s [348, 349]. Here, the interface is meant as an extension and moreover shall restore humans, for instance with restricted motor abilities, to a certain extent with more autonomy. Beside the control of limbs [39, 369] or robots [309], BCIs are also investigated and applied in communication [39, 115, 317], environmental control [129], leisure and information [249], rehabilitation training [271, 272] and mobility [68, 211]. At this, the implementation of the control signals reaches from detection of direct imagined movements (low-level control right up to the control by a set of control signals (high-level control) that representatively stand for a selection inside a menu. The underlying ideas is that each thought is accompanied by a spatio-temporal pattern. Then, the BCI has the task to read the brain signals and find exactly these patterns to categorize them. The triggered signals patterns for the executions are not inevitably equal for the same person, even less when distinct individuals fulfill the task. Thus, it is necessary to learn a more or less general pattern that enables a recognition of different orders. The learning techniques can be summarized as *machine learning* and are treated explicitly in Sec. 2.5. A schematic description of a BCI system is shown in Fig. 2.5.

2.3.1 Neural Control Signals

There are several types of signals that can be extracted from the human brain recordings to control a device. The two following paragraphs will give a short introduction of two groups of control signals that were analyzed in this thesis. For further control signals and a more detailed description see [368].

Event-Related Potentials

Event-related potentials (ERPs) denote a transient signal pattern that is provoked according to the appearance of an event or stimulus [220]. The stimulus can appear in various kinds: visual [179, 316], auditory [90], tactile [141] or electrical [262], whereas the amplitude of the response typically covers ranges of $1 - 20\,\mu V$, appearing approximately $< 500\,ms$. The response is phase- and time-locked to the event and distinct samples show pretty much the same temporal course for reoccurring events. Usually, several time-locked responses are recorded and averaged, to determine a good estimate for the underlying

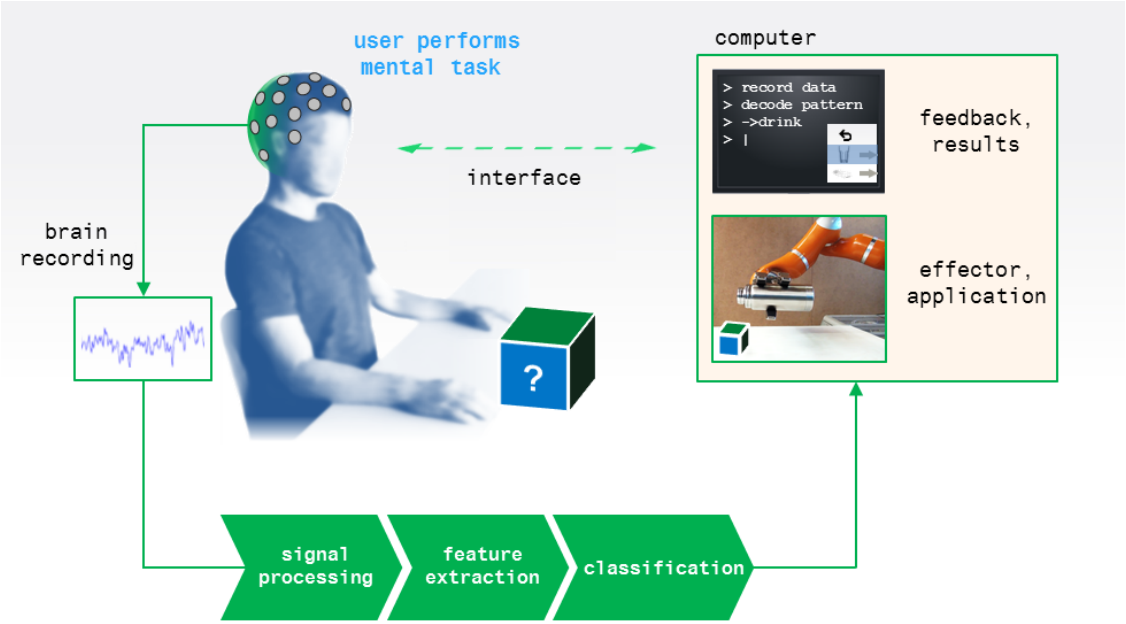

Figure 2.5: Concept of a Brain Computer Interface. The user performs a mental task with the intention to initiate a certain execution. The brain recordings are pre-processed before features are extracted and patterns are classified. The decoding results feed either a planner for a robotic effector or any other application, or both of it. Likewise, the user gets a feedback about the results of the classification.

pattern. By this means, activity that is non event-related can be averaged out and the signal-to-noise ratio increases.

Early components of the event-related potential are assigned to physical processing of the signal [276], while later components are rather connected to cognitive processes, like attention [162] or expectation [354]. Based in intracranial data the ERP could be connected to linguistic syntax processing [300] or heart cycle-related effects [187]. In addition to it, it could be shown that ERPs serve as outstanding control signals for BCIs [45, 115, 200].

Oscillatory Activity

In general, the brain exhibits oscillations without any external influence. Prominent appearances are for example the increase of δ-activity ($0.5 - 4\,Hz$) during deep sleep or the typical α-activity ($8 - 12\,Hz$) during states of consciousness and relaxation [246], which is strongest when the eyes are closed. Modulations of the μ rhythm, also lying in the α band, is connected to the level of relaxation of the motor system. The power of the μ rhythm, likewise of the β-activity ($12 - 30\,Hz$), is decreased during observation [17], execution or imagination [269] of motor activities. This phenomenon is called *event-related desynchronization* (ERD) [269] and is used for BCI applications, such as spellers [42]. Surprisingly the desynchronization also appears up to $2\,s$ before voluntary, self-initiated movements [18, 268]. Furthermore, there are several cognitive and sensory processes that can be associated with oscillatory power of different frequency bands.

Alertness [182], memory encoding [180, 192], perception [224, 306, 333], work load [136, 165] or attention [29, 157, 193] are just some of the examples. Typically, the power spectrum of the brain recordings appears in a characteristic 1/f shape [264], while several peaks indicate the mentioned changes in particular frequency bands [66]. Tab. 2.1 gives an overview of frequency bands and associated rhythms.

Table 2.1: Frequency bands and typical activities [329]

band	frequencies	activity
δ	$0.5 - 4\,Hz$	increase during deep sleep
θ	$4 - 8\,Hz$	increase during light sleep
α	$8 - 12\,Hz$	increase during state of relaxation; μ rhythm
β	$12 - 30\,Hz$	decrease during motor activities
γ	$> 30\,Hz$	increase during high-level tasks

2.3.2 Online Brain Computer Interface

So far, many of the cited scientific papers on the subject of BCI especially discuss the post-hoc analysis of data, that is without the direct intent to implement an applicable BCI. Basically, the objective of extracting control signals and typecasting them as useful for BCI applications is followed. The brain signals can be recorded under ideal conditions and the requirements in a laboratory significantly simplify the experiments. However, in an online BCI the task is to implement a real-time control, what renders an elaborate pre-processing and decoding impossible. Similarly, when it comes to real-life applications, many facilities that suppress disturbing signals and deliver a widely noiseless signal are inapplicable. The challenge persists in developing a robust and reliable system under the aggravated conditions. Preferably, the system should work adaptive and improve the performance with each repetition. For such a system a real-time error recognition framework is fundamental. It allows a fastest possible correction or standstill in case of occurring errors and progressive adaption to an optimal process. In the following, the neuronal basics for the error processing and the according pattern will be explained.

2.3.3 Error-related Patterns

For cooperative scenarios involving both robots and humans, especially when both are sharing the same workspace, a safe and smooth human-robot interaction is required. For example when robot behaviour disagrees with the user's intention, a system has to ensure the humans safety and that the user's commands are executed correctly. Autonomous and intelligent robotic systems constitute a great possibility to e.g. extend human activities or replace lost abilities, but still are vulnerable for malfunction of several components of the collaborative system. While it would be optimal to prevent such robot errors entirely,

this is unlikely to become feasible soon. Thus, detection of robot errors and correction of their consequences remains a relevant problem. Furthermore, for an adaptive movement control, the system has to be called attention to faulty realizations of a task. The awareness of errors is essential for learning, thus, it builds the base for refining motor skills and adaptive behaviour. Especially when it comes to online implementations this can play a fundamental role. Overall, a proper error detection system can lead to substantial improvements in the performance of an (online) BCI system.

Error-related potentials (ErrP) constitute a central brain pattern, that is connected to the processing of erroneous process, either observed or executed. The ErrP forms a component of the ERP that reliably gives a signal during the first $300\,ms$ after a response. Since the work of *Falkenstein et al.* [113] and *Gehring et al.* [133], numerous research has been done on the field of error-related potentials. Immediately after the response a sharp negative deflection in fronto-central scalp regions can be measured, called *error-related negativity* (ERN) [133] or *error negativity* (Ne) [113]. The deflection holds for period of about $(50 - 100\,ms)$ and is suggested to be a direct correlation with behaviour adaption following the error-related responses [330]. In addition, the amplitude of the deflection is modulated by both the subjective importance of the error and the subjective perception of it [365]. Likewise the amplitude seems to be the larger the less an error indeed appears [7, 8, 152]. Dependent on the set task the ERN/Ne component is followed by a positive deflection with centro-parietal expansion. The positive deflection is called *error positivity* (Pe) and is recognizable up to $400\,ms$ after the response. It exhibits an early, sharp propagation in frontal regions, which translates to parietal regions as time goes by. The subsequent, blurred Pe is modulated by conscious error perception and can be connected to the user's awareness of errors [364].

The dorsal part of the anterior cingulate cortex (ACC), the anterior insular cortex (AIC), the prefrontal cortex (PFC) as well as the pre-supplementary motor area (preSMA) are involved in the cerebral error processing [27, 56]. However, it is not entirely understood which parts of the error processing are realized by the the different areas. In recent times, intracranial measurements give a deeper insight into the operating principles of deeper cerebral structures, such as action monitoring or error processing [27, 351]. Beside the conventional error-related patterns, changes of the spectral power in high-γ band ($\approx 50 - 200\,Hz$) could serve as a hint to cortical error processing. Compared to conventional scalp EEG, intracranial EEG shows the potential to reveal fundamental structures of cerebral error-related processes and provides the possibility to realize direct neurophysiological examinations in determining areas of error processing, particularly if those can merely be covered by the scalp EEG with difficulties. The knowledge about the temporal propagation of error processing could give early indication to erroneous actions and therefore would be a suitable control signal for early detection in BCI applications. Likewise in this context, the spatial coverage of the intracranial EEG clearly becomes noticeable. Here, research with non-invasive EEG can support, where for instance erroneous events of a preparatory attention peak could be proven around $100\,ms$ prior to the event [257] or where increased activity in α band could be observed about $20\,s$ before a visual target was missed [253]. In summary it can be said that error-related signal allow a fast and direct recognition of errors and indeed qualify for real-time application.

Basically, the information of the error-related signals can be basically utilized for two different implementation types in BCI applications. Considering a first implementation, a certain command is transmitted to an effector or robot, but performed erroneously. A classical example is the spilling of liquid in a pouring task [31]. In such cases the system is installed to recognize the error, to correct it and to optimally avoid such an error in subsequent attempts. Ultimately, the error detection serves as a training for adaptive behaviour, also accounting for erroneous decoding of a command. Indeed, BCI applications are still error-prone [285] and actual performances are far from practical application, first of all when it comes to implementations where safety of e.g. the user is a fundamental requirement for the application. Nonetheless, there are already several studies that could improve the performance of a BCI by implementing an error detection system. For example in a P300 speller [318], during observation of robot action [173] or in a motor task based on intracranial data [111, 232], the applications could be improved considerably. A second implementation comprises the prevention of errors at an early point in time. Here, the idea is to use the brain data to predict whether an error could possibly occur, for instance caused by lack of attention. Most notably, sectors like the motor industry are interested in such applications and scientific interest in the topic of EEG-based early detection of tiredness during car driving becomes remarkable [4, 170, 356]. Likewise in aviation, scientific work is benefiting from error detection, for instance for the application in flight simulators [216].

2.4 Spectral Decomposition

Not exclusively time series of voltages can be utilized to describe electroencephalographic data, but also oscillatory components play a significant role in the characterization of information processing inside the brain, see 2.3.1. To determine the impact of these components on our thinking and acting, the informative characteristics of the oscillations have to be extracted and the spectral representation investigated. The underlying idea of spectral decomposition is that the signal itself expresses a superposition of a spectrum of frequencies and according phases. In other words, a signal can be composed by a combination of finite number of frequency components. Thus, frequency analysis describes the dynamic properties of an oscillating system by separating the signal into the individual frequency components and analyzing them apart. Many different approaches are able to manage the issue, whereof the analysis based on the FOURIER transform probably represents the most popular and widespread method of all, by far. In the following, the FOURIER analysis will be introduced, while afterwards a more elaborate method, the *multitaper method*, will be established. The multitaper method plays a central role in the spectral analysis in this thesis.

2.4.1 FOURIER Transform

A quite popular tool in signal analysis is the FOURIER transform, or continuous FOURIER transform, a mathematical description from the FOURIER analysis. The transformation defines a mathematical rule to decompose a continuous, aperiodic signal into a continuous spectrum of frequencies. Let $x(t) \in L^1(\mathbb{R}^2)$ be an integrable function, for convenience in a two-dimensional space, then, the FOURIER transform is defined by

$$\mathcal{F}_{wt}\{x(t)\} = \tilde{x}(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t)\, \mathrm{e}^{-iwt} \mathrm{d}t, \tag{2.12}$$

where L^1 describes the LEBESGUE space of onefold integrable functions and i the imaginary unit. The normalization constant is not equally defined in literature and depends on the problem-specific conventions. Here, for a two-dimensional space, it is defined as $1/\sqrt{2\pi}$ so that the inverse transform analogously can be described as

$$\mathcal{F}_{wt}^{-1}\{\tilde{x}(w)\} = x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{x}(w)\, \mathrm{e}^{iwt} \mathrm{d}w. \tag{2.13}$$

If we now assume discrete, equidistant values of the function, for example time steps, the discrete FOURIER transform (DFT) can be applied, which decomposes the discrete signal into a discrete, mirrored spectrum of frequencies. Thus, the transform for a complex vector $\mathbf{x} = (x_0, x_1, ..., x_{N-1}) \in \mathbb{C}^N$ with N elements can be regarded as:

$$\tilde{x}_k = \sum_{j=0}^{N-1} x_j\, \mathrm{e}^{-2\pi i \frac{jk}{N}}, \tag{2.14}$$

for $k = 0, ..., N-1$. This means that for each frequency component the contribution of each single data point to this component is to be determined. The sum gives an estimate of how strong the frequency component is represented in the signal. The inverse transform can be written as

$$x_j = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{x}_k\, \mathrm{e}^{2\pi i \frac{jk}{N}}. \tag{2.15}$$

The DFT is a linear operation and can also be written as a matrix multiplication, $\tilde{\mathbf{x}} = \mathbf{F}\mathbf{x}$, where $\mathbf{F}$ is a unitary $N \times N$ matrix. In physics, this implies that the transformation preserves energy and that the inverse transformation exists and is defined as the adjunct matrix, $\mathbf{F}^{-1} = \mathbf{F}^H$. Generally, to speed up the calculation of the transform, the fast FOURIER transform (FFT) is used, which provides an algorithm that needs substantially less calculation steps for the determination of the FOURIER coefficients as in the direct implementation. According to the algorithm of *Cooley* and *Tukey* [81] the coefficients can be written separately for even ($x_j^+ = x_{2j}$) and odd ($x_j^- = x_{2j+1}$) indices, and for the

transform follows:

$$\tilde{x}_k = \sum_{j=0}^{n-1} x_{2j}\, \mathrm{e}^{-2\pi i \frac{(2jk)}{2n}} + \sum_{j=0}^{n-1} x_{2j+1}\, \mathrm{e}^{-2\pi i \frac{(2j+1)k}{2n}} \tag{2.16}$$

$$= \sum_{j=0}^{n-1} x_j^+\, \mathrm{e}^{-2\pi i \frac{jk}{n}} + \mathrm{e}^{-\pi i \frac{k}{n}} \sum_{j=0}^{n-1} x_j^-\, \mathrm{e}^{-2\pi i \frac{jk}{n}} \tag{2.17}$$

$$= \begin{cases} \tilde{x}_k^+ + \mathrm{e}^{-\pi i \frac{k}{n}}\, \tilde{x}_k^- & \text{if } k < n \\ \tilde{x}_{k-n}^+ - \mathrm{e}^{-\pi i \frac{k-n}{n}}\, \tilde{x}_{k-n}^- & \text{if } k \geq n \end{cases}, \tag{2.18}$$

with $n = N/2$. The direct implementation of the defined FFT halves the calculation time and can be expressed in form of a recursive algorithm, see Alg. 1. The fast FOURIER transforms in this thesis are almost entirely calculated with an extension of the built-in MATLAB function *fft*, based on implementations of *Frigo* and *Johnson* [125].

Algorithm 1 recursive FFT code

Require: input **x** with N elements
if N=1 **then**
  return x
else
  $\tilde{\mathbf{x}}^+ = \mathrm{fft}\left(\frac{N}{2}, (x_0, x_2, ..., x_{N-2})\right)$
  $\tilde{\mathbf{x}}^- = \mathrm{fft}\left(\frac{N}{2}, (x_1, x_3, ..., x_{N-1})\right)$
  for $k = 0, ..., \frac{N}{2} - 1$ **do**
    $\tilde{x}_k = \tilde{x}_k^+ + \mathrm{e}^{-\pi i \frac{k}{n}}\, \tilde{x}_k^-$
    $\tilde{x}_{k+N/2} = \tilde{x}_k^+ - \mathrm{e}^{-\pi i \frac{k}{n}}\, \tilde{x}_k^-$
  end for
  return $\tilde{\mathbf{x}}$
end if

2.4.2 Multitaper Method

Despite the ubiquitous application of the nonparametric FOURIER transform its implementation involves certain limitations [16]. The assumption, that the FOURIER coefficients in a spectral decomposition constitute a reliable representation of amplitude and phase of a frequency component, is not necessarily given. The power spectral density, determined by the FOURIER transform, rather represents a biased estimate of the true spectral composition. Moreover, the resulting periodogram exhibits a high variance. These drawbacks can be addressed by using the multitaper method [16, 332], proposed by *Thompson* in 1982. At this, the method of tapering handles the trade-off between broadband and narrowband bias of spectral estimates in an efficient manner [41]. As well, the utilization of numerous tapers takes care of a reduction of the variance. The underlying idea was initially proposed by *Bartlett* [24] and *Welch* [359]. The multitaper method makes use of multiple

reciprocal orthogonal tapers that build a local basis of eigenvectors in frequency space for finite data pieces [265]. As a consequence they provide a statistically independent estimate of the underlying spectrum. The tapers are convoluted with the argument of the FOURIER integral, see Eq. (2.12), whereas the final estimated spectrum is determined by averaging over all individual tapered spectra. Originally, *Thompson* suggested to choose the so-called *slepian sequences* or *discrete prolate spheroidal sequences* (DPSS) [315] as tapers.

Given a vector $\mathbf{x} = (x_0, x_1, ..., x_{N-1})$, a stochastic time-discrete process with N discrete time steps is assumed, without any loss of generality. Then, the direct multitaper spectral estimate $S_{MT}(k)$ is defined by

$$S_{MT}(k) = \frac{1}{M} \sum_{m=1}^{M} |\tilde{x}_{k,m}|^2. \tag{2.19}$$

The $S_{MT}(k)$ calculates the average over individual tapered spectral estimates $\tilde{x}_{k,m}$ for frequency k and taper m, which can be described as

$$\tilde{x}_{k,m} = \sum_{j=0}^{N-1} h_{jm}\, x_j\, \mathrm{e}^{-2\pi i \frac{jk}{N}}. \tag{2.20}$$

The taper h_{jm} represents the m^{th} discrete prolate spheroidal sequence for data point j. Generally, an adaptive and more sophisticated method is used, where the individual tapers are weighted to prevent an increase of broadband leakage in tapers of higher orders [265]. Nevertheless, multitaper methods are not as extensively used as it could possibly be [16]. Still, conventional methods are preferred and widely used e.g. in the spectral decomposition of human brain signals. However, the multitaper method has entered several fields and already has been applied in multiple papers where brain data is analyzed [49, 235, 279]. In this thesis, when spectral decompositions refer to multitaper methods, an extended version of the built-in MATLAB function *pmtm* is applied, employing the adaptive weighted tapers.

2.4.3 The Spectrogram

In neuroscience, the spectrogram is a typical tool for the analysis of effects that are reflected in (human) brain signals. The idea is to determine the contribution of a certain frequency (range) to the signal recorded by an electrode at a certain moment. For this purpose, the spectral decomposition is calculated for a sliding time window that is each time assigned to a particular point in time. To work out event-related information, the results are often compared to a predefined baseline that contains no information according to an event and are color coded by the relative power of each frequency and point in time. The results of such an analysis type is shown in Fig. 2.6.

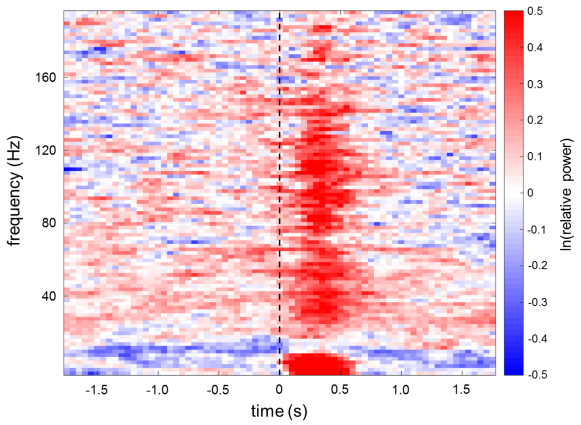

Figure 2.6: Spectrogram of the signal of an exemplary human intracranial electrode contact. According to a certain event ($t = 0\,s$) the spectral decomposition was calculated for frequencies $< 200\,Hz$ and a sliding time window of $0.05\,s$. The relative power values are determined according to an event-independent baseline.

2.5 Machine Learning

Since the classical antiquity, humans chase the desire of creating intelligent machines, machines that are able to generate thoughts and have the ability to act autonomously. The fundamental property to fulfill such tasks is the ability to learn. With the invention of programmable computers this quest had a concrete object to apply the earlier formed theories. Machine learning (ML) describes the artificial generation of knowledge based on experience [6, 104]. Generally speaking, an implemented system has to learn certain attributes by studying a set of examples to be able to generalize on not yet experienced situations or objects. During the learning phase, patterns and regularities are stored to build a knowledge base for further examples. Since the the beginning of machine learning around 1950, the field has grown immense and meanwhile has an intangible impact on our everyday life. Fig. 2.7 shows schematically how a machine learning algorithm generally proceeds. Hereby, the implementations of their framework can differ in the composition of operators. Fig. 2.7A separates feature selection and classification, as it is done e.g. in the *Common Spatial Pattern* algorithm, see paragraph 2.5.2. In contrast *artificial neural networks*, that will be discussed later, learn both stages at the same time, see Fig. 2.7B.

Computers enable to store and process big amounts of data locally, but also from distant servers. Many types of problems, especially when they can be formulated based on precise mathematical rules and required large computational resource, computers have no difficulties in solving the problem while humans have trouble dealing with it. But as soon as it comes to recognition tasks of e.g. speech or visual inputs, humans intuitively and easily handle the problem. In contrast, it is difficult to find mathematical definitions that can help computers to learn those capabilities. Deep learning methods, whose architectures are inspired and borrowed from human neural processing, can be used to solve problems that can not be formulated explicitly. Their structure is formed by many

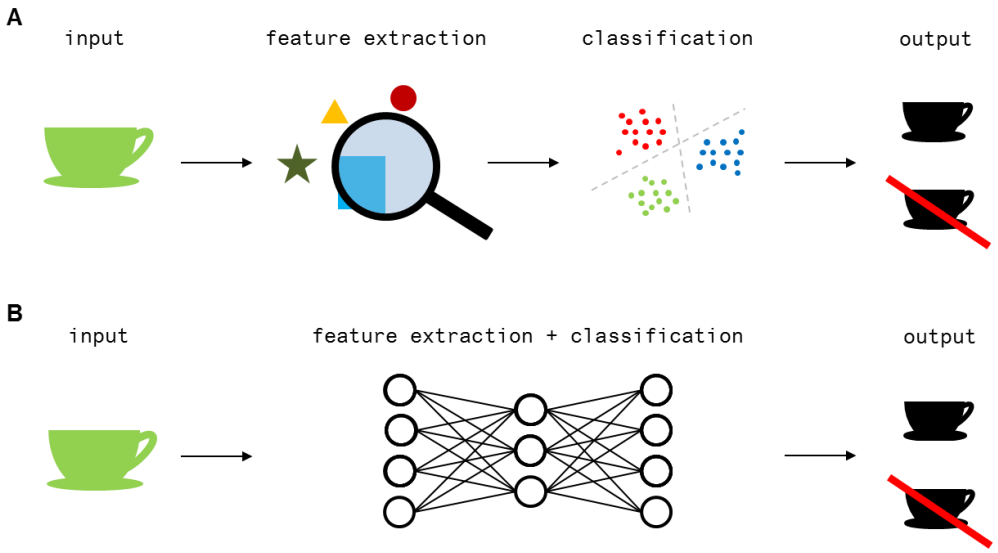

Figure 2.7: General machine learning approach. A To train the model, the input is analyzed for features and then passed to the classifier, which distinguishes between classes. Here, feature extraction and classification are separated. **B** Machine learning approach where the classifier is performing feature extraction and classification jointly, like e.g. in artificial neural networks. *Inspired by https://www.xenonstack.com.*

consecutive modules, called layers, motivating the naming *deep learning*. *Convolutional Neural Networks* (CNNs) form a prominent interpretation of the deep learning methods [40, 145, 304] and build an essential tool for the analysis in this thesis . Nowadays, an incredible amount of general and personal information is stored everyday and can give an detailed insight into behaviour and patterns of all kinds of things. Tab. 2.2 shows a rather bold comparison of a brains and a computers properties.

Table 2.2: A (lame) Comparison of Brain and Computer [376]

	Brain	**Computer**
# arithmetic units	$\approx 10^{11}$	$\approx 10^{9}$
type of arithmetic units	neurons	transistors
type of calculation	massively parallel	generally serial
data retention	associative	address-based
switching time	$\approx 10^{-3}s$	$\approx 10^{-9}s$
theoretical switching processes	$\approx 10^{13}\frac{1}{s}$	$\approx 10^{18}\frac{1}{s}$
actual switching processes	$\approx 10^{12}\frac{1}{s}$	$\approx 10^{10}\frac{1}{s}$

In this thesis three different classification implementations are realized, namely *regularized Discriminant Analysis* (rLDA), the *Filter Bank Common Spatial Pattern* (FBCSP) and the *deep Convolutional Neural Network* (deep CNN). Firstly, in this section the

two more conventional and common used techniques for decoding of brain signals are described. Linear Discriminant Analysis (subsection 2.5.1), often applied in form of a regularized Linear Discriminant Analysis, is a linear transformation technique to reduce dimensions in a problem and to allow an easier separation of classes in order to avoid overfitting. The idea of the Common Spatial Pattern (CSP) algorithm (e.g. [198, 270]) is the decomposition of a set of signals, followed by a transformation into a pseudo-signal space formed by additive subcomponents of the original space, maximizing the differences in variance. Combined with frequency filtering, [14] could generate a more efficient way to solve the problem of frequency band specifity. The idea of the Common Spatial Pattern will be described in subsection 2.5.2, including the FBCSP approach. In the next subsection 2.5.3 Convolutional Neural Networks are introduced, whereas this method is quite novel referred to the examination of EEG. The basic concepts of *transfer learning* and *regularization* complete the section.

The derivations of methods and algorithms in this section are kept quite general according to data type, whereas the data itself is specified to be human brain recordings. The data sets are considered to be derived from healthy subjects or patients (suffering from epilepsy), which will be all in all called participants hereafter. For the human (intracranial) EEG recordings a rather general generative model is assumed,

$$\underline{\mathbf{X}} = \mathbf{V} + \boldsymbol{\zeta}. \tag{2.21}$$

$\underline{\mathbf{X}} \in \mathbb{R}^{e \times t}$ represents the raw, unprocessed brain recordings of e electrodes or channels, and t discrete time steps per trial. $\mathbf{V} \in \mathbb{R}^{e \times t}$ is the underlying activity at the e channels (derived in section 2.1, see Eq. (2.11)), while $\boldsymbol{\zeta} \in \mathbb{R}^{e \times t}$ considers any kind of disturbing signals, be it signals originating from the sensors or correlated noise caused e.g. by artifacts. Only $\underline{\mathbf{X}}$ is considered to be observable and further builds the base for the preprocessed data $\mathbf{X}$. The processed set $\mathbf{X}$ serves as input for the classifiers f.

It is assumed that for each participant a single data set of brain recordings is given. The data sets are considered to be cut into time segments according to an event, whereas those segments are called samples or trials. The data set per participant is defined as $\mathbf{D}_i = \{(\mathbf{X}_{1,i}, l_{1,i}), ..., (\mathbf{X}_{N_i,i}, l_{N_i,i})\}$, where N_i denotes the total number of trials for participant i, and $l_{j,i}$ the class label of trial j and participant i, referring to one of the classes c_k, with k as the number of classes. In this thesis decoding will be performed to distinguish either between robot type, or between erroneous and correct conditions of a task e.g. $c_j \in \mathcal{C} = \{c_1 = "erroneous", c_2 = "correct"\}$

The parametric classifiers f are designed to project from the electrode space to the space of classes, assigning labels to the individual trials, depending on the set of parameters $\boldsymbol{\theta}$,

$$f(\mathbf{X}_j, \boldsymbol{\theta}) : \mathbb{R}^{e \times t} \rightarrow \mathcal{C}. \tag{2.22}$$

It is assumed that the mathematically defined classifiers f perform the feature selection as well as fulfill the function of a classifier.

2.5.1 Linear Discriminant Analysis

Linear discriminant analysis is a supervised linear transform technique to discriminate multiple classes or reduce dimensions. The problem was initially described by *Fischer* in 1936 [123] as a two dimensional problem and later extended to multiple dimensions by *Rao* in 1948 [287]. Hereby the goal is to project data onto a lower-dimensional space where classes are better separable, meaning a maximal distance of class means while keeping the within class variance minimal, see Fig. 2.8. Formally, the transformation is performed from an n dimensional to a k dimensional space with $k \leq n-1$, but maintaining the information discriminating the classes. The dimensionality reduction helps to reduce computational costs and furthermore can be helpful to avoid *overfitting* (see subsection 2.5.3) by minimizing the error in parameter estimation (curse of dimensionality). The question arises of how a suitable subspace can be found. The determined *eigenvectors* give an answer to this question, where the length of the vector (*eigenvalue*) defines how informative the vectors are, with values close to 0 showing less informative entity. An eigenvector refers to a certain transformation and doesn't change its direction after the transformation. If the determined eigenvectors exhibit a similar magnitude, it is be an indicator that the data is already projected on a suitable feature space.

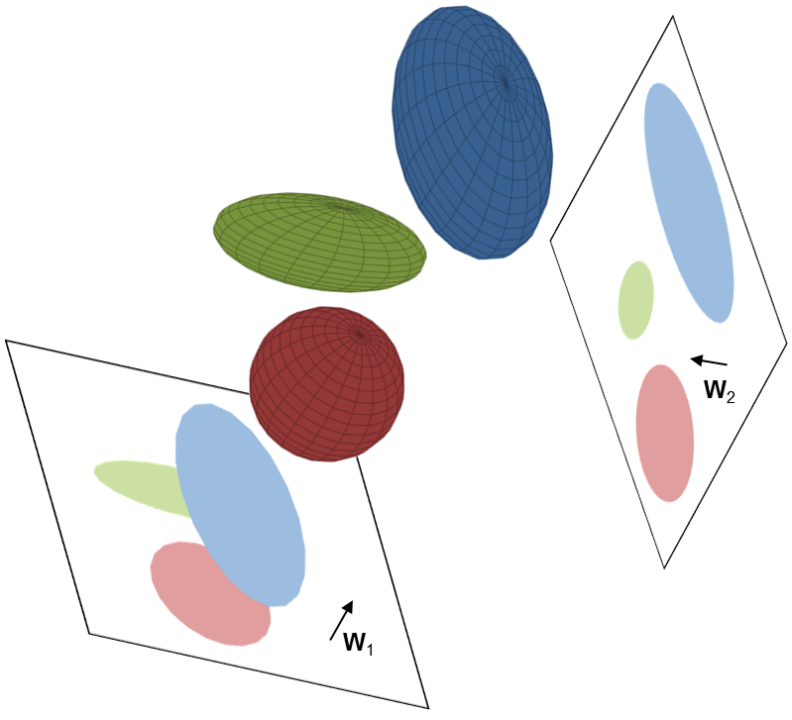

Figure 2.8: Exemplaric projections of a 3D classification problem. The three-dimensional distributions of the classes are projected onto a two-dimensional subspace, according to the hyperplane normals $\mathbf{W}_1$ and $\mathbf{W}_2$. LDA searches for an optimal projection to maximize the distance between the distributions and to minimize the within-class variance. In this case, the projection onto the hyperplane defined by $\mathbf{W}_1$ represents a rather poor decision for a projection to distinguish the classes. In contrast, the second projection yields in a good separation of classes while keeping the variances minimal. *Inspired by [104].*

Generally, LDA requires GAUSSIAN distributed data, statistically independent features and identical covariance matrices for each class. Furthermore, it assumes that the class distribution is known. However, although the required distribution as well as a joint covariance matrix is not ensured, LDA often can achieve good classification results, e.g., in face and object recognition tasks [104, 217].

As already indicated, the idea behind the linear discriminant analysis is to maximize the difference between the classes (distance between the class means) while minimizing the variance within the classes. Suppose that $\mathbf{X} \in \mathbb{R}^{e \times t}$ represents a sample of preprocessed brain signals and $\mathbf{\Xi}_i$ the space of all samples labeled with class index i, then

$$\mathbf{M}_i = \frac{1}{n_i} \sum_{\mathbf{X} \in \mathbf{\Xi}_i} \mathbf{X} \tag{2.23}$$

describes the mean of class i with n_i elements and

$$\mathbf{M} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{X}_j = \frac{1}{N} \sum_{i=1}^{k} n_i \mathbf{M}_i \tag{2.24}$$

the total mean over all k classes with the total number of trials $N = \sum_i n_i$. We need to find a projection

$$\hat{\mathbf{X}}_j = \mathbf{W}\mathbf{X}_j \tag{2.25}$$

for trials j that helps us to discriminate the classes as good as possible. This problem can be addressed by maximizing FISHER's criterion

$$\mathcal{J}(\mathbf{W}) = \frac{\sum_{i=1}^{k} \sum_{\mathbf{X} \in \mathbf{\Xi}_i} (\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^2}{\sum_{i=1}^{k} \hat{\mathbf{S}}_i} \tag{2.26}$$

where $\hat{\mathbf{M}} = \mathbf{W}\mathbf{M}$ is the projection of the total mean and $\hat{\mathbf{S}}_i = \sum_{\mathbf{X} \in \mathbf{\Xi}_i} (\hat{\mathbf{X}} - \hat{\mathbf{M}}_i)^2$ the scatter of class i after the projection. We can now define the scatter matrix in the native electrode space for each class:

$$\mathbf{S}_i = \sum_{\mathbf{X} \in \mathbf{\Xi}_i} (\mathbf{X} - \mathbf{M}_i)(\mathbf{X} - \mathbf{M}_i)^\top \tag{2.27}$$

what leads us to the within-class scatter matrix

$$\mathbf{S}_w = \mathbf{S}_1 + ... + \mathbf{S}_k. \tag{2.28}$$

The numerator of FISHER's criterion Eq. (2.26) quantifies the distance between the class means and can be transformed as follows

$$\sum_{i=1}^{k} \sum_{\mathbf{X} \in \mathbf{\Xi}_i} (\hat{\mathbf{M}}_i - \hat{\mathbf{M}})^2 = \sum_{i=1}^{k} n_i (\mathbf{W}^\top \mathbf{M}_i - \mathbf{W}^\top \mathbf{M})^2 \tag{2.29}$$

$$= \sum_{i=1}^{k} \mathbf{W}^\top (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^\top \mathbf{W} \tag{2.30}$$

$$= \mathbf{W}^\top \mathbf{S}_b \mathbf{W} \tag{2.31}$$

whereas

$$\mathbf{S}_b = \sum_{i=1}^{k} n_i(\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})^\top. \tag{2.32}$$

defines the between-class scatter matrix. We are now interested in the projection $\mathbf{W}$ that maximizes the between scatter and at the same time minimizes the within scatter of the classes. A measure for the spread of scatter matrices, or covariance matrices, is the determinant, which is a product of the eigenvalues. Here, the eigenvalues reflect the variance along his eigenvector. We can now reformulate FISHER's criterion in form of the RAYLEIGH coefficient

$$\mathcal{J}(\mathbf{W}) = \frac{|\mathbf{W}^\top \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^\top \mathbf{S}_w \mathbf{W}|}. \tag{2.33}$$

Eq. (2.33) can be solved as a generalized eigenvalue problem, assuming $\mathbf{S}_w$ to be non-singular:

$$\mathbf{S}_b \mathbf{W} = \lambda \mathbf{S}_w \mathbf{W}. \tag{2.34}$$

Here, λ denotes the eigenvalue(s). Based on the derived matrix $\mathbf{W}$, the data can be projected on the according hyperplane of the original data space and separated into different classes. The LDA classification can be broke down into the steps defined in Alg. 2.

Algorithm 2 LDA classification

Require: input $\mathbf{X}$

1: $\mathbf{S}_w$, $\mathbf{S}_b \leftarrow$ compute scatter matrices
2: $\mathbf{v}$, $\lambda \leftarrow$ compute eigenvectors and according eigenvalues
3: sort the $\mathbf{v}$ by decreasing λ
4: select k vectors with largest λ
5: **for** all $\mathbf{X}_j \in \mathbb{R}^{e \times t}$ **do**
6: $\quad \hat{\mathbf{X}}_j \leftarrow$ compute projected data
7: **end for**
8: discriminate classes on hyperplane

2.5.2 Spatial Filtering: the Common Spatial Pattern

Common Spatial Pattern is an algorithm to construct optimal spatial filters to maximize variances and e.g. discriminate several classes. Thereby, multivariate signals are separated into additive subcomponents. For instance, it is highly efficient in spatial filtering for the detection of localized neural rhythmic activity, *Event-Related Synchronization* and *Event-Related Desynchronization*, which is elicited by performed and imagined motor activity [42]. Hence, spatial CSP filtering admits inference from motor action to the underlying regions that drive the motor action. For the functionality of the CSP algorithm, an oscillatory process is required. Furthermore, the implementation assumes that frequency band and time window of the wanted effect are known. The band-passed signal is expected

to be jointly GAUSSIAN and the source activity constellation is expected to differ between classes.

Common Spatial Pattern is the approach to find a linear transform of signals that makes them more discriminative by minimize the correlation between the signals or classes, respectively. This is expected to yield in a better classification. Again, assuming $\mathbf{X}$ as input signal, a linear transform or projection can be defined,

$$\hat{\mathbf{X}} = \mathbf{V}^\top \mathbf{X}, \tag{2.35}$$

where $\mathbf{V} \in \mathbb{R}^{e \times e}$ represents the mixing matrix that projects the data from the electrode space onto a so-called surrogate electrode space, e being the number of channels. The column vectors $\mathbf{v}_j \in \mathbb{R}^e$ of this projection matrix are called (spatial) filters and create the surrogate electrodes out of the original input channels. However, the rows of the mixing matrix $\mathbf{V}$ determine the influence of the according channel on the newly created surrogate electrode. Moreover, the matrix $\mathbf{A} = (\mathbf{V}^{-1})^\top \in \mathbb{R}^{e \times e}$ is the demixing matrix, whereas its column vectors $\mathbf{a}_j \in \mathbb{R}^e$ are referred to as (spatial) patterns [43].

Assuming that the data is centered and scaled, the estimates of the covariances can be written as

$$\mathbf{\Sigma}_i = \frac{1}{||\boldsymbol{\xi}_i||} \sum_{j \in \boldsymbol{\xi}_i} \mathbf{X}_j \mathbf{X}_j^\top \tag{2.36}$$

with $\boldsymbol{\xi}_i \, (i \in \mathcal{C})$ as the vector of indices of the according trials pertaining to class i. For two classes the purpose is to maximize the covariance of the spatially filtered signal for one class, while minimizing it for the other one, which once more can be reformulated as a maximization problem of the RAYLEIGH coefficient

$$\mathcal{J}(\mathbf{W}) = \frac{|\mathbf{V}^\top \mathbf{\Sigma}_1 \mathbf{V}|}{|\mathbf{V}^\top \mathbf{\Sigma}_2 \mathbf{V}|}. \tag{2.37}$$

As in Eq. (2.34), the simultaneous diagonalization of the two covariance matrices can be solved as a generalized eigenvalue problem, assuming $\mathbf{\Sigma}_2$ to be non-singular:

$$\mathbf{\Sigma}_1 \mathbf{V} = \lambda \mathbf{\Sigma}_2 \mathbf{V} \tag{2.38}$$

Often, this can be found in an alternative formulation, when discriminating activity for one class and common activity for both classes $\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$

$$\mathbf{\Sigma}_1 \mathbf{V} = \lambda (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \mathbf{V}. \tag{2.39}$$

As a consequence of this procedure, large positive eigenvalues correspond to a large response for one of the classes, while large negative values correspond to another class. Therefore one should consider filters from both sides of the eigenvalue spectrum. Now, based on the eigenvalues a feature selection can be performed, e.g. taking the m best filters [43, 286]. Commonly, merely a small amount of spatially filtered signals is used as features [286]. There are several ways to make decisions on the filters, e.g. taking the mutual information into account by calculating the *Mutual Information based*

Best Individual Feature (MIBIF). An overview of different approaches is given in [14]. However, in this thesis the selection of the 3 first and the 3 last filters for decoding the recorded signals is consequently used. Generally speaking, after the selection of the filters, the according rows of the projected signals, $\hat{\mathbf{X}}_{m,:}$, $m \in \mathbf{\Phi}$, form the feature vector

$$\mathbf{Z}_m = \log \frac{\mathrm{var}(\hat{\mathbf{X}}_m)}{\sum_{i \in \mathbf{\Phi}} \mathrm{var}(\hat{\mathbf{X}}_i)}, \tag{2.40}$$

which serves as in input to the classifier. Here $\mathbf{\Phi}$ represents the index space of selected filters. To approximate normal distribution the resulting ratio is log-transformed.

Filter Bank Common Spatial Pattern

The efficiency of the CSP algorithm is strongly dependent on its operational frequency band. Unfiltered data or unsuitable processing yields in low performances. However, the frequency band is participant-specific and broad band or manual selection can either lead to bad classification or can be annoying. To address this issues, several ideas have led to an improved handling of the CSP algorithm. For example the *Common Spatio-Spectral Pattern* (CSSP) optimizes filters using a one time-delayed sample [212] and finds an improvement in the *Common Sparse Spectral Spatial Pattern* (CSSSP) [102]. Another idea, the *Sub-Band Common Spatial Pattern* (SBCSP), addresses the fundamental problem by decomposing the data into several sub-bands, calculating a score for each of the spatial filters and fusing the score after the classification. A final classification discriminates the different sub-band scores. However, the algorithm used in this thesis is based on the generation of a filterbank. The *Filter Bank Common Spatial Pattern* was built following the theory of [14].

Fig. 2.9 shows the architecture of the FBCSP algorithm that is used in this thesis. Here, the preprocessed data is separated into 35 different, non-overlapping frequency bands between $0.5\,Hz$ and $144\,Hz$, based on a previously defined filter bank. The bandwidth of power modulations in EEG appears lower for small frequencies than it is for higher frequencies [66]. To consider this concept and to obtain an optimal spectral coverage, in the used implementation a bandwidth of $2\,Hz$ is applied to frequencies up to $30\,Hz$, while choosing a bandwidth of $6\,Hz$ for higher frequencies ($> 30\,Hz$). Each of the frequency bands passes the spatial CSP filtering and delivers several spatial filters. In general, only a well-defined selection of spatial filters is sufficient to yield good performances, whereas too much filters often result in overfitting [14, 75]. Therefore, a predefined algorithm or selection rule subsequently selects the features which will be applied to the data. As already mentioned, in this thesis the first and last three filters are consequently selected (highest discriminative properties), resulting in six spatial filters that are entering into the classification. The rLDA classifier finally cares for the categorization. Similar implementations [14, 43, 303] show that this architecture is suitable for decoding patterns in physiological EEG. Furthermore, the filter bank common spatial pattern is a standard method in EEG classification like e.g. motor decoding. This encourages the implementation of this method in this thesis for decoding and also strengthens the

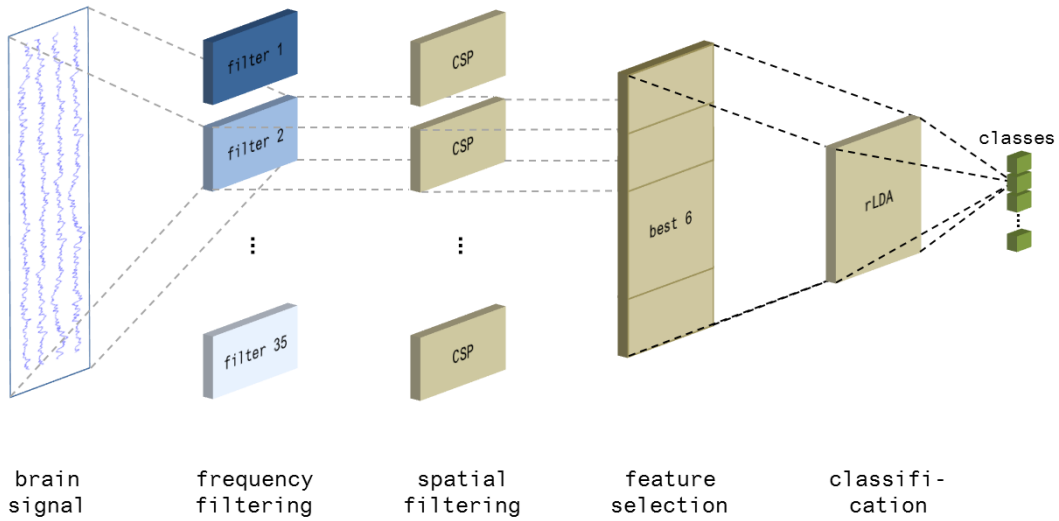

Figure 2.9: Filter Bank Common Spatial Pattern architecture. The architecture in this thesis is built according to the recommendations in [14]. The brain signals are filtered into different frequency bands. The CSP algorithm extracts the spatial filters per band, sorted by variance. The decision on the features is made either by an algorithm or a decision rule. In this thesis, the final classification is performed by an rLDA unit.

utilization as a baseline for the evaluation of deep CNN performances. The FBCSP algorithm implemented in this thesis is described in the following Alg. 3.

Algorithm 3 FBCSP classification

Require: input $\mathbf{X}$
1: compute band pass of $\mathbf{X}$
2: **for** all classes c_i **do**
3: $\quad \mathbf{\Sigma}_i \leftarrow$ compute class-wise covariance matrix
4: **end for**
5: $\mathbf{v}, \lambda \leftarrow$ compute eigenvectors and according eigenvalues
6: select filters $\mathbf{v}_j$
7: **for** all selected filters $\mathbf{v}_j$ **do**
8: $\quad$ calculate spatial filtering of band-passed $\mathbf{X}$
9: **end for**
10: form feature vector
11: perform rLDA classification

2.5.3 Artificial Neural Networks

Inspired by biological models, artificial neural networks (ANNs) commenced their path in the early 1940s, nearly at the same time when programmable electronic computer were invented. The underlying idea was to imitate the information processing of natural neurons, connecting nerve cells in the brain and the spinal cord. 1943 *McCulloch* and *Pitts*

described a first neuron-based neurological network and proved its capability to solve logical and arithmetical functions [229]. Quickly, they discovered that those networks could be used for recognition of (spatial) patterns. After pioneering work of *Minsky* and *Rosenblatt* first neurocomputers where designed and implemented. Not until detailed analysis of the perceptron showed that important problems could not be solved [234], the triumphal march of artificial neural networks was stopped for the time being. Thanks to propelling work of e.g. *Kohonen* [195, 196, 197], *Werbos* [363] and *Hopfield* [168] the research area experienced a revival in the 1970s and early 1980s. Since then this field has developed incredibly fast and problems that were not linearly separable could be solved by means of multilayer perceptrons, trained by the *backpropagation* algorithm. Most of all in the last decade, artificial neural networks won a huge popularity and dominated in several pattern recognition contests.

ANNs are an arrangements of many interconnecting neurons, exchanging information, whereas their connections inhibit certain numeric values called *weights* that are adapted during learning processes. A greater "parallel" arrangement of neurons is called layer. Convolutional neural networks are a particular implementation of artificial neural networks, using discrete convolutions to form the input for succeeding layers. Neurons only receive input from preceding layers, what makes them feedforward networks. In particular, convolutional neural networks were directly inspired by first discoveries on the visual system [171]. According to *Hubel* and *Wiesel*, basic visual features like oriented edges elicited a response in neurons in the primary visual cortex. So-called *complex cells* obtained more spatial invariance by pooling over inputs from several *simple cells*, which responded to preferential orientations at small, spatially localized receptive fields. The increasing spatial invariance reached by feedforward connections and the selectivity to specific patterns made for the design of convolutional neural networks. In 1983, *Fukushima* introduced the neuronal model *neucognitron*, an artificial visual system that was able to recognize handwritten symbols or patterns, respectively [127]. *Yann LeCun*, who is considered to be the father of CNNs, credits the root of his work to the neucognitron [207].

The following subsection shall give a brief introduction to the theoretical background of artificial neural networks, starting with the concepts of the perceptron, the loss function and multilayer perceptrons. Then, before introducing the theory of CNNs, an essential learning algorithm called backpropagation is described, building the base for the supervised learning used for analysis in this thesis.

The Perceptron

Artificial neural networks are based on interconnection of several neurons. One of the most simple representation of such a network is called *perceptron* and was introduced in 1958 by *Rosenblatt* [293]. It is one of the basic concepts for artificial neural networks and builds the fundamental processing unit by handling with diverse inputs. Each input $x_j \in \mathbb{R}$, $j = 1, ..., d$ is connected to an according weight w_j, where w_0 defines the bias of system, see Fig. 2.10. For the most simple case the output is a weighted aggregate of the

inputs,

$$y = \sum_{i=1}^{d} w_j x_j + w_o = \mathbf{w}^\top \mathbf{x}. \tag{2.41}$$

This equation describes a hyperplane in two-dimensional space and therefore divides the space into two parts, what can be used to discriminate classes. If we want the perceptron

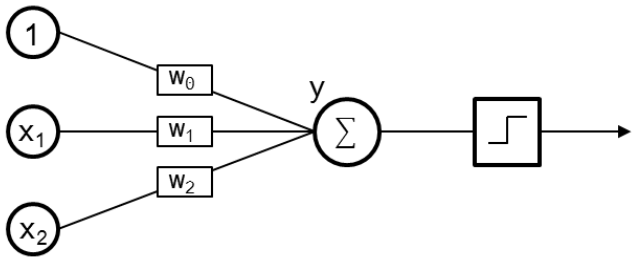

Figure 2.10: The perceptron. The input units $x_j, j = 1, ..., d$ are weighted by the according weights w_j, where w_0 stand for the bias. y is given by the summed and weighted inputs. The output of the perceptron is compared to e.g. a step function.

to discriminate classes, we can check the output of the function and compare it to some well defined threshold ϑ. Suppose that the classes are linear distinguishable, we can e.g. define the classifier

$$f(y) = \begin{cases} c_1 & \text{if } y > \vartheta \\ c_2 & \text{else} \end{cases} \tag{2.42}$$

that assigns the input to a certain class according to the relation of the output y and the threshold ϑ. If we now consider that there is more than one output, $m > 1$, we can calculate the individual projections by

$$y_i = \mathbf{w}_i^\top \mathbf{x}, \tag{2.43}$$

whereas the set of outputs is defined by

$$\mathbf{y} = \mathbf{W}\mathbf{x}. \tag{2.44}$$

$\mathbf{W} \in \mathbb{R}^{m \times d}$ represents the weight matrix and the rows stand for the weight vectors of the m perceptrons. Then, the largest output defines the underlying class,

$$f(\mathbf{y}) = c_i \,|_{i,\, y_i = \max_k y_k}. \tag{2.45}$$

Up to now, our consideration releases the results of the classification as information about the class, not about the probability. Thus, the output of the last layer is usually given after applying the *softmax function* [57], converting the output into a categorical probability distribution. In this approach, the normalized exponential of the weighted sums is calculated

$$p_i = \frac{\mathrm{e}^{y_i}}{\sum_k \mathrm{e}^{y_k}}, \tag{2.46}$$

which can be used for various multiclass classification tasks. Given predefined weights $\mathbf{W}$ and a threshold ϑ, training a perceptron would be accomplished be applying an input vector $\mathbf{x}$ to the network, calculating the output $\mathbf{p}$ and optimizing the weights according to some *criterion function* $\mathcal{L}(\mathbf{W}, \mathbf{p})$ that is minimized if the weights are optimal and the probabilities map the targets as good as possible. For training step $n + 1$ the weights are adjusted the following way:

$$\mathbf{W}(n+1) = \mathbf{W}(n) - \eta(n)\vec{\nabla}\mathcal{L}(\mathbf{W}, \mathbf{p}). \tag{2.47}$$

η is a positive scalar that defines the size of the adjustment steps and is called *learning rate*. Hereby, the criterion function estimates the goodness of the weight choices by comparing the (softmax) output of the perceptron with the class target value t. Hence, we can formulate the perceptron training for an input set $\boldsymbol{\chi} = \{(\mathbf{x}_1, l_1), ..., (\mathbf{x}_d, l_d)\}$ with d elements in algorithmic form, using stochastic gradient descent, see Alg. 4.

Algorithm 4 perceptron training with stochastic gradient descent

```
initialize W, η, threshold ϑ
while |η∇⃗L| ≥ ϑ do
    for (x_i, l_i) ∈ χ do
        y ← Wx
        p ← e^y / Σ_k e^{y_k}
        W ← W − η∇⃗L
    end for
end while
return W
```

The Loss Function

The previously introduced criterion function often appears in form of a *loss function* and is part of the statistical decision problem. According to the decisions made by the classifier for the elements of the data set, the loss function assigns a well defined score that originates from the difference of output y or probability p, respectively, and target t. Both output and target are valuation of the decision on the same underlying set of events. Based on the score, the parameters of the classification process can be adapted, leading to an optimization in discriminating classes. It is not obvious to make a decision on a specific loss function related to an existing classification problem. At this point, some of the most popular functions will be introduced.

The *0-1 loss* ascribes decisions drastically, punishing wrong decisions equally while punishing correct decisions not at all,

$$\mathcal{L} = \begin{cases} 0 & \text{if y = t} \\ 1 & \text{else.} \end{cases} \tag{2.48}$$

The *quadratic loss function* is a more resilient version of a criterion function. Its symmetric construction ensures that errors to both sides of the target are scored equally. This technique is often used in regression analysis and can be formulated using a constant b,

$$\mathcal{L} = b(t-p)^2. \tag{2.49}$$

Other approaches are based on the natural logarithm of the decisions, so is the *negative log likelihood loss*, evaluating the calculated probability of classifier:

$$\mathcal{L} = -\log p. \tag{2.50}$$

The *cross entropy loss* originates from information theory and provides a measure of difference between two probability distributions. In our case this is to evaluate the quality of our classifying model. The following equation shows the formulation in case of discrete values, while the second equality holds for two classes:

$$\mathcal{L} = -\sum_i t_i \log p_i = -t_1 \log p_1 - (1-t_1)\log p_2. \tag{2.51}$$

The Multilayer Perceptron

The problem of linear separability, for example in the often cited XOR-problem, shows the limited applicability of singlelayer perceptrons. They are not able to estimate non-linear discriminants and can only approximate linear functions of the input. This constraint does not hold for feedforward networks with at least three layers, one input and ouput layer each and an additional layer. In such an architecture, non-linear discriminants can be implemented, enabling the user to approximate nonlinear functions of the input. Networks with multiple layers of perceptrons are called *multilayer perceptrons* (MLPs), where layers that are not input or output layers are called *hidden layers*. Fig. 2.11 shows

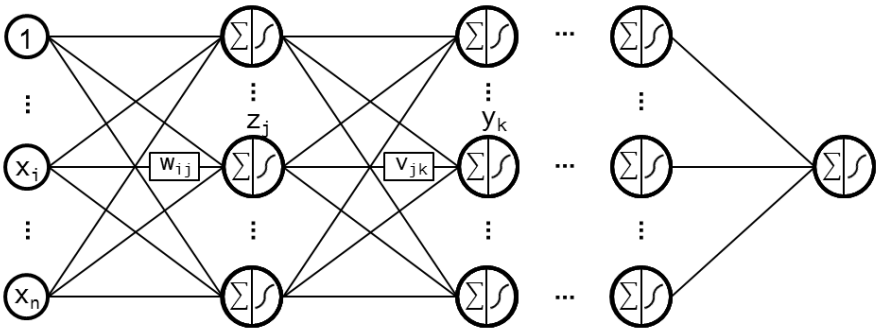

Figure 2.11: Structure of a multilayer perceptron. The weights w_{ij} establish a weighted connection between the input neurons x_i and the neurons z_j. The $z_j, y_k, ...$ represent the neurons of the hidden layers, whereby each neuron receives a linear combination of preceding neurons. The activation function transforms the linear combination before passing it to succeeding neurons.

a schematic arrangement of an MLP with several hidden layers. Each neuron is fed by

a linear combination of preceding neurons. To model the relative firing frequency of the action potential in the cell, the processed input of each neuron is converted by the so-called *activation-function*. Among others this is typically accomplished by the use of a logistic function, a *rectified linear unit* (ReLU) [243] or an *exponential linear unit* (ELU) [78]. Activation functions can implement non-linear features to the network and therefore build an essential module. Complicated, non-linear relations between input and target can hereby be learned. In general those functions have to be differentiable, as the learning algorithms used here are based on gradients (see paragraph **Backpropagation**). In the following the set of all neurons j for which a connection $j \rightarrow k$ exists will be called $pred(k)$. Given the activation $a_j = act(net_j)$, the input net_k for neuron y_k can be calculated as

$$net_k = \mathbf{v}_k^\top \mathbf{a} = \sum_{j=1}^{J} v_{jk}\, act(net_j) \tag{2.52}$$

with preceding neurons z_j (see Fig. 2.11) having an input net_j, $k \in pred(k)$. Please note that here and in the following considerations the $z_j, y_k, ...$ stand just for the names of the neurons, whereas the $net_j, net_k, ...$ are defined as the sums of the weighted inputs. Furthermore, connections can also hop over one or more layers, what is called a *shortcut*. However, this shall not be further explained in detail. Alg. 5 drafts the code for the forward pass of an MLP.

Algorithm 5 forward pass of MLP

Require: input **x**, MLP, topological naming and indexing of neurons
1: **for** input neurons i **do**
2: $a_i \leftarrow x_i$
3: **end for**
4: **for** hidden and output neurons i in topologocial order **do**
5: $net_i \leftarrow \sum_{j \in pred(i)} w_{ij} a_j + w_{i0}$
6: $a_i \leftarrow act(net_i)$
7: **end for**
8: **for** output neurons i **do**
9: $\mathbf{y} \leftarrow (a_1, ..., a_i)$
10: **end for**
11: **return y**

Backpropagation - Training the MLP

Training a MLP is similar to the training of a single perceptron, but in contrast the output is non-linear. The idea is to optimize the output according to the target by propagating stepwise from the output back to the input. This can be done by adapting the weights and minimizing the error or criterion function, respectively. Thus, for a given set of training

data $\mathbf{D} = \{(\mathbf{x}_1, l_1), ..., (\mathbf{x}_N, l_N)\}$ we have to minimize

$$\mathbf{L} = \sum_{i=1}^{N} \mathcal{L}(\mathbf{W}; \mathbf{x}_i, l_i). \tag{2.53}$$

Now, the influence of the individual weights on the criterion has to be determined

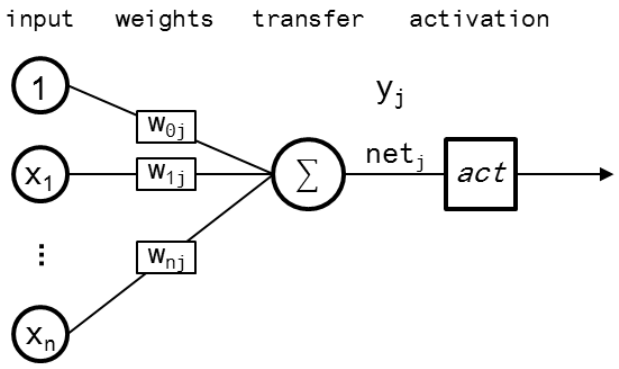

Figure 2.12: Schematic overview of a single neuron. The summed and weighted predecessor inputs for neuron y_j are defined as net_j. The net_j are passed to the activation function before contributing to the input for preceding neurons.

and optimized. In Fig. 2.12 a single artificial neuron and its predecessors is shown schematically. To minimize the error, we have to calculate the partial derivative of the criterion function and use the chain rule. For weight w_{kl} in the simplified example in Fig. 2.12 we get

$$\frac{\partial \mathbf{L}}{\partial w_{kl}} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}(\mathbf{W}; \mathbf{x}_i, l_i)}{\partial w_{kl}} \tag{2.54}$$

$$= \sum_{i=1}^{N} \frac{\partial \mathcal{L}(\mathbf{W}; \mathbf{x}_i, l_i)}{\partial a_l} \frac{\partial a_l}{\partial w_{kl}} \tag{2.55}$$

$$= \sum_{i=1}^{N} \frac{\partial \mathcal{L}(\mathbf{W}; \mathbf{x}_i, l_i)}{\partial a_l} \frac{\partial a_l}{\partial net_l} \frac{\partial net_l}{\partial w_{kl}}. \tag{2.56}$$

This reversed learning method is called *backpropagation* [61, 186, 363] and is widely used to train artificial neural networks. Furthermore, it could be shown that for the input the algorithm can lead to useful internal representations in deeper layers, building the base for deep learning [297]. Backpropagation belongs to the group of supervised learning algorithms. Based on this method, we can propagate through a deeper and more complex network and minimize the errors for each single pattern by adapting the weights. The idea is schematically shown in Alg. 6.

Algorithm 6 MLP learning

```
initialize W, minimization approach
while error diverges do
    for (x_i, l_i) ∈ D do
        apply x_i, generate output
        calculate all ∂L(x_i)/∂w_kl
    end for
    calculate all ∂L(D)/∂w_kl
    update weights
end while
```

The Convolutional Neural Network

Convolutional neural networks are a special type of feedforward neural networks [145, 209, 304], designed for the processing of grid-like topology data, e.g. time series or image data with pixels or voxels. Thus, the architecture partially exhibits 2D or 3D assemblies of neurons, see Fig. 2.13B. CNNs are especially useful if the input has an intrinsic hierarchical structure, as e.g. images are built of basic components like edge and lines that form simple shapes that again are combined to form more complex objects and so forth. Besides the networks are able to learn non-linear features, whereas the lower level features are combined to higher level features. The activity of a neuron is computed by discrete convolutions, before serving later for succeeding layers as an input. The input is convoluted with a comparatively small convolution matrix or tensor called *filter kernel*, see Fig. 2.13A. As a consequence not every output unit interacts with every input unit, known as *sparse connectivity*. The information transfer is realized according to the *receptive field* in biological models, which is defined by the set of sensory receptors that translates information to one single neuron. This corresponds to the expression $pred(k)$ for MLPs, being the set of predecessors (neurons) that contribute to the input for a neuron k. The weights for the neurons of one layer are identical, as a certain filter kernel is applied to each local information carrier, or receptive field. This concept is called *parameter sharing* or *tied weights*. Another property of CNNs is its equivariance to translations. Fig. 2.13A depicts an example of two receptive fields being convoluted by a filter kernel and mapped onto an output feature space. Hereby, the kernel sweeps the positions of the input step by step, where the stepsize ≥ 1 is called *stride*. Thus, it can appear that neighbouring neurons are partly activated by a same subset of preceding neurons. A common method also adds zeros at the edges of the input to allow the kernel to sweep a larger area. This is referred to as *zero padding*, but shall not be deepened incidentally.

Basically, best CNN architectures are built on a basic structure of layers (blocks) based a convolutional layer, followed by an activation and a pooling layer. Containing at least of one convolutional layer, which will be examined more precise in this paragraph, usual networks are a stack of several of the so-called blocks. The use of multiple filters per convolution results in an increase in some of the dimensions. The pooling layer (in

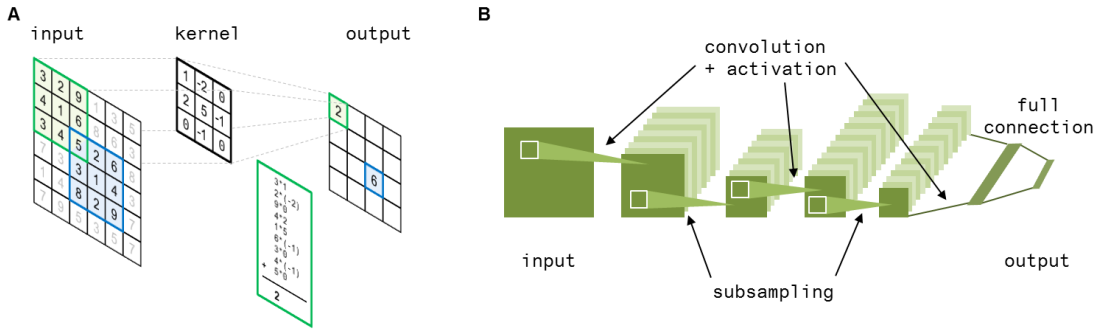

Figure 2.13: Basic function of a convolutional neural networks. A Exemplary description of convolution of a 2D grid, using a 3×3 kernel without zero padding and a stride of 1. The convolution of the green square with the filter kernel delivers a scalar value. **B** Structure of a typical CNN. Per convolutional layer, the network exhibits several filters to extract different features, resulting in an increase in some of the dimensions. The pooling layer is represented by the subsampling. The output is given after the fully connected (dense) layer. *Inspired by [310]*

Fig. 2.13B referred to as subsampling) has the function to discard dispensable information and to reduce computation costs, as e.g. the exact positions of patterns in grid-like structures is often not that important. The pooling layer assigns local information to a single score, e.g. by an average or a max operator. Beyond the space-saving property, it allows to build deeper networks, solving more complex tasks. Incidentally, it helps to work against overfitting, which will be explained later in this paragraph. Regularly one or more fully connected layers complete the CNN architecture. The number of output neurons corresponds to the number of distinguishable classes and therefore gives an output for each class. The output can either be given as a probability for each class or as a *one hot* vector, which gives a binary value for each class whether it was selected or not for classification. For sufficient convolutional layers the CNN is called deep convolutional neural network and is considered as a part of deep learning.

The basic mathematical operation behind the CNN is the convolution. For two functions $g(x)$, $h(x)$ the convolution $(g * h)(x) : \mathbb{R}^n \to \mathbb{C}$ is defined as

$$(g * h)(x) = \int_{\mathbb{R}^n} g(\tau)h(x - \tau)d\tau. \tag{2.57}$$

Descriptively, the hereby defined convolution averages function g, weighted by function h, for a continuous set of drifts between the two functions. In case of convolutional neural networks the first argument g of the convolution is represented by the input, while the second argument h can be described by the filter kernel. In our case the data is not continuous but it is available in discrete values e.g. the pixels of an image or the discrete time steps of an EEG channel. So if we now assume that our functions $g, h : \mathbb{D} \to \mathbb{C}$ are defined on top of an discrete space $\mathbb{D}$, the convolution becomes

$$(g * h)(x) = \sum_{m \in \mathbb{D}} g(m)h(x - m). \tag{2.58}$$

At last, if the convolution is to be applied simultaneously on more than one dimension, the kernel can also be chosen to have multiple dimension, e.g. for two dimensions:

$$(g * h)(x_1, x_2) = \sum_m \sum_n g(m, n)h(x_1 - m, x_2 - n) \tag{2.59}$$

$$= \sum_m \sum_n g(x_1 - m, x_2 - n)h(m, n). \tag{2.60}$$

The last equation holds because convolution is commutative, defining the convolution $(h * g)(x_1, x_2)$ by flipping the kernel. Many ANN libraries make use of the *cross-correlation*, but call it convolution. It is similarly defined, but without flipping the kernel:

$$(g * h)(x_1, x_2) = \sum_n g(x_1 + m, x_2 + n)h(m, n). \tag{2.61}$$

However, the methods in this thesis are built on top of PYTORCH, where the cross-correlation (Eq. (2.61)) is used to implement the convolutional layers for the CNNs. The contribution of the convolutional layers is to detect local patterns of the provided input and to map them onto a feature map [209]. The convolution of each receptive field of the input generates the translated information for a succeeding neuron, whereby the number of succeeding neurons per filter usually decreases, Fig. 2.13B. The succeeding map stores the information about the position of certain features (defined by the filter) and how strong the features are represented in this area. To generate layers containing comprehensive features, the input is applied to several different filter kernels, which defines the depth of the volume of output feature maps. Hence, the convolution of layer k for feature map i is given by

$$\mathbf{Y}_i^k = \mathbf{B}_i^k + \sum_j \mathbf{Y}_j^{k-1} \mathbf{K}_{i,j}^k, \tag{2.62}$$

where $\mathbf{B}_i^k$ denotes the bias matrix of layer k and $\mathbf{K}_{i,j}^k$ the filter kernel that connects feature map j in layer $k-1$ with feature map i in layer k. This concept leads to the fact that images are classified in the same way as in the visual system, going from simple features like edges to more complex structures [208].

In this thesis, the CNN decoding is based on the algorithms provided in the BRAINDECODE toolbox, an open-source deep learning toolbox for decoding of raw time-domain EEG. The toolbox was built according to the theory in [303] and includes several models of CNN implementations. This thesis reverts to the model DEEP4NET[2], which constitutes the deep learning delegate provided by the toolbox and is applied using trial-wise training. The models architecture with the thesis-specific parameters is schematically shown in Fig. 2.14. In short, the model's framework is formed by 4 convolutional blocks and a final linear classification. The network is provided with a two-dimensional input consisting of discrete time steps and channels. Initially, the first block stepwise executes a temporal convolution and a spatial filtering over all channels, what already serves as a regularization (see paragraph 2.5.5), without an activation in between. This generates maps of

[2]https://github.com/robintibor/braindecode/blob/master/braindecode/models/deep4.py

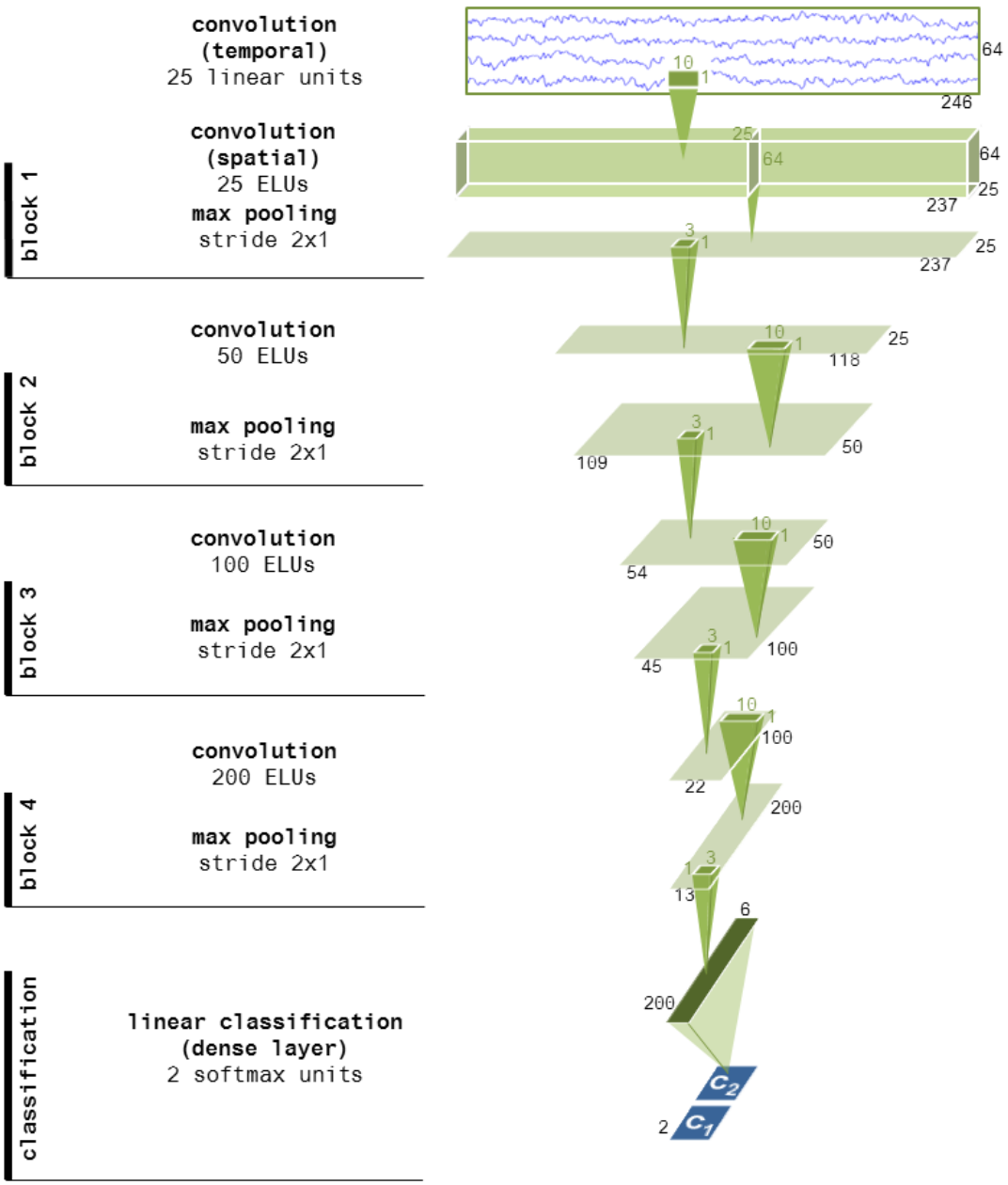

Figure 2.14: Deep convolutional neural network architecture as used in this thesis. The basic structure consists of four blocks, containing convolution, an activation and a max pooling each. The first block contains two convolutional layers, performing subsequently a temporal convolution followed by a linear activation and a spatial convolution followed by an exponential linear unit. The classification is done by the final dense layer, discriminating between two classes. The light green rectangles are the layers inputs while the dark green rectangles represent the filter kernels. Consider that some of the parameters in this specific scheme model depend on the given input. The number of time points and channels varied for different paradigms and analyses. In this example, the samples consisted of 246 time points and 64 channels.

informative units composed of filtered time scraps and channels, which later enter the first max pooling layer. Remember that the measurable physics behind the EEG signals is an overlay of electrical fields of locally generated dipoles, predominantly from the cortical layers [251]. This fact suggests a decomposition of the recorded signals by using spatial filters, as already applied in other successful approaches (see paragraph 2.5.2). In addition, the informative properties of the EEG range over several temporal scales and exhibit local and global temporal modulations. Based on this considerations, the network is conceived to learn at first the spatial filters that care for a decomposition, meanwhile temporal hierarchies are learned in deeper structures of the network. The later blocks consist of an apposition of convolutional layer, an activation by a ELU ($f(x) = x, \forall\, x > 0$ and $f(x) = e^x - 1, \forall\, x < 0$) and finally a max-pooling subsampling. After the last block, a dense softmax classification layer provides the output.

The model is designed to enable the learning of a broad spectrum of features. Moreover a general structure helps to find optimal solutions for different types of control signals and allows a fast adjustment by new methods and extensions. For regularization reasons among others, the network made use of *batch normalization* and performed a *dropout* for all convolutional layers but the first, with a probability of 0.5 (see below). The backward computation of the gradients in the utilized CNNs was based on the output of the categorical cross-entropy loss and optimized using *adam* [190].

Overfitting is a noticeable problem that often occurs in machine learning when a model is trained to classify unseen data. The term designates the fact that a learned model describes a particular set of data too exact e.g when containing to much explanatory variables. Assume that a model learns to classify patterns according to irrelevant features, for example the exposure of photographs. As a consequence, the model doesn't serve to classify unseen data. Overfitting can make itself noticeable e.g. when the error on the training data over trained epochs further decreases while the error on the test data increases, indicating that there was too much training. Several techniques can help to counteract overfitting. Common provisions make use of the fact that more training data can prevent overfitting, therefore data augmentation can be helpful. But also on the side of the model a number of modifications can lead to a solution. A use of a well-generalizing architecture, *regularization* (see subsection 2.5.5) or a reduction of architecture complexity are popular examples to address this problem. The implementation in this thesis also refers to another method called dropout which regulates the models specialization [320]. Hereby, random activations are thrown away with a certain probability, so that the model has to learn these activations once more and doesn't get too fitted. It is advisable to gradually increase the dropout with the depth of the network, because otherwise less information gets translated to the later layers and information gets lost.

Covariate shift denotes a change in the underlying distribution of a functions domain, e.g. the input of an artificial neural network [312]. Thus, due to the change in the distribution, the model is no longer able to generalize on unseen data. Imagine a model that has learned to classify the gender of persons based on passport photos, but e.g was only fed with female samples exhibiting black hair and male samples exhibiting blond hair. Then, the model might have difficulties in generalizing on a larger variety of hair color. A change of the distribution can also appear on the level of internal nodes and is called *internal covariate shift*. With an increasing number of layers in a neural network this effect becomes more severe, since the output of one layer gets translated as an input for the next layer. Batch normalization is an approach to address this issue and was implemented in the employed networks in this thesis. A *batch* refers to the number of training samples that are transferred through the network in one iteration. In a nutshell, in the procedure the output of an activation layer is normalized by subtracting the batch mean and dividing by the batch standard deviation. Now, mean and variance stay the same, the values become more stable and the covariate shift can be reduced; for details see [172]. As a nice side effect, batch normalization accelerates the computation time and has a slight regularization effect, thus, it also helps to avoid overfitting.

2.5.4 Transfer Learning

The general idea of machine learning is the right prediction of unknown instances, based on the learned features of a set of already seen instances. This can also be referred to as *generalization*. During the learning process the quest is to minimize the training error, but as an overall goal the error on the unknown instances, the generalization error, has to be made small. Until now the derivations were based on the assumption that training data and test data originate from the same distribution P. *Transfer learning* refers to classification tasks where the training instances arise from a distribution $P(X)$ but the instances to be predicted are drawn from a different distribution $P(Y)$. Examples for a setting where distributions are distinct can easily be derived from computer vision. A model can be trained to distinguish between certain types of cats, but then also be applied to a task discriminating several types of flowers. It can not be assumed that the generated samples are drawn from a common distribution. But a fundamental assumption for the functioning of a transfer of knowledge is that some factors, that explain the variety in $P(X)$, are relevant for the variety given in $P(Y)$. This can be low-level characteristics that describe the general appearance of the data. So the aim is to utilize data with a certain distribution $P(X)$ to extract information that might be beneficial for predicting on data with a second distribution $P(Y)$. Especially when few data is given, transfer learning can help to increase performances of the used models. Lately, transfer learning methods have gained increased entry into deep learning [101, 130, 373].

For now, machine learning competitions are rejoicing a grand popularity in computer science, likewise using deep learning for transfer learning. A remarkable insight gained from a transfer learning competition [231] is the fact that the learning curve of new categories get much better if the architecture deploys deep representation. For those representations, fewer samples are needed to reach an asymptotic behaviour of the gener-

alization performance [145]. There are different approaches to implement the underlying ideas of transfer learning. Just to name two (extreme) appearances in this context, *zero-shot* and *one-shot* learning make use of either no or only one sample for re-training a model after the transfer before using it for classification. Other approaches use more samples to fine-tune the model.

2.5.5 Regularization

As repeatedly mentioned before, the objective of machine learning approaches is the generalization of learned features on an unknown set of data. Strategies to minimize test errors are called *regularization*, providing a large base for research in the field of machine learning. It is a method to solve ill-posed problems or prevent overfitting by an intended algorithm modification. In general the features are kept, but the influence of the parameters is controlled. One approach is to limit the capacity of models by adding a parameter norm penalty $\boldsymbol{\Omega}(\boldsymbol{\theta})$ to the objective (or loss) function $\mathcal{L}$, resulting in the regularized objective function

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}; \mathbf{X}, y) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, y) + \alpha \boldsymbol{\Omega}(\boldsymbol{\theta}). \tag{2.63}$$

Hereby, $\alpha \in [0, \infty)$ is a hyperparameter weighting the penalty term $\boldsymbol{\Omega}$, where 0 means no regularization and high values for α mean high regularization. The minimization of the regularized objective function decreases the actual objective $\mathcal{L}$.

Ridge regression, also known as TIKHONOV or L^2 regularization, uses one of the most simple and common parameter norm penalties, the *weight decay*. The regularization term $\boldsymbol{\Omega}(\boldsymbol{\theta}) = 1/2||\mathbf{W}||^2$ is thereby added to the objective function to ease the weights $\mathbf{W}$ closer to the origin. If we assume no bias term ($\boldsymbol{\theta} = \mathbf{W}$), the regularized objective function becomes

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}; \mathbf{X}, y) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, y) + \frac{\alpha}{2} \mathbf{W}^\top \mathbf{W}. \tag{2.64}$$

In several cases, e.g. when applying LDA or CSP, it is inevitable to estimate the class means and the class-wise covariance matrices. If only a small number of observations is available, compared to the number of variables, this might become an essential problem regarding the estimate of the empirical covariance matrix. In this case, the empirical covariance matrix can be modified based on a *shrinkage* parameter $\gamma \in [0, 1]$ to generate a well estimated matrix. Introducing this regularization term, the empirical covariance matrix $\boldsymbol{\Sigma}$ can be reformulated:

$$\tilde{\boldsymbol{\Sigma}}(\gamma) = (1 - \gamma)\boldsymbol{\Sigma} + \gamma \nu \mathbf{I}. \tag{2.65}$$

Hereby, ν is a scaling parameter, representing the average eigenvalue of $\boldsymbol{\Sigma}$, and $\mathbf{I}$ the identity matrix. The hyperparameter γ regulates the shrinkage of the covariance towards a spherical covariance, whereas $\gamma = 0$ illustrates the case with no regularization. There are time-consuming ways to optimize γ via cross-validation [124], while other approaches find an optimal hyperparameter using an analytical path [210]. In this formulation, large

sample-to-sample variances are penalized with higher shrinkage values. By minimizing the *expected mean squared error* the estimate of the covariance matrix is determined,

$$\gamma^* = \arg\min_{\gamma} E\left[\sum_{i,j}(\tilde{\mathbf{\Sigma}}_{i,j}(\gamma) - \mathbf{\Sigma}_{i,j})^2\right] \tag{2.66}$$

$$= \frac{\sum_{i,j}\left[\mathrm{var}(\mathbf{\Sigma}_{i,j}) - \mathrm{cov}(\mathbf{\Sigma}_{i,j}, \nu\mathbf{I}_{i,j})\right]}{\sum_{i,j} E[(\mathbf{\Sigma}_{i,j} - \nu\mathbf{I}_{i,j})^2]}. \tag{2.67}$$

2.5.6 Visualization

In addition to striving for the best possible performance, it is enormously important to understand what and how the algorithms learn. Assuming that numerous features are intended to provide information on whether a patient has pathological medical findings or not. Then it would be useful to understand which of these features, e.g., collected by medical tests, can lead to the determination of a disease. For example random forests or linear support vector machines show the importance of single features and the decision of the system is explainable. This does not necessarily apply to neural networks in this form and one of the most common criticisms is that there is no satisfactory explanation of their decision-making behaviour and the importance of individual features. Therefore it is essential, especially in medicine for example, to make further progress in understanding deep neural networks. There are a number of approaches that try to visualize intermediate results and learned properties. *Zeiler* and *Fergus* underline the importance of this question by the example of images and demonstrate how much one can learn from it about the functioning of CNNs [375]. However, for EEG data there are hardly any efforts. In this work the approach of *Schirrmeister et al.* is used to extract information about the learned features [303]. Here, the correlation of changes in network predictions with perturbation changes in input spectral amplitudes are used to obtain information about what the deep networks learned from the data.

Training trials are transformed into frequency domain using the FOURIER transformation (see Eq. (2.12) and Alg. 1) and randomly perturbed by adding GAUSSIAN noise G, while keeping the phases steady. The GAUSSIAN noise exhibits a zero mean and unit variance, so the probability density function of a GAUSSIAN random variable z can be defined as

$$p_G(z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}. \tag{2.68}$$

The perturbed signals are then retransformed to time domain using the inverse FOURIER transformation. Both, the unperturbed and the inverse-FOURIER transformed signals are used to feed the network. The output of the network before the softmax activation is extracted and the difference of the predictions for the perturbed ($\mathtt{pr2}$) and original network ($\mathtt{pr1}$) signal are correlated with the perturbation itself, resulting in the *input-perturbation network-prediction correlation map*

$$\mathsf{M} = corr(G, \mathtt{pr2} - \mathtt{pr1}). \tag{2.69}$$

In some situation, an investigation of the time domain signal may be advantageous. Then, the perturbations can be applied to the time domain voltage signal as well. Again, the network output changes are correlated with the GAUSSIAN perturbation of the signal.

2.6 Statistics

Whenever a systematical connection between empirical evidence (or experience) and a theoretical structure has to be established, there is no way around statistics. Originating from the Latin word *statisticum* (concerning the state), statistics initially described the teaching of data concerning the state, but nowadays denotes the more general field of collecting and evaluating data. The subdomain inductive statistics deduces from a random sample properties of a basic population. In this section two focal areas are described, which later will be used for the evaluations in this thesis.

2.6.1 Statistical Testing

Particularly in basic research, statistical testing is an absolutely essential procedure. With the collection of empirical data it has to be decided whether the actual evidence is based on scientific regularities or rather originates from random processes. Statistical hypothesis testing addresses this interrogation, which is however not always simple and straight forward [334]. In general, the *null hypothesis* is established, assuming that there is no underlying systematic effect. Based on the distribution of raised data it can be deduced how probable the null hypothesis is, what is regularly quantified by the *p-value*. The p-value states the probability to get the generated test statistic (or a more extreme one), if the null hypothesis is true. The p-value is compared to a previously defined *significance level* and if the probability falls below this level, the null hypothesis is rejected. In this case, the alternative hypothesis is accepted and the observed effects are called *significant*.

Basically, there are several methods to test a statistical hypothesis, depending on the sample, the distribution of the sample and the parameter that is going to be examined [103, 358]. Without going too much into detail, in this thesis only an assessable amount of tests is used. The *binomial sign test* is a nonparametric test [164] for one or two paired samples with an extent n (MATLAB: *signtest*). For two samples X, Y the null hypothesis assumes that $P(X < Y) = P(X > Y) = 0.5$, while the test checks the sign of the difference of the two samples. It is a version of a binomial test, but not as strong as e.g. the *Wilcoxon test* or the *Wilcoxon ranksum test* [367]. The WILCOXON test (MATLAB: *signrank*) or WILCOXON signed-rank test checks whether the expectation value μ of a sample can be reconciled with a given desired value μ_o. Likewise as for the sign test, the WILCOXON test requires that the random variable is continuously distributed, but additionally demands a symmetric distribution. However, the WILCOXON ranksum test (MATLAB: *ranksum*) or MANN-WHITNEY U-test can handle two samples exhibiting distinct extent and also demands a continuous distribution. The test validates the accordance of the expectation values of the two samples ($\mu_1 = \mu_2$?).

For testing significance of the decoding accuracies on the level of participants, in this thesis a random permutation test [122, 275] was applied. The random permutation test assists in situation where an underlying distribution cannot be deduced. A vector **c** consisting of the true distribution of class labels is compared to $n = 10^6$ vectors of randomly shuffled labels to generate a realistic distribution of possible outcomes of the classification. Often it appears that the number of trials per class is highly unbalanced. The problem arising from the imbalance can be solved by defining the label matches per vector separately for each class, then averaging over classes and comparing the outcome to the decoded accuracy to estimate the p-value relating to the underlying distribution,

$$p = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \Psi} \delta_{ij}. \tag{2.70}$$

Here, Ψ denotes the subspace of permutations for which the average matches over classes are greater or equal than the decoding accuracy. A schematic implementation of the random permutation test is given in Alg. 7.

Algorithm 7 random permutation test

Require: input **c**
1: **for** n permutations **do**
2: $\bar{\mathbf{c}} \leftarrow$ randomly permute labels of **c**
3: $\boldsymbol{\mu_i}, \boldsymbol{\mu_j}, ... \leftarrow$ classwise mean of label matches
4: $\hat{\boldsymbol{\mu}} \leftarrow$ mean of classwise means
5: **end for**
6: compute p
7: **return** p

2.6.2 Evaluation Metrics

To judge the quality of an algorithm there are numerous different metrics. The evaluations in this thesis are based on the *accuracy*, which gives a measure of how many of the trials have been classified correctly. For binary classification the accuracy is formally defined as

$$ACC = \frac{TP + TN}{P + N}, \tag{2.71}$$

where TP represents the number of true positives, TN the number of true negatives, P and N the total number of positives and negatives, respectively. There are situation where the number of trials per class is highly unbalanced, e.g. $P >> N$. In such cases the accuracy tends to approximate to 1, independent of the classifiers performance. In such a case it makes sense to use the *balanced accuracy*,

$$BACC = \frac{TP}{2P} + \frac{TN}{2N}. \tag{2.72}$$

Chapter 3

Brain Responses During Robot-Error Observation

Brain-controlled robots are a promising new type of assistive device for severely impaired persons. Little is however known about how to optimize the interaction of humans and brain-controlled robots. Information about the human's perceived correctness of robot performance might provide a useful teaching signal for adaptive control algorithms and thus help enhancing robot control. This chapter interrogates the fundamental question whether watching robots perform erroneous vs. correct action elicits differential brain responses that can be decoded from single trials of electroencephalographic (EEG) recordings, and whether brain activity during human-robot interaction is modulated by the robot's visual similarity to a human. To address these topics, two different paradigms have been designed. In a first experiment, participants watched a robot arm pour liquid into a cup. The robot performed the action either erroneously or correctly, i.e. it either spilled some liquid or not. In a second experiment, participants observed two different types of robots, humanoid and non-humanoid, lifting a ball. The robots either managed to lift the ball or not. High-resolution EEG during the observation tasks in both experiments was recorded to train a Filter Bank Common Spatial Pattern (FBCSP) pipeline on the multivariate EEG signal and decode for the correctness of the observed action, and for the type of the observed robot. The findings show that it was possible to decode both correctness and robot type for the majority of participants significantly, although often just slightly, above chance level. Furthermore, the findings suggest that non-invasive recordings of brain responses elicited when observing robots indeed contain decodable information about the correctness of the robot's action and the type of observed robot. This chapter also indicates that, given the relatively low decoding accuracies of this study, further improvements analysis and decoding techniques or the utilization of intracranial measurements of neuronal activity will be necessary for practical applications.

Autonomous technical systems are increasingly accessing our everyday life: The industry has been using robots for construction and assembly for years, autonomous cars are under development, and first robots especially designed for private users or social interaction (e.g., NAO (TM), Aldebaran Robotics, Paris, France or PARO Therapeutic Robot (TM),

Intelligent System Co., Japan) already entered the open market. There is no reason to assume that this trend should lose momentum: Especially healthcare is a very promising field for robotic development with possible applications including robot-assisted surgery, motor analysis, rehabilitation, mental, cognitive and social therapy as well as robot-based patient monitoring systems. Numerous approaches in science try to develop intelligent systems for these purposes, for example when autonomous robotic assistants enable intake of fluids [305]. Besides the process of optimizing the intelligent autonomous drinking [98, 99, 100], there are also more holistic systems, which for example enable a user to communicate with an intelligent robotic service assistant via conscious brain signals by means of a high-level framework [62]. In short, robotic devices are among the key effectors in present and future applications of Brain Computer Interface (BCI) systems [374]. Relying on learning algorithms, BCIs allow controlling the behavior of external devices, such as computers or exoskeletons [48, 126].

There has been more than a decade of research on human-robot interaction (see [146] for a review), mostly in the fields of robotics and psychology, during which great importance has been assigned to the question of how to make interaction with robots most intuitive and *natural* for the human user [89]. Literature on human-robot interaction identifies, among others, two important issues to be addressed in order to enhance a natural user experience: One question is how to enable robots to *read* human signals, both for control [245] and to detect errors in their own performance [218]. The second point is to assess the influence of a robot's appearance and/or behavior on the user's cooperation and feelings towards the robot [25].

This chapter addresses both of these issues from the perspective of neuroscience, which is so far only weakly represented in the research on human-robot interaction. On the one hand, the idea was to investigate which brain signals can be detected and thus be *read* by a robot (aided with machine learning techniques) to optimize robot behavior. On the other hand, the influence of the robot's visual similarity to a human on such error-related brain activity was investigated. For this purpose, two experiments were conducted, in which participants watched different kinds of robots perform correct and erroneous actions. This chapter lays the foundation for further work in this thesis. Hence, it investigates and verifies whether the decoding of error-related brain signals is possible in principle, especially with regard to collaborative human-robot interaction, and also present the current situation in the literature.

3.1 System and Experimental Design

The series of experiments used in this chapter was designed in such a way that participants more or less passively fulfilled a task and only had to react for attention tasks. The aim was to measure signals caused by observation of faulty execution. In both experiments, participants observed a set of short videos. The videos were presented repeatedly in randomized order.

3.1.1 Observation Tasks

Pouring Observation Task (POT)

In the *Pouring Observation Task* (POT), participants were shown videos of a robotic arm (LBR iiwa, KUKA Roboter GmbH, Augsburg, Germany) pouring orange juice from a non-transparent container into a cup, see Fig. 3.1A. There were two classes of videos: The juice was either correctly poured into the cup, or incorrectly spilled over the table. Movement of the robotic arm and position of the cup were the same for all videos and conditions. Different outcomes were accomplished by varying the amount of juice in the container. The participants were thus unable to predict the outcome of the pouring action before it started. There were ten different video stimuli (five correct, five incorrect) of a $7.6\,s$ length, with a frame rate of $30\,fps$ (frames per second). The orange juice became first visible between $2.6\,s$ and $3.23\,s$ after the start of the video.

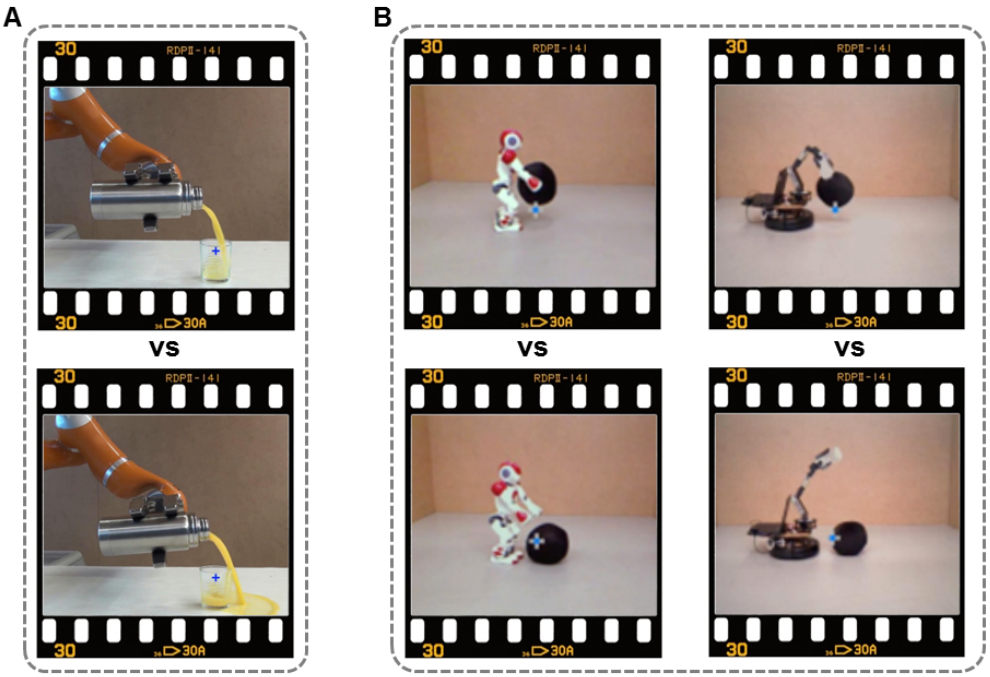

Figure 3.1: Visual stimuli, showing a correct and an incorrect condition. A In the first set a robotic arm performed a pouring task, either hitting or missing the vessel. **B** In a second set either a humanoid robots (NAO) or a non-humanoid robot (NoHu) performed a grasping task, either managing or failing to lift a ball from the ground. *Slide mount by pixelio*.

Lifting Observation Task (LOT)

Two different robots, either a small humanoid (NAO - Aldebaran Robotics, Paris, France) or a non-humanoid (custom-built, referred to as NoHu), approached a ball and tried to grab and lift it. In non-erroneous trials, the robots managed to grab and lift the ball and in

erroneous trials, they failed to do so (Fig. 3.1B). The videos were invariant concerning starting position of the robot, initial position of the ball, and visual properties of the surrounding. To generate clips showing the robots approaching from left and right, the existing ones were vertically flipped. There were 40 different video stimuli each of a $7\,s$ duration, also with a frame rate of $30\,fps$ (ten for each of the four conditions NAO - correct, NoHu - correct, NAO - incorrect, NoHu – incorrect; 5 of each set of 10 videos with the robot approaching from the left and 5 from the right, respectively). This experiment will be referred to as *Lifting Observation Task* (LOT).

General Paradigm

Before the video stimuli were presented, participants fixed their gaze on a white fixation cross on gray background for baseline recording ($3\,s$ in the POT, $2\,s$ in the LOT). Then, the video was initiated ($7.6\,s$ in the POT, $7\,s$ in the LOT). In the lifting observation task, $1\,s$ of post-baseline activity was recorded to exclude preceding motor artifacts generated by answering the attention task (see Fig. 3.2). Up to $10\,s$ of time between trials followed, allowing the participants to move, blink, swallow and answer a simple attention control question (Action correct? – Yes/No). The fixation cross was shown on top of the video display in the area of the main events of interest: For the pouring observation task, this was the area of the cup in the lower right part of the screen and for the lifting observation task, it was the initial place of the ball in the center of the screen. The control question was displayed at the same position. It was answered by pressing a key on a keypad-controller. Respective keys for answers ”yes” and ”no” were switched every 40 trials. After self-paced answering of the control question, the subsequent trial was initiated. The experiment was conducted in sessions of 30 trials in the pouring observation task and of 40 trials in lifting observation task, respectively. Trigger pulses containing an unambiguous ID were generated with the onset of video presentation and recorded via the EEG amplifiers. An additional optical trigger for post-hoc reconstruction in combination with a photo diode was embedded in the video.

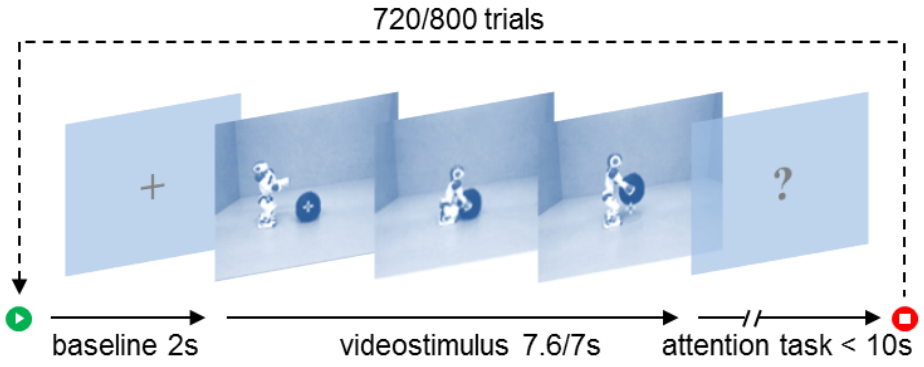

Figure 3.2: Timing structure of the experiments. Each trial consisted of a $2\,s$ fixation period, video stimulus of $\sim 7\,s$ and an attention control task. Altogether $\geq 720/800$ trials per participant.

In the pouring observation task, at least 720 trials per participant (360 trials per condition, 72 trials per video) were recorded. In participants 1 to 3, the number of trials per class was unbalanced (60% correct to 40% incorrect). Therefore, a weighing-mechanism was included in the analysis (see below). In the lifting observation task, at least 800 trials per participant (200 trials per condition, 20 trials per video) were recorded. For each participant, the experiment including preparation and pauses lasted about between 5 to 6 hours for the POT and between 5 to 7 hours for the LOT.

3.1.2 Participants

The participants were included in the study upon their informed written consent. All were healthy adults, either students or PhD students (22 to 31 years old). In the first experiment (POT), 6 participants (6 male) took part; one participant was excluded due to insufficient number of trials. In the second experiment (LOT), a total of 12 participants took part (6 female); one participant was excluded due to insufficient number of trials. The study was approved by the Ethics Committee of the University Medical Center Freiburg.

3.1.3 Data Acquisition

Experiments were conducted in an electromagnetically shielded cabin ("mrShield" – CFW Trading Ltd., Heiden, Switzerland). All electric devices in the cabin were powered by DC batteries. Information between inside and outside of the cabin was exchanged only via fiber-optic cables. High-precision EEG amplifiers with a 24-bit digital resolution and low noise (NeurOne - Mega Electronics Ltd., Kuopio, Finland) were used to record EEG from 128 scalp positions according to the *five percent* electrode-layout (Waveguard 128 - ANT Neuro, Netherlands). The gel-filled electrodes were prepared to reach impedances below $5\,k\Omega$ if possible. Sampling rate was $5\,kHz$; electrode Cz was used as reference, the ground was located between AFz- and Fz-position.

Besides, electrooculograms (EOGs), electrocardiograms (ECGs) and electromyograms (EMGs) of arms and legs of the participants were recorded and additionally an infra-red eye-tracker to monitor eye movements (EyeLink 1000+ - SR Research Ltd., Canada) was used. Eye-tracking data and EOG was used to inspect whether participants looked at the stimuli. EMG was to verify that participants remained still, though ECG recordings were only used for analyses that do not contribute to this thesis.

3.2 Pre-Processing, Classifier Design and Statistics

Data was down-sampled to $500\,Hz$, and then high-pass filtered with a cut-off frequency of $0.5\,Hz$ using a stable 4^{th} order Butterworth filter. Noisy channels were determined by visual inspection first and post hoc by using an automatic cleaning algorithm optimized to detect muscle-artifacts based upon the variance of the signal (BBCI-Toolbox [44]). To identify noisy trials, data were analyzed in intervals corresponding to decoding intervals

plus a preceding $500\,ms$ and the trials were rejected if the difference between the maximum and minimum value exceeded $600\,\mu V$. Rejected trials were only excluded from the training sets but kept in the test sets of the cross-validation (see below). Then, common-average re-referencing was performed and trials were cut according to the decoding intervals described below.

A Filter Bank Common Spatial Pattern (FBCSP) algorithm was implemented as described in Subsec. 2.5.2 (see Fig. 2.9 and Alg. 3). The data was bandpass-filtered in 35 non-overlaying frequency bands between $0.5\,Hz$ and $144\,Hz$. Between $0.5\,Hz$ and $30\,Hz$ a filter with a bandwidth of $2\,Hz$ was applied and between $30\,Hz$ and $144\,Hz$ with a bandwidth of $6\,Hz$, since band power modulations in low frequencies typically occur in narrower bands than in high frequencies Buzsáki and Draguhn [66]. CSP analysis was then performed on each of these frequency bands in a 10-fold cross-validation: The feature selection was set to choose the first 3 and the last 3 filters ordered according to their eigenvalues, i.e. the most discriminative six filters (see [43] for more details on this heuristic), to maximize between-class variance in the training set. These spatial filters were then applied on the trial data. The logarithm of the variance of the resulting signal was used as features. Then, as a first step, rLDA classifiers (see Subsec. 2.5.1), were trained on the training features and evaluated on the test features, which resulted in frequency-resolved decoding accuracies. In a second step, the stored features from either all frequency-bands or two different subsets (all frequency bands below $20\,Hz$, and all frequency bands above $60\,Hz$) were taken together to train FBCSP classifiers (in a 10-fold cross-validation analogous to the above). To account for unbalanced numbers of trials in the different classes, the mean over decoding accuracies per class was used instead of the overall decoding accuracy.

For the first experiment (POT), binary FBCSP for the classes correct vs. incorrect was performed in the decoding intervals $0 - 7.6\,s$ (full interval) and $3.3 - 7.5\,s$ (late interval) relative to the start of the video display. In addition, the interval from -0.5 to $3\,s$ relative to the point in time when the liquid first became visible was extracted. This interval (intermediate interval) differed depending on the video displayed and accounted for the fact that the frame where liquid first become visible varied among the stimuli; liquid in incorrect trials appeared between 0.4 to $0.6\,s$ earlier than in correct trials. For the second paradigm (LOT), binary FBCSP for the classes NAO vs. NoHu as well as for the classes correct vs. incorrect was performed in the intervals $0 - 7\,s$ (full interval), $5.1 - 6.9\,s$ (late interval) and $4 - 7\,s$ (intermediate interval), relative to the start of the video display, to cover the different phases of the stimuli. P-values for FBCSP decoding accuracies for each participant were estimated by a permutation test (Alg. 7), as described in Sec. 2.6.

3.3 FBCSP Filters and Activation Patterns

Fig. 3.3 exemplarily shows both filter and corresponding activation patterns calculated for CSP decoding in the pouring observation task (Fig. 3.3A), the error condition in the lifting observation task (Fig. 3.3B) and the robot condition in the lifting observation task (Fig. 3.3C). The visualization shows the filters and patterns of participant 1 of the pouring

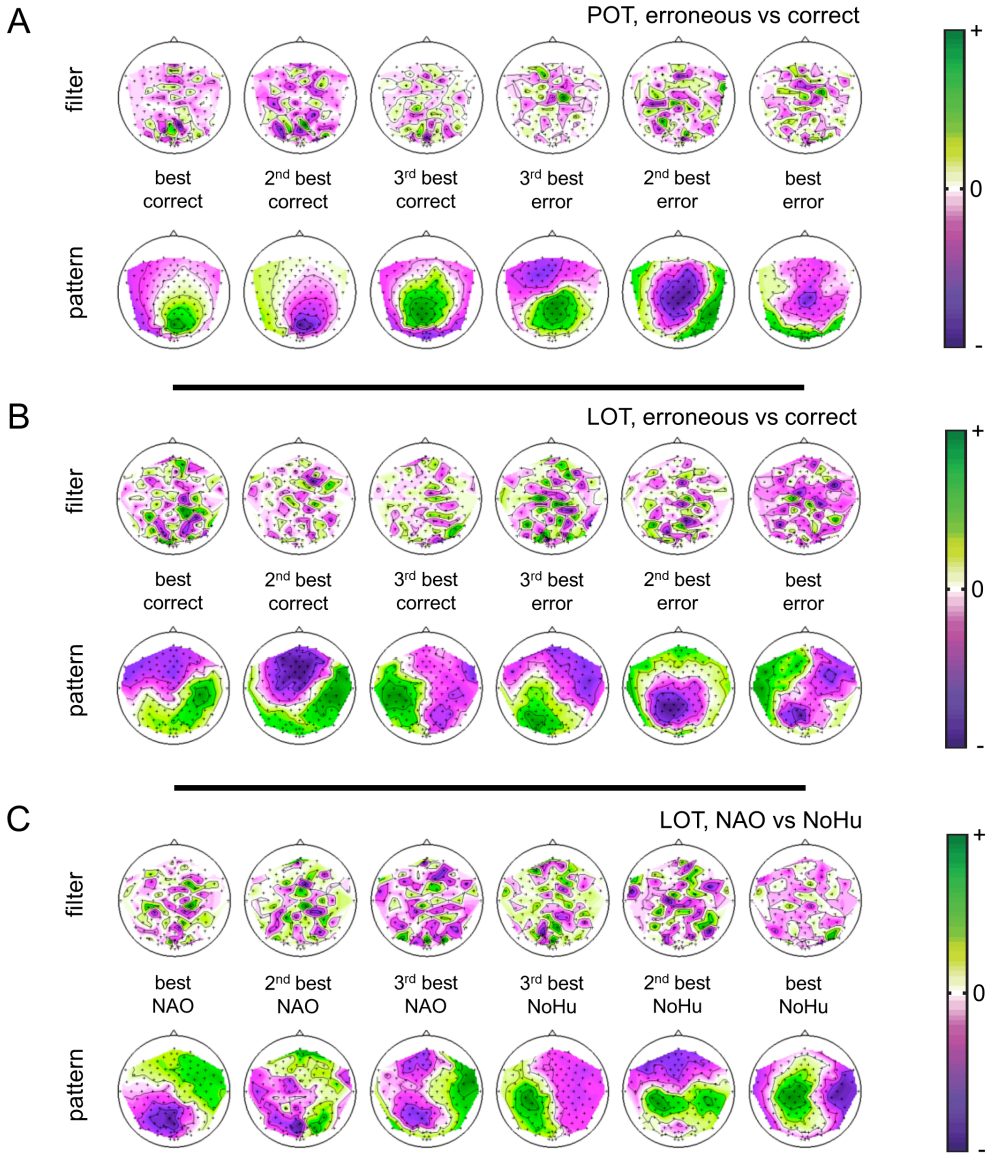

Figure 3.3: Exemplary CSP-filters and activation patterns. A Error condition in the pouring observation task, **B** the error condition in the lifting observation task, and **C** the robot-type condition in the lifting observation task.

observation task and of participant 2 of the lifting observation task, which reached the highest decoding accuracies among all frequency bands below $20\,Hz$. Here, filters depict the vectors $\mathbf{v}_j$ of channels, represented as a columns of the mixing matrix $\mathbf{V} \in \mathbb{R}^{e \times e}$ and e being the number of channels, that perfor the linear transformation from the electrode onto the surrogate electrode space (for details see Subsec. 2.5.2). This projection ensures that the signals are distinguished in the best possible way with regard to variance, see Eq. (2.37). The activation patterns visualize the vectors $\mathbf{a}_j$, which represent the column vectors of the demixing matrix $\mathbf{A} = (\mathbf{V}^{-1})^\top \in \mathbb{R}^{e \times e}$. The visualization makes clear

how the presumed signal sources project onto the head surface, which, according to theory, serves to validate neurophysiological associations and underlying processes. In Fig. 3.3, the sign of the vectors is irrelevant and simply illustrates the contrast between the behavior of the filter with respect to a channel and the contrast in the activation patterns, respectively.

The filters show a rather disorganized distribution over the scalp, but resulting in clear activation patterns. For the selected participants, the depicted activation patterns suggest bipolar, sometimes multipolar, generators. The generators are separated differently by sagittal and coronary axes as well as by transverse axes. Besides, within a condition some of the activation patterns practically seem to be quite the negative of others.

3.4 Decoding Errors and Robot Type

To obtain pretty much the distribution of the information content over the whole frequency range, in a first step decoding was performed on 35 frequency bands using the CSP algorithm. For the pouring observation task, the results of the decoding after applying the filters is shown in Fig. 3.4.

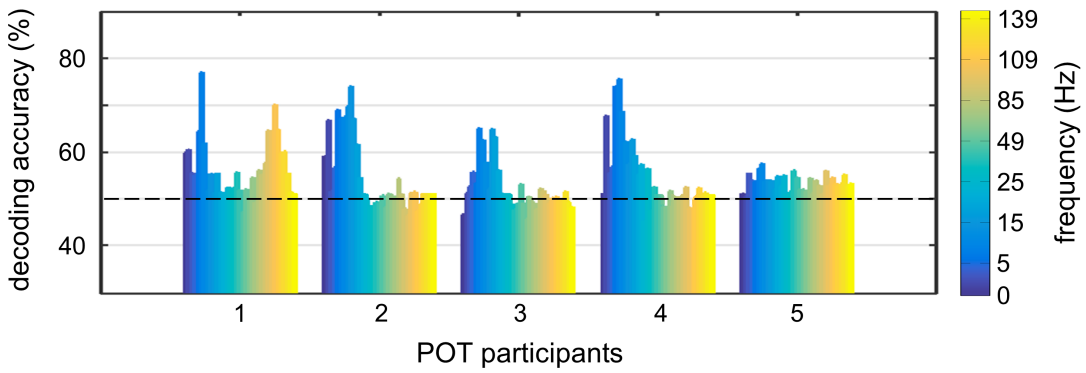

Figure 3.4: Frequency-resolved CSP-decoding results in the pouring observation task. Accuracies of 35 frequency bands in the range between $0.5 - 144\,Hz$ for 5 participants, using the decoding interval $3.3 - 7.5\,s$ relative to video stimulus onset.

Fig. 3.4 shows mean CSP decoding accuracies in the pouring observation task of all 35 frequency bands in the range between $0.5 - 144\,Hz$ for participants 1 to 5 (decoding interval $3.3 - 7.5\,s$ relative to video stimulus onset). Decoding accuracies were mainly above chance, and especially in participants 1 to 4 the accuracies were generally higher for frequency ranges below $20\,Hz$. This trend with respect to frequency ranges was also found in the other decoding intervals and somewhat weaker for the lifting observation task.Participant 1 also showed above-average decoding accuracies in frequency bands beyond $60\,Hz$. Maximal decoding accuracies reached up to around $75\,\%$ in participants 1 and 5. To yield high efficiencies be selecting an optimal operational frequency band, the

FBCSP algorithm was used. Fig. 3.5 compares decoding accuracies of the FBCSP implementaion for different frequency ranges, broadband from 0.5 to $144\,Hz$, low frequencies from $0.5 - 19\,Hz$ and high-gamma from $61 - 144\,Hz$.

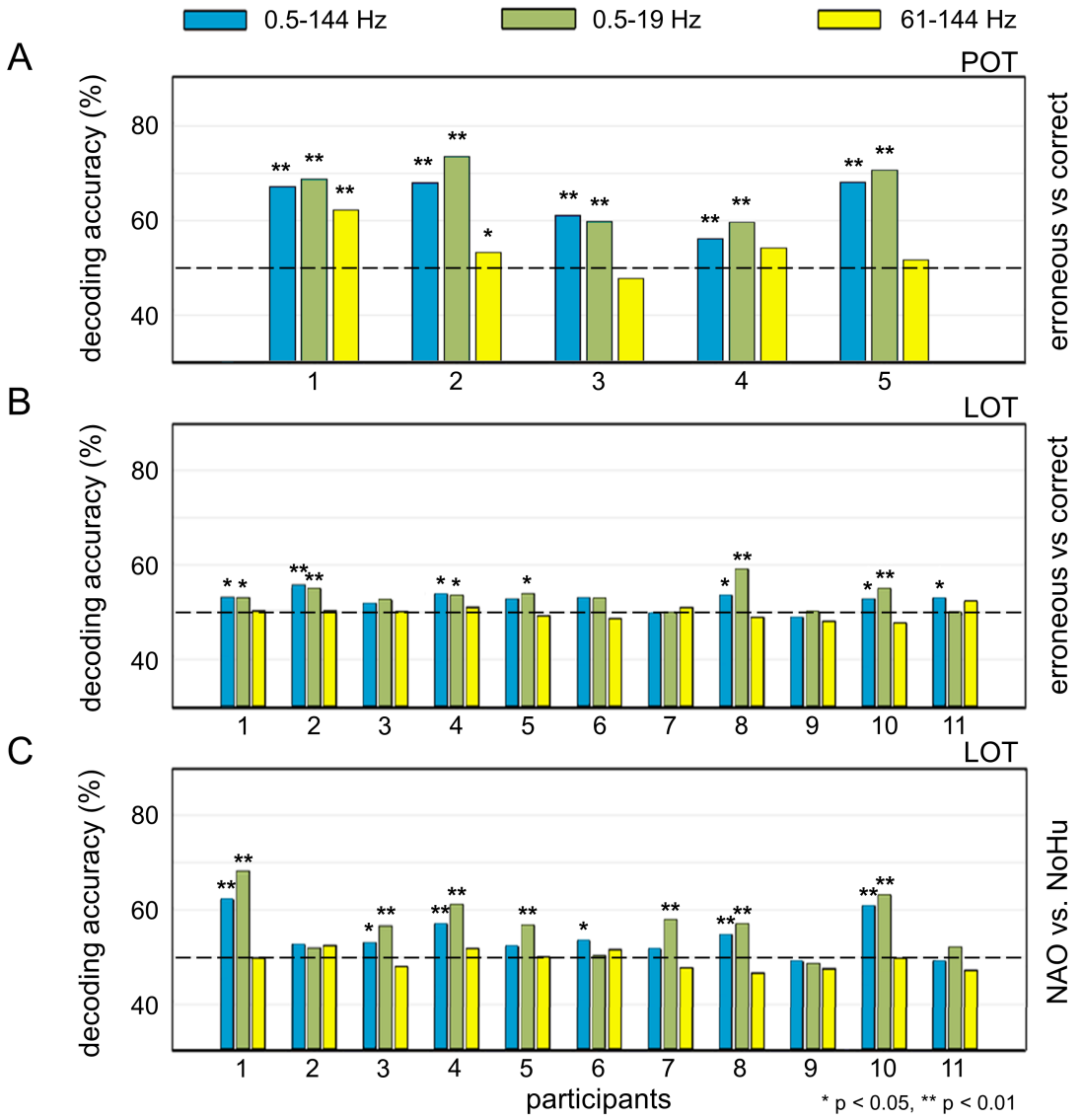

Figure 3.5: FBCSP decoding results for three different frequency ranges. A Pouring observation task (POT) for the interval $3.3 - 7.5\,s$. **B** Error condition of the lifting observation task (LOT) for the interval $4 - 7\,s$. **C** Robot condition of the lifting observation task (LOT) for the interval: $0 - 7\,s$. Significance is indicated by asterisks, * $p < 0.05$ ** $p < 0.01$.

Fig. 3.5A shows the results for participants 1 to 5, decoding errors in the pouring observation task. The presented results originate from the decoding interval which yielded the highest mean decoding accuracies, in this case represented by the interval $3.3 - 7.5\,s$. The results show mainly significant decoding accuracies clearly above chance level. Furthermore, decoding on the high-gamma frequencies obviously can not compete with

performances reached by decoding with broadband and low-frequency ranges, the low-frequency band performs best overall. Fig. 3.5B and C compare FBCSP classifiers over the different frequency ranges for participants 1 to 11 for the lifting observation task. Again, results originating from the decoding intervals which yielded the highest mean decoding accuracies are displayed. For the error condition the interval $4-7\,s$ was selected and for the robot condition the interval $0-7\,s$. Here, the performances are not that good as for the pouring observation task. For the error condition accuracies exceed chance level, partly significant, and show values up to $\sim 60\,\%$. The decoding on the robot type slightly performs better, yielding accuracies up to $\sim 70\,\%$.

Mean decoding accuracies over all participants for the different frequency ranges and decoding intervals can be found in Tab. 3.1. For the different paradigms and condition, the upper part of the table contains decoding accuracies extracted from the best decoding interval in each case (POT-error: late interval $3.3-7.5\,s$; LOT-error: intermediate interval $4-7\,s$; LOT-robot: full interval $0-7\,s$). It becomes clear that decoding happens to be most efficient if only low frequencies enter the FBCSP algorithm. The lower part of Tab. 3.1 shows the decoding results of FBCSP for frequencies $< 20\,Hz$. In this case, no clear rule seems to be found and the performance of the respective intervals for the recognition of the conditions varies individually.

Table 3.1: Mean FBCSP accuracies for different frequency ranges using the best performing interval (top) and mean FBCSP accuracies for different decoding intervals for frequencies $< 20\,Hz$ (bottom)

	POT error	**LOT error**	**LOT robot**
0.5-144 Hz	(60.2 ± 5.3) %	(52.6 ± 1.9) %	(54.2 ± 4.3) %
$<$ 20 Hz	(62.1 ± 5.7) %	(53.2 ± 2.7) %	(56.7 ± 5.8) %
$>$ 60 Hz	(54.0 ± 5.3) %	(49.6 ± 1.4) %	(49.4 ± 2.0) %
full interval	(55.2 ± 3.6) %	(51.1 ± 3.2) %	(56.7 ± 5.8) %
late interval	(62.1 ± 5.7) %	(53.0 ± 2.5) %	(53.4 ± 3.2) %
intermediate interval	(58.0 ± 5.3) %	(53.2 ± 2.7) %	(56.4 ± 4.3) %

3.5 Discussion

The findings show that both conditions investigated in the present study were decodable from the recorded EEG signals: observation of erroneous vs. correct robot actions in the POT and, although at very low accuracies and not in all participants, also in the LOT, as well as humanoid vs. non-humanoid robot type in the LOT. FBCSP decoding accuracies varied in a range from around chance level to around $70\,\%$ (Fig. 3.5), Tab. 3.1). The fact that erroneous vs. correct robot action can be decoded from human brain activity is in line with prior findings by [176]. Taken together, the results indicate that observation of erroneous vs. correct robot action and observed robot type are encoded in human brain

activity and that related brain signals can be detected, at least in some circumstances, in the non-invasively recorded EEG.

In contrast to promisingly high error decoding accuracies in the POT, the accuracies for the broadband and low-frequency components of the error condition of the LOT were mostly little above chance level (Fig. 3.5). This renders a generalizing of the results difficult and suggests that an unknown factor may have played a role. For example, one possible explanation could be that errors in the POT were indirectly decoded by differences in the visual properties of the stimuli during correct and incorrect trials, which were more prominent in the POT. However, differences in visual properties were also present in the LOT. As a matter of fact, they are an inevitable consequence of errors in everyday settings as investigated in the present study. Another explanation could be that different error types may elicit different affective responses, which in this case were possibly more pronounced in the POT (liquid spilling) than in the LOT (unsuccessful ball lifting). Follow-up studies will be necessary to further investigate these observations and generally the factors modulating robot-error recognition by humans as well as the underlying neurophysiology.

Besides, the results indicate an effect of the EEG frequencies used for decoding. FBCSP taking into account only frequencies below $20\,Hz$ yielded generally higher decoding accuracies than broadband or high-frequency components (Tab. 3.1). This is in line with previous studies which suggested an involvement of the mirror neuron system (MNS, [291]) and motor system, manifested in mu- and beta-band modulations [194, 252], as well as frequency power modulations in response to erroneous action execution mainly found in lower frequency bands such as delta, theta, alpha and beta [69, 199, 347, 372]. High frequency-components in the gamma range were previously also proposed to be related to error processing [69, 350]. Yet, in this study, FBCSP using frequency components of the high-gamma range ($> 60\,Hz$) yielded comparatively low decoding accuracies. This effect, however, seemed to be participant-dependent (see Fig. 3.4), reminiscent of inter-participant variability of movement-related high-gamma EEG responses as they have previously been observed [21, 93]. Further investigations are necessary to elucidate the role of different EEG frequency bands, including gamma, and of their dynamics in the context of robot-error observation.

The results also indicate a possible effect of the time intervals used for decoding. Decoding for correctness in both experiments, intervals starting after the stimulus condition had become evident to the observer (late and intermediate intervals) appeared to yield the best results (Tab. 3.1). This suggests that the timespan preceding the error, which was designed to contain minimal visual differences, was indeed uninformative for decoding. In contrast, decoding for the robot type in the LOT, the full video interval yielded the best overall decoding results, while the shorter intervals resulted in lower accuracies, consistent with the fact that the robot type difference was present throughout the trials.

Robots were investigated during naturalistic tasks, namely liquid pouring and object grabbing. These tasks were designed to approximate application scenarios of autonomous or semi-autonomous robots under high-level control or surveillance via brain-computer interfacing. An important consequence of these naturalistic conditions is that the exact time point of error events is less clearly defined than in other experimental paradigms that

have previously been used to elicit error-related brain responses (e.g., in forced choice visual discrimination tasks, such as the Erikson flanker task) or to investigate low-level BCI control. Some of these paradigms are able to yield much higher decoding accuracies than observed here (up to $91\,\%$ for single-trial classification; see [261]). Yet, with respect to future prospects of robotics and shared-control BCI applications, unpredictable, asynchronous errors are an important, complementary topic. Another, related perspective would be the detection and evaluation of action consequences that are not *objectively* right or wrong, but rather depend on the intention of the user (for example, picking up an apple vs. picking up a piece of chocolate).

The achieved accuracies in all experiments are not as high as would be required for improving BCI-applications. *Chavarriaga et al.* [74] suggest $80\,\%$ accuracy as a benchmark for decoding of error-related potentials that has been shown to be sufficient to improve information transfer rate in most BCI applications. Improvement of decoding accuracy in the experimental paradigm could be reached by resorting to other types of electrophysiological recordings: Even though a highly optimized EEG setup was used, intracranial recordings can be expected to provide more reliable error-related signals if recorded from informative brain areas. Despite the low accuracies reached here, non-invasive studies as in the present study combined with source localization approaches may be useful for guiding such studies and selecting the most promising target areas involved in the perception and cognitive evaluation of robot action.

The future will bring many major developments in the fields of robotics and shared-control BCIs: With growing relevance of robots in everyday life, new types of interaction between humans and machines will evolve. The current practice of surveying robots (e.g., [89]) could profit by elaborated, empirically supported theories. In this context, it may be fruitful to join efforts across scientific disciplines and, for example, include concepts and theories from philosophical action theory into empirical studies, as previously proposed by [331]. Action theory conceptualizes human agency by analyzing agency-related phenomena like intention and planning (for an overview see [254]). A central topic of the emerging interdisciplinary field of *action science* [277] is the integration of philosophical concepts with related empirical findings, e.g., from the field of neuroscience [64, 151, 256], into a comprehensive theory. In the past years, so-called *shared agency* among cooperating human agents [55, 256, 345, 346] became a focus of interdisciplinary studies in this area [362] and very recent experimental work investigated shared agency during human-robot cooperation [120, 163, 323] as well as the agentive properties of robots in general [188]. Such collaboration across disciplines has already proven successful in the field of action theory, where the application of concepts from philosophy to the field of intelligent systems has led to the development of belief-desire-intention architectures [135].

3.6 Related Work

Error-Related Brain Responses

Prior research found that human brain activity is modulated by both performed and observed erroneous action in other humans. When humans observe other humans committing errors or when they err themselves, their brains show a specific activation pattern in response to these errors (see [371] for a review).

With respect to the time domain of human electroencephalography (EEG) signals, there are well-documented event-related potential components (ERPs), which are linked to the processing of errors (mainly investigated in erroneous motor-execution): the error-related negativity (ERN), consisting of a negative deflection (Ne) and sometimes followed by a positive deflection (Pe). While the Pe seems to appear exclusively in conscious error processing [248], the Ne can be measured when participants do not report to have committed an error and a small negativity often even appears in correct trials [347]. Ne and Pe do not share the same scalp distribution: The Ne is maximal over frontocentral areas while the Pe is usually recorded with a parietal maximum [114].

In the frequency domain of EEG signals, several studies demonstrated frequency-specific power modulations in response to erroneous action-execution in different motor tasks: Effects were mainly found in lower frequency bands, such as delta ($< 4\,Hz$), theta ($4-8\,Hz$) [199, 372], alpha ($8-12\,Hz$) [69], and beta ($12-30\,Hz$) bands [347]. *Carp and Compton* [69] also suggested error-related spectral power changes in frequencies higher than $30\,Hz$. In recent EEG studies, our group found first evidence for modulations in the high-gamma range ($50 - 150\,Hz$) related to erroneous execution of the ERIKSEN flanker motor task [350, 351], a response inhibition test. *Koelewijn et al.* [194] demonstrated an effect of the correctness of observed human actions on beta power-modulation over sensorimotor areas: This effect, however, was weaker in the observation settings than in a corresponding execution task.

In the context of BCIs for directly controlled prostheses, *Milekovic et al.* [232] used electrocorticographic recordings obtained while participants where engaged in a simple videogame, for which they controlled a cursor with an analogue joystick. The experimenters were able to detect execution errors (i.e. when motor commands resulted in an unexpected movement) and outcome errors (i.e. when participants failed to reach the intended goal) from the neural activity in real-time significantly above chance level.

[176] used EEG-measured error-related potentials in order to teach neuroprosthetics suitable behaviors in scenarios of varying complexity. In three experiments, participants were asked to monitor a device as it tried to reach a goal, that only the participant was aware of, and assess whether the actions of the device were incorrect or correct (i.e., whether the actions brought the device closer to the intended goal or further away). In a first experiment, participants observed a cursor on a screen as it moved either right or left in order to reach the target. In a second experiment, a simulated virtual robotic arm, that could perform four different actions (moving left, right, up or down) to reach the target, was displayed on a screen. For a third experiment, the simulated robotic arm was replaced by a real robotic arm. All experiments were divided into a training phase, during which the

classifier that should detect the error-related potentials was built, and an online operation phase, during which the decoded information on correctness of the device's action was used as a reward for a reinforcement learning algorithm. *Iturrate et al.* [176] were able to show that in 11 out of 12 participants, classification performance was significantly above chance level and that the user-controlled device reached the goal significantly more often as compared to a device following a random control policy. These findings demonstrate that error-related potentials are an adequate reward signal for reinforcement learning algorithms with the purpose of neuroprosthetics control.

Following the same line of research, [299] investigated the role of EEG-measured error-related potentials for robot control during an object selection task in four conditions. In the online closed-loop condition, participants observed the robot perform binary object selection. If the EEG classifier detected an error-related potential, the robot's behavior was corrected, which in turn was immediately observed by the participant. In the offline closed-loop condition, the EEG classifier was trained using the data from all closed-loop trials of each participant. In the offline open-loop condition, participants observed the robot perform object selection correctly or incorrectly and no feedback was given to the robot. In the fourth condition, secondary errors, which occur in response to real-time misclassification (i.e., when the EEG signals are misclassified leading to incorrect robot behavior), were additionally considered in the online closed-loop condition. Performance in the online closed-loop condition was around chance level and on average above in the offline closed-loop and the offline open-loop condition. Interestingly, [299] found that taking into consideration secondary errors improved performance significantly.

Brain Responses During Observation of Correct Human and Robot Actions

One of the most striking findings in recent neuroscience was the discovery of mirror neurons: [292] observed that certain neurons in the macaque brain fire both when the monkey performs an action and when it observes the same action performed by an experimenter. These cells were termed mirror neurons. The mirror neurons distributed across various brain regions together form the MNS. Findings from neurophysiological and brain-imaging studies indicate that a MNS also exists in the human brain, possibly even spatially more extended than in monkeys, and that the MNS is reliably activated when humans observe other humans perform meaningful actions [239, 291].

Until now, there are very few published neurophysiological experiments on the perception of robotic action by a human observer. If it was processed similar to human movement, the human MNS should be involved during the observation of robot action. An EEG study by [252] suggests that the human MNS is indeed not selective for biological movement but can also be activated by robotic movement: Observation of a grasping action (target-directed and non-target-directed) performed by a robotic arm lead to suppression of mu-band activity ($8 - 13\,Hz$) in left and right sensorimotor cortex (scalp positions C3 and C4), which has been linked to MNS activity. The study also found a significant hemispherical effect, as mu-band suppression was stronger on electrode C3 (left) than on C4 (right). Mu-band suppression also occurred when observing human motion. There was no significant difference in the strength of mu-band suppression during

human vs. robot motion observation [252]. *Bates et al.* [28] and *Oberman et al.* [252] suggested that the MNS is involved in the error-observation-related brain responses.

Humanoid Robots

So far there is hardly any research that directly compares the perception of humanoid robots to the perception of non-humanoid robots, with respect to the underlying brain responses. Many previous studies focused on how humanoid robots are perceived by humans with respect to facial features. *Disalvo et al.* [97] found that the presence of certain features in a robot's face, such as eyes, nose and mouth, the dimensions of the robot's head as well as the total number of facial features play a key role for the perceived humanness of a robot. Assuming the findings on facial features of robots can be transferred to general body features of robots, it appears likely that the higher the number of individual humanoid body features of a robot, the more humanoid it is perceived as a whole.

Observation Task

Based upon the findings by [252] and [344], the hypothesis was stated that watching a robot perform erroneous compared to correct action differentially modulates the observer's brain activity: Given that watching other humans performing erroneous actions triggers an automatic cognitive evaluation reflected in an error-related brain response [344] and that observation of robot movement is processed similar to the observation of human movement [252], it can be assumed that watching robots commit errors also elicits error-related brain responses. Information about the perceived correctness of robot performance decoded from the EEG seems to provide a useful teaching signal for adaptive control algorithms in order to optimize robot control, in particular for so-called shared-control BCIs. Based on the results by [97], who were able to show that the number of humanlike features in a robot affects the perceived humanness of a robot, this chapter should give information about whether this degree of perceived humanness would also be reflected in the brain responses that occur when observing a robot perform an action (both for actions where the robot commits an error and where it did not).

3.7 Conclusion

In this chapter, two experiments were presented that were designed to investigate the possibility to decode erroneous action from robots performances and whether the number of humanlike features in a robot affects the perceived humanness of a robot. In a first experiment, participants watched a robot arm pour liquid from a non-transparent container into a cup. The robot performed the action either incorrectly or correctly, i.e. it either spilled some liquid or not. In a second experiment, a 2x2 factorial design was employed. Participants observed two different kinds of robots, a humanoid and a non-humanoid, grabbing and lifting a ball. Similar to the first experiment, each of the robots was either

successful at the action, i.e. managed to lift the ball, or not. The approach was to decode the correctness of observed robot actions from the EEG signals recorded during the two different passive observation tasks and for the second experiment, additionally, the aim was to decode the type of the observed robot from the EEG signal.

Decoding was implemented using the common spatial pattern (CSP) approach for feature extraction, as applied to multi-channel EEG data by *Müller-Gerking et al.* in 1999 for decoding motor tasks [240]. Regularized linear discriminant analysis (rLDA) was used a classifier on these features. Since the the original studies, CSP has continuously been adapted and become a standard method in EEG classification tasks, especially for motor behavior or motor imagery [43]. To validate the reliability of CSP decoding results, the spatial filters and corresponding activation patterns computed by the CSP algorithm were assessed.

As described above, previous studies have focused on brain responses during observation of correct and incorrect actions in humans but only few studies investigated the brain responses to watching robot action. So far, brain responses to perception of correct and incorrect robot performance and with different types of robots have never been assessed in previous experiments. Hence, the present study aimed to add to this field by investigating these aspects of robot observation.

The findings presented in the present chapter indicate that it is possible to decode the correctness of at least some kinds of observed robotic actions as well as the type of observed robot from non-invasively recorded human EEG. These findings add to relevant topics in the research on human-robot interaction, such as enabling robotic systems to *read* human signals or the influence of a robot's appearance and/or behavior on the user's perception of the robot.

Accessing error recognition in robot performance might be helpful for EEG-based BCIs; the observation tasks in this study were designed to approximate future application scenarios of autonomous or semi-autonomous robots under high-level control or surveillance via brain-computer interfacing (self-feeding, go-and-fetch tasks). There are several perspectives for follow-up investigations which derive from the present study, and which could be addressed with similar methods. Given that the achieved accuracies are likely not yet sufficient for practical applications, it would be helpful if alternative machine learning approaches such as artificial neural networks reached higher decoding accuracies. Another question to be addressed in the future would be which kind of robot errors are generally suitable for decoding of the user's perceived correctness and how they differ from non-decodable errors. Closely related to this, would be the investigation of how visual, affective, and movement-related brain systems are involved in the generation of the differential responses to robot action.

Chapter 4

Decoding and Visualization Using Deep Convolutional Neural Networks

The importance of robotic assistive devices grows in our work and everyday life. Cooperative scenarios involving both robots and humans require safe human-robot interaction. One important aspect here is the management of robot errors, including fast and accurate online robot-error detection and correction. Analysis of brain signals from a human interacting with a robot may help identifying robot errors, but accuracies of such analyses have still substantial space for improvement. This chapter evaluates whether a novel framework based on deep convolutional neural networks (deep CNNs) could improve the accuracy of decoding robot errors from the EEG of a human observer, both during an object lifting and a pouring task. It can be shown that deep CNNs reached significantly higher accuracies than both regularized Linear Discriminant Analysis (rLDA) and filter bank common spatial patterns (FBCSP) combined with rLDA (see Chap. 3), both widely used EEG classifiers. Deep CNNs reached mean accuracies of $(75 \pm 9)\,\%$*, rLDA* $(65 \pm 10)\,\%$ *and FBCSP + rLDA* $(63 \pm 6)\,\%$ *for decoding of erroneous versus correct trials. Visualization of the time-domain EEG features learned by the CNNs to decode errors revealed spatiotemporal patterns that reflected differences between the two experimental paradigms. Across participants, CNN decoding accuracies were significantly correlated with those obtained with rLDA, but not CSP, indicating that in the present context CNNs behaved more "rLDA-like" (but consistently better), while in a previous decoding study with another task but the same CNN architecture, it was found to behave more "CSP-like". The findings thus provide further support for the assumption that deep CNNs are a versatile addition to the existing toolbox of EEG decoding techniques, and the steps how CNN EEG decoding performance could be further optimized are discussed.*

Chap. 3 has shown that it is possible to detect errors when observing robots, based on human brain signals. However, the performances seem to be far from practical applicability. There are now several approaches to get this problem under control. For example, the so far used paradigms do not seem to give an exact time, a discrete *event*, for the occurrence of the error and it is therefore difficult to define an optimal time interval, since the perception of the error can also be very subjective. An experiment that gives time-discrete errors, such as the ERIKSEN flanker task [110], could help here.

Especially for analyses based on surface EEG, the improvement of any process is target-oriented and an optimization of the signal quality can lead to features having a clearer characteristic and thus also a better decoding effect. Another approach would be the use of intracranial measurement data. Here, the brain signal is decisively stronger and artifacts, caused by any kind of movement, hardly need to be considered. In addition, the measurement is carried out directly at the tissue, which makes it possible to measure the signals directly in the respective brain areas and keeps the potentially unwanted overlapping of several different signals to a minimum.

However, this chapter is about improving the decoding on the classifier side. The problem is addressed by applying deep learning to a naturalistic decoding task where participants observed a robot performing different assistive actions either successfully or failing to do so. In EEG research, architectures including deep convolutional neural networks (CNNs) have recently been used to explore their applicability in brain-signal decoding [303, 328], but not yet to robot-error decoding from EEG.

4.1 System and Experimental Design

As a follow-up to the previous chapter, the analyses in this study are based on the same data set. Just to mention the fundamental aspects, in both paradigms participants had to observe robot fulfilling instructed tasks, either managing or failing their mission. The visual stimulus was presented in form of short video clips, which were repeatedly played in a randomized order. The videos for different conditions were as invariant as possible concerning starting position of the robot, initial position of the ball/glass, robot movement and visual properties of the surrounding. To the participants the task was more or less passive, as they only had to observe the execution by the robot, only acting actively when it came to the attention tasks. In this chapter, the two paradigms will be referred to as follows:

- Pouring Observation Task (POT), see Fig. 3.1A
- Lifting Observation Task (LOT), see Fig. 3.1B

Further setup details are described in Sec. 3.1, while the timing structure of the experiments can be seen in Fig. 3.2. Sec. 3.1 also provides information about the participants and the EEG acquisition. In the LOT an extra condition was implemented additionally, where in 10% of the trials the participants were instructed to press a button if and exactly when they perceived a robot error. This allowed an estimation of the approximate time point of error perception.

4.2 Pre-processing, Classifier Design and Statistics

Since this chapter targets the comparison of different algorithms, an equal mutual pre-processing would be desirable. However, it would make no sense to change the working

concepts of the FBCSP implementation. The CNN is practically designed to learn a broad spectrum of features and generally is intended for the use in real-life applications, what justifies the exclusion of some pre-processing steps. Moreover, this change rather tends to make the CNN worse off. The recorded EEG data were re-referenced to a common average (CAR) and resampled to $250\,Hz$. To compute exponential moving means and variances for the CNNs, an electrode-wise exponential moving standardization with a decay factor of 0.999 was applied [303], while the rLDA implementation reached higher accuracies without the standardization. Based on predefined decoding intervals, the data was cut into trials according to the stimulus onset. Data analyses employing rLDA (see Subsec. 2.5.1 and Alg. 2) and deep CNNs (see Subsec. 2.5.3 and Fig. 2.14) are based on python implementations. Pre-processing and implementation of the FBCSP algorithm is discussed in detail in Subsec. 2.5.2, but the cleaning was slightly improved in comparison to Chap. 3. According to the results in Chap. 3 only frequencies $< 20\,Hz$ contributed in decoding the errors in the FBCSP implementation. All underlying architectures are described in Sec. 2.5, training and classification was only done within each participant. Thereby, the architecture of the rLDA classifier complies with the theory of [124] and is leant on the realization of [44], for the shrinkage regularization the *LedoitWolf* estimator [210] was used, see Eq. (2.66) in Subsec. 2.5.5.

Significance for individual decoding results was estimated using a permutation test, see Alg. 7. Mean differences of accuracies between decoding methods were evaluated by WILCOXON signed-rank tests. Significance of correlation coefficients was evaluated by randomizing the order of one of the input vectors of the correlation. The number of guesses that resulted in higher coefficients than the true correlation coefficient was compared to the total number of guesses.

4.3 Comparison of Decoding Performance

In this chapter, three different decoding algorithms (CNNs, rLDA and a combination of FBCSP and rLDA) were implemented and the outcome of the error decoding was compared. The decoding intervals $3.3 - 7.5\,s$ (POT) and $4 - 7\,s$ (LOT) were selected according to the results in chapter Chap. 3. Additionally, for the POT the data between $2.5 - 5\,s$ was analyzed, since this seemed as an intuitive interval in which the error became obvious. As described before, in the LOT paradigm an extra condition was integrated to serve as an error perception feedback. If an error occurred, the participants had to press a button as soon as they became aware of it. This extra condition comprised both correct and erroneous trails, and was only used for the analysis of the error perception feedback, not for later analyses. In average, the participants became aware of the error at around $(5.4 \pm 0.5)\,s$. The interval was then selected as the 5-fractile range of the button press moments of all participants, resulting in the decoding interval $4.8 - 6.3\,s$, see Fig. 4.1.

Tab. 4.1 shows the mean decoding accuracies of the error classification for the different time intervals and the respective paradigms and for all classification methods. Comparing the performances of the different methods, apparently one difference is striking: for both paradigm and time intervals, the CNNs clearly yielded the highest mean decoding

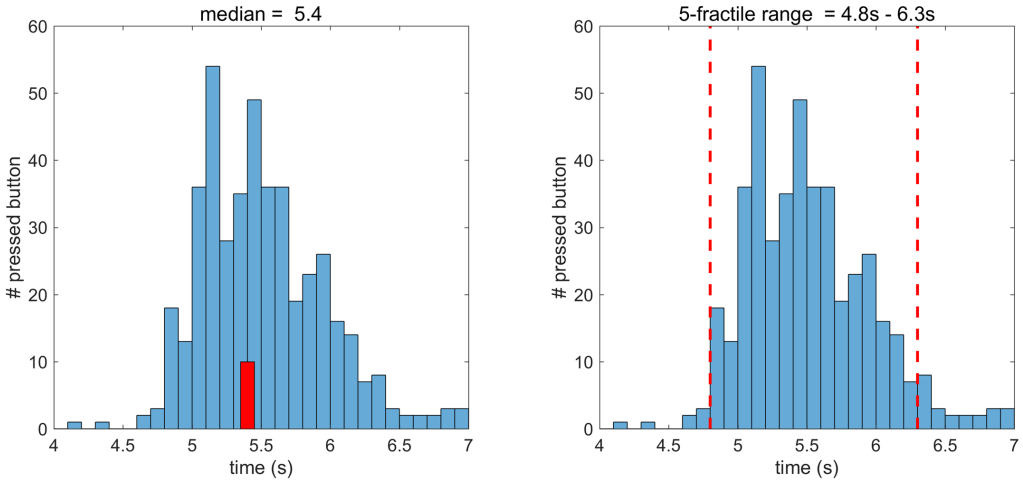

Figure 4.1: Analysis of participants button press according to error appearance. (Left) Evaluation of average moment of error awareness of around $(5.4 \pm 0.5)\ s$. (Right) 5-fractile range of the overall button press time for all participants.

accuracies and beats both of the other implementations. Also the rLDA manages to perform in three of the four decoding interval better than the FBCSP.

Table 4.1: Comparison of mean decoding accuracies

paradigm	interval	CNN	rLDA	FBCSP
POT	2.5-5s	(78.2 ± 8.4) %	(67.5 ± 8.5) %	(60.1 ± 3.7) %
POT	3.3-7.5s	(71.9 ± 7.6) %	(63.0 ± 9.3) %	(67.1 ± 5.4) %
LOT	4.8-6.3s	(59.6 ± 6.4) %	(58.1 ± 6.6) %	(52.4 ± 2.8) %
LOT	4-7s	(64.6 ± 6.1) %	(58.5 ± 8.2) %	(53.1 ± 2.5) %

Fig. 4.2 shows the pairwise comparison of the error decoding accuracies obtained in the individual participants for all three classification methods. Here, one panel comprises the accuracies for both decoding intervals of one paradigm. As already indicated by the average performances, the dominance of the CNN likewise becomes clear on the level of participants. Fig. 4.2A demonstrates that in the POT the CNN decoding accuracies significantly exceeded those of the other two decoding methods for each single participant, and on the group level was significantly better compared to both rLDA and FBCSP. There was no significant difference between the two latter methods in POT on the group level, however, there were significant differences between rLDA and CSP on the individual level which were almost always in favor of rLDA (Fig. 4.2A, bottom panel).

In the LOT, comparing CNNs to rLDA, significant differences were also nearly in all cases in favor of the CNNs. In contrast to POT, in part of the participants there was no significant performance difference detectable, and there was also no significant difference on the group level. Compared to FBCSP, CNNs were however again significantly better

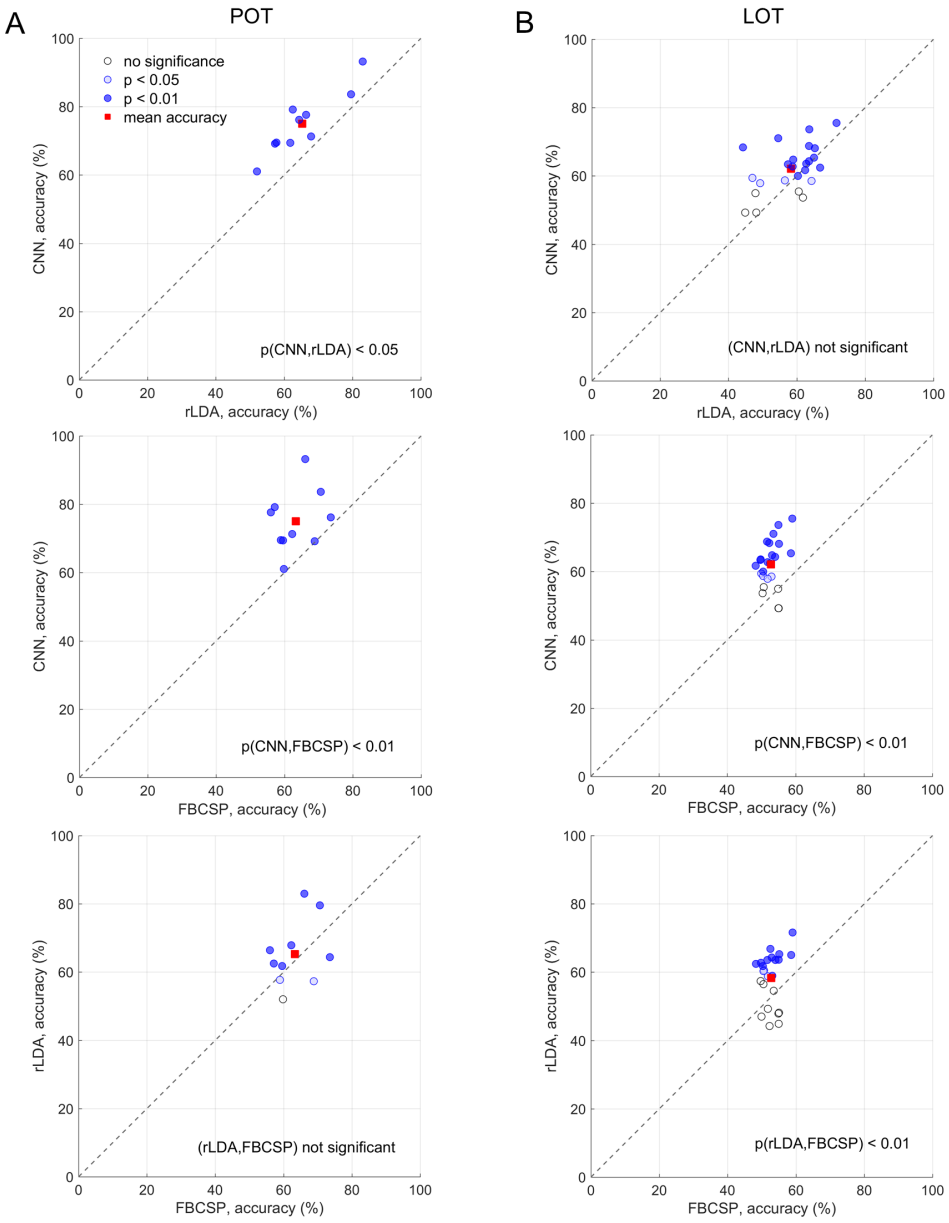

Figure 4.2: Pairwise comparison of decoding performance. **A** Decoding accuracies of CNN vs rLDA vs FBCSP for the pouring observation task. **B** Decoding accuracies of CNN vs rLDA vs FBCSP for the lifting observation task.

in nearly all individual participants and also on the group level. In the LOT (but not in the POT), rLDA significantly outperformed FBCSP.

To quantify the relationship between the methods, the linear correlation between decoding accuracies over participants was calculated pairwise for the different methods. Fig. 4.3 shows the correlation for the comparison of the CNN and the rLDA performances for POT and LOT error decoding. Particularly for the error decoding in POT, there was a highly significant linear correlation. There was no significant correlation with FBCSP

performance. The comparison of all methods with one another is summarized in Tab. 4.2. The correlations of CNN with FBCSP and rLDA with FBCSP behave in a similar way.

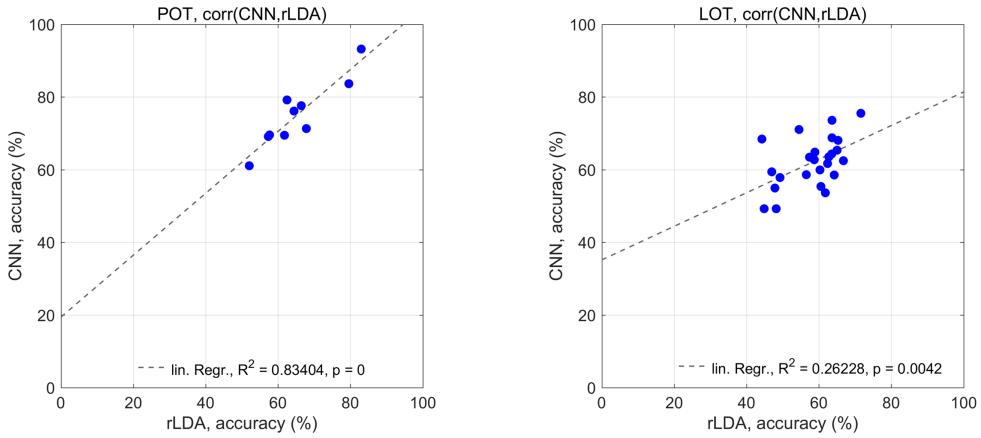

Figure 4.3: Correlation of CNN and rLDA results for both paradigms.

Table 4.2: Linear correlation coefficients & p-values

paradigm	CNN/rLDA	CNN/FBCSP	rLDA/FBCSP
POT	0.913 (<0.001)	0.292 (0.213)	0.375 (0.152)
LOT	0.512 (0.004)	0.277 (0.095)	0.162 (0.223)

4.4 Visualization of Error-related Correlations

To make the behaviour of the CNN more understandable, the visualization method described in Subsec. 2.5.6 was applied. Firstly, the correlation of changes in CNN predictions with perturbation changes in input spectral amplitudes was used to obtain information about what the deep CNNs learned from the data. Training trials were transformed into frequency domain using the FOURIER transformation and randomly perturbed by adding GAUSSIAN noise ($\mu = 0$, $\sigma = 1$), while keeping the phases steady. Both, the unperturbed and the inverse-FOURIER transformed signals were used to feed the deep CNN. The output of the CNN before the softmax activation was extracted for 30 iterations and the difference of the perturbed and original CNN predictions were correlated with the perturbation itself, see Eq. (2.69), resulting in spatial correlation maps. Secondly, perturbations to the time domain voltage signal were also applied. Again, the perturbation was correlated with the CNN output changes. The algorithm covering both implementations is described in Alg. 8, the maps M_t for the time domain visualization are shown in Fig. 4.4.

Algorithm 8 input-perturbation network-prediction correlation map $\mathsf{M}_{t/f}$ calculation

Require: data set $\mathbf{D}$, GAUSSIAN probability density p_G with mean μ and std σ

for all trials $\mathbf{X}_i \in \mathbf{D}$ **do**
  $\mathbf{G}_{t,i} \leftarrow$ GAUSSIAN noise $(\mathbf{X}_i, \mu, \sigma, p_G)$
  $\mathbf{X}_{dist,i} \leftarrow \mathbf{X}_i + \mathbf{G}_{t,i}$
  $\tilde{\mathbf{X}}_i \leftarrow \text{FFT}(\mathbf{X}_i)$
  $\mathbf{G}_{f,i} \leftarrow$ GAUSSIAN noise $(\tilde{\mathbf{X}}_i, \mu, \sigma, p_G)$
  $\tilde{\mathbf{X}}_{dist,i} \leftarrow \tilde{\mathbf{X}}_i + \mathbf{G}_{f,i}$
  $\mathbf{X}'_{dist,i} \leftarrow \text{inverseFFT}(\tilde{\mathbf{X}}_{dist,i})$
end for
for $j = 1$ **to** number of iterations **do**
  $\mathtt{pr}_j \leftarrow$ CNN output of $\mathbf{X}$ after softmax activation
  $\mathtt{pr}_{j,t} \leftarrow$ CNN output of $\mathbf{X}_{dist}$ after softmax activation
  $\mathtt{pr}_{j,f} \leftarrow$ CNN output of $\mathbf{X}'_{dist}$ after softmax activation
  $\mathsf{M}_{j,t} \leftarrow corr(\mathbf{G}_t, \mathtt{pr}_{j,t} - \mathtt{pr}_j)$
  $\mathsf{M}_{j,f} \leftarrow corr(\mathbf{G}_f, \mathtt{pr}_{j,f} - \mathtt{pr}_j)$
end for
$\mathsf{M}_t \leftarrow$ iteration mean of $\mathsf{M}_{j,t}$
$\mathsf{M}_f \leftarrow$ iteration mean of $\mathsf{M}_{j,f}$
return $\mathsf{M}_f, \mathsf{M}_t$

In this study, the focus lies on the visualization results in the time domain, as rLDA trained on the time-domain EEG signals outperformed FBCSP. The latter is designed to exploit band-specific spectral power differences, however, the CNN behaved more rLDA-like (Fig. 4.3 and Tab. 4.2) and in the LOT experiment FBCSP decoding was at chance level. This suggests that in the present decoding problems, band-specific spectral power differences did not play the dominant role as a source of decodable information. Accordingly, frequency-resolved CNN visualizations (not shown) were rather noisy-looking. In Fig. 4.4A, the averaged time-resolved input-perturbation network-prediction correlation maps for voltage features of the two decoding classes in the error decoding of the POT paradigm are shown. Video frames shown below the maps were selected according to the specific point in time of each map. The patterns of the correct and error classes showed two times windows with high correlation, first around $3.1\ s$ and then again around $3.7\ s$ (time relative to the onset of the video stimuli). In both instants the network appears to learn a similar occipitally-pronounced EEG pattern. The comparison of maps from both conditions expectedly shows opposite patterns. The occipital predominance of correlation effects in these time windows would suggests that the participants' brains differentially processed visual aspects distinguishing correct and incorrect robot action as presented in the stimuli. As a first step to investigate which visual features carried the error-specific information, the L1 distance between temporally corresponding frames in both conditions was calculated, as well as between the frame-wise change (black curves in Fig. 4.4A and B). At least with these simple features, there was no obvious relation

between the time course of changes in them and the time points where the EEG was most informative for CNN error decoding. Analogous visualizations for error decoding in the LOT experiment showed spatially more widespread effects (Fig. 4.4B). Temporally, however, these effects had a remarkable sharp onset at approximately $4.6\,s$, around the time when success versus failure became first evident, but long before the obvious consequences of error versus success became visible (ball being lifted from the ground or not). Again, there was no obvious time relation to the two low-level measures of image similarity (Fig. 4.4B, bottom panel).

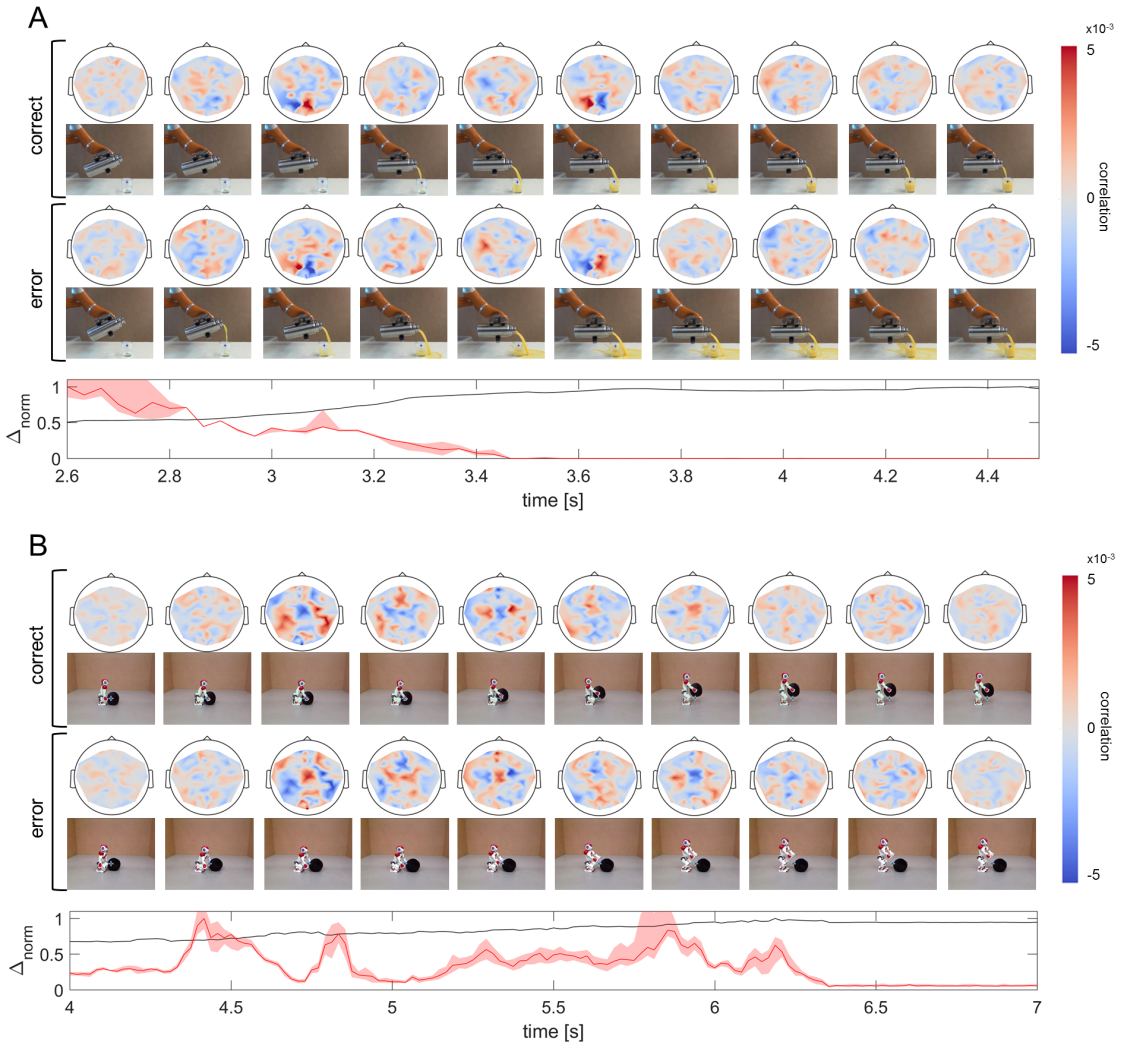

Figure 4.4: Time-resolved voltage feature input-perturbation network-prediction correlation maps. **A** Error decoding in the POT, averaged over 30 iterations and all POT participants (top). Time-resolved normalized L1 distance Δ_{norm} between (1) video frames for both conditions (bottom, black) and of (2) sequential pairs of video frames for both conditions (bottom, red). **B** Visualizations for LOT error decoding, all conventions as in A.

4.5 Related Work

Essentially, the related work in this chapter is based on that of the previous Chap. 3. The main focus lies on the detection of robot errors by reference to brain signals of human observers. Assistive robotic solutions in health care and in non-medical applications depend on a proper error detection and error management. In the last years, several studies investigated decoding of robot errors from brain signals of a human observer [176, 299, 360]. Error-related potentials recorded with EEG have been used, e.g., to teach neuroprosthetics suitable behaviors in scenarios of varying complexity [176] or to investigate their role for robot control during an object selection task [299].

Accuracies however leave space for improvements, which would be desirable to optimize the practical usefulness of error-related brain signals. The preceding chapter confirms this lack of practical and daily life applicability, which encourages the need to optimize decoding or to find better technical solutions to address the problem. In image recognition, deep neural networks in particular have contributed to boosting performances and enabling machine vision in real time. Several implementation could demonstrate best classifications on, e.g., the NORB and CIFAR-10 dataset [77] or the PASCAL VOC dataset [142]. Above all, *Krizhevsky et al.* were able to convince with their groundbreaking success at the ImageNet 2012 clasification benchmark [202]. Especially Convolutional Neural Networks have proven their potential. However, for this kind of method the classification on human brain data is quite new territory and first experiments transfer their qualities into this field [13, 73, 158, 288, 303].

4.6 Conclusion

The tricky issue of low decoding classification performances according to error detection is addressed by applying deep learning to a naturalistic decoding task where participants observed a robot performing different assistive actions either successfully or failing to do so. In EEG research, architectures including deep CNNs have recently been used to explore their applicability in brain-signal decoding [303, 328], but not yet to robot-error decoding from EEG.

The findings of the present study can be seen from two sides: first, decoding the success of robot action from brain signals is a problem with potential practical relevance and, hence, has been investigated in a number of previous studies [176, 299, 360]. More specifically, improving the decoding accuracy in this context is a topic with practical relevance, particularly under complex, real-life-like conditions. Thus the video stimuli was designed to mimic such conditions. Second, CNNs are still relatively new in EEG decoding, and the findings from the present decoding problems also contribute some new facts to the growing methodological literature on this topic. The results show that, compared with 2 other widely-used classifiers, the deep CNNs performed consistently better. The same CNN as applied here yielded mean accuracies of 93% for classification of 4 different movements in [303] and 85% in discriminating normal and pathological EEG in [302], and was in all cases at least as good or significantly better than the baseline

comparison methods.

In the present study mean accuracies of $(75 \pm 9)\,\%$ (POT) and $(62 \pm 7)\,\%$ (LOT) for error decoding were yielded. In a previous study, using a combination of rLDA and reinforcement learning, decoding of actions that participants evaluated as either erroneous or correct [176] resulted in a mean EEG decoding accuracy of 75%. For one of the paradigms (POT) a similar mean accuracy was reached here. In some of the participants, accuracies were above 90%, but overall still better accuracies are needed. Among other recent advances in the field of deep learning research, automatic hyperparameter optimization and architecture search, including recurrent and residual network architectures, data augmentation, using 3D convolutions, or increasing the amount of training data all have the potential to further increase CNN performance. CNNs were systematically better but in their accuracies over participants linearly correlated with those of the rLDA, but not with those of the FBCSP (Fig. 4.3). So in the present examples the CNNs behaved "rLDA-like". Interestingly, in a previous study where the same CNN architecture and training strategy as here was used for movement decoding from EEG, accuracies over participants were highly correlated with those of FBCSP [303], and it was shown that CNNs indeed used frequency-specific spectral power changes (rLDA was not evaluated there). This points to the possibility that CNNs might become more "CSP-like" or more "rLDA-like" (or even more similar to other decoding methods) depending on what features are informative in the EEG signal.

The results as discussed so far indicated an important role of time-domain EEG signal changes for the decodability of errors in the tasks, thus for their visualization the perturbation-based technique as described in [303] for spectral changes to time-domain voltage features was adapted. Resulting maps confirmed that the CNNs learned to use time-domain EEG responses to distinguish between classes. Maps also indicated that specific time windows and scalp regions were informative, with different patterns in the two tasks (Fig. 4.4). Particularly for errors in the pouring task (POT), perturbation maps pointed to the occipital/visual areas as important sources of information learned by the CNNs. This kind of decoding could be practically helpful in situations where robot errors would be visually distinct, such as in the example of liquid spilling to a table. Further it would be interesting to investigate in how far the decodability of such differential visual input depends on its subjective interpretation as an error. Maps visualizing which EEG signals CNNs learned to decode errors in the lifting task (LOT) showed a spatially more widespread pattern, but also with a relatively sharp onset around the time when failure and success became first evident from the stimuli (Fig. 4.4B). Speculatively, observation of the reaching-grasping-lifting task might activate the human mirror neuron system (MNS) [156, 239, 291, 292]. The human MNS involves widespread frontal and parietal regions as involved in the maps in Fig. 4.4B. The engagement of the MNS might be modulated by the degree of humanoid appearance of the robot. Thus as a next step, differences related to the two robot types (more and less humanoid) used in the reaching-grasping-lifting experiment could be analyzed.

Chapter 5

The Role of Robot Design in Decoding Error-related Information

For utilization of robotic assistive devices in everyday life, means for detection and processing of erroneous robot actions are a focal aspect in the development of collaborative systems, especially when controlled via brain signals. Though, the variety of possible scenarios and the diversity of used robotic systems pose a challenge for error decoding from recordings of brain signals such as via EEG. For example, it is unclear whether humanoid appearances of robotic assistants have an influence on the performance. In this chapter, a paradigm, in which two different robots executed the same task both in an erroneous and a correct manner, is used to differentiate robot types. The error-related EEG signals of human observers indicate that the performance of the error decoding is independent of robot design. However, it can be shown that it was possible to identify which robot performed the instructed task by means of the EEG signals. This chapter demonstrate that deep convolutional neural networks (deep CNNs) could reach significantly higher accuracies than both regularized Linear Discriminanat Analysis (rLDA) and filter bank common spatial patterns (FBCSP) combined with rLDA. The findings indicate that decoding information about robot action success from the EEG, particularly when using deep neural networks, may be an applicable approach for a broad range of robot designs.

The role of robot design in the decodability of error-related information has rarely been investigated, although robots come in a broad range of different designs. Beside specifying properties like mobility or autonomy, a robot can be classified regarding its grade of being humanoid. For example, in scenarios in which humans and robots collaborate, this characteristic may have an influence on the human's perception of the robots behaviour. In a comparison between a human and a robotic agent performing several movement tasks, an activation of the human mirror neuron system (MNS) [292] could be shown [131]. Already in Chap. 3 it was discussed how the observation of robots affects the human MNS in general, and that it is not selectively activated on biological movements and thus provides comparable signals for robot movements [252]. In fact, there was no significant difference in mu-band suppression when observing human movements and robot movements. Besides, [28] and [252] suggested that the MNS is involved in the error-observation-related brain responses.

The previous chapter showed that deep CNNs significantly improved the decoding of robot errors from the EEG of a human observer. The methods are also used in this chapter to make the distinction between robot types even more efficient, and are compared to the results from rLDA and FBCSP+rLDA classification. Also, the question of whether decoding errors depends on the design of the robot, and thus on the human perception, will be clarified by means of the networks. In particular, the robot types in this chapter differ in their general similarity to humans and therefore the number of human-like features.

5.1 System and Experimental Design

In this chapter, only recordings from the lifting observation task (LOT) were used, where participants were instructed to observe two robots performing a naturalistic task either in a correct or an erroneous manner. Details about system, experimental design and participants can be found in Chap. 3, the timing structure of the experiments is depicted in Fig. 3.2. However, this chapter concentrates on the differences in the brain recordings of human observers according to the robot type, see Fig. 5.1, not the error-related information. In the analyses, trials exhibiting erroneous action for both robot types were coupled for decoding, as well as correct action for both robot types. In this way, it could be ensured that no error-related information from the stimulus falsified the results.

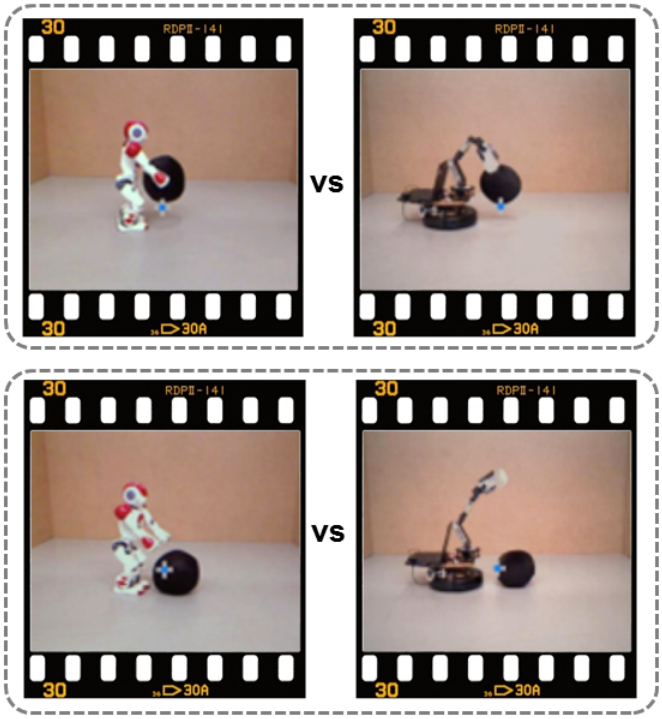

Figure 5.1: Visual stimuli showing different robot types during lifting task. For both conditions, correct and incorrect, there were stimuli with two different robot types. Both robots try to approach, grasp and lift the ball, either managing or failing to lift a ball from the ground. *Slide mount by pixelio*.

5.2 Pre-Processing, Classifier Design and Statistics

The recorded EEG signals were re-referenced to a common average reference (CAR) and resampled to $250\,Hz$. An electrode-wise exponential moving standardization was applied to normalize the data by exponential moving means and variances. Then, according to stimulus onset and predefined decoding intervals the data were cut into trials and used as classification features. The design of the deep CNN was the same as used for classification in Chap. 4 and is described in Subsec. 2.5.3 (see also Fig. 2.14). The data were split into two sets, $80\,\%$ were used for training while the remaining $20\,\%$ served as a test set. For the rLDA the pre-processing was similar except the fact that no standardization was implemented, resulting in higher accuracies than with standardization. Exactly as in Chap. 4, the rLDA algorithm comprised a shrinkage parameter estimation, based on the *LedoitWolf* estimator [210], see Eq. (2.66). The results of the FBCSP algorithm were calculated using the implementation of Chap. 3, but with a slightly improved cleaning, and are described in Subsec. 2.5.2, see also Fig. 2.9 and Alg. 3. According to the results in Chap. 3 the best performing frequency range $< 20\,Hz$ contributed in decoding the errors in the FBCSP implementation. Relative to the video onset, two decoding intervals were selected. A long decoding interval of $0 - 7\,s$ covered the full length of the video, while a late interval of $4 - 7\,s$ was selected since it was covering the actual process of grasping and lifting the ball. For both intervals decoding performances were calculated using the CNN, the rLDA and the FBCSP+rLDA implementation.

For the decoding results on the level of participants, a permutation test was applied, see Alg. 7, while group significances (e.g. differences between decoding methods) were estimated by means of the single trials and using a sign test. For the correlation coefficients, significance was tested by randomizing the order of one of the input vectors of the correlation. Guesses that exhibited higher coefficients than the true correlation coefficient were counted and contrasted to the total number of guesses. In this way, the significance of the correlation could be estimated.

5.3 Decoding Errors of Different Robot Types

Firstly, the influence of robot design on the performance in an error-decoding scenario was investigated. Therefore, the trials were sorted by robot type and the decoding analysis of erroneous vs. correct trials was performed separately on the two data sets. In this case, only the deep CNN implementation was used for error detection. Fig. 5.2 shows the outcome of this analysis, showing the distribution of the accuracies for both robot conditions for all participants. Even though the decoding of trials presenting the NoHu robot executing the task showed a broader range, there was no significant difference between the two conditions (sign test, $p = 1$). The CNNs achieved median accuracies of $(64.8 \pm 6.8)\,\%$ for the NAO condition and $(64.0 \pm 8.7)\,\%$ for the NoHu condition, yielding almost the same performance taking the errors into account.

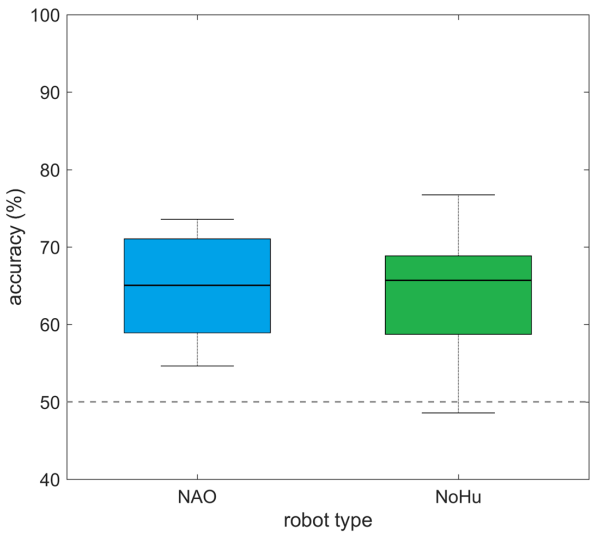

Figure 5.2: Robot-related error decoding. Accuracies for error decoding using only stimuli with one type of robot each.

5.4 Distinction Between Robot Types

Three different decoding methods were used for classification, CNNs, rLDA and FBCSP+rLDA. For each participant, decoding accuracies were determined for both decoding intervals which were defined according to the video onset. An overview of mean decoding accuracies is given in Tab. 5.1. The results indicate that a distinction between the two kinds of robots was decodable with all of the applied techniques. Furthermore, for both decoding intervals the deep CNN architecture performed consistently and significantly better (sign test, $p < 0.01$).

Table 5.1: Mean accuracies for different decoding intervals

	0 - 7s	**4 - 7s**
CNN	$(78.3 \pm 8.1)\,\%$	$(73.8 \pm 7.5)\,\%$
rLDA	$(68.3 \pm 8.0)\,\%$	$(64.7 \pm 7.4)\,\%$
FBCSP	$(55.7 \pm 4.5)\,\%$	$(56.8 \pm 3.9)\,\%$

Fig. 5.3A shows the pairwise comparison of performances for the different classifiers, including the results of the analysis gained with both of the decoding intervals. Significance is indicated by blue color while the red squares represent the mean accuracies, the diagonal indicates equal performance. For each single participant, the decoding accuracies gained by the deep CNN implementation significantly exceeded those of the other methods, i.e., compared to rLDA and FBCSP, the CNN performed significantly

better on the group level. Clearer as in the previous chapter for distinction of robot types, the rLDA performances on the level of participants exceeded those of the FBCSP, except for one participant, almost exclusively significant with $p < 0.01$.

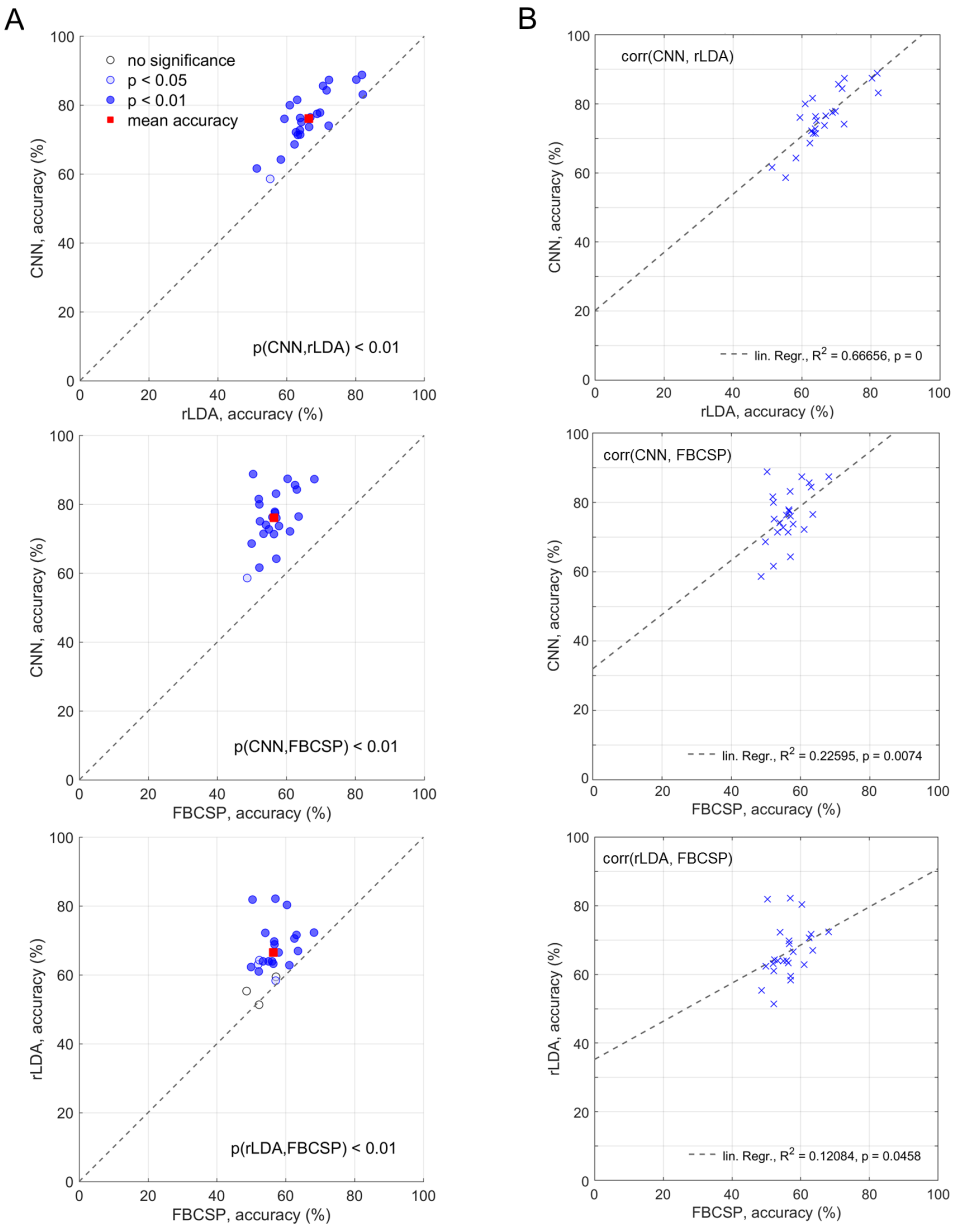

Figure 5.3: Pairwise comparison of decoding performance and correlation of these. **A** Decoding accuracies of CNN vs rLDA vs FBCSP for distinction between the two robots. **B** Pairwise linear regression of the participants performances.

To gain a measure of the relationship between the decoding results of the different decoding techniques, the pairwise linear correlation between decoding accuracies over subjects was calculated. Again, the idea was to figure out whether the net behave more

like an rLDA algorithm, learning the temporal features or rather basing the decision on spectral features. Fig. 5.3B shows the correlation for all pairs; the dotted line represents the result of the linear regression. The coefficient of determination R^2 indicates a highly significant linear relationship between deep CNN and rLDA performances, but also for both of the other combinations a significant ($p < 0.05$) correlation can be found.

5.5 Visualization of Correlations Related to Robot Type

The visualization used in this chapter was applied as described in Subsec. 2.5.6, see also Alg. 8, visualizing EEG features that the deep CNN learned from the data when used to distinguish between the different robot designs. To this aim, the correlation between changes of the predictions made by the CNN and perturbation changes in time domain voltage amplitudes were calculated. By adding GAUSSIAN noise, training trials were randomly perturbed. After that, the perturbed signal and the unperturbed signal were both fed into the deep CNN. The two outputs were extracted before the softmax activation, and their difference was correlated with the perturbation itself. The results are visualized in a channel-resolved manner, see Fig. 5.4. The decision to use time domain signals as an input for the perturbations instead of spectral amplitudes was made according to the fact of lower decoding performances of the CSP algorithm, which relies on spectral features. This circumstances suggest that spectral information was likely not that prominent in the underlying decoding problem. The correlations shall indicate at which moment changes in certain channels might contribute to the classifiers decision causally.

In Fig. 5.4 the (participant- and iteration-) averaged time-resolved input-perturbation network-prediction correlation maps for voltage features of the robot type decoding are shown. The maps are depicted from $2 - 6\,s$, whereby each map illustrates the correlation of a $0.2\,s$ time bin. For each bin, the two maps (NAO and NoHu) are shown together with their corresponding video frame. As expected, pairs of maps for the two different conditions exhibit opposed correlations. Fig. 5.4 is divided into a section were the robots approach the ball (top) and a section were the robots try to lift the ball (bottom). At the bottom of each section, the time-resolved normalized L1 distance Δ_{norm} of sequential pairs of video frames for both conditions is illustrated.

The perturbation maps in Fig. 5.4 exhibit increasing, prominent correlation patterns for signals around $3.2\,s$ and $5.0\,s$ according to video onset. At the first time point, the effects show spatially more widespread correlations with a remarkable, centered peak in frontal areas accompanied by an occipital symmetric effect. The later time window around $5.0\,s$ shows similar but less symmetric patterns in occipital regions, and a more pronounced frontal peak. The occipital effects for both time points might indicate different cerebral processing of the visual characteristics in the robots execution of the programmed task.

As a first step, to examine visual features which might have led to differences in brain signals for the two robot conditions, the time-resolved normalized L1 distance Δ_{norm} was calculated. The corresponding curve in Fig. 5.4 indicates a rather small difference between sequential frames and a steady difference between the two conditions, as the curve varies only little but with consistently high values. Furthermore, with this method

no obvious correlation between the time course of visual changes and informative features extracted by the perturbation analysis can be suggested.

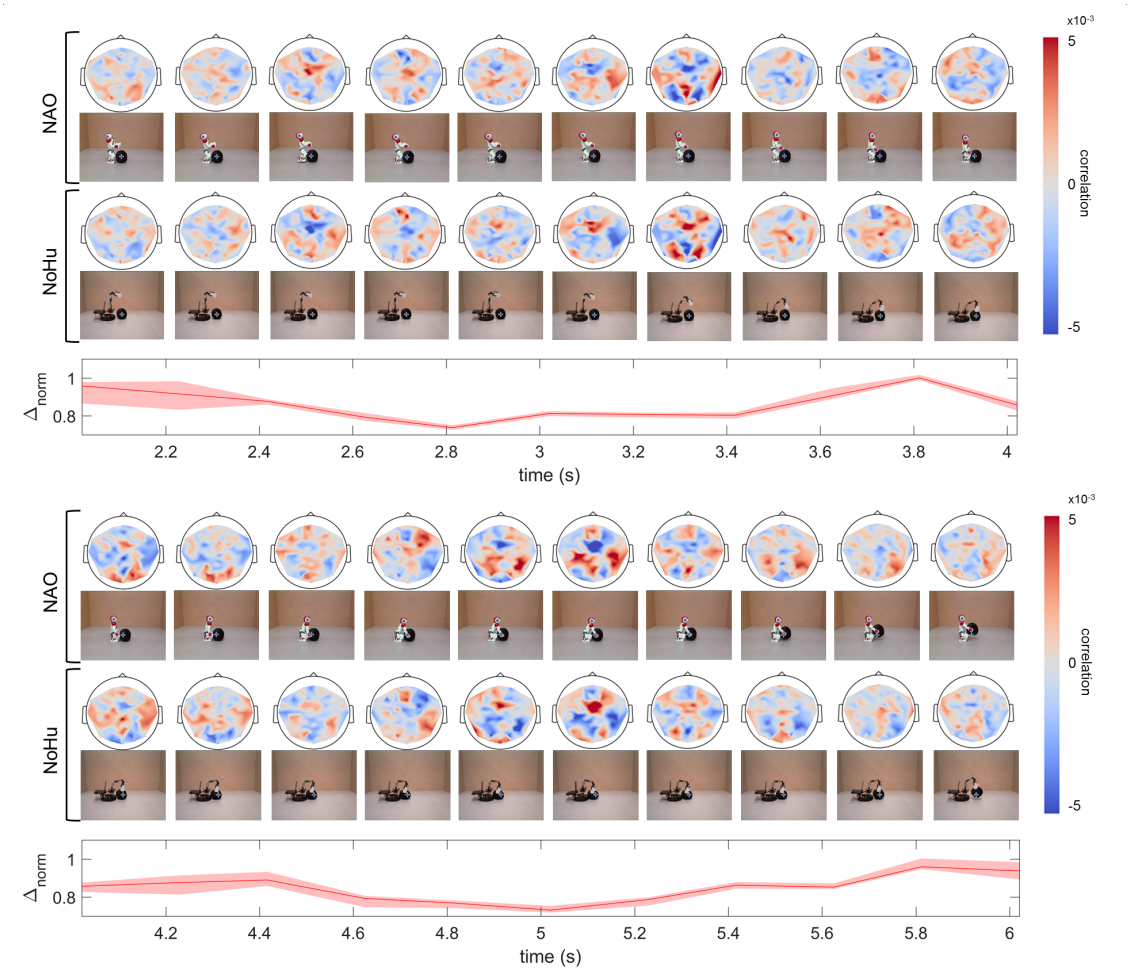

Figure 5.4: Time-resolved voltage feature input-perturbation network-prediction correlation maps. Robot type decoding, averaged over participants and 30 iterations, and corresponding video frames (top rows). Time-resolved normalized L1 distance of sequential pairs of video frames for both conditions (bottom row).

5.6 Related Work

Error-related brain activity has the potential to support the management and implementation of BCI systems in various scenarios. Error detection systems have been embedded to improve real life tasks, e.g., to detect and correct erroneous actions performed by an assistive robot online, using primary and secondary error-related potentials (ErrPs) [299]. Other approaches benefit from ErrPs to extract incorrect maneuvers in a real-world driving task [377] or use error-related negativity (ERN) to correct erroneous actions in object moving tasks [261]. In addition, error-related brain signals have been utilized to teach

neuroprosthetics appropriate behaviour in situations with changing complexity [176]. The previous chapter particularly addresses this topic in more detail.

In cooperation between humans and robots, the question of how the robot is perceived by humans and how this affects the behaviour of a collaborator can play an important role. However, there has been little research into this so far, especially with regard to the distinction between humanoid and non-humanoid robots. Some studies have already dealt with how robots are perceived by humans in terms of facial features. It was shown, for example, that the appearance of certain features in the face of a robot, such as eyes, nose and mouth, the dimensions of the robot's head as well as the total number of facial features, plays a major role in how human-like it is perceived [97]. According to these results, the effect triggered by humanoid features can be generalized to the whole body. In Chap. 3 the distinction between robot types has already been investigated by means of the FBCSP algorithm. The results show that it is possible to decode the type of observed robot from human surface EEG, and contribute to the understanding of how and whether the appearance and/or behaviour of a robot affects the perception of the user.

In the last chapter it has been shown that deep CNNs significantly improved the decoding of robot errors from the EEG of a human observer. However, the role of robot design in the decodability of error-related information in this scenario has rarely been investigated, although robots come in a broad range of different designs. Beside specifying properties like mobility or autonomy, a robot can be classified regarding its grade of being humanoid. Especially when collaborative interactions with humans take place, this characteristic may have an influence on the human's perception of the robots behaviour. In a comparison between a human and a robotic agent performing several movement tasks, an activation of the human mirror neuron system [292] could be shown [131].

5.7 Conclusion

The potential of error detection for practical BCI applications has recently led to several studies in EEG research, e.g. [176, 261, 299, 377]. However, aspects of robot design has not been part of investigations on error-related brain signals so far. In this chapter, this aspect was examined and it could be shown that at least the two different robot designs used in the current study did not have any significant negative impact on the performance of error decoding, indicating that in principle error decoding is applicable to various robot types.

The distinction between robot types based on the participants brain-signals yielded in mean accuracies of $(78.3 \pm 8.1)\,\%$ for the CNN, $(68.3 \pm 8.0)\,\%$ for the rLDA and $(55.7 \pm 4.5)\,\%$ for FBCSP implementation when decoding on the whole time-window covered by the video stimulus. Hence, the robot type indeed could be classified on the basis of the EEG data and certainly improved compared to Chap. 3. Moreover, the CNN implementation could reach systematically better results than the two other methods. Deep CNNs have already entered the field of EEG research and proven their applicability [328], but not yet for distinction of robot types. Overall, the accuracies reached here are

however still far from the requirements for practical application. E.g. intracranial signals or further improved non-invasive methods will be necessary to reach better performance.

The pairwise comparison of the participants decoding results for the different methods showed a significant linear correlation for all cases. Particularly the behaviour of CNN und rLDA accuracies were apparent: the net appears to act more ”rLDA-like” for the given problem, in line with the assumption that time-domain features were mainly informative.

For the visualization, the time-domain EEG signal changes were correlated with decoding results determined after a GAUSSIAN perturbation. Based on the correlation maps as shown in Fig. 5.4 it appears that the CNN learns time-domain information to distinguish between classes, as specific time windows reveal more pronounced effects than others. Thus, a prominent focus of the distribution of informative signals lies on occipital brain regions. This might reflect the difference in the processing of the visual input for the two robot types. Visualizations also show a distinct medio-frontal peak. This effect appeared around the same time when the robot approached to the target and grasped it. It is possible that in this cases the mirror neuron system might get activated [156, 239, 291, 292], which involves widespread frontal and parietal regions. According to this, the different robot types could have lead to a differential activation of this system.

Chapter 6

ELAS: a Toolbox for Assignment and 3-D Visualization

Intracranial electroencephalography (iEEG) plays an important role in pre-neurosurgical epilepsy diagnostics and is increasingly used in neuroscientific research. However, the individual position of the electrode contacts varies greatly between patients, which makes group analyses particularly difficult and thus restricts the interpretation of the iEEG results. In general, an assignment procedure is required that enables the neuroanatomical information of the underlying brain areas to be obtained for each individual electrode contact. Such a neuroanatomical atlas system is already successfully used for analysis of neuroimaging data, it enables the probabilistic assignment of individual voxels in the MNI space to cytoarchitectonically defined brain areas. But until now, it was unclear if and how exactly this probabilistic atlas can be utilized in the growing field of iEEG studies. This chapter presents the electrode assignment algorithm ELAS for iEEG electrode contacts, implemented in a MATLAB-based open source interface, that allows a hierarchical probabilistic assignment (HPA) of individual electrode contacts to cytoarchitectonically-defined brain areas. Beside a cortical projection, the here presented ELAS consists of two major steps: (I) a pre-assignment to the cerebral lobes (frontal, parietal, occipital or temporal) based on the position of the individual electrode contacts with respect to the anatomical landmarks and (II) a following probabilistic assignment to cytoarchitectonically-defined brain areas based on lobe-specific probability maps of the SPM Anatomy Toolbox. This assignment procedure is so far the first approach that combines both individual macro-anatomical and cytoarchitectonic probabilistic information, yielding in relevant improvements to anatomical assignments in iEEG. To evaluate the method, ECoG data obtained in 14 epilepsy patients with a total of 781 intracranial electrode contacts from a wide range of cortical areas was analyzed. Assignment was possible in $81.8\,\%$ *of the electrode contacts and due to integration of information of individual anatomical landmarks derived from the patients' MRIs, the ELAS approach avoided incorrect assignments in approximately* $8\,\%$ *of electrode contacts. The presented hierarchical probabilistic assignment is freely available in the open source toolbox ELAS, including a 3D visualization of the assignment results and an object wavefront OBJ file export for use in virtual reality setups.*

In Chap. 3, the problem of error detection based on human brain signals was already examined in detail and it became apparent that the performances achieved there by far do not provide what is desired for a successful human-robot cooperation. The question was also discussed to what extent the design of the robot influences human perception and whether the differentiation between different robot types is possible in principle. Likewise the differentiation of robots could be done by the algorithm inserting the brain signals of a human observer, but here, too, the accuracy was rather moderate. In the following chapters, the problem of efficiency was addressed on the decoders side, in order to significantly improve the performances by means of convolutional neural networks. In both cases, this was successful and the network, which had been kept relatively general so far, could be brought even closer to the problem by fine tuning in future.

In addition to the possibility of adjusting the decoder, an enhancement of the signal quality can also contribute in improving the decoding. One possibility is to work with data that originates from invasive recordings and therefore provides a much better signal quality. Also, these types of data are not so burdened by artifacts and paradigms can often be performed without restricting the participant's mobility. In some circumstances, the latter can lead to the participant becoming more involved in the task and thus the signals are clearer defined, e.g. [343]. However, invasive measurements are not an option for everyday, realistic scenarios and therefore do not fall within the scope of applicability. Experiments with invasive methods can though identify general peculiarities of e.g. error-related brain signals and the error processing can be better understood. In addition, implanted electrodes can be used to make precise statements about which areas of the brain are involved in the processing of error processes, when and how.

This chapter deals with the basic problem of assigning electrode contacts to specific brain areas. Only through precise allocation the areas involved in processes can be isolated and reliable answers to where, when and how can be provided. The new approach presented in this chapter uses a probabilistic but general model of assigning MNI (Montreal Neurological Institute) coordinates to possible areas. This method extends the model for use on individual brains on the basis of anatomical landmarks and, in addition, reverses the displacement of ECoG electrode contacts due to deformation during the implantation with respect to the standard maps by cortical projection.

The method presented is embedded in the user-friendly *MATLAB*-based toolbox ELAS[1] (electrode assignment) and is intended to enable intuitive use, especially for users, for example from the medical field, without any programming experience. The results can be visualized directly within the interface in 3D together with individually selectable brain areas on a semitransparent standard brain. An export for use in virtual reality is also included in the software package. Localization of iEEG electrodes is already covered by several approaches (e.g. *BioImage Suite* [260] or *FieldTrip* [255]) and can be individually combined with the method. In that way, the toolbox ELAS provides a complete processing from *magnetic resonance imaging* (MRI) data right up to the visualization of electrode contacts and (assigned) anatomical areas.

[1] https://github.com/joosbehncke/elas

6.1 Methods

First, the developed methods for the assignment of intracranial electrode contacts will be presented and explained before the method is validated in the later sections. This analysis was restricted to ECoG electrodes, without loss of generality. The general process from imaging techniques like MRI to visualization in 3D or virtual reality is described in Fig. 6.1, including all individual processing steps. The dotted area represents the essential contribution of the ELAS interface, whereby other intermediate steps, such as image processing in SPM, can also be processed directly and in parallel inside the toolbox.

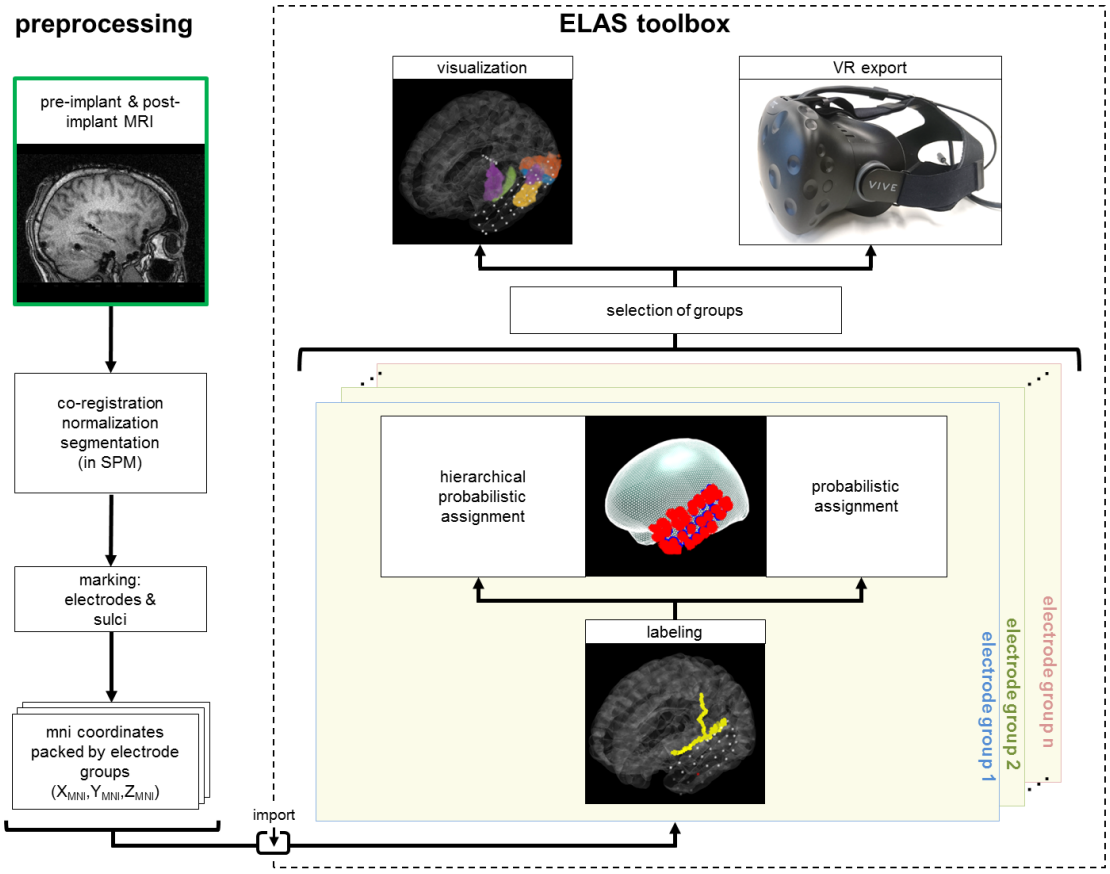

Figure 6.1: Flowchart of the ELAS electrode assignment and visualization procedure. The normalized pre- and post-implant MRIs serve as an input and basis for the electrode marking. The procedure can be started at any intermediate step, assumed that the required intermediate results already exist. The ELAS toolbox provides the possibility to label electrodes according to imported MNI (Montreal Neurological Institute) coordinates. In a second step, the electrode contacts are assigned to cytoarchitectonically defined brain areas. Finally, the results can be visualized, exported as a *MATLAB* file and/or transformed into wavefront OBJ files for visualization in virtual reality.

6.1.1 Patients and Implantations

The iEEG-adapted hierarchical probabilistic assignment (HPA) technique was evaluated in a sample of 14 epilepsy patients who underwent ECoG electrode implantation for presurgical evaluation (Tab. 6.1). All electrodes had a $4\,mm$ diameter, with an exposed area of a $2.3\,mm$ diameter. For further information on the implanted subdural electrodes, see Tab. 6.1. Written informed consent was obtained for all patients, stating that the electrophysiological data might also be used for scientific purposes. For P1-P10, positions of a total of 687 subdural grid electrode contacts over brain areas of the lateral convexity were analyzed. For P2, as well as P11-P14, positions of a total of 94 subdural strip electrode contacts overlapping with occipital brain areas were analyzed.

Table 6.1: Meta data for P1 to P14

	Age	**Sex**	**Subdural electrodes**	**Inter-contact distance**	**Contact material**
P1	40	m	64-contact grid L fronto-parietal	10 mm	steel
P2	41	f	64-contact grid L fronto-lateral 1*4-contact interh. strip L parieto-occipital	10 mm	steel
P3	49	f	64-contact grid L fronto-parietal	10 mm	steel
P4	15	f	64-contact grid R fronto-parietal	10 mm	steel
P5	38	f	64-contact grid L fronto-parieto-temporal	10 mm	steel
P6	41	f	64-contact grid L fronto-parieto-temporal	10 mm	steel
P7	25	m	64-contact grid L frontal	10 mm	steel
P8	45	f	112-contact grid L fronto-parieto-temporal	7.1 mm	platinum
P9	27	m	64-contact grid L frontal	10 mm	steel
P10	21	f	64-contact grid L fronto-parieto-temporal	10 mm	steel
P11	14	f	2*6-contact strips occipito-polar 2*6-contact strips temporo-basal	10 mm	steel
P12	33	m	1*6-contact interh. strips R parieto-occipital	10 mm	steel
P13	28	f	1*4- & 1*6-contact interh. strips L parieto-occipital 3*4- & 1*6-contact strips L occipito-basal 2*4-contact strips L occipito-polar	10 mm	steel
P14	48	m	1*6- & 1*4-contact strips L occipito-basal 2*4-contact strips left occipito-polar 1*6-contact strip L temporo-basal	10 mm	steel

6.1.2 Normalization of the Post-operative MRI

A post-operative T1-weighted magnetization-prepared rapid-acquisition gradient-echo (MPRAGE) data set was acquired for each patient 1-2 days after implantation in a $1.5\,T$ magnetic resonance imaging scanner (Vision, Siemens, Erlangen, Germany) with an isotropic image resolution of $1\,mm$, which was sufficient for identification of the individual electrode contacts.

Previous studies applied a range of techniques for normalizing intracranial electrode contacts to MNI space, such as based on MRI-CT coregistration [46, 105, 161] or utilizing post-implantation MRI data [201, 370]. Here, post-implantation MR images, which were normalized to the MNI space using combined affine and non-linear basis-function-based normalization for modeling local distortion in *SPM12*, were also used. Spatial normalization of patients' data is a topic of research in itself [52, 91, 184, 308, 339], but was not in the focus of this chapter. The normalization procedure was verified by visual comparison of the normalized images with the MNI template and observed good normalization accuracy in the post-operative MRI data analyzed in the present study (see Fig. 6.2A).

After the visual inspection of the normalized post-operative MRI, error of normalization was calculated in the following way: the normalized post-operative MRIs including electrode contacts and the respective artifacts was compared to the normalized pre-operative MRIs (without implanted electrodes) of the very same patients (Fig. 6.2B). Since the normalization procedure in *SPM12* is optimized for MRIs without artifacts, as they occur in post-operative MRIs due to the implanted electrode contacts, the normalized pre-operative MRI were used as a ground truth for the respective patient. Thus, differences in the spatial location of individual voxels between the normalized pre- and post-operative MRIs were assumed to be errors in normalization of the post-operative MRI. In the following paragraph, it is described how the individual voxels of the post-operative MRI that are assumed to be the counterpart of the respective voxel of the pre-operative MRI were detected.

In a first step, for each reference voxel of the pre-operative MRI of a specific patient a reference cuboid with 5 voxels edge length was defined, with the reference voxel in the center of the cuboid (Fig. 6.2B left column). Then, the respective post-operative MR image of the very same patient was used to extract the search cuboids, also with 5 voxels edge length, from the same position in MNI space as the reference cuboid, but also shifted in 1-voxel steps up to 5 voxels in each spatial direction (Fig. 6.2B right column). For each of the resulting 1331 ($11 \times 11 \times 11$) "search-cuboids", the respective vector of grey values (125 values; $5 \times 5 \times 5$ voxels) was used to calculate the correlation coefficient (SPEARMAN's correlation) to the respective grey values of the reference cuboid of the pre-operative MRI. The search cuboid with the highest correlation to the reference cuboid gives a good estimate of the true MNI coordinates of the respective voxel in the pre-operative MRI.

The maximum correlation values, i.e., the correlation values between the cuboids of the pre-operative MRI and the best-fit cuboids of the post-operative MRI are shown exemplarily for a section of the brain of Patient 14 (Fig. 6.2C). In this example, it can be

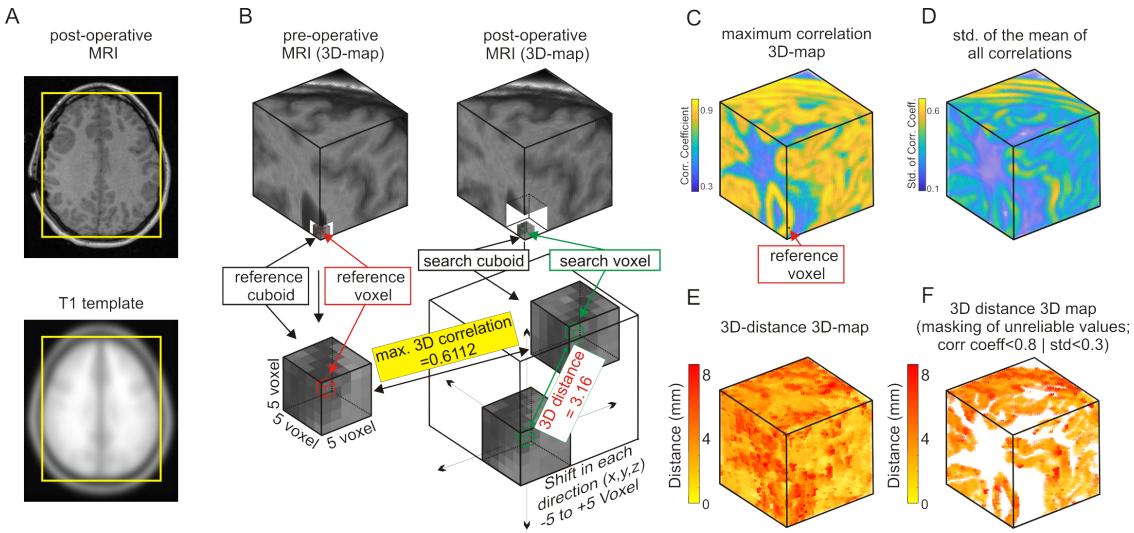

Figure 6.2: Volumetric normalization of post-operative MRI data and analysis of the error in normalization. A Horizontal slice of normalized MRI of P4 and of the T1 template used for normalization. Yellow box encompasses the normalized brain and the T1 template at the same position, showing a good spatial correspondence of the anterior/posterior as well as of the lateral extent of the brain in the normalized image to the template image. **B** A sector ($41 \times 41 \times 41$ voxels) of the normalized pre-operative MRI and the same sector, i.e., with the same MNI coordinates, of the normalized post-operative MRI is shown exemplarily for P14. In the lower corner of the pre-operative MRI sector, one reference voxel (red) and the respective reference cuboid ($5 \times 5 \times 5$ voxels) is shown exemplarily. In the same corner of the post-operative MRI sector, the respective search voxel (green) and the search cuboid, which was shifted in one-voxel steps from -5 to $+5$ voxels in each spatial direction, is shown. The search cuboid (of the post-operative MRI) with the highest correlation to the reference cuboid (of the pe-operative MRI) was used to estimate the *ground truth* with regard to the position of the respective search voxel. **C** Correlation values of the search cuboid with the highest correlation to the reference cuboid are shown color-encoded for all voxels of the MRI sector. **D** Standard deviation of the mean of all correlation values of each reference cuboid are shown color-encoded for all voxels of the MRI sector. **E** 3D distance between the original position and the "true position" of all voxels of the MRI sector is shown color-encoded. **F** 3D distance as in e), but voxels with a correlation value smaller than 0.8 and/or a standard deviation smaller than 0.3 are masked

seen that high correlation values are present in regions of the brain with high contrast, e.g., nearby sulci, while only low correlation values were observed within homogeneous (in term of color) regions, e.g., within white matter. The standard deviation of the mean of all 1331 correlations was calculated for each reference voxel (Fig. 6.2D) to investigate the width of the respective distribution. Comparable to the maximum correlation 3D-map, the 3D-map of the standard deviation shows higher values for regions of the brain with high contrast while lower values were observed within homogeneous (in terms of color) regions.

Then the distance between the reference cuboid and the search cuboid with the best fit (highest correlation) was used to determine the 3D distance between the reference voxel of the pre-operative MRI and the respective search voxel of the post-operative MRI (see Fig. 6.2B; bottom right). The resulting 3D distances are shown color encoded for the same section of the brain of Patient 14 (Fig. 6.2E). Since low maximum correlation values (< 0.8) point to low similarities between the pre- and post-operative MRI and low

standard deviation values point to small differences between all tested search cuboids to a specific reference cuboid, resulting 3D distance of these best fits are debatable. Thus, also the 3D distance 3D map were calculated where voxels with low maximum correlation values (< 0.8) or low standard deviation values (< 0.3) were disregarded. For each patient investigated, the averaged (median) correlation values, averaged standard deviation values and the averaged 3D distance with and without masking unreliable values are shown in Tab. 6.2. The MNI-normalized MRI data sets were used for localization of both the ECoG electrodes and the central sulci (CS) and lateral sulci (LS), see below.

Table 6.2: Comparison of pre- and post-implantation MRIs

	Average maximal correlation	**Average std of maximal correlation**	**Average distance pre-post**	**Average distance pre-post (corr>0.8 & std<0.3)**
P1	0.823	0.327	3.000	2.450
P2	0.680	0.266	4.123	3.606
P3	0.790	0.312	3.606	2.828
P4	0.804	0.317	3.317	3.140
P5	0.781	0.311	4.123	3.317
P6	0.680	0.232	3.000	2.236
P7	0.822	0.322	3.740	3.162
P8	0.680	0.266	4.123	3.606
P9	0.806	0.323	3.162	2.450
P10	0.815	0.327	4.359	4.123
P11	0.692	0.270	5.385	5.196
P12	0.657	0.233	4.243	3.601
P13	0.790	0.310	3.742	3.606
P14	0.782	0.303	4.123	3.317

6.1.3 Localization in MRI Data Sets

The positions of the ECoG electrode contacts in the normalized post-implantation MRI data sets were determined, which were further processed using custom programs implemented in *MATLAB*. The centers of electrode artifacts were identified and marked using views of horizontal, coronal, and sagittal MRI slices (Fig. 6.3A, B). This procedure allowed obtaining MNI coordinates of all electrode contacts (Fig. 6.3E). The course of the individual central and lateral sulcus was also determined in the same MRI data sets (Fig. 6.3E, yellow dots). These sulci were used as anatomical landmarks that indicate the borders between the frontal, parietal and temporal lobes in the first step of the hierarchical probabilistic assignment (see below).

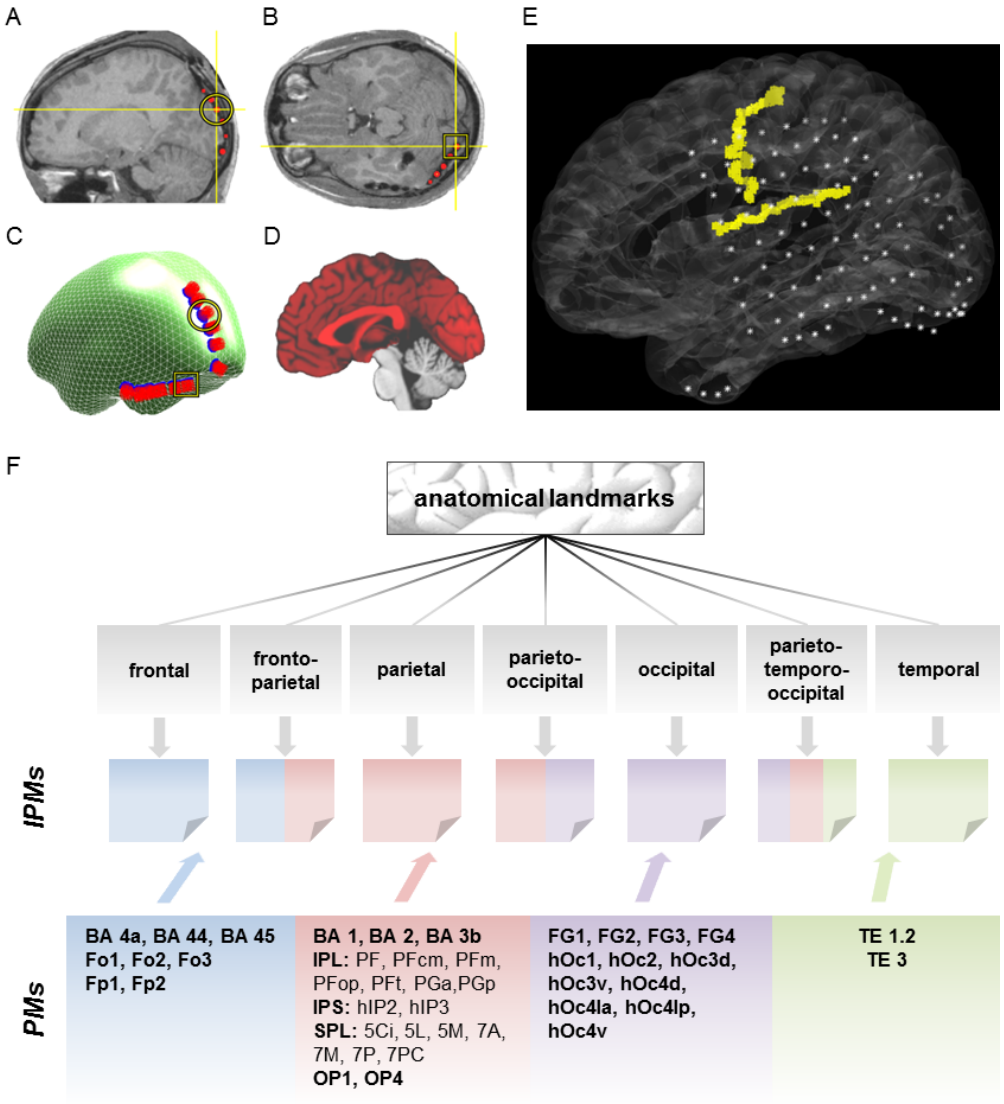

Figure 6.3: Hierarchical probabilistic assignment. The cerebellum was removed to allow assignment of occipito-basal electrodes (examples from P11). **A** and **B**: Horizontal and sagittal views of an electrode void artifact in a post-implantation MRI. Yellow crosshairs: center of one void artifact. **C** Cortical projection using orthogonal vectors (red) on the cortical hull. **D** Median-sagittal slice of the Colin standard brain (red) with the cerebellum removed (grey). **E** Electrode positions (white stars) in MNI space and with respect to the central sulcus visualized in yellow. **F** Operating principle of the HPA (BA: BRODMANN Area, all other abbreviations as in the *SPM Anatomy Toolbox v2.2c*). According to anatomical landmarks derived from the individual post-implantation MRI, the electrodes are assigned to lobes or lobar poolings. According to the performed assignment, exclusive MPMs are generated that are subsequently used for the probabilistic assignment.

6.1.4 Assignment Procedures

The iEEG electrode assignment method presented here is a modified version of the probabilistic assignment (PA) method of SPM for iEEG. In contrast to this standard PA the hierarchical method uses anatomical landmarks, which are higher in the assign-

ment hierarchy than the probabilistic maps and is therefore a hierarchical probabilistic assignment. Note that whenever PA or standard PA is mentioned, this includes both the standard assignment method according to *SPM Anatomy Toolbox* but *including* preceding cortical projection, which is a specially developed method here, see below. This is done to investigate the influence of the addition of hierarchy to the method. Fig. 6.3 gives an overview of the main steps of the electrode assignment procedure ELAS that is presented in this chapter. The decisive steps of this method are explained in detail in the following section:

Step 1, individual pre-assignment: Since the anatomical landmarks central sulcus (CS) and lateral sulcus (LS) are definite boundaries for the adjacent anatomical areas, the here presented assignment method ELAS was designed with the aim of using the information on the exact individual course of the lateral sulcus and the central sulcus hierarchically first. Both sulci are derived from the individual normalized post-implantation MRI data (Fig. 6.3E) and enable the lobar allocation of the electrode contacts according its position relative to the individual CS and LS (Fig. 6.3F), whereby for regions posterior to the LS the horizontal ramus is used as a landmark [242].

According to this lobar pre-assignment, specifically composed individual probability maps (IPMs) are then used to complete the hierarchically organized probabilistic assignment. In Fig. 6.4 an IPM is shown exemplary for two areas: Area 1 (Fig. 6.4A) and area IPC (Fig. 6.4B) are used to calculate an IPM (Fig. 6.4C) that contains both areas. These IPMs contain only the areas of the brain that come into question for the respective lobe. An IPM based on a combination of parietal, temporal and occipital areas is used if electrode contacts are located posterior to the posterior end of the horizontal ramus. All areas without contact to the outer cortical surface are also excluded when subdural implanted ECoG electrode contacts should be assigned.

Electrode contacts positioned directly on the CS or the LS likely cover several distinct areas located in both banks of these sulci, e.g., Area 4 and Area 3a in case of the CS [378], but can still be assigned to the most probable area. For each electrode contact the probabilities of all areas in question are still available and thus, allow the probability for each individual electrode contact to be estimated for each area in question.

Step 2, cortical projection: ECoG electrodes are implanted in the subdural space directly on the cortical surface, while the probabilistic anatomical information refers to the brain itself. To take this into account, positions of the electrode contacts are projected onto the cortical surface, which will be described hereafter. Generally, the *mesh_shrinkwrap.m MATLAB* function [357] was used to generate a smooth cortical hull. The cerebellum was removed before generation of the hull, as ECoG strips could also be implanted on the posterior basal surface of the brain, between the cortex and the cerebellum (Fig. 6.3D). The average side length of triangles of the final hull is $0.726\,mm$. On this surface, a patch with a $5\,mm$ radius around each electrode position is determined. This radius is chosen due to the expected error associated with electrode localization after implantation, which lies in the range of $5\,mm$ [238, 273, 370]. This $5\,mm$ fits also very well with the results regarding the differences in the spatial location between the normalized pre- and post-operative MRIs (see Tab. 6.2). Vectors orthogonal to the surface of the hull are determined at all hull vertices within the $5\,mm$-radius patch ($\sim 10^3$ vertices, Fig. 6.3C).

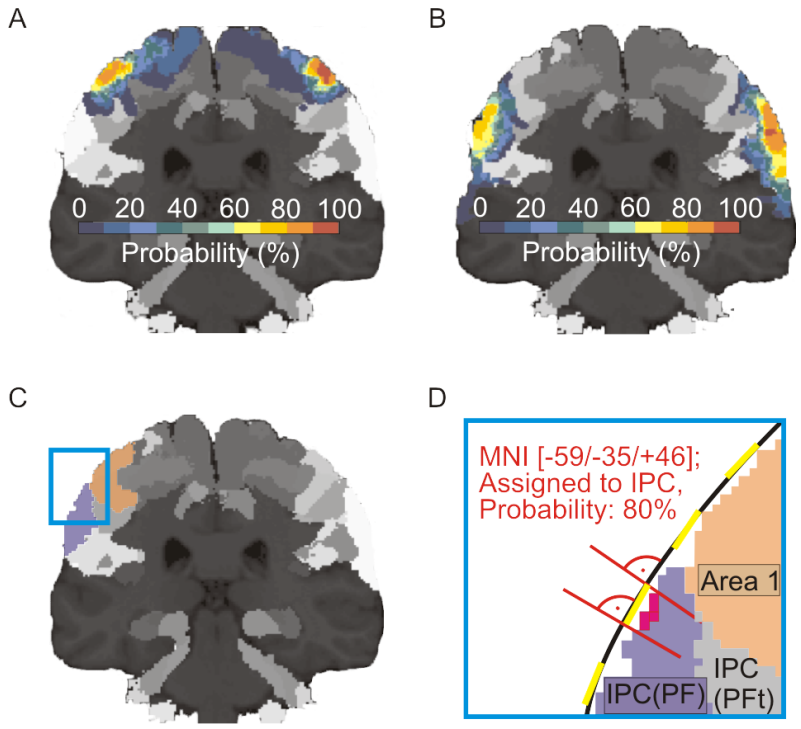

Figure 6.4: Probabilistic cytoarchitectonic assignment of ECoG electrodes. **A** Cytoarchitectonic probability map of Area 1 visualized on a standard brain. Probability of each individual voxel to be located in Area 1 is color-coded. **B** Cytoarchitectonic probability map of the parietal area IPC (PF); conventions as in (A). **C** The resulting map, derived from the probability information presented under (A) and (B), shows at which positions these areas are more likely $> 50\,\%$than other areas. Orange: Area 1; blue: IPC (PF). The cyan box indicates the region magnified in the following panel D. **D** To assign ECoG electrode contacts using the IPMs, a method is described based on surface orthogonals (red) of a smoothed 3D cortical hull (black) fitted through the ECoG electrode positions (yellow). This allows the definition of cortical voxels beneath the individual electrode contacts (magenta) and assignment of electrodes to the most likely brain areas according to the IPMs.

The intersections of these orthogonal vectors with the cortical surface are used to assign electrode contacts to the outer brain surface of the lateral convexity (Fig. 6.4D). However, search along the surface normal is restricted to a maximal depth of $10\,mm$, as the hull surface was within this distance in all patients investigated.

6.1.5 Application Examples

High-Gamma Mapping

For a detailed description of the time-frequency analysis methods used for topographic high-gamma mapping (Fig. 6.6D-F), see [296]). Summarized, time-resolved spectral magnitude was calculated using a multitaper method [266]. Trial-averaged magnitude changes in the gamma frequency range ($60 - 400\,Hz$ for P3; $60 - 128\,Hz$ for P1; dependent on respective sampling rate) were computed for a time window comprising the first $500\,ms$ after the start of a movement or speech production. Spectra were calculated

relative to the baseline activity of the first $200\,ms$ in the pre-event period, i.e., $1\,s$ or $2\,s$ before the onset of speech production or movement, respectively. Saccade-related spectral magnitude changes (Fig. 6.9) were calculated using the same multitaper method with a $100\,ms$ sliding window, $20\,ms$ time steps and 5 SLEPIAN tapers. Spectra were calculated relative to baseline activity between 250 and $50\,ms$ before saccade onset.

Electrical Stimulation Mapping

In one patient (P2), the hierarchically organized probabilistic electrode assignments are compared to location and extent of functional areas as identified during electrical stimulation mapping (ESM). Data of one patient was used in order to exemplify the potential use of individual macro-anatomical information and probabilistic cytoarchitectonic information to investigate cortical reorganization. ESM was performed in a clinical context in order to delineate eloquent cortices prior to resection of epileptogenic foci. The ESM procedure is described in [296].

6.2 ELAS Toolbox

The ELAS toolbox is an open source graphical user interface (GUI) developed in *MATLAB* and comprises the methodology utilized and described in this chapter. ELAS is built on *SPM12* (Statistical Parametric Mapping, October 2014) and the *SPM Anatomy Toolbox*, with *SPM* being initialized parallel to the toolbox directly after its function call. The installation of *SPM* and *SPM Anatomy Toolbox* is required for the functionality of the ELAS algorithm. If needed, the *SPM Anatomy Toolbox* can be started separately inside the GUI and utilized simultaneously. The ELAS toolbox is compatible with any version of the *SPM Anatomy Toolbox*, including possible future updates.

The main emphasis of the open source package ELAS lies on the hierarchical probabilistic assignment of iEEG electrodes, especially subdural ECoG electrodes, to the underlying brain areas in MNI (Montreal Neurological Institute) space as well as their visualization. The ELAS toolbox provides the export of a header file in *MATLAB* format, containing all information about the individual electrode contacts and their neuroanatomical assignment. Furthermore, ELAS enables a three-dimensional visualization of the electrodes and relevant brain areas, mapped on the ICBM152 standard brain [228], as well as an individual export of all mentioned objects as a wavefront OBJ file for the use in virtual reality. The interface is designed for an intuitive use by providing online help for each step of the assignment procedure. Beginning with the normalization of pre- and post-implant MRI, the workflow of the entire procedure is depicted in Fig. 6.1.

The process can be started at each individual step of the workflow, if the intermediate results of the prior step are available. *SPM* is used for co-registration, normalization and segmentation of the images, which serve as a basis for the marking of the individual electrode contacts and sulci. After importing the MNI coordinates into ELAS, the toolbox provides an interface to label the electrode contacts and to perform a lobar pre-assignment (see Fig. 6.1 and Fig. 6.5A). Latter information is used for the (hierarchical) probabilis-

tic assignment according to the cytoarchitectonically defined brain areas from the *SPM Anatomy Toolbox*. For the assignment of ECoG electrodes, the novel hierarchical probabilistic algorithm is utilized, whereas for SEEG electrodes the conventional probabilistic algorithm without projection is applied.

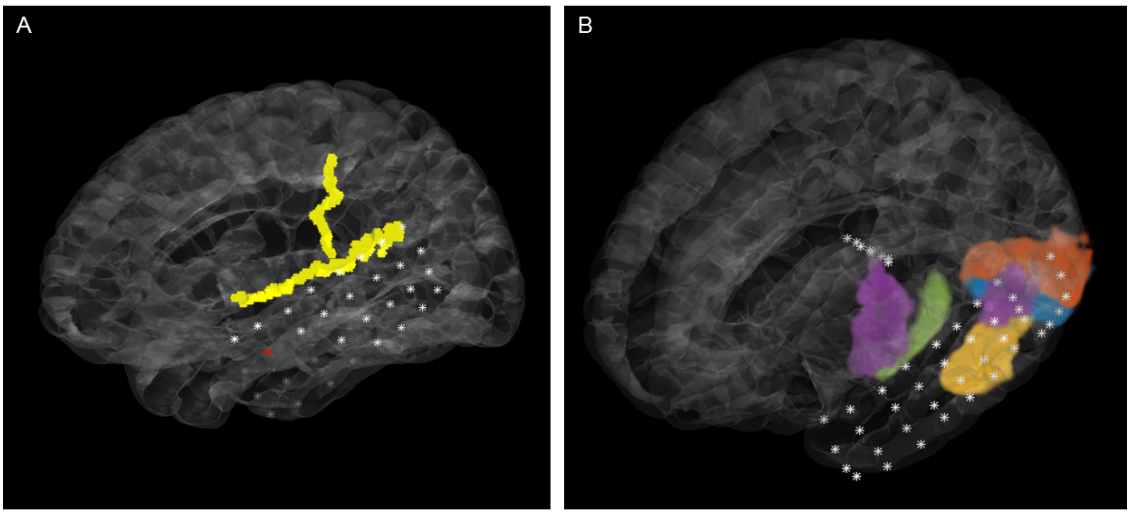

Figure 6.5: Visualization in ELAS. A Interface for labeling of intracranial electrodes. **B** The standard brain, the cytoarchitectonically defined brain areas, and the electrode contacts can be visualized in 3D.

Basically, the MNI coordinates of the electrode contacts and the marked anatomical landmarks are needed for the first step of the computation. They are used to perform the lobar pre-assignment, that ensures an area assignment based on the brain's individual characteristics. The matrix of MNI coordinates $\mathbf{C}$ and the lobar pre-assignment are then utilized to feed the HPA algorithm, shown in Alg. 9. After the cortical projection of an electrode contact, surface normals of all hull vertices, lying within a distance $d < d_{max} = 5\,mm$ to the cortical projection, are calculated. Then, the most probable areas are determined according to areas in the closest vicinity of the surface normals and according to the individually computed probability maps IPMs. The IPMs contain coordinate wise possible areas and the according probabilities. For each electrode contact i the algorithm outputs the probable areas $\mathbf{A}_i$ that might be recorded by the electrode contact as well as the according probabilities $\mathbf{p}_i$.

The visualization makes use of the ICBM152 standard brain and enables a display of preselected areas jointly with different electrode groups. The preselection of areas can be limited to either all most probable areas of a patient, all cortical areas or the entire set of areas provided by the *SPM Anatomy Toolbox*. Inside the three-dimensional, rotatable display, each of the preselected areas and electrode groups can be turned on and off interactively (see Fig. 6.5B). In the visualization mode, the shown electrode contacts are based on the real MNI coordinates, not on the projected ones. Furthermore, the brain template as well as the brain areas and the electrode contacts can be individually exported as a wavefront OBJ file and used in virtual reality setups. The exported objects then rely on the (smoothed) surface extraction of each selected object.

Algorithm 9 hierarchical probabilistic assignment algorithm

Require: electrode contact coordinates $\mathbf{C}$, lobar pre-assignment of $\mathbf{C}_i$, cortical vertices $\mathbf{F}$, maximal distance d_{max}

for all electrode contacts **do**
  request lobar pre-assignment for contact i
  $\mathcal{M} \leftarrow$ compute individual probability map (IPM)
  $\mathbf{V}_0 \leftarrow$ vertex $\mathbf{F}_m$ closest to coordinate $\mathbf{C}_i$
  $\mathbf{V} \leftarrow$ all vertices $\mathbf{V}_n \in \mathbf{F} \mid d(\mathbf{V}_n, \mathbf{V}_0) < d_{max}$
  for all vertices $\mathbf{V}_j$ **do**
    $\mathbf{n} \leftarrow$ compute surface normal on cortical hull towards center of brain
    if $\mathbf{n}$ hits any area in $\mathcal{M}$ **then**
      $\mathbf{W}_j \leftarrow$ coordinate of area strike closest to $\mathbf{V}_j$
    end if
  end for
  $\mathbf{W} \leftarrow$ vector of area coordinates
  for all $\mathbf{W}_k \in \mathbf{W}$ **do**
    $\mathcal{B}_k \leftarrow$ area names of $\mathbf{W}_k$ according to $\mathcal{M}$
    $\mathbf{Q}_k \leftarrow$ probabilities of coordinate $\mathbf{W}_k$ to lay in areas $\mathcal{B}_k$
  end for
  $\mathbf{A}_i \leftarrow$ vector of unique areas in $\mathcal{B}$ for electrode contact i
  $\mathbf{p}_i \leftarrow$ probabilities of unique areas, mean of $\mathbf{Q}$ over all $\mathbf{W}_k$
end for
return $\mathbf{A}, \mathbf{p}$

6.3 Assignment of ECoG Electrodes I

In this section, the results are described concerning the assignment of ECoG electrodes to some of the major areas of interest in ECoG studies on the outer surface of the frontal, parietal, temporal and occipital lobes. In the following section, as an application examples it is shown how probabilistic assignments can improve the interpretation of ECoG responses during hand movements, speech production and saccadic eye movements. Finally, results of a case study on cortical reorganization are presented as a further application example of probabilistic analysis of electrode locations.

In the large majority of cases ($> 80\,\%$ of all 687 grid electrode positions on the lateral convexity), the hierarchical probabilistic assignment to a brain area on the outer surface of the frontal, parietal, and temporal lobes was possible. The remaining approximately $20\,\%$ of electrodes were either in regions not yet included in the probabilistic atlas (approximately $12\,\%$) or directly above the CS or the LS (approximately $8\,\%$), hence making their assignment to one particular area problematic. Thus, although current probabilistic maps are not yet available for the entire cortex, a hierarchically organized probabilistic assignment is suitable to provide a neuroanatomical framework for the interpretation of ECoG responses. Especially electrode contacts located over the sensorimotor areas of the lateral convexity can already be successfully assigned with the approach (Fig. 6.6).

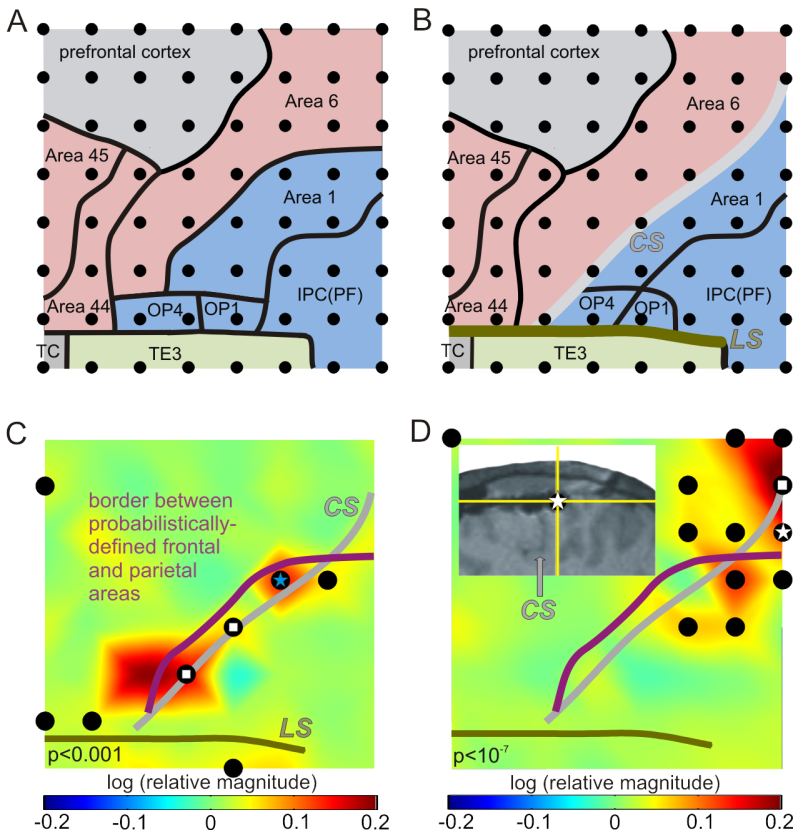

Figure 6.6: Standard PA (A) and hierarchical PA (B) results for a 64-contact ECoG grid (P3). Black dots: electrode contacts; black lines: borders between areas; blue: frontal areas; yellow: parietal areas; red: temporal areas; grey: areas not covered by the currently-available probability maps. TC: temporal cortex; all remaining abbreviations as in the *SPM Anatomy Toolbox v1.8*. Grey and olive lines: central sulcus (CS) and lateral sulcus (LS) derived from individual post-implantation MRI data and used in the hierarchical PA of ELAS (B). **C** High-gamma ($60 - 400\ Hz$) brain responses during speech production of P3. Black dots: significant responses (see [296] for further details). Grey and olive lines: CS and LS derived from individual structural MRI as in B; purple line: fronto-parietal border resulting from standard PA (i.e., only using probabilistic atlas information but not the individual course of the CS and LS). Cyan star: electrode with significant response located pre-central in the individual MRI, but post-central according to the standard PA. White squares: electrodes with significant response located on the CS in the individual MRI, but pre- or post-central according to the standard PA. **D** as (C) for contralateral arm movements (data as in [296]). The individual MRI (insert) clearly shows a postcentral position of the electrode marked by the white star in the activity map, in conflict to a precentral assignment according to the standard PA.

The influence of anatomical information in the form of the major sulci CS and LS was investigated, obtained from individual post-implantation MRI data, on the accuracy of the anatomical mapping. To this end, the hierarchically organized probabilistic assignment with ELAS was performed where the course of the CS and LS to pre-assign electrode contacts to the frontal, parietal, or temporal lobes was used and the results were compared to the ones of the standard probabilistic assignment (PA) procedure from SPM. For instance, an electrode contact was wrongly assigned to the postcentral region with the standard probabilistic assignment (Fig. 6.6A), but was clearly located precentral and consequently assigned to the precentral region Area 6 by using the ELAS approach

(Fig. 6.6B). At this electrode contact, speech production-related gamma band responses were recorded (Fig. 6.6C). Similar examples for arm movements are shown in Fig. 6.6D, respectively. In the latter case, two electrode contacts located in the frontal lobe were wrongly assigned to the parietal cortex with standard PA. The post-implantation MRI of this patient clearly showed that these electrodes were positioned on the precentral *hand knob*, which indicates the hand motor cortex [325] and consequently they were correctly assigned to the motor cortex with the ELAS method.

In Fig. 6.7 the results of a systematic analysis of the impact of the inclusion of individual macro-anatomical landmarks of the ELAS approach on lobe assignments for all 687 electrode contacts on the outer surface of the frontal, parietal, and temporal lobes are shown. Compared to standard PA, with the ELAS method more electrode contacts were assigned to the frontal lobe, while fewer electrodes were assigned to the parietal lobe (Fig. 6.7A). Approximately 10 % of electrodes differed in their lobe assignment when using the ELAS method instead of standard PA. These results indicate that inclusion of individual anatomical information is important for the correct lobe assignment of a substantial portion of ECoG electrode contacts. Fig. 6.7B illustrates the direction of changes in lobe assignments when using ELAS instead of standard PA. Such changes most frequently happened between the frontal and parietal lobes, followed by changes between parietal and temporal lobes, while only a few electrodes changed assignments from frontal to temporal lobes and vice versa.

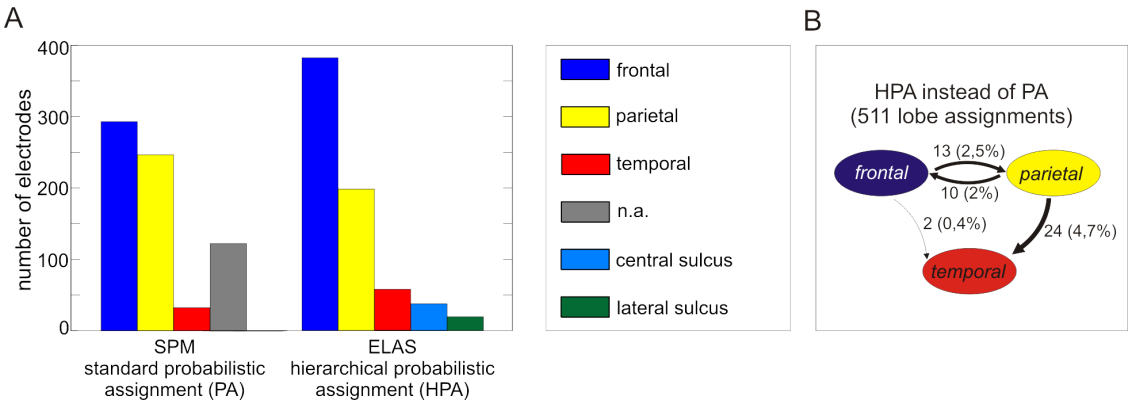

Figure 6.7: Impact of hierarchy. A Impact of of the inclusion of individual macro-anatomical landmarks of the ELAS approach on lobe assignment. Using standard PA, some contacts are not assigned to any lobe, as the probabilistic atlas (cf. [107]) is not yet complete. With the ELAS approach, all contacts could be assigned to a lobe based on the position of the CS and LS in the individual MRIs, except for those located directly on the CS or LS. **B** Direction of lobe assignment changes between lobes for both assignment methods are shown.

Results so far indicate that the presented hierarchically organized probabilistic assignment ELAS can be applied to typical ECoG grids above the sensorimotor cortex and may provide useful information for the interpretation of brain responses in this region. The electrode assignment approach for iEEG differs to the standard probabilistic assignment method for functional imaging data in two decisive steps: first, the inclusion of individual anatomical information and second, a projection onto the cortical surface as ECoG elec-

trodes are implanted outside the brain directly on the surface of the cortex. The impact of these two additional components on the anatomical assignment of iEEG electrode contacts was analyzed (Fig. 6.8). To this end, the results of the ELAS method were compared to the results of standard PA method and with a direct probabilistic assignment (dPA), that corresponds to the standard PA method without using the cortical projection.

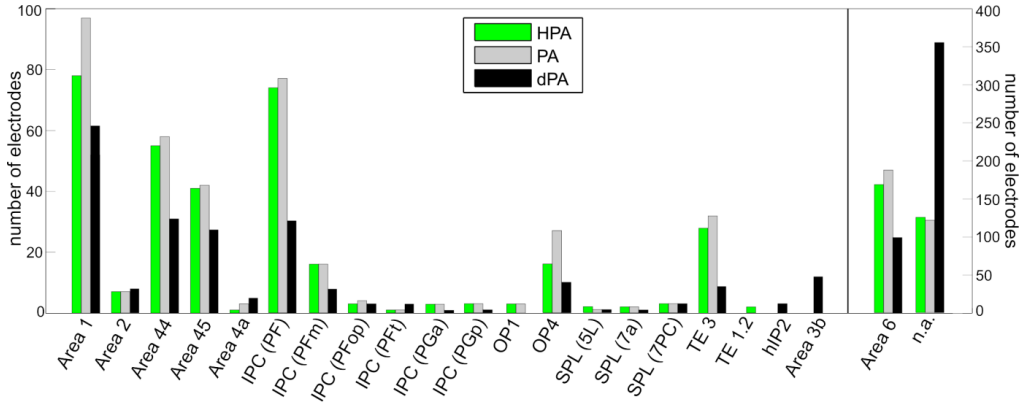

Figure 6.8: Results of HPA, PA and dPA of 687 **grid electrodes.** Abbreviations for areas as in the *SPM Anatomy Toolbox*.

Differences between ELAS and standard PA were particularly evident close to the CS and LS, such as in area 1, area 6, IPC (PF), OP1 and OP4. In all of these areas, the total number of electrode assignments was higher for standard PA compared to ELAS, owing to the fact that with ELAS a total of 37 and 19 electrode contacts were lying directly on the CS and LS, respectively. A large percentage of electrode contacts ($51.4\,\%$) was not assigned to any cortical area when using the dPA method (compared to $18.2\,\%$ using ELAS and $17.6\,\%$ using standard PA), due to the fact that they were located outside of the brain. This example demonstrates that the cortical projection is a crucial step in the application of probabilistic assignments to ECoG data.

6.4 Normalization of Post-operative MRI

To validate the accuracy of the normalization to MNI space of the post-operative MRIs (with electrode artifacts), the spatial difference was compared to the corresponding images of normalized pre-operative MRIs (without implanted electrodes), see Fig. 6.2B. The normalization tools from *SPM12* are best suited for the normalization of MRIs without implanted electrodes, thus the mismatch between these two kinds of normalized MRIs is defined as inaccuracy of the normalization of post-operative MRIs. Since the coordinates of the electrode contacts are taken from the post-operative MRIs, the inaccuracy of the normalized post-operative MRI is consequently also a localization error of the electrode contacts.

The averaged (median) correlation values, standard deviation values and 3D distance with and without masking unreliable values (corrcoeff<0.8; std<0.3) are shown in Table 6.2. The average 3D-distance between pre- and post-operative MRI is between 3 and $5\,mm$ (mean over patients: ≈ 3.9) and lower (≈ 3.3) when unreliable values were disregarded. It should be noted that the highest localization errors were observed in the direction perpendicular to the cortex surface, which should be automatically leveled out by the cortical projection of the electrode contacts.

6.5 Assignment of ECoG Electrodes II

So far, the assignment of ECoG electrodes on the outer surface of the frontal, parietal, and temporal lobes was described. Recordings in these regions have been utilized in a large number of previous ECoG studies, such as on motor and sensory processing [84, 85, 187, 214, 296, 340]. Recordings from occipital regions, however, have also been utilized in previous ECoG studies, particularly on visual processing in occipital brain areas [149, 206, 230, 241]. Probability maps of early visual areas are also available in the *SPM Anatomy Toolbox*, which makes the electrode assignment approach of electrode contacts to these visual areas an interesting tool for ECoG research related to visual processing. The HPA method was used in five patients with implanted strip electrodes over the occipital cortex. In Fig. 6.9 the results of the HPA method on two electrode strips of P11 is shown with electrode contacts assigned to the early visual areas V1-V4.

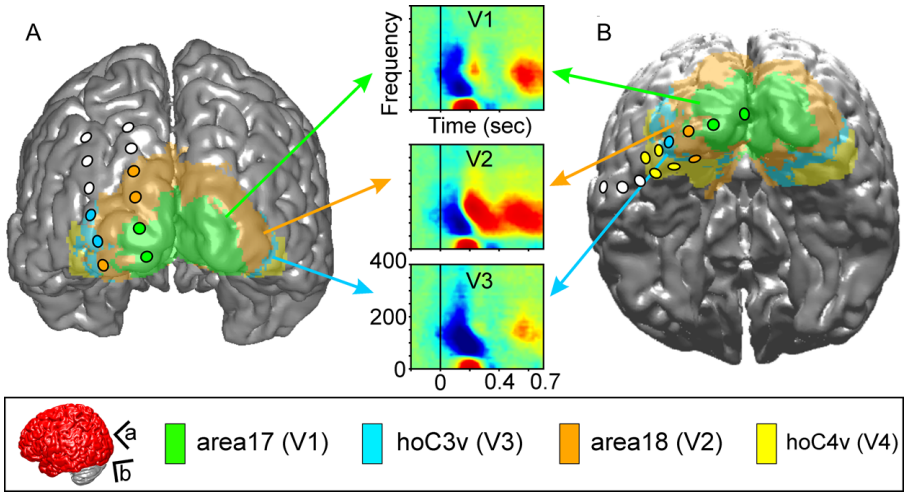

Figure 6.9: Assignment of an occipital subdural strip electrode to visual areas and time-frequency spectra of saccade-related brain activity (P11). A Surface extent of MPMs of areas 17, 18, hoC3v (V3) and hoC4v (V4) are shown on a standard brain surface. Electrode positions of two occipital subdural strips are marked as dots in the same color as the corresponding probabilistically-defined areas. White dots show electrode positions not assigned to any area. **B** Same as (A) but for two occipito-basal electrode strips. The viewing angle used in (A) and (B) is illustrated in the insert (bottom). Middle panels: Average time-frequency spectra of brain activity recorded at electrodes illustrated in (A) and (B) during saccades. Relative magnitudes were averaged over all recordings of electrode contacts located in the respective area. Using the probabilistic method HPA, saccade-related responses are shown for each of the areas V1-V3.

Saccade-related brain responses were used to show the usefulness of these probabilistic assignments. The saccade-related activity pattern averaged over electrodes of each visual area (V1-V3) consisted of an onset-related magnitude decrease in high-gamma frequencies, followed by an increase in magnitude (also see [340]). The comparison between visual areas enabled by the probabilistic assignment of individual electrode contacts revealed differential regional patterns, e.g., while the high-gamma decrease at saccade onset was most pronounced in area V3, the subsequent increase in magnitude was most prominent in V2.

6.6 Investigating Cortical Reorganization

Both conventional non-probabilistic and probabilistic brain atlases refer to the anatomy of the healthy brain, while ECoG is recorded from patients, mostly in the context of the pre-surgical evaluation of epilepsy. This fact must be considered in the interpretation of the resulting assignments (see Conclusion), as epilepsy can lead to reorganization of functional brain areas [2, 76, 203, 322]. Fig. 6.10 exemplifies for P2 how comparison between electrical stimulation mapping (ESM) results and the HPA method can be used to analyze potential consequences of such cortical reorganization. Here, an unusually large upper-extremity motor representation was revealed by the ESM. By comparing these ESM findings with the HPA results, it became evident that the ESM-defined motor area clearly extended into the prefrontal cortex, possibly due to reorganization induced by a focal cortical dysplasia in the superior premotor area (shaded area in Fig. 6.10).

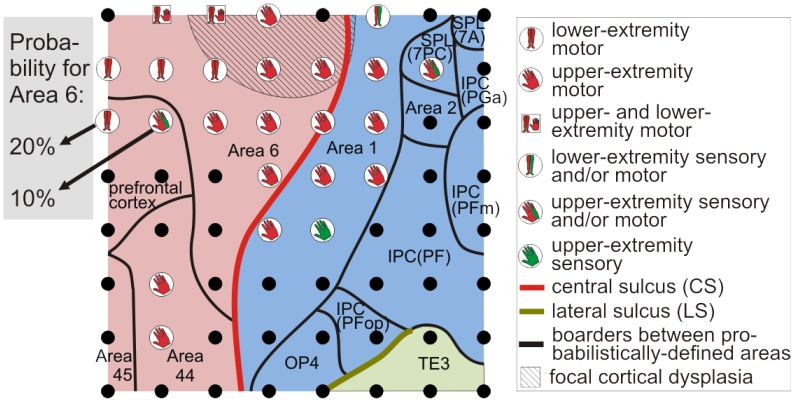

Figure 6.10: Functional brain regions revealed by ESM and electrode assignments to cytoarchitectonic areas using HPA (P2). For convenience, some areas are labeled with abbreviations (7A: SPL(7A), 7PC: SPL(7PC), PGa: IPC(PGa), BA 2: Area 2 and PFop: IPC(PFop)). Hand and leg symbols indicate ESM effects. Frontal areas are illustrated in red, parietal areas in blue and the temporal areas in green; abbreviations as in Fig. 6.8. HPA assignment clearly showed that sensorimotor responses extended into the prefrontal cortex, possibly due to reorganization induced by a focal cortical dysplasia in the superior premotor region (shaded area).

6.7 Related Work

Intracranial electroencephalography (iEEG) including Electrocorticography (ECoG) and stereo EEG (SEEG) plays an increasingly important role for exploration of human brain function. ECoG provides detailed information about cortical activity with good spatial and very high temporal resolution [109]. Such data have been used to study cortical function with respect to motor [47, 84, 233, 296], visual [121, 204, 227], language [53, 72, 95, 213, 314, 336], auditory [59, 85] and also viscerosensory [187] processing. Of late, such data are analyzed in the context of social cognitive neuroscience, and they also provide a potential control signal for brain-machine interfaces [20, 215, 274, 301]. Many current iEEG studies use atlas-based neuroanatomical assignment to interpret the position of electrode contacts. Atlas-based neuroanatomical assignments are used to map the individual electrode contacts onto cortical areas that are associated with specific tasks, such as primary motor cortex, premotor cortex, auditory cortex, BROCA's area, etc., and thus to establish a neuroanatomical context of the iEEG results.

Several methods have been recently proposed for atlas-based neuroanatomical localization and conventional assignment of iEEG electrodes [46, 88, 148, 155, 161, 233, 370]. They combine structural imaging data (X-rays, intraoperative photography, pre- and/or post-implantation magnetic resonance imaging (MRI), computed tomography (CT)) and standard brain atlases, such as the TALAIRACH atlas [327]. Other studies have reported on techniques to visualize iEEG electrodes relative to the individual cortical anatomy derived from pre- or post-implantation imaging data [51, 105, 150, 201, 223, 238, 273, 307, 324, 361]. However, a shortcoming of standard brain atlases, such as the TALAIRACH atlas, is that they do not provide information on the variability in the position and extent of cytoarchitectonically-defined brain areas between individuals [11]. This shortcoming cannot be resolved by techniques that visualize iEEG electrodes relative to the individual cortical anatomy, as the position and extent of cytoarchitectonically-defined brain areas varies relative to the available macroscopic cortical landmarks within the cortical lobes [9, 10, 378].

A probabilistic neuroanatomical atlas system has been developed to address these issues [107]. This system has already been highly successful in the area of functional neuroimaging, particularly in assigning functional MRI (fMRI) peak coordinates to cytoarchitectonically-defined brain areas. It relies on the histological processing and the resulting microanatomical definition of cortical areas in 10 human post-mortem brains. Up to the present day, this analysis has been completed for a large set of brain regions (Fig. 6.3), including the primary motor cortex [138], somatosensory areas 3a, 3b, 2, and 1 [139, 140, 147], BROCA's region [10, 12], the auditory cortex [237, 282], the premotor cortex [137], the parietal operculum [106] and visual areas [11, 225, 295]. Such cytoarchitectonic maps of individual areas can be combined in a maximum probability map (MPM), a non-overlapping representation of probabilistically- defined areas (Fig. 6.4C)[108]. Probabilistic maps and MPMs are freely available in the *SPM Anatomy Toolbox* [107]. Despite the success of these methods in the field of functional neuroimaging, it is currently unclear how they can be utilized in the growing field of iEEG studies.

6.8 Conclusion

This chapter introduces the novel open source *MATLAB* toolbox ELAS for assignment of intracranial electrodes, making use of cytoarchitectonically-defined brain areas as well as individual macro-anatomical landmarks. The included visualization tool allows for flexibly generating 3D displays of electrode contacts with respect to the relevant neuroanatomy. It is described how probabilistic neuroanatomical assignment procedures that were previously developed for analysis of neuroimaging data can be adapted to iEEG. The results show that the iEEG-adapted probabilistic assignment can be successfully used with recordings from a wide range of cortical areas in the frontal, temporal, parietal, and occipital lobe. More than $80\,\%$ of the several hundred individual electrode positions analyzed on the lateral cerebral convexity could be probabilistically assigned to the respective underlying brain area. Probabilistic neuroanatomical assignment has several advantages over conventional, non-probabilistic methods: First, it allows assignment to cytoarchitectonically-defined brain areas that are not reliably delineated by macroanatomical landmarks of the cortex [9, 10, 378]. Second, it takes into account the inter-individual variability of position and extent of these areas [107]. Thus, the advantages of probabilistic assignment as previously used in neuroimaging studies are adapted for iEEG research. Still, not all electrode contacts were successfully assigned with PA and HPA to cytoarchitectonically defined brain areas for the following reasons: First, electrode contacts positioned directly above the CS or the LS (e.g., see Fig. 6.6) were not probabilistically assigned, as in these cases electrode contacts are expected to record brain signals from both banks of the sulci. However, the probabilisitc information is still available, taking areas from both banks of the sulci into account. Second, not the whole lateral convexity is mapped in the currently available probabilistic atlas system (e.g., see Fig. 6.3F and Fig. 6.8). New areas, however, are constantly added to the atlas and thus, the number of unassigned electrode contacts will gradually decrease over time.

The question of how structural information derived from individual post-implantation MRI data could be combined with the probabilistic information was investigated and whether this approach further improved the reliability of the outcomes. To combine individual structural and general probabilistic information, a hierarchical procedure has been designed, where in the first step, electrode contacts were pre-assigned to cerebral lobes based on their position relative to the LS and CS in the respective post-operative MRIs and in the second step, probabilistic information was used for within-lobe assignments (cf. Fig. 6.4, Fig. 6.3). The CS is the borders between the frontal and parietal lobe, the LS the border between the fronto-parietal lobes and the temporal lobe, respectively [87, 366, 378]. Importantly, both the position of electrode contacts and the position of the LS and CS were determined in the same MRI data set. Thus, the effects of normalization inaccuracies could be avoided since the electrode positions relative to the sulci obtained in this manner are generally considered as *ground truth*. HPA avoided wrong lobar assignments in approximately $10\,\%$, compared to standard PA (Fig. 6.6, Fig. 6.7) and also affected the assignments to individual cortical areas in the vicinity of the LS and CS (Fig. 6.8), reflecting the well-known high inter-individual variability of major sulci [378]. It can be concluded that inclusion of individual structural information is feasible

and further improves the neuroanatomical assignment in iEEG.

A general limitation of atlas-based localization of iEEG electrode contacts is that the available atlases represent the healthy brain, while iEEG is measured in neurological patients, mostly in epilepsy patients. Epilepsy may be secondary to a brain tumor or a focal cortical dysplasia (FCD, [38, 58, 116, 117, 118, 294]) with pronounced pathological changes of cortical anatomy. Therefore, it is important to keep in mind that atlas-based localization, both conventional and probabilistic ones, can only provide information about the brain structures that can be expected at a given location in standard anatomical space in the healthy brain. It is noteworthy, however, that this limitation does not apply to the lobar assignments in HPA as described in the present study, since here the individual patient's MRI brain morphology is used. It is also noteworthy that application of a neuroanatomical framework, derived from the healthy brain, to iEEG data from epilepsy patients is not necessarily just a limiting factor, but also opens up possibilities to investigate reorganization of cortical function (Fig. 6.10). Both lesional and non-lesional epilepsy can lead to reorganization of functional areas [2, 76, 203, 322]. Lesion-related reorganization has been investigated previously, combining anatomical and functional MRI with cytoarchitectonic probabilistic maps [321]. Along the same lines, here findings of the electrical stimulation mapping (ESM) in a patient with an unusually large ESM-defined motor area are shown, similar to previous reports [203, 226, 322, 340]. A combination of ESM findings and HPA showed that ESM motor responses were with a high probability located in the prefrontal cortex (and a low probability of being located in the premotor cortex; Fig. 6.10).

Atlas-based localization requires normalization to the atlas space [290], in this case the MNI space. Previous studies applied a range of techniques for normalizing iEEG electrodes to MNI space, such as based on MRI-CT coregistration [105, 161] or utilizing post-implantation MRI data [201, 370] as in the present study. Spatial normalization of patients' post-operative MRI is a topic of research in itself [52, 91, 184, 308, 339] but it was not in the focus of the present study. However, by using the normalized pre-operative MRI as ground truth spatial localization errors of the electrode contacts from the post-operative MRIs between 2 and $5\,mm$ were observed, which is in line with results from previous related studies [238, 273, 370]. This indicates that if atlas-based anatomical analysis of iEEG data in general, and probabilistic methods in particular, would find a more widespread use in iEEG research, this would also spur further interest in optimizing the spatial normalization of structural imaging data, particularly of post-operative MRIs.

In summary, this chapter demonstrates how probabilistic assignment procedures that were previously developed for analysis of neuroimaging data can be adapted to the iEEG and used for the neuroanatomical interpretation of iEEG recordings. With the continuous extension of the available probabilistic atlas system by new areas, the applicability of probabilistic maps will further increase. Thus, the here presented probabilistic anatomical assignment method might be a valuable addition to iEEG research. The results in the following chapter profit, among other things, from the methods of the Toolbox. The classification into specific areas comes from the hierarchical probabilistic assignment and is used as a basis for the neurophysiological interpretations. The ELAS toolbox is freely available under GITHUB (*https://github.com/joosbehncke/elas*) for non-commercial and academic use and comes with a detailed documentation for installation and application.

Chapter 7

Spectral attributes of Neural Error-related Patterns in iEEG

The recognition and processing of errors forms the basis for all human learning processes and has been the subject of neuroscience for some time. The complex relationships and interactions in the human brain are difficult to decipher and the question of origin and dynamics of the underlying signal keeps this area in suspense. Both temporal and spectral responses play a decisive role in this process. This chapter compares spectral error-related brain activity of participants performing both an ERIKSEN *flanker task and a car driving task, and benefits from measurements with intracranial electrode contacts. The 1552 electrode contacts exhibit comprehensive coverage of a wide range of brain areas and thus provide a global insight into brain areas involved in error-processing. It turns out that simultaneously activated regions lie mainly in frontocortical areas, but also in the anterior cingulate cortex (ACC), and several so far not to error-processing related areas exhibited a spectral response. Furthermore, a power increase in high gamma band (HGB) could be demonstrated, and in addition spawning the error-related power increase as a dominant feature.*

The first chapters dealt with the question of whether and how errors or robot types can be detected by means of human EEG. Based on convolutional neural networks a framework could be implemented, which tops conventional methods used for comparison. The convolutional neural network was generally formulated and proved that it can learn different features, such as temporal characteristics (Chap. 4 and Chap. 5) or features based on spectral responses [303]. Beside the possibility to improve the decoder, the signal can also be optimized. One approach would be to use invasive measurements, as it will be discussed in the following chapters. Beside the neurophysiological basics (this chapter) also the decoding will be in the foreground (Chap. 8). In Chap. 6 a new method for assigning electrode contacts was introduced, completed by an intuitive toolbox interface for non-programming users. The aim of this method is to optimize the assignment of electrode contacts especially to brain areas for ECoG electrodes in subdural space and to include the individual characteristics of the brain, based on anatomical landmarks, in the calculation. The method was used to interpret the data underlying this chapter. For the development of an understanding of cerebral processes, the most exact possible assignment to the involved areas is essential.

As already mentioned, many studies are mainly based on results of non-invasive measurements. However, these allow above all the investigation of the near-surface regions. For a holistic understanding of the generation and transmission of signals as well as the network interaction of different areas, a global view of the problem is indispensable and intracranial measurement is a key to deeper insights into the underlying processes. In addition, the question arises to what extent the modality of the stimulus influences the spectral activity of error-related signals.

To eliminate these problems, data based on intracranial recordings was analyzed for their spectral responses. The unique opportunity existed to collect data from 15 epilepsy patients who gave their voluntary and informed consent to participate in two different paradigms. The generated data set comprises recordings of a total of 1552 electrode contacts for two different paradigms. Both paradigms aim for the incorrect execution of a task instructed to the participant, but differ in their proximity to everyday situations. In order to obtain a comprehensive overview, the results were not only evaluated for each paradigm isolated, but also put in relation to each other, and checked for their spatial significance. The study shows that spectral error processing also seems to be apparent in higher frequency bands up to around $115\,Hz$ and that the error-related power increase turns out to be a dominant feature. This could also be manifested as similarities for different modalities of an error stimulus. Furthermore, the broad coverage of electrode contacts revealed the multitude of areas that contribute to error-processing and subprocesses. As a result of the investigation of spatial distribution of the simultaneously contributing areas for both paradigms, predominantly frontocortical regions but also the anterior cingulate cortex (ACC) exhibited significant activity.

7.1 System and Experimental Design

To create the considerable set of intracranial EEG (iEEG) recordings, in which error-related brain activity is accessible, two different paradigms were designed. This time, the focus of the paradigm's concepts rather lay on erroneous performance of the user himself than on the observation of faulty execution of an instructed task by a robotic effector. Thereby, intracranial methods allow tasks in which movements and muscular artifacts don't affect the quality of the data, as the signals are extracted directly from the brain's tissue. This enables a more active contribution of the participants in the task and a higher grade of empathy referred to occurring errors. In a first paradigm, a flanker task is supposed to elicit error signals under strictly experimental conditions, while in a second paradigm a car game-like environment simulated a more real-life situation. Each participant took part in both experimental paradigms, what led to the exceptional possibility to compare two paradigms in iEEG. In general this is quite difficult, as different patients exhibit distinct implantations covering diverse brain areas. The paradigms are depicted in Fig. 7.1.

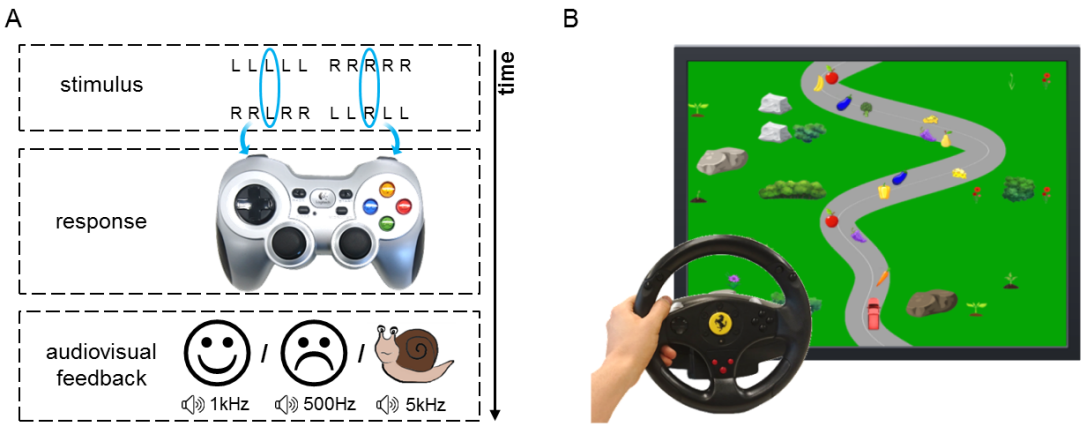

Figure 7.1: Two different paradigms to elicit error-related responses. A A schematic sketch of the paradigm using an ERIKSEN flanker task, adapted from [133]. **B** Modified screen shot of the car driving task, in which the participant has to collect rewards and avoid collisions with obstacles (here represented by fruits and vegetables) while keeping the car on the road.

7.1.1 ERIKSEN flanker task (EFT)

This task was designed according to [110]. The participants had to respond to the middle character (either R or L) of a set of letters, acting under time pressure. The audiovisual feedback was given according to a right or a wrong button press, or a reaction slower than a predefined time limit (see Fig. 7.1A). The time limit was set individually to the mean reaction time of a training phase. For details see [351]. An error was defined as a wrong button press, while a right button press was cited as correct. What is called an event in the following, however, is the time of the feedback that was given after about $2\,s$ regarding the button press. In order to avoid the influence of the cerebral response to the button press on the feedback response, the feedback was given with that delay.

7.1.2 Car driving task (CDT)

The second paradigm consisted of a car driving task in which participants were instructed to stay on a road while avoiding certain obstacles (e.g. bombs) punished with a negative score and collecting beneficial objects (e.g. coins) rewarded with plus points (see Fig. 7.1B). As the speed of the game was fixed, the participant's goal consisted of achieving a highest possible score when reaching the finish line. In this task, an error event was traced when an obstacle was hit; when a beneficial object was hit, the event was declared correct.

7.1.3 Participants and Data Acquisition

In this experiment the data was raised from intracranial recordings, based on intracranial EEG. The group of voluntary participants was exclusively formed by patients suffering

from epilepsy. Altogether 47 participants fulfilled the car driving task of which 15 also accomplished the ERIKSEN flanker task, each having different implantation and number of electrode contacts. The recordings of these 15 participants (7 female) were employed for the analyses and comparisons of this study, comprising altogether 1552 electrode contacts. All participants provided their informed consent for the study, which was approved by the local ethics committee. Furthermore, the experiments were accomplished under clinical supervision in the epilepsy center in Freiburg, Germany, as well as in the epilepsy center of the Motol University Hospital in Prague, Czech Republic. At the epilepsy center in Freiburg the signals were recorded with the COMPUMEDICS amplifier (Singen, Germany) at a sampling frequency of $2\,kHz$, while the epilepsy center of the Motol University Hospital in Prague made use of the SCHWARZER EPAS amplifier (Munich, Germany) and the NICOLET EEG C-series amplifier (Pleasanton, USA), recording at a sampling rate of $512\,Hz$. A more detailed description of the underlying methods and used hardware for the inquiry of the intracranial brain data can be found in subsection 2.2.2.

7.2 Pre-processing & Statistics

According to unique trigger pulses, generated during each experiment, the acquired data were aligned to the event-related meta information. The aligned data were re-referenced to a common average and high-pass filtered using a 3$^{\text{rd}}$ order Butterworth filter with a lower cutoff frequency of $0.1\,Hz$. Following, the recorded event stamps determined the different trials and the data were cut and re-sampled to $250\,Hz$. Previous to the analysis in frequency domain, spectral decomposition of the trials was performed by reference to the multitaper method, see Sec. 2.4 and Eq. (2.19). Hereby, a window size of $0.5\,s$ and a step size of $0.05\,s$ was chosen. In order to test the significance of correlations, a permutation test was applied, see section 2.6 and Alg. 7. In any other case, e.g. when referring to the significance of an activity, a WILCOXON ranksum test (MATLAB: *ranksum*) over trials was used. For both methods the significance level was set to $\alpha = 0.01$. Assignment to anatomical areas was accomplished using the methods according to Chap. 6, also described in [33]. Subsequent in this study, recordings will be parsed for temporal development of activity. Hereafter, $t = 0\,s$ represents the moment of the error-related event and temporal placement of any activity will always refer to this event.

7.3 Error-related Activity in iEEG

To investigate the influence of error-related action on the recorded brain signals, time-frequency analysis was performed for each participant and both paradigms. Fig. 7.2A shows the responses to the committed mistakes in the time-frequency domain for both paradigms, depicted for two exemplary electrode contacts I4 in the Insular Cortex and S15 in the Postcentral Gyrus. The color code represents the log of the relative power (erroneous vs. correct action) and can be seen for a frequency range up to $200\,Hz$.

For this exemplary participant P1, the patterns shortly after the appearance of an error ($t = 0\,s$) exhibit a remarkable similarity for some electrode contacts (here as an example represented by contact I4), while others showed observable responses for only one (e.g. S15) or none of the two paradigms.

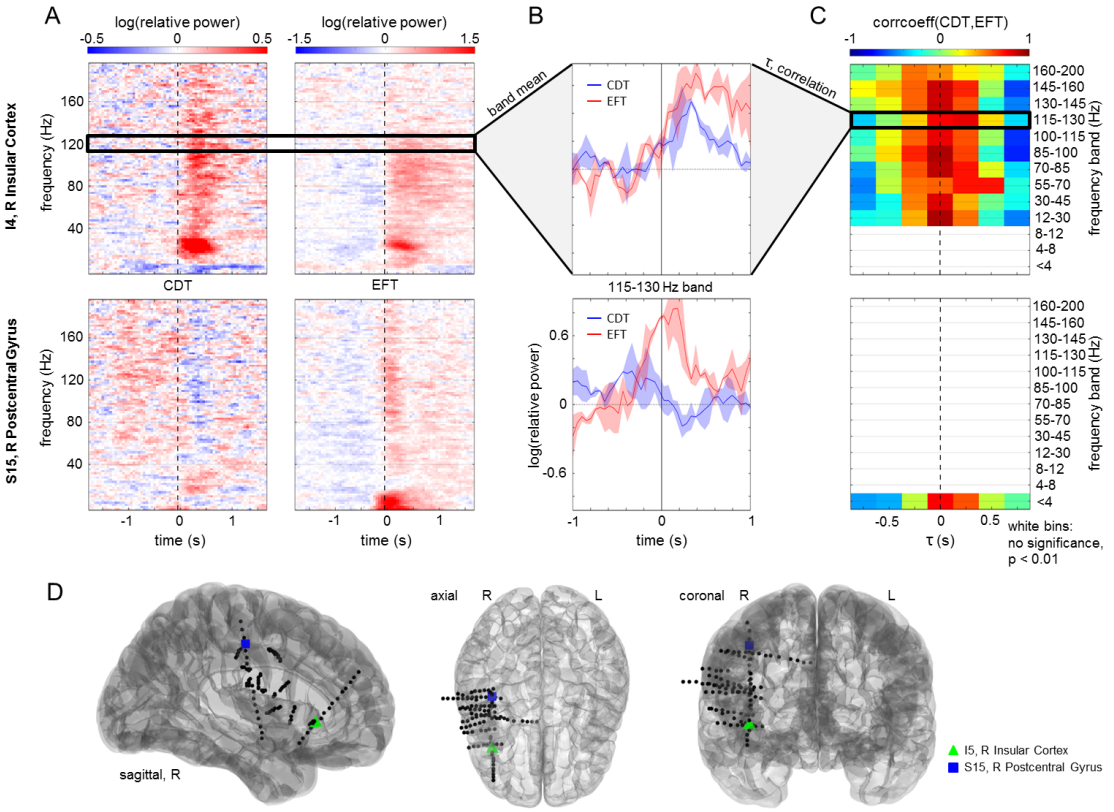

Figure 7.2: Error-related activity of intracranial electrode contacts for an exemplary participant P1. **A** Time-frequency analysis (see Sec. 2.4) for both paradigms (CDT: left, EFT: right) and for two selected electrode contacts where either (1) similar frequency modulation across several bands can be seen in both paradigms (top: I4, R Insular Cortex) or (2) characteristic activity is only observable in one paradigm (bottom: S15, R Postcentral Gyrus). $t = 0\,s$ marks the appearance of an error. **B** Mean time course of the logarithm of the relative power of frequency range $115 - 130\,Hz$ (blue line: CDT, red line: EFT). Standard error of the mean is represented in light coloured areas. The contacts are arranged as in A. **C** Correlation of frequency band dependent power time course of both curves, CDT and EFT. The color value of a bin is obtained by correlation with regard to a certain time offset τ of the two curves, white bins indicate no significance ($p < 0.01$). The contacts are arranged as in A. **D** Position of the exemplary electrode contacts according to the *ICBM152* standard brain [228].

In order to extract a measure for resemblance of the frequency responses for the different paradigms, the time course of previously defined frequency bands were averaged over the band and compared pairwise, see Fig. 7.2B. This made a comparison on the level of frequency bands available. At this, the standard error of the mean was calculated based on bootstrapping methods [236]. These methods are used if the theoretical distribution of underlying statistics is unknown. Moreover, the extracted time courses for the different paradigms were correlated for several time offsets defined by the variable τ. The results

for the two electrode contacts can be seen in Fig. 7.2C. Again, the similarity in I4 becomes obvious and especially above $12\,Hz$ the time courses show a comprehensive high correlation around the error event. However, the two signals of S15 exhibit rather poor similarities, what reflects the gained impression of the spectrograms in Fig. 7.2A quite well. White bins indicate that the power didn't show significance for both of the paradigms (WILCOXON ranksum test, $p < 0.01$). The spatial distribution of all electrode contacts of participant P1 over the *ICBM152* standard brain [228] for 3 different views is illustrated in Fig. 7.2D. The exemplary electrode contacts from Fig. 7.2A-C are marked as a green triangle (I4) and a blue square (S15).

7.4 Common Error-related Spectral Patterns

To obtain a more profound insight into the distribution of common patterns inside the spectra, the spatial information of the data was analyzed on the basis of each time-frequency bin. For each bin, the vector over all channels was correlated for the two data sets. Again, the correlation was not only performed for simultaneous data points but also for shifted values obtained by the time offset τ. An example of the results of the analysis for participant P1 is illustrated in Fig. 7.3A. White areas indicate values with no significance ($p < 0.01$), which was determined by the means of a permutation test, permuting the channel vector randomly for 1000 times and correlating each of the result to the true distribution of channels. This elucidates how often the correlation of the random channel compilation exceeds the correlation for the real data. Especially for no offset and small offsets a strong correlation in lower frequency bands is distinct during the first $500\,ms$ after the event. The statistical values for all participants enter the analysis that leads to Fig. 7.3B. Here, the sum over the participants of all statistical significances is calculated for each time-frequency bin and applied in color code. Likewise in this comprising visualization a comprehensive correlation for frequencies up to $80\,Hz$ is obvious for small time offsets, beginning with the appearance of an error.

The temporal trend of the significant time courses during an interval of -0.5 to $1\,s$ was investigated per band and channel, evaluating whether a similar significant increase or decrease is present in both paradigms, see Fig. 7.4. A significant increase or decrease occurred when, within a running time window of $20\,ms$, the absolute value of the average power exceeded the standard deviation of the total interval of -0.5 to $1\,s$. Besides, the analysis of this section is accomplished to give an impression of how the trends for the two paradigms relate to each other. Channel by channel and band by band this information is broken down for an exemplary participant in Fig. 7.4A. Red positions inside the grid indicate an increase and a decrease for both data sets, dark orange points to a common increase and light orange points to a common decrease. In case of no common trends, the yellow coloration stands for a situation where the paradigms exhibit an opposing trend, purple stands for an increase or decrease in only one of the sets and white for no significance (WILCOXON ranksum test, $p < 0.01$).

In addition, the way of presentation provides information about the spatial distribution as well as about the frequency bands that are affected by error-related trends. Similarly

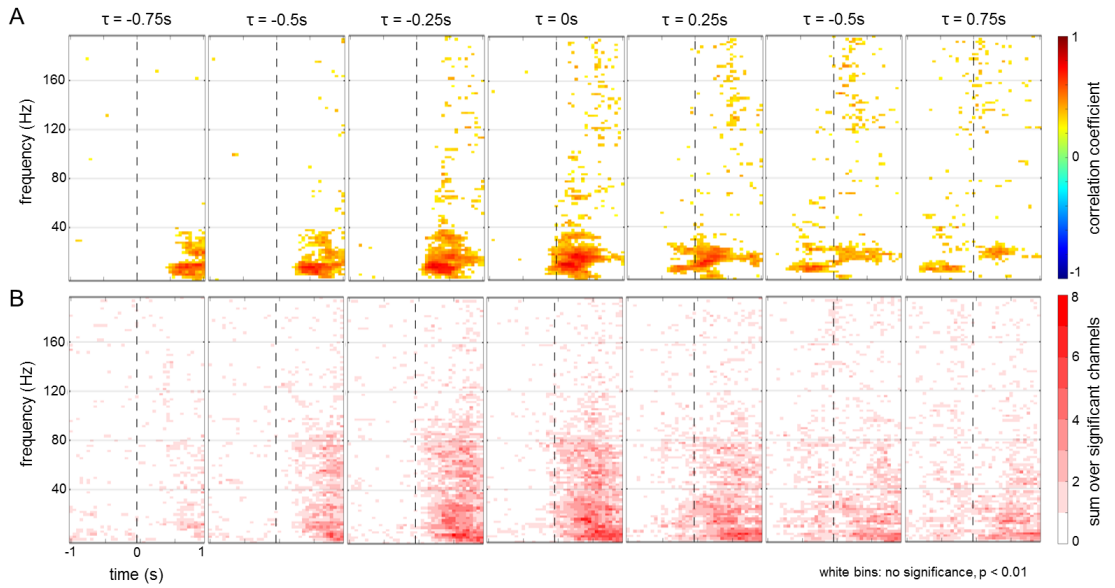

Figure 7.3: Spatial time-frequency similarities between CDT and EFT. A Correlation of significant time-frequency bins of all channels between CDT and EFT for an exemplary participant P1. White bins indicate no significance ($p < 0.01$). The correlation of the different depictions are calculated with regard to a certain time offset τ between CDT and EFT data. **B** Sum over significant time-frequency bins over all participants. Significance is determined by values extracted from A. The sum of the different depictions is calculated with regard to a certain time offset τ between CDT and EFT data. **A & B**: $t = 0\,s$ marks the appearance of an error.

to the results presented in Fig. 7.3, the significant channels show salience in the lower frequency bands below $80\,Hz$. Fig. 7.4B gives a visualization of this information on the level of participants. For each participant and condition the appearances of trends were summed up over all channels (Fig. 7.4A, right) and entered color coded into the according plot (black box in condition *increase both* of Fig. 7.4B). The illustration reveals the dominance of similar increase over similar decrease of both paradigms. Thereby, the values per condition are normalized by the maximal number of counts over all participants and conditions, to obtain a more relative view on the results. This has the advantage that one can judge how relatively strong the effects are with the respective conditions. Furthermore, the comparison of number of significant effects per participant becomes more obvious. Hereby, the information about the averaged significant activity per channel is suppressed though. While for the other conditions activity can be mainly observed for lower frequencies, increase can be found quite comprehensively, but particularly between 55 and $130\,Hz$.

So far, the results particularly indicate one similarity in the appearances elicited by the two paradigms: a power increase over several frequency bands. Fig. 7.5 puts the results into the focus of the spectral distribution of significant trends. For each of the 13 frequency bands, the total number of appearances of significant power increases over all electrode contacts is gathered for each paradigm. For means of comparison, the total numbers are normalized by the highest representative within each condition. Obviously,

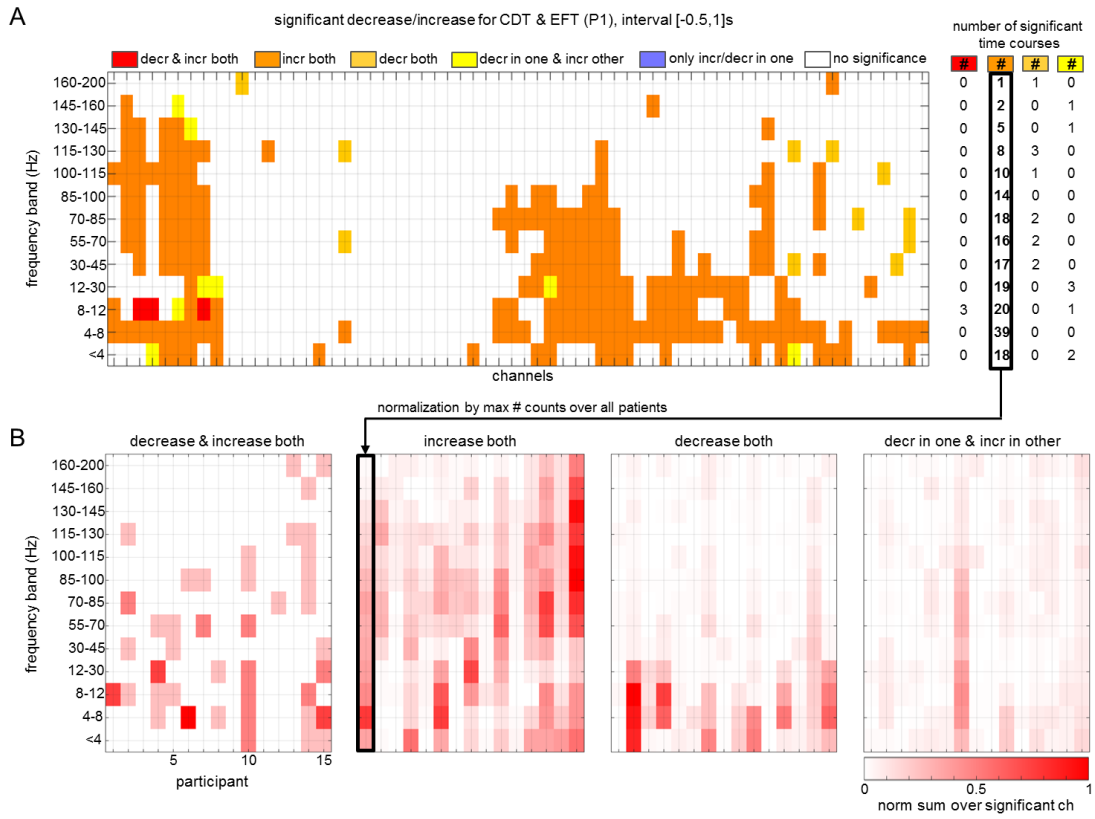

Figure 7.4: Significant increase and decrease of the relative power. A Significant trend of the time course of the logarithmic relative power for an exemplary participant, per frequency band and channel. Red indicates decrease and increase in both paradigms, dark orange increase in both, light orange decrease in both, yellow an increase in one but a decrease in the other paradigm, purple increase or decrease in only one paradigm and white indicates no significant trend. Right: Sum of significant appearances over all channels. **B** Normalized sum over channels of significant appearances per frequency band and participant. The Values per depiction are normalized by the maximal number of counts over all participants and conditions.

both paradigms elicit error-related power increase in delta ($< 4\,Hz$), theta ($4 - 8\,Hz$) and alpha band ($8 - 12\,Hz$). While in gamma range around $55 - 130\,Hz$ a moderate activity can be registered for the EFT recordings, the signals gained in the CDT paradigm show highest activities. By contrast, error-related increase in beta band ($12 - 30\,Hz$), low-gamma band ($30 - 45\,Hz$) and the higher gamma bands ($> 130\,Hz$) is rather sparsely represented. The question arises when the occurrence of a power increase can be observed according to a committed error.

In order to not only investigate the distribution of frequencies but to consider the temporal aspect, the results were examined with regard to the appearance of the power increases related to the moment of the error-related event, confirming the already observed characteristics. Fig. 7.6 shows both the temporal breakdown to frequency band and condition, and the aggregated appearances over all bands per condition. At this, the condition *CDT* (Fig. 7.6, left column) refers to all occurred significant power increases (all participants and channels) in the recordings of the car driving task, *EFT* (Fig. 7.6,

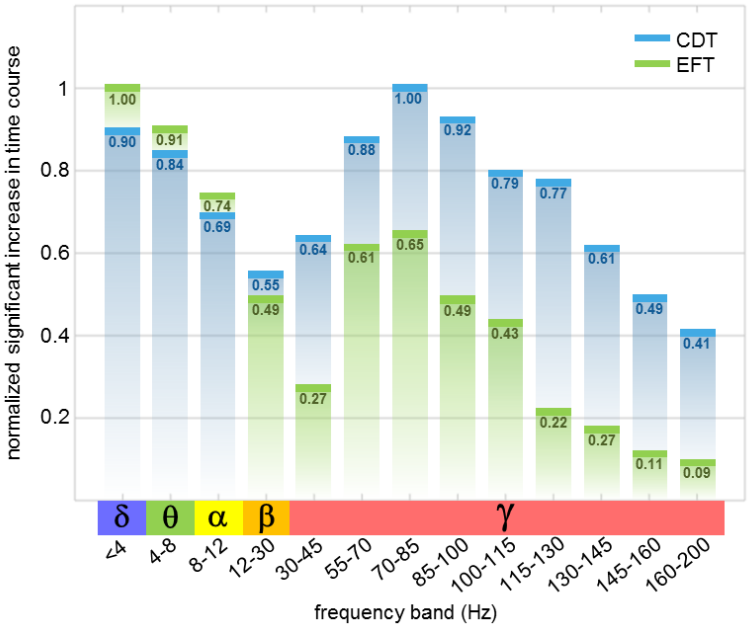

Figure 7.5: Normalized significant power increase. For both paradigms, CDT and EFT, the total number of appearances of significant power increases is depicted. The values are normalized by the maximal occurring value within each condition.

middle column) refers to ones of ERIKSEN flanker task recordings and *both* (Fig. 7.6, right column) describes the phenomena of coincident significant power increases in both paradigms for the same channel and frequency band. The maximal appearing number of increasing trends per condition is in each case represented by the maximal value of the orange highlighted panel ($max_{\mathrm{CDT}} : 190$, $max_{\mathrm{EFT}} : 71$, $max_{\mathrm{both}} : 27$). To allow a comparison per condition across frequency bands, the light blue values are normalized by the maximum of the condition max_i. On the other hand, the normalization by the maximum within a panel enables the examination of the time course per frequency band and condition. Those values are kept in dark blue. The lowest row in Fig. 7.6 shows the sum of the respective discrete bars, belonging to a certain time bin, over all frequency bands for one condition, normalized by the maximal value of the aggregates ($max_{\mathrm{CDT}} : 1772$, $max_{\mathrm{EFT}} : 403$, $max_{\mathrm{both}} : 171$).

A global time-wise consideration of the results of the analyses for the different paradigms especially points to an appearance of increasing trends of the power in the lower frequency bands delta ($< 4\,Hz$) and theta ($4 - 8\,Hz$). Likewise, for condition CDT this effect can be observed in middle gamma bands ($55 - 130\,Hz$), whereas for EFT the activity in those frequency ranges appears comparatively smaller. This relation is also reflected by the examination of the common activity in both paradigms (see Fig. 7.6, right column). For CDT the wanted activities in alpha ($8 - 12\,Hz$), beta ($12 - 30\,Hz$) and low-gamma ($30 - 45\,Hz$) band are comparatively moderate, which likewise can be increasingly seen for EFT. The appearance of significant power increase for CDT can be observed predominantly with begin of an event and is strongest represented right up

to approximately $0.5\,s$. For frequencies $> 12\,Hz$ this is also apparent for EFT, while here activity before and during the event can be increasingly observed for frequencies $< 12\,Hz$. In both conditions, a drop of the appearance of increases right before the event around $0.1\,s$ becomes evident. For the case of simultaneous activity in both paradigms the temporal distribution is more focused and exhibits highest activity in the range of $0 - 0.25\,s$. Taken as a whole, the observations likewise can be deduced from the sum over all frequency bands, see lowest row in Fig. 7.6.

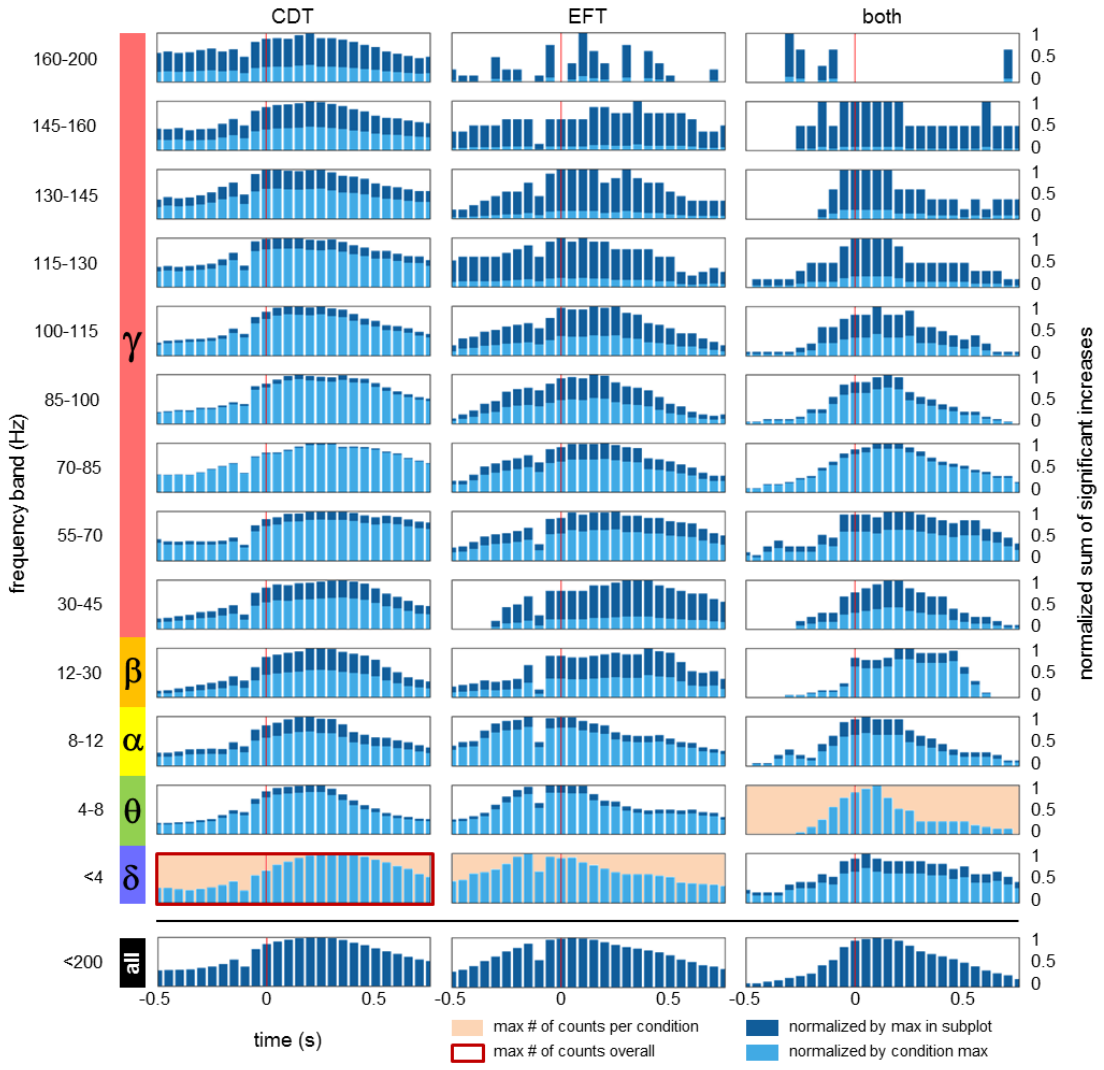

Figure 7.6: Temporal distribution of significant power increase for distinct frequency bands. Normalized sum (over participants and channels) of significant increases per band and condition CDT (left), EFT (middle) and both (right). Dark blue values are normalized by the maximal value within each panel, light blue values are normalized by the maximal value of all frequency bands within each condition (CDT: 190, EFT: 71, both: 27). The occurring of the maximal value per condition is designated by the orange background, the red frame marks the overall maximum. Bottom: depicted is the sum over participants, channels and frequency bands, normalized by the individual maximum (CDT: 1772, EFT: 403, both: 171). $t = 0\,s$ marks the appearance of an error.

7.5 Spatial Distribution of Error-related Power Increase

Aside from information about the temporal behaviour of error-related power changes, the spatial distribution of the activity over the brain constitutes an informative detail. Here again, the analysis addresses the error-related information concerning only increase in frequency band power. The results are concluded in Fig. 7.7, presenting brain areas with the significant increase for each paradigm or both simultaneously.

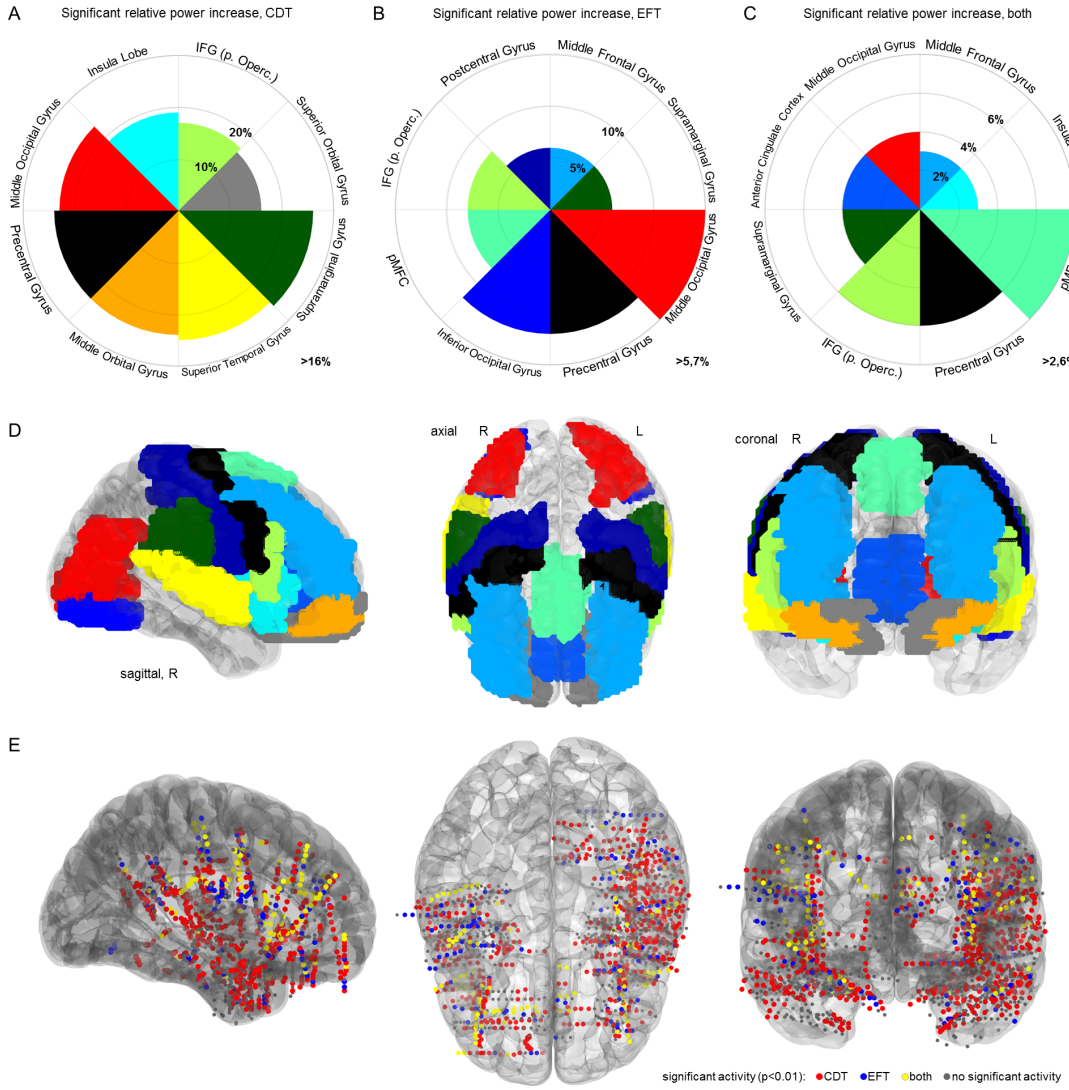

Figure 7.7: Spatial distribution of significant power increase for all participants. A-C Significant power increase in frequency-band-channel bins per area, in relation to the total amount of electrode contacts in this area, for the CDT (A), the EFT (B) and for the case that in both paradigms a significant power increase was observed in the equal frequency band and channel at the same time (C). A shows areas exhibiting proportions $> 16\,\%$, B areas exhibiting proportions $> 5.7\,\%$ and C areas exhibiting proportions $> 2.6\,\%$. **D** Point clouds of areas listed in A-C, exhibiting equal color code as in A-C. **E** Spatial distribution of all electrode contacts over the *ICBM152* standard brain.

Fig. 7.7A-C present in each case the eight areas with highest relative numbers of significant increase appearance for CDT (A), EFT (B) and both paradigms (C). The percentage is given relatively to the total number of electrode contacts that were located in this areas. The analysis was performed for 13 frequency bands, thus, a significant increase could have been observed in each band per electrode contact. As a consequence, the maximal amount of 100% is reached as soon as for every electrode contact in a certain area all 13 bands show a significant power increase in the interval of interest, $N_{max,A} = 13 \times N_A$, where N_A is the total number of electrode contacts in area A. The percentage displayed in Fig. 7.7A-C is then calculated by

$$\mathcal{N}_{rel,A} = 100 * N_{incr,A}/N_{max,A}, \tag{7.1}$$

where $N_{incr,A}$ represents the total number of increase appearances in a certain brain area A. Here, Fig. 7.7A shows areas exceeding proportions $> 16\,\%$, B areas exceeding proportions $> 5.7\,\%$ and C areas exceeding proportions $> 2.6\,\%$. Fig. 7.7D gives an overview of all areas listed in Fig. 7.7A-C for three different views. The colors of the different areas match the colors that represent distinct areas in Fig. 7.7A-C. Finally, Fig. 7.7E shows the spatial distribution of all channels over the brain (view angels equal to Fig. 7.7D), whereby the different conditions are color coded. The priority lies on cases where an electrode contact exposed increase in equal bands for both paradigms, coloring the location in yellow. Locations of electrode contact with an increase for CDT are colored in red, those with an increase for EFT are colored in blue, while contacts with no significant increase at all are displayed in grey. The depiction reveals an extensive coverage of the brain, with exception of occipital and parietal regions of the right hemisphere and around the occipital pole of the left hemisphere, due to implantation sites.

Tab. 7.1 gives an overview of all assigned areas that showed any significant increase in the frequency response for any of the conditions, according to the error-related stimulus. For columns *CDT*, *EFT* and *both*, the value indicates the number of channel-frequency-band pairs that exhibited a significant increase. As already mentioned, N_A represents the total number of electrode contacts in area A. $\mathcal{N}_{rel}$ gives the respective average of the relative power increase appearance over CDT and EFT,

$$\mathcal{N}_{rel} = 100 * \left(\frac{N_{incr,CDT} + N_{incr,EFT}}{2 * N_{max}} \right). \tag{7.2}$$

Bold printed areas are those appearing in Fig. 7.7A-C and Fig. 7.7D. $\mathcal{N}_{rel}$ represents an indicator for the occurrence of effects in relation to the total number of electrode contacts in a certain area. However, the values must be taken with caution especially for small numbers of electrode contacts. For small numbers, outliers carry more weight and values such as $\mathcal{N}_{rel} = 19.2$ for the Middle Occipital Gyrus are more difficult to class.

Table 7.1: Overview of significant power increases per area during the error-related task

Area	**CDT**	**EFT**	**both**	N_A	$\mathcal{N}_{rel}$
all	*2022*	*486*	*240*	*1315*	
Amygdala	4	1	0	15	1.3
Angular Gyrus	15	1	0	11	5.6
Anterior Cingulate Cortex	49	19	13	25	10.5
Cuneus	13	0	0	8	6.3
Fusiform Gyrus	44	0	0	70	2.4
Hippocampus	44	5	1	79	2.4
IFG (p. Opercularis)	58	27	21	27	12.1
IFG (p. Orbitalis)	28	8	1	37	3.7
IFG (p. Triangularis)	79	23	13	42	9.3
Inferior Occipital Gyrus	0	9	0	6	5.8
Inferior Parietal Lobule	8	4	0	9	5.1
Inferior Temporal Gyrus	53	4	0	96	2.3
Insula Lobe	263	54	37	106	11.5
Lingual Gyrus	7	3	0	13	3.0
Medial Temporal Pole	44	6	0	31	6.2
Mid Orbital Gyrus	8	0	0	8	3.8
Middle Cingulate Cortex	7	2	0	6	5.8
Middle Frontal Gyrus	138	53	28	69	10.6
Middle Occipital Gyrus	6	4	1	2	19.2
Middle Orbital Gyrus	53	6	5	17	13.3
Middle Temporal Gyrus	219	10	5	191	4.6
ParaHippocampal Gyrus	26	11	3	59	2.4
Postcentral Gyrus	97	50	16	64	8.8
Posterior Cingulate Cortex	3	1	0	8	1.9
Posterior-Medial Frontal Cortex	4	2	2	2	11.5
Precentral Gyrus	201	103	52	65	18.0
Precuneus	11	1	0	12	3.8
Rectal Gyrus	4	2	0	8	2.9
Rolandic Operculum	40	11	0	46	4.3
Superior Frontal Gyrus	5	3	1	4	7.7
Superior Medial Gyrus	6	1	0	3	9.0
Superior Orbital Gyrus	42	6	3	20	9.2
Superior Temporal Gyrus	255	9	7	80	12.7
Supramarginal Gyrus	168	38	29	50	15.8
Temporal Pole	20	9	2	26	4.3

7.6 Related Work

In order to understand human error recognition and processing, neuroscientists have addressed the questions of where and when, but principally how the error-related information is generated and subsequently transmitted in the brain. Previous studies revealed that human brain activity is modulated by both observation and commission of erroneous action. In both cases the brain signals exhibit specific activation patterns. *Falkenstein et al.* [113] and also *Gehring et al.* [133] observed a negative deflection approximately $50 - 100\,ms$ after an erroneous event, representing a component of the *event-related potentials* (ERP). The deflection is known as *error-related negativity* (ERN) or *error negativity* (Ne), and in the following is simply referred to as ERN/Ne. The appearance of the ERN/Ne is independent of the modality of the stimulus [114] or the modality the response [166]. Furthermore, a positive deflection can be shown, the *error positivity* (Pe) [113]. This positive deflection appears somewhat later as its negative counterpart and is recognizable up to $500\,ms$ after the erroneous event.

Still, the dynamics of the error processing and the functional hierarchy are not consistently defined. The most general interpretation connects the ERN/Ne with the reflection of a comparative process, contrasting a deliberately correct or intentional action to an actual, possibly erroneous action [113, 133]. In another theory the ERN/Ne is described as a reinforcement learning signal that is transmitted to the ACC [167]. According to this idea, the ACC merely receives a signal from the basal ganglia that evaluates action but does not mirror the error itself. The transmitted signal is then used to adapt the response selection process, which reflects the activity of the error processing system. However, the ERN can also be interpreted as a mirror of the evaluation of emotional and motivational importance of an error [63, 132, 258]. *Yeung et al.* [371] state the theory that ERN rather reflects conflict monitoring than an explicit signal that directly informs about the occurrence of an error, representing the continuous evaluation of the response conflict. In addition, investigations of the medial prefrontal cortex (mPFC) have led to a model that reinterprets the error-related effects as learning and outcome-prediction processes [5].

Initial research relates error-related activity to time-courses of the voltage recorded by scalp EEG [113, 133], but also spectral components have gained increasing attention. Spectral analyses of noninvasive EEG recordings, for example, have shown that the error-related activity partly originates from increased phase-locking of frontal midline theta activity [221, 222]. Further, synchronizations in theta band likely leads to network communication of action monitoring and cognitive control network interaction [71]. Apart from theta activity, error-related increase of delta activity could be observed [199, 372]. Here, *Yordanova et al.* [372] indicate that one of the processes reflected by the ERN/Ne represents the error-specific activity of the delta band. An overview of previous studies on spectral activity in error-related human EEG and iEEG recordings is given in Tab. 7.2.

Apart from the research into the nature of error-related signals, numerous studies have dealt with the question of which brain areas are involved in the processing of these signals. For example, the ERN/Ne seems to reflect the activity of the ACC [92], exhibiting a frontocentral distribution symmetric to the midline axis. Based on dipole models, source reconstructions have placed the neural source of the error signal in the medial

Table 7.2: Selection of previous studies about error-related spectral activity in EEG and iEEG. According choices of frequency band ranges are given due to deviations in literature

Publication	Signals	Spectral activity	Paradigm
Yordanova et al., 2004 [372]	64 EEG channels	**Delta (1.5-3.5 Hz)** power increase for ERN/Ne and **theta (4-8 Hz)** power increase for erroneous motor response	Four-choice reaction task, 14 subjects, 4 excluded due to low error rate.
Kolev et al., 2005 [199]	64 EEG channels	**Delta (1.5-3.5 Hz)** power increase and phase locking for ERN/Ne. Error-related **theta (4-7 Hz)** power increase in young subjects.	Four-choice reaction task, 10 young subjects (22.5 ± 1.5 years), 11 old subjects (58.3 ± 2.1 years).
Trujillo et al., 2007 [337]	25 EEG channels	**Theta (4-7 Hz)** power increase from -150 ms to 400 ms according to button press.	ERIKSEN flanker task, 21 subjects.
Cohen et al., 2008 [79]	12 iEEG channels, 28 iEEG channels	Error-related **theta (4-8 Hz)** power increase and **beta (15-30 Hz)** power suppression.	Modified flanker task, 2 epilepsy patients.
Carp et al., 2009 [69]	8 EEG channels	Absence of post-error **alpha (10-14 Hz)** power, compared to correct trials	Stroop task, 81 subjects.
Bastin et al., 2017 [27]	total 559 iEEG channels	Broadband **gamma (50-150 Hz)** activity increase due to erroneous action.	Stop-signal task, 6 epilepsy patients.
Völker et al., 2018 [351]	128 EEG channels	Error-related response in **low-frequency bands (<50 Hz)** and **high gamma (>50 Hz)**.	ERIKSEN flanker task, 35 subjects (4 rejected)
	total 690 iEEG channels	Error-related response in **low-frequency bands (<50 Hz)** and **high gamma (>50 Hz)**, followed by decrease in **high gamma**	ERIKSEN flanker task. 9 epilepsy patients.

frontal cortex (MFC), which might suggest an origin in the ACC [63, 311] or in the supplementary motor areas (SMA) [50, 92, 134, 166]. The assumption that the ACC plays a fundamental role in this is supported by results obtained with functional magnetic resonance imaging (fMRI), verifying error-related activity in the ACC [70]. Likewise in the medial frontal cortex [298], the pre-supplementary motor area [342] or in the prefrontal cortex (PFC) [244] activity regarding processing of errors and error-monitoring could be observed, while signals from the anterior insula might serve as an input for error-monitoring [154, 319] and are related to error-awareness [191, 341].

To a large extent, however, previous studies are based on the work with non-invasive

methods, which allow above all the investigation of the near-surface regions. For a holistic understanding of the generation and transmission of signals as well as the network interaction of different areas, a global view of the problem is indispensable and intracranial measurement is a key to deeper insights into the underlying processes. Furthermore, high gamma band $(50 - 200\,Hz)$ power modulations are more difficult to detect with surface EEG than with intracranial EEG, and it requires highly optimized data acquisition to be able to detect effects in a meaningful way [351]. But it is precisely in this area that activity seems to open up fundamental neural networks on both spatial and temporal scales [60, 84, 86]. In addition, the question arises to what extent the modality of the stimulus influences the spectral activity of error-related signals.

7.7 Conclusion

Error-related Activity in Intracranial Electrode Contacts

Spectral analyses concerning the committing of errors for two different paradigms revealed similarities in the spectral response patterns, partly in single frequency bands but also area-wide over several bands, see Fig. 7.2. The statistical evidence of these patterns was confirmed by correlation of the power time series of the respective bands for the two paradigms. Further similarities also manifested themselves in spectral patterns of spatial arrangements, see Fig. 7.3. For both paradigms, these similarities were simultaneously shown predominantly around the range of occurrence of an error-related event and extended over a broad spectrum of frequencies, predominantly and most strongly, however, in the frequencies $< 80\,Hz$.

Common Error-related Spectral Patterns

Looking at the power time series trends common to both paradigms for the respective electrode contacts and frequency bands, the dominance of increase over decrease becomes apparent (Fig. 7.4B). This connection is also already mentioned by *Völker et al.* [351], where among other things increase and decrease in the high-gamma band were observed for EEG and iEEG recordings. In the present study, a simultaneous increase in both paradigms can be observed in all frequency ranges, but this is most evident between $55\,Hz$ and $130\,Hz$. This temporal positive trend of power thus seems to reveal a comparatively high conspicuity with regard to the similarities of the two stimuli.

Depending on where the respective electrode contact is located in the brain, the two paradigms have in common an increase in power. On the other hand, a decrease in power is not so predominant, neither in the simultaneous consideration of both paradigms (Fig. 7.4B) nor in the separate analysis for the individual paradigms. Within the condition *decrease* it is noticeable, however, that for both paradigms activity is most frequently observed in the low frequencies ($< 30Hz$). So far, this effect has hardly been investigated but at least mentioned in some studies, e.g. [337]. However, this study focuses on the time and error related increases in spectral power.

The increase in power with the occurrence of the event was examined more closely, see Fig. 7.6. Looking at the individual frequency bands for each paradigm, the temporal occurrence of the increases shows a comparatively broad distribution around the event, especially in delta, theta and parts of high gamma ($55\,Hz - 115\,Hz$) band, whereas the occurrence in the higher frequencies is not as strong for the EFT, see also Fig. 7.5. Previous studies of error processing correlates show that an error commission is reflected by an increase in power in the delta [36, 37, 372] and theta [221, 222, 337] band. However, the current results are in line with *Kolev et al.* [199, 289], among others, that rather suggest both an error-related power increase in delta *and* theta. Theta is suggested to reflect more general response monitoring while delta represent an error-specific subcomponent [372]. Here the representative bands which show response to error-related events are extended by the high-gamma band, which was also recently described by *Völker et al.* [351]. In particular, the late components of high-gamma activity could indicate processes such as behavioral adaptation and motor learning, which follow the recognition of errors over time. Alpha and beta show comparatively lower, but nevertheless obvious activity regarding the error commission. In previous studies, however, a suppression of the activity in alpha [69] and beta [79] regarding error-related events is reported. A drop in power could also be observed here, but this remains to be investigated. For low-gamma ($30 - 45\,Hz$) and frequencies $> 115\,Hz$ increases are comparatively little to extremely rare represented, but still present in isolated cases.

When comparing the two paradigms, it becomes clear that the CDT is experiencing far more activity than the EFT (Fig. 7.6). This could be explained by the fact that different errors are perceived individually and the response is modulated both by subjective error significance [343] and by conscious error perception [364]. The activity in EFT seems to start earlier (especially for $< 12\,Hz$), even before the event, starting around $-0.25\,s$. The occurrence of activity just before the error is not uncommon [114] and the easier the error is to detect, the earlier the activity on the error is noticed. In addition, the participants fulfilling the EFT have often already noticed the error before, namely with the button press. This leads on the one hand to the fact that the error may no longer be evaluated so strongly and on the other hand to a kind of awareness of the upcoming error feedback [280, 281]. Likewise, this effect was already seen for lower high-gamma power in pre- and postcentral gyri, supramarginal gyri and insular cortex [351], indicating a pre-activation to the error.

One of the innovations of this study is the comparison of intracranial, error-specific spectral responses in human brain recordings of different paradigms. Especially with regard to the generalization of response patterns across different error types, the analysis of simultaneous activity provides information about similarities. At the frequency band level, simultaneous modulation for both paradigms shows comprehensive activity in the low frequencies ($< 30\,Hz$) and in high-gamma ($70 - 100\,Hz$), which is rather moderate for gamma between 30 and $70\,Hz$. The temporal occurrence of the spectral responses is mainly concentrated in the range of about $0 - 0.25\,s$.

Spatial Distribution of Error-related Power Increase

Reorganization and phase-resetting of oscillatory brain activity according to the stimulus happen to occur in several brain regions and at different times. At first glance, activity seems to occur comprehensively, see Fig. 7.7F, but this does not reflect the actual strength and abundance of the activity. The ACC in particular is frequently mentioned when it comes to the processes involved in error processing [27, 56, 355] and is supposed to handle reinforcement learning signals [167]. Further areas like the orbitofrontal cortex [338], the prefrontal cortex including the medial frontal cortex [27, 56, 79], the supplementary motor area [27, 50, 56] and the supramarginal gyrus [56] are reported to contribute in the error-monitoring network, while signals from the anterior insula serve as in input to this network [154, 319] and represent error awareness [191, 341].

In this study, there is a large overlap with the areas given in the literature and in addition the set of involved areas is extended (see Fig. 7.7A-D and Tab. 7.1). Also here the multi-layeredness of the underlying processes is confirmed. However, as already mentioned, especially the values of areas with a low number of electrodes in Tab. 7.1 should be treated with caution. The number of areas showing activity reveals the complexity of error processing in the human brain, in which many upstream and downstream processes, such as error detection, error evaluation, adjustment behavior or transfer to memory, play leading roles. Moreover, error processing does not simply seem to exist per se, but also to result in a series of physiological changes [83, 152, 153]. Further analyses can provide information here and also the investigation of the temporal occurrence can contribute enormously.

The high participation of occipital areas in EFT is conspicuous. This suggests that in this paradigm the visual processing played a greater role than in the CDT. In contrast, for electrode contacts in the orbitofrontal cortex, the response for the CDT was much more frequent, see also Tab. 7.1. It seems that in the more real-life game the immersion and thus the emotionality concerning the performance of the task is greater than in the more experimental task of the EFT, which can lead to activity in orbitofrontal regions among others [119, 128].

The number of common significant activity in both paradigms is relatively low and not that prevalent as for the single paradigm analysis, see Fig. 7.7E. However, a clear preference for frontocortical regions cannot be disregarded here, which can be derived from Fig. 7.7C and Fig. 7.7D. It is possible that these regions are involved in a general processing of the error-related signals, while other areas tend to respond to the individualities of the stimuli. The often mentioned ACC is also present for the simultaneous activity in both paradigms, what supports this thesis.

Outlook

In the rarely investigated field of intracranial, error-related spectral responses, the study confirms previous results and comes to new conclusions. Non-cortical regions such as hippocampus or insula cortex, which are difficult to measure by surface EEG, can hardly be measured by non-invasive methods [250]. Also, cortical processes are far more

complex than what can be derived by surface EEG [351]. The broadness of the study thus contributes to knowledge gains for regions that could hardly be and have hardly been investigated with regard to error processing.

The remarkable data set generated in this study includes measurements of 47 different patients who completed the car driving task, of which 15 additionally performed the ERIKSEN flanker task. This enormous set of intracranial recordings can be used to study motor and cognitive signals, such as here to analyze error-related responses. In this study, error-related responses of two paradigms were investigated and correlated. Common peculiarities and patterns have been worked out and their cerebral distribution investigated.

The essential addition to previous discoveries is the extension of spectral error processing to higher frequency bands, especially as a common feature for different stimuli. The error-related power increase of the different frequency bands appears to be a dominant feature in the processing. In addition, the broad coverage of the implanted electrode contacts has contributed to the collection of areas involved in error processing and confirmed the present status. This can extend the understanding of error processing in the brain, among other things. The common areas for both paradigms refer, apart from the ACC, to predominantly frontocortical regions.

It would be interesting to investigate an extended temporal development of error-related spectral increase over space, broken down to frequency bands. Likewise the decrease was left out, because it was comparatively rare represented. Nevertheless, a decrease of the power with the error-related event could be observed as well and can give additional information about the temporal development of the signal via the brain. Thus further analyses can contribute to a model of the propagation of an error signal in the brain and the relevant areas. In addition, the results of temporal and spatial distribution of the signal may contribute to the recognition of when and where decoding can be useful. This is especially advantageous when no exact time has been determined and an error is to be decoded from a continuous signal.

Chapter 8

Cross-paradigm Pretraining of Convolutional Neural Networks

When it comes to the classification of brain signals in real-life applications, the training and the prediction data are often described by different distributions. Furthermore, diverse data sets, e.g., recorded from various subjects or tasks, can even exhibit distinct feature spaces. The fact that data that have to be classified are often only available in small amounts reinforces the need for techniques to generalize learned information, as performances of brain-computer interfaces (BCIs) are enhanced by increasing quantity of available data. In this chapter, transfer learning is applied to a framework based on deep convolutional neural networks (deep CNNs) to prove the transferability of learned patterns in error-related brain signals across different tasks. The experiments described in this chapter demonstrate the usefulness of transfer learning, especially improving performances when only little data can be used to distinguish between erroneous and correct realization of a task. This effect could be delimited from a transfer of merely general brain signal characteristics, underlining the transfer of error-specific information. Furthermore, similar patterns in time-frequency analyses in identical channels could be extracted, leading to selective high signal correlations between the two different paradigms. Classification on the intracranial data yields in median accuracies up to $(81.5 \pm 9.5)\,\%$*. Decoding on only* 10% *of the data without pretraining reaches performances of* $(54.8 \pm 3.6)\,\%$*, compared to* $(65.0 \pm 0.8)\,\%$ *with pretraining.*

Chap. 3 to Chap. 5 have addressed the improvement of the decoder for error-related responses of a human observer. The recordings were obtained non-invasively, which has both advantages and disadvantages (a more detailed discussion can be found in Sec. 2.2). However, in this chapter the insights gained so far are applied to intracranial data. The methods from Chap. 6 were used to assign the electrode contacts to the underlying areas. Yet, this information was initially omitted from the analyses. Similarly, the insights gained in Chap. 7 were not included in the approaches of this chapter, which, however, forms a promising basis for future analyses. This shall be discussed at a later point in time.

Furthermore, this chapter examines the relationship between pretraining of a CNN and the performances achieved. The underlying idea is to transfer information the CNN learns to another classification task. Specifically, the network will be trained with the trials of

one paradigm in order to evaluate the influence on the decoding of another paradigm. In this study it is a prerequisite that the paradigms aim at similar cerebral responses, which can then be transferred.

In situations where little data is available, classifying different conditions is not necessarily successful. If one imagines an everyday scenario, then a possible application should be applied as quickly as possible and not have to go through an elaborate calibration procedure. A classifier that already "knows" the problem and already has a certain information advantage would therefore be desirable. The aim of the underlying analyses in this study is to contribute in this vision, and it is to be investigated in principle whether such a transfer is possible.

In this chapter, the impact of transfer learning in intracranial brain recordings across two different error tasks is determined. The paradigms may differ slightly in their way to elicit errors, but basically target the same reaction of perceiving self-caused mistakes in instructed tasks. The classification performances are analyzed separately for both data sets and are compared to those gained by algorithms including transfer learning implantations. It becomes apparent that under conditions with few available data, pretraining and subsequent transfer can improve decoding in error-related classifications tasks.

8.1 System and Experimental Design

The data on which the analyses of this chapter are based correspond to those of the previous chapter Chap. 7. In order to recall the most important facts, the two paradigms are summarized briefly. The paradigms consisted of an active task, which the participant had to fulfill under pressure with the aid of a laptop. Likewise both paradigms aimed at provoking faulty executions of the participants, but differed in their modality. An ERIKSEN flanker task had a much more experimental design, while a car driving task allowed more proximity to lifelike situations. In this chapter, the two paradigms will be referred to as follows:

- ERIKSEN Flanker Task (EFT), see Fig. 7.1A
- Car Driving Task (CDT), see Fig. 7.1B

More detailed information can be found in chapter Chap. 7, where Sec. 7.1 provides additional information about patients and data acquisition. Likewise, in Sec. 2.2 the different recording techniques of the EEG and the iEEG are distinguished against each other and advantages and disadvantages, including complications of the different methods, are discussed. Above all, it should be mentioned that the participants in the measurements in this chapter were patients suffering from drug-resistant epilepsy. More about this in Sec. 2.2.

8.2 Pre-processing, Decoding & Statistics

As already mentioned, the data were obtained by intracranial recordings collected in experiments with 15 patients suffering from epilepsy, who gave their informed consent. According to unique trigger pulses, generated during each experiment, the acquired data were aligned to the event-related meta information. The aligned data were re-sampled to $250\,Hz$ and re-referenced to a common average, subsequently an electrode-wise exponential moving standardization [303] with decay factor 0.999 was applied. The data were cut into trials and divided into test and training set according to the specific decoding intervals. For classification, the *Deep4Net* model of the CNN described in Subsec. 2.5.3 was used, see also Fig. 2.14. The CNN made use of batch normalization and dropout, exponential linear unit (ELUs) served as activation functions for the different layers. The backward computation of the gradients was based on the output of the categorical cross-entropy loss and optimized using *adam* [190]. Further details of the basic implementation and decisions according to design of the network are discussed in Subsec. 2.5.3.

A random permutation test [275] was applied to determine significances per participant, see Alg. 7. A vector consisting of the true distribution of class labels was compared to $n = 10^6$ vectors of randomly shuffled labels to generate a realistic distribution of possible outcomes of the classification. It appeared that the numbers of trials per class were highly unbalanced for all participants. To overcome this problem when creating batches during the training, a class balanced batch size iterator related the samples per batch with the inverse relation to the distribution of the actual trials. For the significance, the imbalance was solved by defining the label matches per vector separately for each class, then averaging over classes and comparing the outcome to the decoded accuracy to estimate the p-value relating to the underlying distribution. Significance was tested for each participant and set of decoding parameters. Single sets exceeding a value of $p = 0.05$ were disregarded in further analysis and did not contribute to final results. The significances of the group differences in Fig. 8.3 were determined on the level of trials, using a sign test [164].

8.3 Decodability of Error-related Signals

For the two data sets, the deep CNN was used to determine the two-class decodability of perceived erroneous/correct events in intracranial human brain recordings. Here and in the following, the available data were split into two sets with a proportion of $80\,\%$ for training and $20\,\%$ for testing. For each participant, the decoding accuracy was calculated for different intuitive decoding intervals, which are defined according to the appearance of an event. Fig. 8.1 shows the comparison of the single accuracies for different intervals in blue symbols contrasted for the two data sets and depicts in addition the median accuracy over all participants per interval in form of filled red symbols. In this illustration, only participants are considered who showed significant classification results for both paradigms. The classification yielded in median performances of $(78.2 \pm 7.5)\,\%$ for the car driving task and $(79.4 \pm 9.7)\,\%$ for the ERIKSEN flanker task using the decoding interval

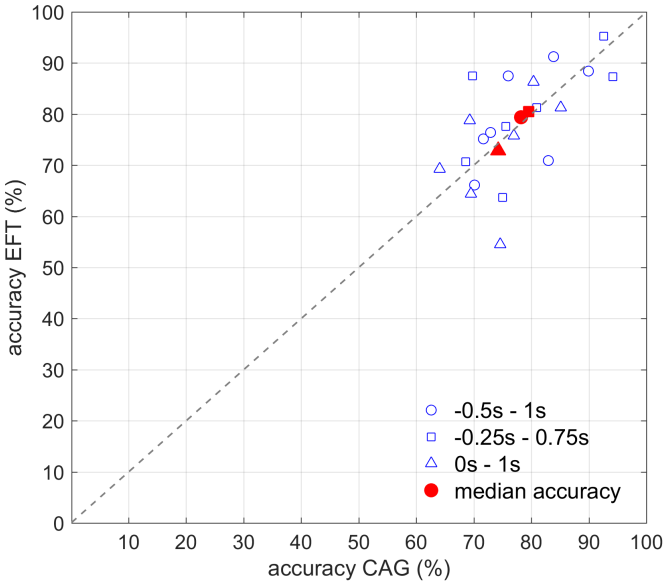

Figure 8.1: Single participant deep CNN accuracies contrasted for the two paradigms and different decoding intervals. Median accuracies per interval are depicted by filled red symbols.

-0.5 to $1\,s$, $(79.5 \pm 10.3)\,\%$ and $(80.5 \pm 10.8)\,\%$ for the interval -0.25 to $0.75\,s$ and finally $(74.3 \pm 7.3)\,\%$ and $(72.9 \pm 10.9)\,\%$ for the interval 0 to $1\,s$. Here, the time points refer to the occurring event. For both tasks, the interval -0.25 to $0.75\,s$ outperformed the others and was therefore used predominantly for the later implementations to transfer learned features. Tab. 8.1 gives an overview of decoding on various intervals and different number of training epochs.

Table 8.1: Median accuracies for different decoding intervals

CDT			
epochs	**-0.5 - 1s**	**-0.25 - 0.75s**	**0 - 1s**
10	(72.1 ± 3.3) %	(67.6 ± 3.1) %	(71.3 ± 6.3) %
50	(72.9 ± 2.2) %	(73.4 ± 5.0) %	(73.4 ± 1.9) %
200	(75.4 ± 3.1) %	(76.9 ± 2.2) %	(74.0 ± 2.0) %
EFT			
epochs	**-0.5 - 1s**	**-0.25 - 0.75s**	**0 - 1s**
10	(73.7 ± 4.2) %	(82.1 ± 7.9) %	(73.0 ± 6.4) %
50	(70.3 ± 5.6) %	(78.5 ± 6.0) %	(70.3 ± 5.6) %
200	(81.2 ± 11.1) %	(81.5 ± 9.5) %	(73.6 ± 9.5) %

8.4 Responses in the Frequency Domain

The single data sets in the frequency domain were investigated using a multitaper method to estimate the power spectral density, see Eq. (2.19). Optical inspection and comparison of time-frequency spectra for identical electrodes but different tasks revealed obvious similarities for several electrodes. Fig. 8.2A depicts one example where a resemblance is unambiguous, showing the response for error vs correct in electrode I2 located in the right insular cortex. The dotted line marks the onset of the event. Nevertheless, other electrodes did not show any effects, or effects could only be seen strongly for one of the tasks, as illustrated in Fig. 8.2C. A global overview of all electrodes for this exemplary participant is given in Fig. 8.2B. The blue and green markers refer to the electrodes selected for figures 8.2A and 8.2C.

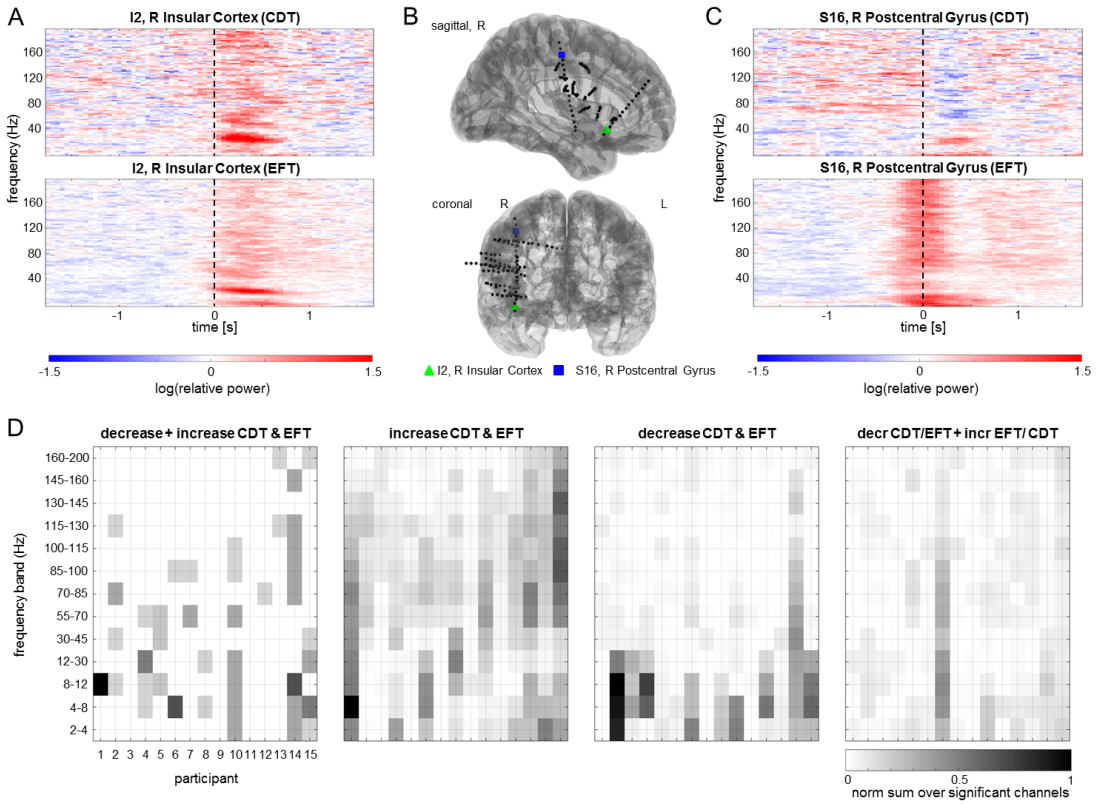

Figure 8.2: Responses in the frequency domain: A Trial-averaged time-frequency spectra for electrode I2 located in R insular cortex, for error vs correct in CDT (top) and EFT (bottom). **B** Saggital (top) and coronal (bottom) view of the implanted electrodes for an exemplary participant, plotted on the ICBM152 brain [228]. **C** Trial-averaged time-frequency spectra for electrode S16 located in R postcentral gyrus, for error vs correct in CDT (top) and EFT (bottom). **D** Normalized sum over significant channels per frequency band and participant, itemized into decrease and increase.

Moreover, the behaviour of frequency-band power time-series of significant channels was analyzed. Decrease and increase of the power were tagged for both paradigms, CDT and EFT, and compared among themselves, to get an estimation of similarities in temporal

developments of the frequency power. Fig. 8.2D illustrates the outcome of this type of analysis, dividing the figure into four conditions of overlapping tags for the two paradigms. The color code in each panel refers to the sum of significant channels, normalized to the number of channels per participant and to the maximal value of significant channels, exhibiting the specific tag indicated by the panel title. The individual color values are broken down to frequency band and participant. Significant decrease for both paradigms as well as a significant increase in the lower frequency bands ($< 30\,Hz$) can be seen in the data for most of the participants. However, for all participants an increase in the gamma band is prominent, covering the bands from $55\,Hz$ to $130\,Hz$, as already discussed in Chap. 7. For some of the participants the manifestation is present in more channels than for others, according to the specific implantation and the adjacent brain area.

8.5 Compilation of Different Transfer Approaches

Initially the output of three different transfer approaches was examined, choosing a small number of post-training epochs compared to the number of epochs ($n = 200$) in the pretraining with the first data set. An assembly of the results is given in Tab. 8.2, showing the median accuracy over the participants. Errors were estimated by selecting the interquartile range of the bootstrapped samples per interval and technique. For each of the three implementations, the network was pretrained on a given data set $\mathbf{D}_i$, while a then unknown set $\mathbf{D}_j$ was used for testing or fine tuning, respectively. The whole data were processed in a way that the feature space remained the same for the two sets. Therefore an adjustment of the input layer was not necessary.

Table 8.2: Median accuracies for different transfer approaches

fine tuning on $\mathbf{D}_{CDT}$ (network pretrained on $\mathbf{D}_{EFT}$)				
layers	**epochs**	**-0.5 - 1s**	**-0.25 - 0.75s**	**0 - 1s**
all	0	(50.5 ± 1.1) %	(49.3 ± 0.7) %	(48.8 ± 2.0) %
all	10	(67.5 ± 1.4) %	(66.8 ± 10.2) %	(69.8 ± 3.3) %
last	50	(61.3 ± 2.9) %	(63.0 ± 1.5) %	(63.0 ± 5.8) %
fine tuning on $\mathbf{D}_{EFT}$ (network pretrained on $\mathbf{D}_{CDT}$)				
layers	**epochs**	**-0.5 - 1s**	**-0.25 - 0.75s**	**0 - 1s**
all	0	(54.0 ± 4.9) %	(57.5 ± 6.7) %	(54.2 ± 3.6) %
all	10	(73.4 ± 7.9) %	(72.1 ± 7.9) %	(76.8 ± 13.5) %
last	50	(66.7 ± 4.0) %	(68.9 ± 2.9) %	(59.5 ± 6.6) %

The first approach consisted of the pretraining and subsequently classification on the second unseen set $\mathbf{D}_j$ based on the predefined weights without fine tuning. Generalizing

from EFT to CDT, the deep CNN was not able to predict the true classes of the tasks and presented poor performances around chance. For the transfer from the CDT to the EFT data set accuracies were slightly better, exceeding chance level and showing a peak performance of $(57.5 \pm 6.7)\,\%$ for the interval -0.25 to $0.75\,s$.

Secondly, the pretrained network was fine tuned by training on a then unknown data set $\mathbf{D}_j$ for $n = 10$ epochs with a smaller learning rate. Here indeed the network learns informative features and obtains accuracies around $70\,\%$ for both of the paradigms. However, comparison with the performances given in Tab. 8.1 indicates that there is no enhancement when using the pretraining. To the contrary, the accuracies do not yield the high values obtained by training directly on the classification data set training for $n = 10$ epochs.

The third implementation was inspired by techniques from computer vision, where networks are pretrained by a huge training set and only a few last layers are trained again by a smaller set of similar data to fine tune the weights in the deeper layers. This idea was captured and all layers after pretraining were frozen and only the weights of the last classification layer were adjusted. In both data sets performances yielded accuracies of $60\,\%$ and higher, but not reaching the values obtained when fine tuning the whole network, even with less epochs.

8.6 Performance Dependency on the Amount of Data

Again, the network was pretrained on a given data set $\mathbf{D}_i$ to implement the weights. To draw a comparison between conditions with only few data and situations where more data are available, the second data set $\mathbf{D}_j$ was selected and the amount of data used for fine tuning was gradually increased from $10\,\%$ to $100\,\%$ of the available training data ($80\,\%$ of the entire data), once more employing a smaller learning rate as in the pretraining. Median accuracies and the underlying distributions are presented in Fig. 8.3A (top) for $\mathbf{D}_j = \mathbf{D}_{CDT}$ and Fig. 8.3B (top) for $\mathbf{D}_j = \mathbf{D}_{EFT}$. The boxes depict the interquartile range, the whiskers extend to the most extreme data points and outliers are drawn as asterisks. The plots at the bottom of each panel reveal the distribution of the intra-participant difference between the two compared decoding accuracies. E.g. to obtain the values for Fig. 8.3A the difference $ACC_{transf} - ACC$ was calculated for each participant, where ACC_{transf} corresponds to the accuracy gained with pretraining while ACC corresponds to the accuracy achieved without pretraining. For decoding on $\mathbf{D}_{CDT}$, Fig. 8.3A, there is no big difference between the two conditions. Even with less data for the final training, the pretraining cannot enhance the performance. In contrast, in Fig. 8.3B the pretraining on $\mathbf{D}_{CDT}$ has the effect that for a decreasing amount of data the performance gradually gets better, exhibiting significant differences of median accuracies up to $10\,\%$ for a fraction of $10\,\%$ of the training data. The distribution of the intra-participant differences for the smaller amount of used data confirm this trend. Median accuracies yielded with pretraining are consistently better than in cases when only the training on the unseen set was performed. Due to the relatively small number of participants, significance was tested on the level of single trials.

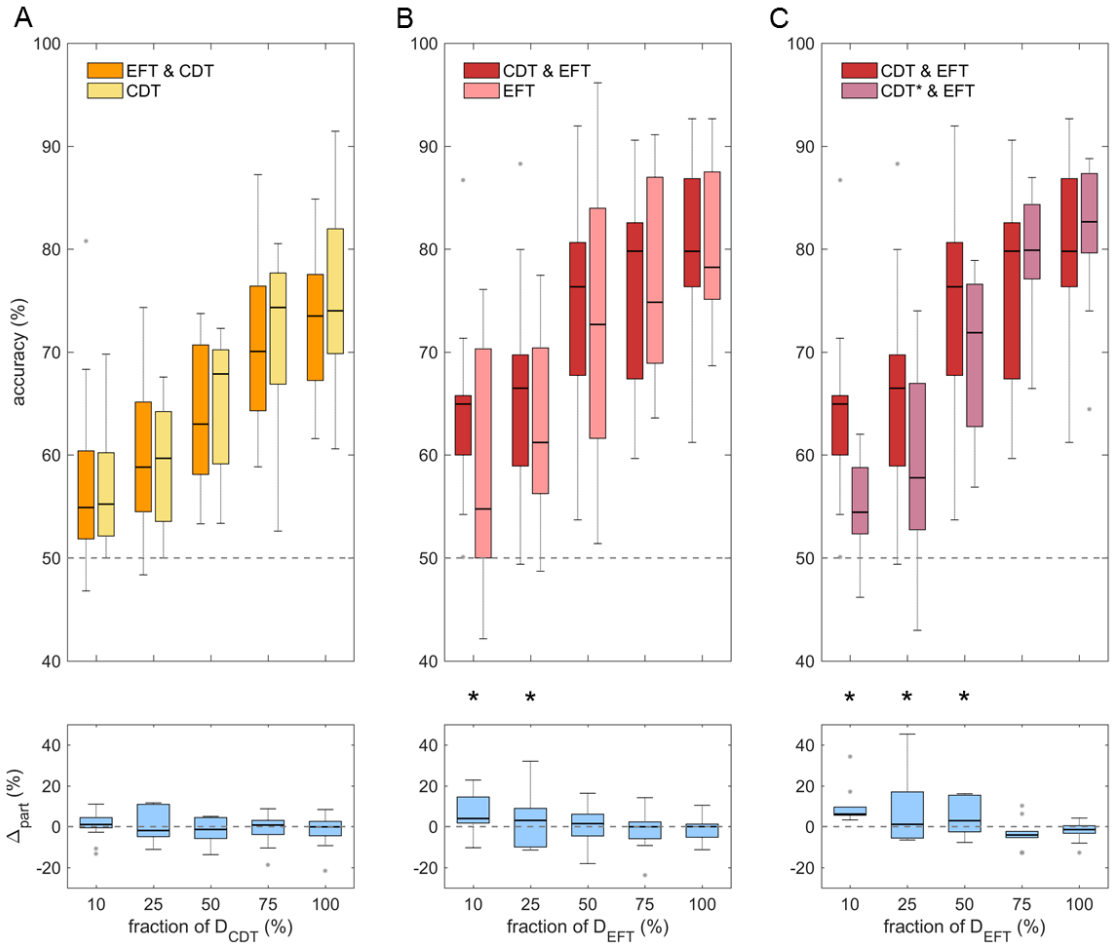

Figure 8.3: Contrast of median accuracies for vanishing data. Accuracies obtained by stepwise reduction of available training data D_{CDT}, comparing (a) the training only on D_{CDT} to (b) pretraining on D_{EFT} and then fine tuning on D_{CDT} (**A**, top) and vice versa for D_{EFT} (**B**, top). **C** (top) pretraining on D_{CDT} and fine tuning on D_{EFT} compared to pretraining on D_{CDT*} with shuffled labels and fine tuning on D_{EFT}. **A**-**C** (bottom) Accuracy distribution of intra participant difference, e.g. $ACC_{transf} - ACC_{CDT}$ for **A**. * indicates significance of the difference with $p < 0.05$.

A last comparison claims to test whether the distinction between the two cases originates from a transfer of more general features of the brain signals and not the true underlying conditions. Therefore, the performed transfer was contrasted to the decoding results of pretraining on $\mathbf{D}_{CDT}$ with randomly shuffled labels and then fine tuned on $\mathbf{D}_{EFT}$. Hereby the network wasn't able to learn the features of the two conditions. Indeed the results show that the decoding using unshuffled labels during the pretraining performs clearly better for decreasing data, as illustrated in 8.3C. The lower plot again shows the distribution of the intra-participant difference, where the values were determined by $ACC_{transf} - ACC_{shuffle}$. Here, too, differences for the fewer data exhibit positive median values and distributions mainly over zero.

8.7 Related Work

After revolutionizing fields like computer vision, deep learning methods have also recently been used to improve classification in applications based on brain computer interfaces (BCIs) [219]. A deep belief network model was used to distinguish motor imagery tasks [13], outperforming support vector machines (SVM) [73], or to extract features of EEG signals [288]. Other approaches to decode EEG data e.g. used deep convolutional neural networks for feature extraction and visualization [158], or built a recurrent convolutional neural network architecture to model cognitive events from EEG data [26], applying multi-dimensional features. Likewise for intracranial EEG data, deep neural networks supported classification of epileptic signals [3, 15, 169].

However, performances of deep learning methods are strongly dependent on the amount of available data. Furthermore, the different methods are mostly restricted to certain conditions when it comes to the design of the data. Assumptions like equal underlying distributions or feature spaces may pertain in classical image recognition tasks, but are mostly not satisfied for real-world applications based on human brain signals. Intra- and inter-individual varieties cause conditions where performances of exactly the same classifier change daily. Also quite similar tasks can exhibit completely different efficiencies in distinguishing classes. In fields such as computer vision, deep learning methods have been enhanced by approaches for transfer learning [259, 313], especially when only small data are given to train a network. Models, pretrained upon extensive databases [94], built the foundation for significant enhancements for example in object categorization or image segmentation [112, 143, 159]. The networks seem to learn the fundamental constitution of the training data to utilize the information for classification in other similar sets. Real-life applications subsist in smooth and fast handling, therefore long training periods are unwanted and collecting substantial real-time data goes beyond the constraints of useful application.

Recently, transfer learning techniques have found their way into the context of BCI implementations [178]. Different approaches are applied e.g. to solve a transfer between different types of error-related potentials [189] using a linear support vector machine or to find a way to deal with deviation in latencies [175] or signal variations [174] in brain controlled interfaces, based on linear discriminant analysis (LDA) [45]. Implementations reverting to deep CNNs already have generalized non-invasive error-related recordings across subjects, without fine tuning the network again [353]. However, there is still little utilization and transfer learning across different error decoding tasks for intracranial human brain data in combination with deep CNNs has not yet been investigated.

8.8 Conclusion

In this chapter, two different issues were analyzed. First, the proof of decodability of error-related signals in the underlying intracranial brain recordings was brought to the fore. This was tested for two paradigms, differing by their affinity to real-life application. Error decoding has been investigated several times using EEG data e.g. when observing

and controlling robots [31, 176, 299], see also Chap. 3 to Chap. 5, or in real interaction simulations [65], but not yet on the basis of intracranial recordings. Here, accuracies up to $(79.5 \pm 10.3)\,\%$ were obtained for the car driving task and $(80.5 \pm 10.8)\,\%$ for the ERIKSEN flanker task. The quite high performances reinforce the use of these data for approaches reverting to transfer approaches. However, the high errors show non-negligible differences of the results, which certainly should be treated with caution. Different patients were equipped with differing implantations, which in turn covered different brain areas. Thus, it cannot be excluded that more or less informative channels were given in the varying data sets, leading necessarily to diverse decoding performances. Because of the different implantations, it was abstained from an inter-subject transfer. The previous chapter analyzed the coverage in detail and gave hints on the contributing areas regarding the error processing. This information may used in future approaches based on intracranial recordings to select more informative features and discard the ones that carry no information about error processes.

The second aspect concerned the similarity of the data sets gained by the different paradigms and their transferability. Time-frequency spectra of same channels revealed striking similarities for some of the channels. More precise examinations of frequency-band dependent time-series of the power spectral density uncovered an extensive increase of significant channels in the gamma band between $55\,Hz$ and $130\,Hz$, as already indicated in [351] and in Chap. 7. Likewise, the results indicate a similarity in the characteristics of the data for the two distinct paradigms.

A comparison of several transfer approaches for the whole extent of data but a lower number of epochs did not lead to improvement of the decoding. When the network was trained directly with the objective data set exclusively, higher accuracies were yielded compared to pretraining the network. As already shown by [353] on EEG data, a direct transfer without further fine tuning did not succeed.

In many cases, acquiring intracranial data is hardly possible and raised data sets are often not extensive. This study illustrates a significant improvement of decoding for decreasing amounts of data when the network is pretrained by a similar set. Interchanging the two data sets led to no enhancements, which might be explained by the fact that in this case the pretraining was performed on the set comprising only few trials and therefor possibly made the generalities of the conditions not sufficiently or hardly learnable. Instead the question arises whether, for a transfer, the relation of the amount of data used for pre- and post-training plays a determining role for the applicability of this technique. Certainly, a degree of similarity between the data sets has to be given, also with respect to the manifestation of the two conditions, which could be shown here by randomization of labels.

Several interesting questions and approaches can be deduced from these results. E.g. a network might be trained on an extensive set of non-invasive data to learn problem-specific characteristics, which subsequently can be fine tuned by a small intracranial data set. Here, a change of network architecture can make a transfer possible, assuming data in different feature spaces. Likewise, data augmentation can contribute to advance classification in rather small data sets.

Chapter 9

Conclusions

9.1 Summary

In the context of this thesis, new concepts for control signals to drive and optimize brain computer interfaces were worked out, that might come into application when humans collaborate with intelligent robotic systems. Based on the analysis of invasive and non-invasive recordings, different machine learning approaches were compared and applied for the detection of both observed and committed errors using human brain signals. In order to advance the basic understanding of error processing in the human brain, the extensive data sets were tested for neurophysiological patterns and correlations, using paradigms of different modalities. For these two paradigms, the transfer of error-related information across paradigms was also tested using deep convolutional neural networks and the dependence on the amount of data available was examined. The last essential part of the thesis was the development and application of an algorithm for the assignment of electrode contacts to brain areas, which considers individual characteristics and is based on spatial retransformation. This approach was realized within a user-friendly software, which has further functions besides the basic application. For each of the proposed approaches, the necessary theoretical foundations were created, experiments were carried out and the connection to related literature was established.

First, the fundamental question was examined whether errors made by robotic systems and observed by a human user can in principle be decoded from human EEG. For this purpose, a large data set based on non-invasive EEG measurements was generated, describing phenomena while healthy participants observed faulty performance of pouring and lifting tasks by robots. These possible scenes from a human-robot collaboration were deliberately selected as tasks in order to mimic such collaborative situations. In order to suppress any interference signals, the experiments were carried out inside an active-shielding FARADAY cage and all electrical components were decoupled from the electricity grid. Also, muscular and saccadic artifacts were largely absent. The decoding was realized using a conventional filter bank common spatial patterns (FBCSP) algorithm for EEG. It turned out that this algorithm was able to distinguish the erroneous from correct execution as well as the type of robots. Before, neither the error detection when observing robots nor the differentiation of robot types on the basis of EEG could have been shown. The fact that the events ”error” and ”robot type” did not occur discretely over time but were rather of a continuous nature made the classification process more

difficult, but also underlines the value of the results achieved.

In a further step, an attempt was made to improve the performance of the decoding. If humans and robots work closely together, the reliability of the automated safety systems is extremely important and the recognition of signals is decisive. In addition to the conventional regularized linear discriminant analysis (rLDA), a deep convolutional neural network (CNN) was installed, which has already proven its strength in other disciplines but is still relatively new in the field of EEG decoding. In this problem with highly practical relevance, the performance on the decoder side can be significantly improved, if the classifier is based on a CNN, compared to rLDA and FBCSP results. The CNN thus revealed its potential for a more efficient implementation of error detection systems. The learned features were visualized to provide information about the relevance of different features for the performance of the decoding. Both visualizations and correlations of the different methods across participants indicated that the network for these processes seemed to use more time-related information for its decisions, and the visualizations could provide information about the cortical distribution of the signals. The CNNs also performed significantly better than the comparison methods in distinguishing between robot types. In this case, the decisions also appear to have been made on features in the time domain, with the visualizations providing hints to the visual processing of the different stimuli.

In addition to measurements based on non-invasive EEG, it is also possible to make recordings directly at the tissue of the brain. However, implantations lead to deformations of the soft tissue and the assignment of electrode contacts to brain areas based on image data is extremely difficult, especially in the cortical regions. The individual characteristics that differ from patient to patient also make this assignment more difficult. A new assignment algorithm is presented which solves this problem. The method makes use of individual landmarks in order to be able to select or exclude brain areas, and uses a cortical retransformation to compensate for the deformations of the tissue. The final output of the probabilities for the participation of an area to the electrode contact signal is calculated based on existing probabilistic, cytoarchitectonically-defined maps. A software environment allows a user-friendly application of the underlying methods and provides a 3D visualization of the results. Especially for users without programming experience, the tool allows an application of the algorithms required especially in the clinical field. The evaluation of the method showed that the consideration of individual anatomical landmarks prevented in about $8 - 10\,\%$ of the cases false lobar assignment and improved the overall neuroanatomical assignment in iEEG.

The question was investigated how the previously non-invasively shown phenomena of error-related brain responses behave in the case of measurements with intracranial EEG, especially if these errors were caused by the user himself. For this purpose, two data sets were generated which provoked both errors in the execution by the participating epilepsy patients, but whose modalities differed in their proximity to everyday life. With a total number of 1552 electrode contacts in different positions, the data set showed a comprehensive coverage of the brains areas. The further investigation of error-related brain signals could provide a broad spectrum of information regarding error processing in the brain and the generated data set provides the basis for extensive neurophysiological

investigations in this area. In addition, the error-related power increase of the different frequency bands turned out to be a dominant feature in the processing. Another advantage of the study was the fact that all patients participated in both paradigms and thus it could be revealed which of the discovered characteristics were common. Besides, it appeared that simultaneously activated regions lie mainly in frontocortical areas, but also in the anterior cingulate cortex, and several so far not to error-processing related areas exhibited a spectral response.

Based on the comprehensive set of intracranial recordings, the errors committed by users could be decoded for both paradigms, using the CNN architecture. It should be mentioned that there was neither a selection of patients nor electrode contacts, which in principle did not protect of little or no information in numerous channels. A look at the time-frequency spectra revealed some astonishing similarities between different paradigms but the same electrode contacts. This finding motivated an approach that transfers information across paradigms and uses it for further classification. Based on a CNN architecture, the network was pretrained using the data of one paradigm and then classified on the second data set with successively increasing availability of data. The method achieved significant improvements for small amounts of available data and can be applied in numerous context, e.g., when transferring knowledge to detection of errors committed by intelligent robotic effectors.

In the scope of this thesis, solutions and answers to the following questions could be developed:

- Can the non-invasive EEG of human observers be used to decode faulty execution of robotic systems?
- Can the classification of error-related signals be improved by using deep learning? Does visualization help to interpret the results neurophysiologically?
- Is it in principle possible to distinguish robot types based on the EEG and how does the type of robot affect the detection of errors? Do criteria such as the number of human similarity characteristics have an influence on the results?
- Can error decoding based on intracranial recordings take place with deep learning when users make errors themselves? Does the performance depend on how strong the error is subjectively perceived or how realistic it seems to be?
- Are convolutional neural networks (CNNs) able to learn generalizing information in one task in order to use it profitably in another task?
- Can the intracranial EEG contribute to an understanding of the temporal and spatial processing of error processes? Is there any indication of decisive features in these processes?
- How can the difficulties in assigning intracranial electrode contacts to underlying brain areas be solved, if for example deformations due to implantation but also individual characteristics complicate the procedure?

Based on the solutions developed, we believe that this work makes a valuable contribution to the detection of error-related signals in the human brain, although the findings do not necessarily have to be limited to this signal type. The work also makes fundamental contributions to the ongoing neurophysiological examination of error signals in particular. The newly developed method ELAS ensures reliable assignment of electrode contact to brain areas and thus correct interpretation of phenomena that are examined on the basis of intracranial recordings.

9.2 Outlook

This thesis has made important contributions to error detection using human brain signals. In addition, it has been shown that CNNs are a promising candidate to realize reliable detection systems in future applications beyond pure research. The studies on which this thesis is based on have been specifically designed to come close to future scenarios in which autonomous robots support via BCI users, as for instance self-feeding or go-and-fetch tasks. Even if the performances could be significantly improved, though, accuracies are still far from practical applicability. However, there are recent advances in deep learning research that have the potential to further improve CNNs, such as automatic hyperparameter optimization and architecture search, including recurrent and residual network architectures, data augmentation, using 3D convolutions, or increasing the amount of training data. Furthermore, in this thesis, when decoding on invasive recordings, it is refrained from profitably using findings from neurophysiological investigations for decoding. Appropriate feature selection could lead to further improvements, especially in the case of iEEG. There are also other promising approaches, such as RIEMANNIAN geometry [23, 80] for decoding on non-invasive EEG, which find other ways to extract classifying features from available data. Overall it particularly would be exciting to verify classification results and further comparisons for online applications.

In an everyday BCI scenario, a system should be designed to be as user-friendly as possible. However, brain signals exhibit individual characteristics and a calibration is needed for new users. CNNs may require a large amount of data, which would result in an undesired long calibration phase. Generalization of relevant features could help here. In this thesis, it could be shown that the approach of generalization can lead to improvements in decoding for little data, but the potential is far from exhausted and numerous other possible applications have to be examined. Several interesting questions and approaches can be deduced from the results in this thesis. E.g. a network might be trained on an extensive set of non-invasive data to learn problem-specific characteristics, which subsequently can be fine tuned by a small intracranial data set. In cases of different feature spaces, a change of network architecture can make a transfer possible. In addition, the results raise the question of whether the relation of the amount of data used for pretraining plays a determining role for the applicability of this technique. The common peculiarities and patterns of error processing that have been worked out for different paradigms could also provide information about the nature of the generalizing features.

In addition to the practical application of the error-related insights, further findings

can also contribute to a better understanding of error processing, but also of the general functioning of our brain. For decoding on non-invasive data, the visualization of learned features showed activity in occipital visual areas, for example. Here the findings could be taken up and investigated to what extent the decodability of different visual inputs depends on the subjective interpretation as actual error. Closely related to this, would be the investigation of how visual, affective, and movement-related brain systems are involved in the generation of the differential responses to robot action. The insights gained in this thesis give first clues and leave room for further investigations of the data, but also in the field in general. A possible approach would be to create a model of the temporal development of an error signal and its propagation via the brain. This relationship can be investigated separately for the different frequency bands. There are also other features besides the error-related power increase that could be investigated, but have been omitted here. Information about the temporal and spatial course of the signal could be helpful if the exact time of a signal to be decoded is not known. Ultimately, the question can be asked to what extent the signals from error observation and those from committing errors show similarities. This question could not be covered in the context of this thesis.

According to faulty robotic execution, another question to be addressed in the future would be which kind of errors are generally suitable for decoding of the user's perceived correctness and how they differ from non-decodable errors. Though, when working with robots, there are other aspects besides efficiency of error detection that can be considered. For example, subtle questions can be examined about how different robot types affect human users, e.g. regarding a sense of security, which might correlate to its degree of humanoid appearance. It could be demonstrated that the appearance of a robot type can be distinguished by the brain signals of a human observer. In the case of the observation of two robots, the visualization of the features discriminating the robot type showed activations in areas attributed to the human mirror neuron system. The engagement of the mirror neuron system might be modulated by the degree of humanoid appearance of the robot. More in-depth investigations of this correlation on the optical differences between robot types with respect to human similarity could provide interesting insights.

The developed interface ELAS is based on maps which are continuously updated and supplemented. These updates ensure that the mapping is always up to date and ELAS can contribute to a better understanding of investigated cerebral phenomena. It is also noteworthy that application of a neuroanatomical framework, that has been derived from the healthy brain, to iEEG data from epilepsy patients is not necessarily just a limiting factor, but also opens up possibilities to investigate reorganization of cortical function. Overall, the evaluation of the novel method indicates that if atlas-based anatomical analysis of iEEG data in general, and probabilistic methods in particular, would find a more widespread use in iEEG research, this would also spur further interest in optimizing the spatial normalization of structural imaging data, particularly of post-operative MRIs.

List of Figures

List of Tables

List of Algorithms

Bibliography

[1] L. Acqualagna. Pushing the boundaries of brain-computer interface technology. *Ph. D. dissertation, Technischen Universität Berlin*, 2017.

[2] J. Adcock, R. G. Wise, J. Oxbury, S. Oxbury, and P. Matthews. Quantitative fmri assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy. *Neuroimage*, 18(2):423–438, 2003.

[3] D. Ahmedt-Aristizabal, C. Fookes, K. Nguyen, and S. Sridharan. Deep classification of epileptic signals. *arXiv preprint arXiv:1801.03610*, 2018.

[4] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun. Exploring neurophysiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous eeg, ecg, and fnirs data. *Frontiers in human neuroscience*, 10: 219, 2016.

[5] W. H. Alexander and J. W. Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature neuroscience*, 14(10):1338, 2011.

[6] E. Alpaydın. *Maschinelles Lernen*. Oldenbourg Verlag, 2008.

[7] D. M. Amodio, J. T. Jost, S. L. Master, and C. M. Yee. Neurocognitive correlates of liberalism and conservatism. *Nature neuroscience*, 10(10):1246, 2007.

[8] D. M. Amodio, S. L. Master, C. M. Yee, and S. E. Taylor. Neurocognitive components of the behavioral inhibition and activation systems: Implications for theories of self-regulation. *Psychophysiology*, 45(1):11–19, 2008.

[9] K. Amunts and K. Zilles. Advances in cytoarchitectonic mapping of the human cerebral cortex. *Neuroimaging Clinics of North America*, 11(2):151–69, 2001.

[10] K. Amunts, A. Schleicher, U. Bürgel, H. Mohlberg, H. B. Uylings, and K. Zilles. Broca's region revisited: cytoarchitecture and intersubject variability. *Journal of Comparative Neurology*, 412(2):319–341, 1999.

[11] K. Amunts, A. Malikovic, H. Mohlberg, T. Schormann, and K. Zilles. Brodmann's areas 17 and 18 brought into stereotaxic space—where and how variable? *Neuroimage*, 11(1):66–84, 2000.

[12] K. Amunts, P. H. Weiss, H. Mohlberg, P. Pieperhoff, S. Eickhoff, J. M. Gurd, J. C. Marshall, N. J. Shah, G. R. Fink, and K. Zilles. Analysis of neural mechanisms

underlying verbal fluency in cytoarchitectonically defined stereotaxic space—the roles of brodmann areas 44 and 45. *Neuroimage*, 22(1):42–56, 2004.

[13] X. An, D. Kuang, X. Guo, Y. Zhao, and L. He. A deep learning method for classification of eeg data based on motor imagery. In *International Conference on Intelligent Computing*, pages 203–210. Springer, 2014.

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

[15] A. Antoniades, L. Spyrou, D. Martin-Lopez, A. Valentin, G. Alarcon, S. Sanei, and C. C. Took. Detection of interictal discharges with convolutional neural networks using discrete ordered multichannel intracranial eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2285–2294, 2017.

[16] B. Babadi and E. N. Brown. A review of multitaper spectral analysis. *IEEE Trans. Biomed. Engineering*, 61(5):1555–1564, 2014.

[17] C. Babiloni, F. Babiloni, F. Carducci, F. Cincotti, G. Cocozza, C. Del Percio, D. V. Moretti, and P. M. Rossini. Human cortical electroencephalography (eeg) rhythms during the observation of simple aimless movements: a high-resolution eeg study. *Neuroimage*, 17(2):559–572, 2002.

[18] O. Bai, Z. Mari, S. Vorbach, and M. Hallett. Asymmetric spatiotemporal patterns of event-related desynchronization preceding voluntary sequential finger movements: a high-resolution eeg study. *Clinical Neurophysiology*, 116(5):1213–1221, 2005.

[19] S. Baillet, J. C. Mosher, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal processing magazine*, 18(6):14–30, 2001.

[20] T. Ball, M. Nawrot, T. Pistohl, A. Aertsen, A. Schulze-Bonhage, and C. Mehring. Towards an implantable brain-machine interface based on epicortical field potentials. *Biomedizinische Technik*, 49(2):756–759, 2004.

[21] T. Ball, E. Demandt, I. Mutschler, E. Neitzel, C. Mehring, K. Vogt, A. Aertsen, and A. Schulze-Bonhage. Movement related activity in the high gamma range of the human eeg. *Neuroimage*, 41(2):302–310, 2008.

[22] T. Ball, M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage. Signal quality of simultaneously recorded invasive and non-invasive eeg. *Neuroimage*, 46(3): 708–716, 2009.

[23] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Riemannian geometry applied to bci classification. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 629–636. Springer, 2010.

[24] M. S. Bartlett. Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2): 1–16, 1950.

[25] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[26] P. Bashivan, I. Rish, M. Yeasin, and N. Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[27] J. Bastin, P. Deman, O. David, M. Gueguen, D. Benis, L. Minotti, D. Hoffman, E. Combrisson, J. Kujala, M. Perrone-Bertolotti, et al. Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex*, 27(2):1545–1557, 2017.

[28] A. T. Bates, T. P. Patel, and P. F. Liddle. External behavior monitoring mirrors internal behavior monitoring: Error-related negativity for observed errors. *Journal of Psychophysiology*, 19(4):281–288, 2005.

[29] M. Bauer, R. Oostenveld, M. Peeters, and P. Fries. Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas. *Journal of Neuroscience*, 26(2):490–501, 2006.

[30] J. Behncke, R. T. Schirrmeister, W. Burgard, and T. Ball. The role of robot design in decoding error-related information from eeg signals of a human observer. *6th International Congress on Neurotechnology, Electronics and Informatics (NEUROTECHNIX)*, 2018.

[31] J. Behncke, R. T. Schirrmeister, W. Burgard, and T. Ball. The signature of robot action success in eeg signals of a human observer: Decoding and visualization using deep convolutional neural networks. In *6th International Winter Conference on Brain-Computer Interface*, pages 1–6. IEEE, 2018.

[32] J. Behncke, R. T. Schirrmeister, M. Völker, J. Hammer, P. Marusič, A. Schulze-Bonhage, W. Burgard, and T. Ball. Cross-paradigm pretraining of convolutional networks improves intracranial eeg decoding. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.

[33] J. Behncke*, M. Kern*, J. Rüscher*, A. Schulze-Bonhage, and T. Ball. Probabilistic neuroanatomical assignment of intracranial electrodes using the elas toolbox. *Journal of Neuroscience Methods*, 2019. *These authors contributed equally.

[34] J. Behncke, J. Hammer, A. Kalina, P. Marusič, A. Schulze-Bonhage, W. Burgard, and T. Ball. A core system for error processing delineated by intracranial eeg. *NeuroImage*, 2020. *submitted*.

[35] H. Berger. Über das elektrenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten*, 87(1):527–570, 1929.

[36] C. Beste, K. Domschke, V. Kolev, J. Yordanova, A. Baffa, M. Falkenstein, and C. Konrad. Functional 5-ht1a receptor polymorphism selectively modulates error-specific subprocesses of performance monitoring. *Human brain mapping*, 31(4): 621–630, 2010.

[37] C. Beste, V. Kolev, J. Yordanova, K. Domschke, M. Falkenstein, B. T. Baune, and C. Konrad. The role of the bdnf val66met polymorphism for the synchronization of error-specific neural networks. *Journal of Neuroscience*, 30(32):10727–10733, 2010.

[38] C. G. Bien, A. L. Raabe, J. Schramm, A. Becker, H. Urbach, and C. E. Elger. Trends in presurgical evaluation and surgical treatment of epilepsy at one centre from 1988–2009. *J Neurol Neurosurg Psychiatry*, 84(1):54–61, 2013.

[39] N. Birbaumer. Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, 2006.

[40] C. M. Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[41] R. B. Blackman and J. W. Tukey. The measurement of power spectra from the point of view of communications engineering—part i. *Bell System Technical Journal*, 37 (1):185–282, 1958.

[42] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. The non-invasive berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

[43] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller. Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56, 2008.

[44] B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. E. Ramsey, I. Sturm, G. Curio, et al. The berlin brain–computer interface: non-medical uses of bci technology. *Frontiers in neuroscience*, 4:198, 2010.

[45] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller. Single-trial analysis and classification of erp components—a tutorial. *NeuroImage*, 56(2):814–825, 2011.

[46] A. O. Blenkmann, H. N. Phillips, J. P. Princich, J. B. Rowe, T. A. Bekinschtein, C. H. Muravchik, and S. Kochen. ielectrodes: A comprehensive open-source toolbox for depth and subdural grid electrode localization. *Frontiers in neuroinformatics*, 11: 14, 2017.

[47] K. Blinowska, N. Crone, P. Franaszczuk, M. Kaminski, R. Kus, and J. Zygierewicz. Electrocorticographical transfer of information during motor task. In *Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on*, pages 619–622. IEEE, 2007.

[48] R. Bogue. Exoskeletons and robotic prosthetics: a review of recent developments. *Industrial Robot: An International Journal*, 36(5):421–427, 2009.

[49] H. S. Bokil, B. Pesaran, R. A. Andersen, and P. P. Mitra. A method for detection and classification of events in neural activity. *IEEE Transactions on Biomedical Engineering*, 53(8):1678–1687, 2006.

[50] F. Bonini, B. Burle, C. Liégeois-Chauvel, J. Régis, P. Chauvel, and F. Vidal. Action monitoring and medial frontal cortex: leading role of supplementary motor area. *Science*, 343(6173):888–891, 2014.

[51] K. Bootsveld, F. Träber, W. Kaiser, G. Layer, C. Elger, A. Hufnagel, J. Gieseke, and M. Reiser. Localisation of intracranial eeg electrodes using three dimensional surface reconstructions of the brain. *European Radiology*, 4(1):52–56, 1994.

[52] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach. Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution. *Neuroimage*, 20(2):643–656, 2003.

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

[54] A. A. Boulton, G. B. Baker, and C. H. Vanderwolf. *Neurophysiological techniques: applications to neural systems*. Springer, 1990.

[55] M. E. Bratman. *Shared agency: A planning theory of acting together*. Oxford University Press, 2013.

[56] M. Brázdil, R. Roman, P. Daniel, and I. Rektor. Intracerebral error-related negativity in a simple go/nogo task. *Journal of Psychophysiology*, 19(4):244–255, 2005.

[57] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[58] C. Brogna, S. Gil Robles, and H. Duffau. Brain tumors and epilepsy. *Expert review of neurotherapeutics*, 8(6):941–955, 2008.

[59] E. C. Brown, R. Rothermel, M. Nishida, C. Juhász, O. Muzik, K. Hoechstetter, S. Sood, H. T. Chugani, and E. Asano. In vivo animation of auditory-language-induced gamma-oscillations in children with intractable focal epilepsy. *Neuroimage*, 41(3):1120–1131, 2008.

[60] N. Brunel and X.-J. Wang. What determines the frequency of fast network oscillations with irregular neural discharges? i. synaptic dynamics and excitation-inhibition balance. *Journal of neurophysiology*, 90(1):415–430, 2003.

[61] A. E. Bryson. A gradient method for optimizing multi-stage allocation processes. In *Proc. Harvard Univ. Symposium on digital computers and their applications*, volume 72, 1961.

[62] F. Burget, L. Fiederer, D. Kuhner, M. Völker, J. Aldinger, R. T. Schirrmeister, C. Do, J. Bödecker, B. Nebel, T. Ball, and W. Burgard. Acting thoughts: Towards a mobile robotic service assistant for users with limited communication skills. In *Proc. of the IEEE European Conference on Mobile Robotics (ECMR)*, Paris, France, 2017.

[63] G. Bush, P. Luu, and M. I. Posner. Cognitive and emotional influences in anterior cingulate cortex. *Trends in cognitive sciences*, 4(6):215–222, 2000.

[64] S. A. Butterfill and C. Sinigaglia. Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, 88(1):119–145, 2014.

[65] A. Buttfield, P. W. Ferrez, and J. R. Millan. Towards a robust bci: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):164–168, 2006.

[66] G. Buzsáki and A. Draguhn. Neuronal oscillations in cortical networks. *science*, 304(5679):1926–1929, 2004.

[67] G. Buzsáki, C. A. Anastassiou, and C. Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407, 2012.

[68] T. Carlson and J. d. R. Millan. Brain-controlled wheelchairs: a robotic architecture. *IEEE Robotics & Automation Magazine*, 20(1):65–73, 2013.

[69] J. Carp and R. J. Compton. Alpha power is influenced by performance errors. *Psychophysiology*, 46(2):336–343, 2009.

[70] C. S. Carter, T. S. Braver, D. M. Barch, M. M. Botvinick, D. Noll, and J. D. Cohen. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364):747–749, 1998.

[71] J. F. Cavanagh, M. X. Cohen, and J. J. Allen. Prelude to and resolution of an error: Eeg phase synchrony reveals cognitive control dynamics during action monitoring. *Journal of Neuroscience*, 29(1):98–105, 2009.

[72] M. C. Cervenka, D. Boatman-Reich, J. Ward, P. J. Franaszczuk, and N. Crone. Language mapping in multilingual patients: electrocorticography and cortical stimulation during naming. *Frontiers in human neuroscience*, 5:13, 2011.

[73] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[74] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán. Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Frontiers in neuroscience*, 8:208, 2014.

[75] Z. Y. Chin, K. K. Ang, C. Wang, C. Guan, and H. Zhang. Multi-class filter bank common spatial pattern for four-class motor imagery bci. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 571–574. IEEE, 2009.

[76] P. Chlebus, M. Brázdil, P. Hluštík, M. Mikl, M. Pažourková, and P. Krupa. Handedness shift as a consequence of motor cortex reorganization after early functional impairment in left temporal lobe epilepsy—an fmri case report. *Neurocase*, 10(4): 326–329, 2004.

[77] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *arXiv preprint arXiv:1202.2745*, 2012.

[78] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[79] M. X. Cohen, K. R. Ridderinkhof, S. Haupt, C. E. Elger, and J. Fell. Medial frontal cortex and response conflict: evidence from human intracranial eeg and medial frontal cortex lesion. *Brain research*, 1238:127–142, 2008.

[80] M. Congedo, A. Barachant, and R. Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3): 155–174, 2017.

[81] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[82] M. Cossu, F. Cardinale, L. Castana, A. Citterio, S. Francione, L. Tassi, A. L. Benabid, and G. Lo Russo. Stereoelectroencephalography in the presurgical evaluation of focal epilepsy: a retrospective analysis of 215 procedures. *Neurosurgery*, 57(4): 706–718, 2005.

[83] H. D. Critchley, J. Tang, D. Glaser, B. Butterworth, and R. J. Dolan. Anterior cingulate activity during error and autonomic response. *Neuroimage*, 27(4):885–895, 2005.

[84] N. E. Crone, D. L. Miglioretti, B. Gordon, J. M. Sieracki, M. T. Wilson, S. Uematsu, and R. P. Lesser. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. i. alpha and beta event-related desynchronization. *Brain: a journal of neurology*, 121(12):2271–2299, 1998.

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

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

[87] D. J. Cunningham. *Contribution to the surface anatomy of the cerebral hemispheres*, volume 7. Academy House, 1892.

[88] S. S. Dalal, E. Edwards, H. E. Kirsch, N. M. Barbaro, R. T. Knight, and S. S. Nagarajan. Localization of neurosurgically implanted electrodes via photograph–mri–radiograph coregistration. *Journal of neuroscience methods*, 174(1):106–115, 2008.

[89] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry. What is a robot companion-friend, assistant or butler? In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1192–1197. IEEE, 2005.

[90] P. A. Davis. Effects of acoustic stimuli on the waking human brain. *Journal of neurophysiology*, 2(6):494–499, 1939.

[91] B. Dawant, S. Hartmann, S. Pan, and S. Gadamsetty. Brain atlas deformation in the presence of small and large space-occupying tumors. *Computer Aided Surgery: Official Journal of the International Society for Computer Aided Surgery (ISCAS)*, 7(1):1–10, 2002.

[92] S. Dehaene, M. I. Posner, and D. M. Tucker. Localization of a neural system for error detection and compensation. *Psychological Science*, 5(5):303–305, 1994.

[93] E. Demandt, C. Mehring, K. Vogt, A. Schulze-Bonhage, A. Aertsen, and T. Ball. Reaching movement onset-and end-related characteristics of eeg spectral power modulations. *Frontiers in neuroscience*, 6:65, 2012.

[94] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[95] J. Derix, O. Iljina, J. Weiske, A. Schulze-Bonhage, A. Aertsen, and T. Ball. From speech to thought: the neuronal basis of cognitive units in non-experimental, real-life communication investigated using ecog. *Frontiers in human neuroscience*, 8: 383, 2014.

[96] A. Destexhe. Spike-and-wave oscillations based on the properties of gabab receptors. *Journal of Neuroscience*, 18(21):9099–9111, 1998.

[97] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler. All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 321–326. ACM, 2002.

[98] C. Do and W. Burgard. Accurate pouring with an autonomous robot using an rgb-d camera. In *The 15th International Conference on Intelligent Autonomous Systems (IAS)*, Baden Baden, Germany, 2018.

[99] C. Do, T. Schubert, and W. Burgard. A probabilistic approach to liquid level detection in cups using an rgb-d camera. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, 2016.

[100] C. Do, C. Gordillo, and W. Burgard. Learning to pour using deep deterministic policy gradients. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018.

[101] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[102] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Muller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE transactions on biomedical engineering*, 53(11):2274–2281, 2006.

[103] J.-B. du Prel, B. Röhrig, G. Hommel, and M. Blettner. Auswahl statistischer testverfahren. *Dtsch Arztebl*, 107(19):343–348, 2010.

[104] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[105] A. R. Dykstra, A. M. Chan, B. T. Quinn, R. Zepeda, C. J. Keller, J. Cormier, J. R. Madsen, E. N. Eskandar, and S. S. Cash. Individualized localization and cortical surface-based registration of intracranial electrodes. *Neuroimage*, 59(4): 3563–3570, 2012.

[106] S. B. Eickhoff, A. Schleicher, K. Zilles, and K. Amunts. The human parietal operculum. i. cytoarchitectonic mapping of subdivisions. *Cerebral cortex*, 16(2): 254–267, 2005.

[107] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, and K. Zilles. A new spm toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, 25(4):1325–1335, 2005.

[108] S. B. Eickhoff, S. Heim, K. Zilles, and K. Amunts. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage*, 32 (2):570–582, 2006.

[109] A. K. Engel, C. K. Moll, I. Fried, and G. A. Ojemann. Invasive recordings from the human brain: clinical insights and beyond. *Nature Reviews Neuroscience*, 6(1): 35, 2005.

[110] C. W. Eriksen and B. A. Eriksen. Target redundancy in visual search: Do repetitions of the target within thedisplay impair processing? *Perception & Psychophysics*, 26 (3):195–205, 1979.

[111] N. Even-Chen, S. D. Stavisky, C. Pandarinath, P. Nuyujukian, C. H. Blabe, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy. Feasibility of automatic error detect-and-undo system in human intracortical brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(8):1771–1784, 2018.

[112] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[113] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke. Effects of crossmodal divided attention on late erp components. ii. error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, 78(6):447–455, 1991.

[114] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein. Erp components on reaction errors and their functional significance: a tutorial. *Biological psychology*, 51(2-3):87–107, 2000.

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

[116] S. Fauser, H.-J. Huppertz, T. Bast, K. Strobl, G. Pantazis, D.-M. Altenmueller, B. Feil, S. Rona, C. Kurth, D. Rating, et al. Clinical characteristics in focal cortical dysplasia: a retrospective evaluation in a series of 120 patients. *Brain*, 129(7): 1907–1916, 2006.

[117] S. Fauser, T. Bast, D.-M. Altenmüller, J. Schulte-Mönting, K. Strobl, B. J. Steinhoff, J. Zentner, and A. Schulze-Bonhage. Factors influencing surgical outcome in patients with focal cortical dysplasia. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(1):103–105, 2008.

[118] S. Fauser, C. Essang, D.-M. Altenmüller, A. M. Staack, B. J. Steinhoff, K. Strobl, T. Bast, S. Schubert-Bast, U. Stephani, G. Wiegand, et al. Long-term seizure outcome in 211 patients with focal cortical dysplasia. *Epilepsia*, 56(1):66–76, 2015.

[119] P. Fettes, L. Schulze, and J. Downar. Cortico-striatal-thalamic loop circuits of the orbitofrontal cortex: promising therapeutic targets in psychiatric illness. *Frontiers in systems neuroscience*, 11:25, 2017.

[120] M. Fiore, A. Clodic, and R. Alami. On planning and task achievement modalities for human-robot collaboration. In *Experimental Robotics*, pages 293–306. Springer, 2016.

[121] L. Fisch, E. Privman, M. Ramot, M. Harel, Y. Nir, S. Kipervasser, F. Andelman, M. Y. Neufeld, U. Kramer, I. Fried, et al. Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron*, 64 (4):562–574, 2009.

[122] R. A. Fisher. The coefficient of racial likeness and the future of craniometry. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 66: 57–63, 1936.

[123] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[124] J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[125] M. Frigo and S. G. Johnson. Fftw: An adaptive software architecture for the fft. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1381–1384. IEEE, 1998.

[126] A. Frisoli, C. Loconsole, D. Leonardis, F. Banno, M. Barsotti, C. Chisari, and M. Bergamasco. A new gaze-bci-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1169–1179, 2012.

[127] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.

[128] J. Fuster. *The prefrontal cortex*. Academic Press, 2015.

[129] X. Gao, D. Xu, M. Cheng, and S. Gao. A bci-based environmental controller for the motion-disabled. *IEEE Transactions on neural systems and rehabilitation engineering*, 11(2):137–140, 2003.

[130] D. Garcia-Gasulla, A. Vilalta, F. Parés, J. Moreno, E. Ayguadé, J. Labarta, U. Cortés, and T. Suzumura. An out-of-the-box full-network embedding for convolutional neural networks. *arXiv preprint arXiv:1705.07706*, 2017.

[131] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers. The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, 35 (4):1674–1684, 2007.

[132] W. J. Gehring and A. R. Willoughby. The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295(5563):2279–2282, 2002.

[133] W. J. Gehring, B. Goss, M. G. Coles, D. E. Meyer, and E. Donchin. A neural system for error detection and compensation. *Psychological science*, 4(6):385–390, 1993.

[134] W. J. Gehring, J. Himle, and L. G. Nisenson. Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological science*, 11(1):1–6, 2000.

[135] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–10. Springer, 1998.

[136] A. Gevins and M. E. Smith. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4 (1-2):113–131, 2003.

[137] S. Geyer. Prologue: Toward the concept of a cortical control of voluntary movements. In *The Microstructural Border Between the Motor and the Cognitive Domain in the Human Cerebral Cortex*, pages 1–8. Springer, 2004.

[138] S. Geyer, A. Ledberg, A. Schleicher, S. Kinomura, T. Schormann, U. Bürgel, T. Klingberg, J. Larsson, K. Zilles, and P. E. Roland. Two different areas within the primary motor cortex of man. *Nature*, 382(6594):805, 1996.

[139] S. Geyer, A. Schleicher, and K. Zilles. Areas 3a, 3b, and 1 of human primary somatosensory cortex: 1. microstructural organization and interindividual variability. *Neuroimage*, 10(1):63–83, 1999.

[140] S. Geyer, T. Schormann, H. Mohlberg, and K. Zilles. Areas 3a, 3b, and 1 of human primary somatosensory cortex: 2. spatial normalization to standard anatomical space. *Neuroimage*, 11(6):684–696, 2000.

[141] E. Gherri and M. Eimer. Links between eye movement preparation and the attentional processing of tactile events: an event-related brain potential study. *Clinical neurophysiology*, 119(11):2587–2597, 2008.

[142] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[143] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2016.

[144] D. E. Goldman. Potential, impedance, and rectification in membranes. *The Journal of general physiology*, 27(1):37–60, 1943.

[145] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[146] M. A. Goodrich, A. C. Schultz, et al. Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275, 2008.

[147] C. Grefkes, S. Geyer, T. Schormann, P. Roland, and K. Zilles. Human somatosensory area 2: observer-independent cytoarchitectonic mapping, interindividual variability, and population map. *Neuroimage*, 14(3):617–631, 2001.

[148] D. M. Groppe, S. Bickel, A. R. Dykstra, X. Wang, P. Mégevand, M. R. Mercier, F. A. Lado, A. D. Mehta, and C. J. Honey. ielvis: An open source matlab toolbox for localizing and visualizing human intracranial electrode data. *Journal of neuroscience methods*, 281:40–48, 2017.

[149] A. Gunduz, P. Brunner, A. Daitch, E. C. Leuthardt, A. L. Ritaccio, B. Pesaran, and G. Schalk. Neural correlates of visual–spatial attention in electrocorticographic signals in humans. *Frontiers in human neuroscience*, 5:89, 2011.

[150] D. Gupta, N. J. Hill, M. A. Adamo, A. Ritaccio, and G. Schalk. Localizing ecog electrodes on the cortical anatomy without post-implantation imaging. *NeuroImage: Clinical*, 6:64–76, 2014.

[151] P. Haggard. Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18 (4):196, 2017.

[152] G. Hajcak, N. McDonald, and R. F. Simons. To err is autonomic: Error-related brain potentials, ans activity, and post-error compensatory behavior. *Psychophysiology*, 40(6):895–903, 2003.

[153] G. Hajcak, N. McDonald, and R. F. Simons. Error-related psychophysiology and negative affect. *Brain and cognition*, 56(2):189–197, 2004.

[154] T. Ham, A. Leff, X. de Boissezon, A. Joffe, and D. J. Sharp. Cognitive control and the salience network: an investigation of error processing and effective connectivity. *Journal of Neuroscience*, 33(16):7091–7098, 2013.

[155] L. S. Hamilton, D. L. Chang, M. B. Lee, and E. F. Chang. Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Frontiers in neuroinformatics*, 11:62, 2017.

[156] R. Hari, N. Forss, S. Avikainen, E. Kirveskari, S. Salenius, and G. Rizzolatti. Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences*, 95(25):15061–15065, 1998.

[157] T. Harmony, T. Fernández, J. Silva, J. Bernal, L. Díaz-Comas, A. Reyes, E. Marosi, M. Rodríguez, and M. Rodríguez. Eeg delta activity: an indicator of attention to internal processing during performance of mental tasks. *International journal of psychophysiology*, 24(1-2):161–171, 1996.

[158] K. G. Hartmann, R. T. Schirrmeister, and T. Ball. Hierarchical internal representation of spectral features in deep convolutional networks trained for eeg decoding. In *Brain-Computer Interface (BCI), 2018 6th International Conference on*, pages 1–6. IEEE, 2018.

[159] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *european conference on computer vision*, pages 346–361. Springer, 2014.

[160] F. A. Heilmeyer, R. T. Schirrmeister, L. D. Fiederer, M. Völker, J. Behncke, and T. Ball. A framework for large-scale evaluation of deep learning for eeg. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.

[161] D. Hermes, K. J. Miller, H. J. Noordmans, M. J. Vansteensel, and N. F. Ramsey. Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of neuroscience methods*, 185(2):293–298, 2010.

[162] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton. Electrical signs of selective attention in the human brain. *Science*, 182(4108):177–180, 1973.

[163] P. J. Hinds, T. L. Roberts, and H. Jones. Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19 (1):151–181, 2004.

[164] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[165] A. Holm, K. Lukander, J. Korpela, M. Sallinen, and K. M. Müller. Estimating brain load from the eeg. *The Scientific World Journal*, 9:639–651, 2009.

[166] C. B. Holroyd, J. Dien, and M. G. Coles. Error-related scalp potentials elicited by hand and foot movements: evidence for an output-independent error-processing system in humans. *Neuroscience letters*, 242(2):65–68, 1998.

[167] C. B. Holroyd, P. Praamstra, E. Plat, and M. G. Coles. Spared error-related potentials in mild to moderate parkinson's disease. *Neuropsychologia*, 40(12): 2116–2124, 2002.

[168] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8): 2554–2558, 1982.

[169] M.-P. Hosseini, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh. Optimized deep learning for eeg big data and seizure prediction bci via internet of things. *IEEE Transactions on Big Data*, 3(4):392–404, 2017.

[170] K.-C. Huang, T.-Y. Huang, C.-H. Chuang, J.-T. King, Y.-K. Wang, C.-T. Lin, and T.-P. Jung. An eeg-based fatigue detection and mitigation system. *International journal of neural systems*, 26(04):1650018, 2016.

[171] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[172] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[173] I. Iturrate, L. Montesano, and J. Minguez. Shared-control brain-computer interface for a two dimensional reaching task using eeg error-related potentials. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5258–5262. IEEE, 2013.

[174] I. Iturrate, L. Montesano, and J. Minguez. Task-dependent signal variations in eeg error-related potentials for brain–computer interfaces. *Journal of neural engineering*, 10(2):026024, 2013.

[175] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. Millán. Latency correction of event-related potentials between different experimental protocols. *Journal of neural engineering*, 11(3):036005, 2014.

[176] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. d. R. Millán. Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control. *Scientific reports*, 5:13893, 2015.

[177] H. Jasper and W. Penfield. Electrocorticograms in man: effect of voluntary movement upon the electrical activity of the precentral gyrus. *Archiv für Psychiatrie und Nervenkrankheiten*, 183(1-2):163–174, 1949.

[178] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

[179] D. Jeffreys and J. Axford. Source locations of pattern-specific components of human visual evoked potentials. i. component of striate cortical origin. *Experimental brain research*, 16(1):1–21, 1972.

[180] O. Jensen and L. L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in cognitive sciences*, 11(7):267–269, 2007.

[181] K. Jerbi, J. R. Vidal, T. Ossandon, S. S. Dalal, J. Jung, D. Hoffmann, L. Minotti, O. Bertrand, P. Kahane, and J.-P. Lachaux. Exploring the electrophysiological correlates of the default-mode network with intracerebral eeg. *Frontiers in systems neuroscience*, 4:27, 2010.

[182] T.-P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski. Estimating alertness from the eeg power spectrum. *IEEE transactions on biomedical engineering*, 44(1): 60–69, 1997.

[183] E. R. Kandel, J. H. Schwartz, T. M. Jessell, D. of Biochemistry, M. B. T. Jessell, S. Siegelbaum, and A. Hudspeth. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[184] H.-O. Karnath, M. Fruhmann Berger, R. Zopf, and W. Küker. Using spm normalization for lesion analysis in spatial neglect. *Brain*, 127(4):e10–e10, 2004.

[185] A. Karpathy. Neural networks part 1: Setting up the architecture. *Notes for CS231n Convolutional Neural Networks for Visual Recognition, Stanford University. http://cs231n. github. io/neural-networks-1*, 2015.

[186] H. J. Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

[187] M. Kern, A. Aertsen, A. Schulze-Bonhage, and T. Ball. Heart cycle-related effects on event-related potentials, spectral power changes, and connectivity patterns in the human ecog. *Neuroimage*, 81:178–190, 2013.

[188] M. Khamassi, B. Girard, A. Clodic, D. Sandra, E. Renaudo, E. Pacherie, R. Alami, and R. Chatila. Integration of action, joint action and learning in robot cognitive architectures. *Intellectica-La revue de l'Association pour la Recherche sur les sciences de la Cognition (ARCo)*, 2016(65):169–203, 2016.

[189] S. K. Kim and E. A. Kirchner. Handling few training data: classifier transfer between different types of error-related potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(3):320–332, 2016.

[190] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[191] T. A. Klein, M. Ullsperger, and C. Danielmeier. Error awareness and the insula: links to neurological and psychiatric diseases. *Frontiers in human neuroscience*, 7: 14, 2013.

[192] W. Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195, 1999.

[193] W. Klimesch, M. Doppelmayr, H. Russegger, T. Pachinger, and J. Schwaiger. Induced alpha band power changes in the human eeg and attention. *Neuroscience letters*, 244(2):73–76, 1998.

[194] T. Koelewijn, H. T. van Schie, H. Bekkering, R. Oostenveld, and O. Jensen. Motor-cortical beta oscillations are modulated by correctness of observed action. *Neuroimage*, 40(2):767–775, 2008.

[195] T. Kohonen. Correlation matrix memories. *IEEE transactions on computers*, 100 (4):353–359, 1972.

[196] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[197] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[198] Z. J. Koles, M. S. Lazar, and S. Z. Zhou. Spatial patterns underlying population differences in the background eeg. *Brain topography*, 2(4):275–284, 1990.

[199] V. Kolev, M. Falkenstein, and J. Yordanova. Aging and error processing: Time-frequency analysis of error-related potentials. *Journal of Psychophysiology*, 19(4): 289–297, 2005.

[200] H. H. Kornhuber and L. Deecke. Hirnpotentialänderungen bei willkürbewegungen und passiven bewegungen des menschen: Bereitschaftspotential und reafferente potentiale. *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere*, 284(1):1–17, 1965.

[201] D. Kovalev, J. Spreer, J. Honegger, J. Zentner, A. Schulze-Bonhage, and H.-J. Huppertz. Rapid and fully automated visualization of subdural electrodes in the presurgical evaluation of epilepsy patients. *American journal of neuroradiology*, 26(5):1078–1083, 2005.

[202] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[203] E. Labyt, E. Houdayer, F. Cassim, J. Bourriez, P. Derambure, and H. Devanne. Motor representation areas in epileptic patients with focal motor seizures: a tms study. *Epilepsy research*, 75(2-3):197–205, 2007.

[204] J.-P. Lachaux, E. Rodriguez, J. Martinerie, C. Adam, D. Hasboun, and F. J. Varela. A quantitative study of gamma-band activity in human intracranial recordings triggered by visual stimuli. *European Journal of Neuroscience*, 12(7):2608–2622, 2000.

[205] J.-P. Lachaux, M. Chavez, and A. Lutz. A simple measure of correlation across time, frequency and space between continuous brain signals. *Journal of neuroscience methods*, 123(2):175–188, 2003.

[206] J.-P. Lachaux, N. George, C. Tallon-Baudry, J. Martinerie, L. Hugueville, L. Minotti, P. Kahane, and B. Renault. The many faces of the gamma band response to complex visual stimuli. *Neuroimage*, 25(2):491–501, 2005.

[207] Y. Le Cun, L. D. Jackel, B. Boser, J. S. Denker, H. P. Graf, I. Guyon, D. Henderson, R. E. Howard, and W. Hubbard. Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27 (11):41–46, 1989.

[208] Y. LeCun, K. Kavukcuoglu, C. Farabet, et al. Convolutional networks and applications in vision. In *ISCAS*, volume 2010, pages 253–256, 2010.

[209] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

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

[211] R. Leeb, D. Friedman, G. R. Müller-Putz, R. Scherer, M. Slater, and G. Pfurtscheller. Self-paced (asynchronous) bci control of a wheelchair in virtual environments: a case study with a tetraplegic. *Computational intelligence and neuroscience*, 2007, 2007.

[212] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller. Spatio-spectral filters for improving the classification of single trial eeg. *IEEE transactions on biomedical engineering*, 52(9):1541–1548, 2005.

[213] E. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenburg, D. Barbour, and G. Schalk. Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task. *Frontiers in human neuroscience*, 6:99, 2012.

[214] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran. A brain–computer interface using electrocorticographic signals in humans. *Journal of neural engineering*, 1(2):63, 2004.

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

[216] F. Li, G. Zhang, W. Wang, R. Xu, T. Schnell, J. Wen, F. McKenzie, and J. Li. Deep models for engagement assessment with scarce label information. *IEEE Transactions on Human-Machine Systems*, 47(4):598–605, 2017.

[217] T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, 10(4): 453–472, 2006.

[218] G. Littlewort, M. S. Bartlett, I. R. Fasel, J. Chenu, T. Kanda, H. Ishiguro, and J. R. Movellan. Towards social robots: Automatic evaluation of human-robot interaction by facial expression classification. In *Advances in neural information processing systems*, pages 1563–1570, 2004.

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

[220] S. J. Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[221] P. Luu and D. M. Tucker. Regulating action: alternating activation of midline frontal and motor cortical networks. *Clinical Neurophysiology*, 112(7):1295–1306, 2001.

[222] P. Luu, D. M. Tucker, and S. Makeig. Frontal midline theta and the error-related negativity: neurophysiological mechanisms of action regulation. *Clinical Neurophysiology*, 115(8):1821–1835, 2004.

[223] M. Mahvash, R. König, J. Wellmer, H. Urbach, B. Meyer, and K. Schaller. Coregistration of digital photography of the human cortex and cranial magnetic resonance imaging for visualization of subdural electrodes in epilepsy surgery. *Operative Neurosurgery*, 61(suppl_5):ONS340–ONS345, 2007.

[224] S. Makeig and T.-P. Jung. Tonic, phasic, and transient eeg correlates of auditory awareness in drowsiness. *Cognitive Brain Research*, 4(1):15–25, 1996.

[225] A. Malikovic, K. Amunts, A. Schleicher, H. Mohlberg, S. B. Eickhoff, M. Wilms, N. Palomero-Gallagher, E. Armstrong, and K. Zilles. Cytoarchitectonic analysis of the human extrastriate cortex in the region of v5/mt+: a probabilistic, stereotaxic map of area hoc5. *Cerebral cortex*, 17(3):562–574, 2006.

[226] P. Marusic, I. M. Najm, Z. Ying, R. Prayson, S. Rona, D. Nair, E. Hadar, P. Kotagal, M. D. Bej, E. Wyllie, et al. Focal cortical dysplasias in eloquent cortex: functional characteristics and correlation with mri and histopathologic changes. *Epilepsia*, 43 (1):27–32, 2002.

[227] N. Matsuzaki, R. F. Schwarzlose, M. Nishida, N. Ofen, and E. Asano. Upright face-preferential high-gamma responses in lower-order visual areas: evidence from intracranial recordings in children. *Neuroimage*, 109:249–259, 2015.

[228] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, J. Lancaster, et al. A probabilistic atlas of the human brain: theory and rationale for its development. *Neuroimage*, 2 (2):89–101, 1995.

[229] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[230] A. R. Merzagora, T. J. Coffey, M. R. Sperling, A. Sharan, B. Litt, G. Baltuch, and J. Jacobs. Repeated stimuli elicit diminished high-gamma electrocorticographic responses. *NeuroImage*, 85:844–852, 2014.

[231] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop-Volume 27*, pages 97–111. JMLR. org, 2011.

[232] T. Milekovic, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Detection of error related neuronal responses recorded by electrocorticography in humans during continuous movements. *PloS one*, 8(2):e55235, 2013.

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

[234] M. Minsky and S. Papert. Perceptrons. 1969. *Cited on*, page 1, 1990.

[235] P. P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical journal*, 76(2):691–708, 1999.

[236] D. Moore, G. P. McCabe, and B. Craig. *Introduction to the Practice of Statistics: w/Student CD*. San Francisco, CA: Freeman, 2012.

[237] P. Morosan, J. Rademacher, A. Schleicher, K. Amunts, T. Schormann, and K. Zilles. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage*, 13(4):684–701, 2001.

[238] K. Morris, T. J. O'Brien, M. J. Cook, M. Murphy, and S. C. Bowden. A computer-generated stereotactic "virtual subdural grid" to guide resective epilepsy surgery. *American journal of neuroradiology*, 25(1):77–83, 2004.

[239] R. Mukamel, A. D. Ekstrom, J. Kaplan, M. Iacoboni, and I. Fried. Single-neuron responses in humans during execution and observation of actions. *Current biology*, 20(8):750–756, 2010.

[240] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filters for single-trial eeg classification in a movement task. *Clinical neurophysiology*, 110(5):787–798, 1999.

[241] T. Nagasawa, C. Juhász, R. Rothermel, K. Hoechstetter, S. Sood, and E. Asano. Spontaneous and visually driven high-frequency oscillations in the occipital cortex: Intracranial recording in epileptic patients. *Human brain mapping*, 33(3):569–583, 2012.

[242] T. P. Naidich, A. G. Valavanis, and S. Kubik. Anatomic relationships along the low-middle convexity: Part i-normal specimens and magnetic resonance imaging. *Neurosurgery*, 36(3):517–532, 1995.

[243] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[244] D. E. Nee, S. Kastner, and J. W. Brown. Functional heterogeneity of conflict, error, task-switching, and unexpectedness effects within medial prefrontal cortex. *Neuroimage*, 54(1):528–540, 2011.

[245] K. Nickel and R. Stiefelhagen. Visual recognition of pointing gestures for human–robot interaction. *Image and vision computing*, 25(12):1875–1884, 2007.

[246] E. Niedermeyer. Alpha rhythms as physiological and abnormal phenomena. *International Journal of Psychophysiology*, 26(1-3):31–49, 1997.

[247] E. Niedermeyer and F. L. da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[248] S. Nieuwenhuis, K. R. Ridderinkhof, J. Blom, G. P. Band, and A. Kok. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, 38(5):752–760, 2001.

[249] A. Nijholt. Bci for games: A 'state of the art'survey. In *International Conference on Entertainment Computing*, pages 225–228. Springer, 2008.

[250] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch. Eeg coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and clinical neurophysiology*, 103(5):499–515, 1997.

[251] P. L. Nunez, R. Srinivasan, et al. *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

[252] L. M. Oberman, J. P. McCleery, V. S. Ramachandran, and J. A. Pineda. Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, 70(13-15):2194–2203, 2007.

[253] R. G. O'Connell, P. M. Dockree, I. H. Robertson, M. A. Bellgrove, J. J. Foxe, and S. P. Kelly. Uncovering the neural signature of lapsing attention: electrophysiological signals predict errors up to 20 s before they occur. *Journal of Neuroscience*, 29 (26):8604–8611, 2009.

[254] T. O'Connor and C. Sandis. *A Companion to the Philosophy of Action*. John Wiley & Sons, 2011.

[255] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:1, 2011.

[256] E. Pacherie. Framing joint action. *Review of Philosophy and Psychology*, 2(2): 173–192, 2011.

[257] M. L. Padilla, R. A. Wood, L. A. Hale, and R. T. Knight. Lapses in a prefrontal-extrastriate preparatory attention network predict mistakes. *Journal of cognitive neuroscience*, 18(9):1477–1487, 2006.

[258] P. E. Pailing, S. J. Segalowitz, J. Dywan, and P. L. Davies. Error negativity and response control. *Psychophysiology*, 39(2):198–206, 2002.

[259] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[260] X. Papademetris, M. P. Jackowski, N. Rajeevan, M. DiStasio, H. Okuda, R. T. Constable, and L. H. Staib. Bioimage suite: An integrated medical image analysis suite: An update. *The insight journal*, 2006:209, 2006.

[261] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda. Response error correction-a demonstration of improved human-machine performance using real-time eeg monitoring. *IEEE transactions on neural systems and rehabilitation engineering*, 11(2):173–177, 2003.

[262] W. Penfield and E. Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443, 1937.

[263] W. Penfield and T. Rasmussen. The cerebral cortex of man; a clinical study of localization of function. 1950.

[264] M. Penttonen and G. Buzsáki. Natural logarithmic relationship between brain oscillators. *Thalamus & Related Systems*, 2(2):145–152, 2003.

[265] D. B. Percival and A. T. Walden. *Spectral analysis for physical applications*. cambridge university press, 1993.

[266] D. B. Percival and A. T. Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge university press, 2006.

[267] K. H. Pettersen and G. T. Einevoll. Amplitude variability and extracellular low-pass filtering of neuronal spikes. *Biophysical journal*, 94(3):784–802, 2008.

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

[269] G. Pfurtscheller and F. L. Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[270] G. Pfurtscheller, C. Guger, and H. Ramoser. Eeg-based brain-computer interface using subject-specific spatial filters. In *International Work-Conference on Artificial Neural Networks*, pages 248–254. Springer, 1999.

[271] G. Pfurtscheller, C. Neuper, G. Muller, B. Obermaier, G. Krausz, A. Schlogl, R. Scherer, B. Graimann, C. Keinrath, D. Skliris, et al. Graz-bci: state of the art and clinical applications. *IEEE Transactions on neural systems and rehabilitation engineering*, 11(2):1–4, 2003.

[272] G. Pfurtscheller, C. Neuper, and N. Birbaumer. Human brain-computer interface. In *Motor Cortex in virtual Movements*. CRC Press, 2005.

[273] T. A. Pieters, C. R. Conner, and N. Tandon. Recursive grid partitioning on a cortical surface model: an optimized technique for the localization of implanted subdural electrodes. *Journal of neurosurgery*, 118(5):1086–1097, 2013.

[274] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Prediction of arm movement trajectories from ecog-recordings in humans. *Journal of neuroscience methods*, 167(1):105–114, 2008.

[275] E. J. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

[276] J. Polich. Frequency, intensity, and duration as determinants of p300 from auditory stimuli. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 6(3):277–286, 1989.

[277] W. Prinz, M. Beisert, and A. Herwig. *Action science: Foundations of an emerging discipline*. MIT Press, 2013.

[278] A. D. Protopapas, M. Vanier, and J. M. Bower. Simulating large networks of neurons. *Methods in neuronal modeling: From ions to networks*, page 461, 1998.

[279] P. L. Purdon, E. T. Pierce, E. A. Mukamel, M. J. Prerau, J. L. Walsh, K. F. K. Wong, A. F. Salazar-Gomez, P. G. Harrell, A. L. Sampson, A. Cimenser, et al. Electroencephalogram signatures of loss and recovery of consciousness from propofol. *Proceedings of the National Academy of Sciences*, page 201221180, 2013.

[280] P. Rabbitt. Consciousness is slower than you think. *The Quarterly Journal of Experimental Psychology Section A*, 55(4):1081–1092, 2002.

[281] P. M. Rabbitt. Three kinds of error-signalling responses in a serial choice task. *Quarterly Journal of Experimental Psychology*, 20(2):179–188, 1968.

[282] J. Rademacher, P. Morosan, T. Schormann, A. Schleicher, C. Werner, H.-J. Freund, and K. Zilles. Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage*, 13(4):669–683, 2001.

[283] J. Raethjen, M. Lindemann, M. Dümpelmann, R. Wenzelburger, H. Stolze, G. Pfister, C. E. Elger, J. Timmer, and G. Deuschl. Corticomuscular coherence in the 6–15 hz band: is the cortex involved in the generation of physiologic tremor? *Experimental Brain Research*, 142(1):32–40, 2002.

[284] W. Rall and G. M. Shepherd. Theoretical reconstruction of field potentials and dendrodendritic synaptic interactions in olfactory bulb. *Journal of neurophysiology*, 31(6):884–915, 1968.

[285] R. A. Ramadan and A. V. Vasilakos. Brain computer interface: control signals review. *Neurocomputing*, 223:26–44, 2017.

[286] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

[287] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

[288] Y. Ren and Y. Wu. Convolutional deep belief networks for feature extraction of eeg signal. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 2850–2853. IEEE, 2014.

[289] A. Riesel, A. Weinberg, T. Moran, and G. Hajcak. Time course of error-potentiated startle and its relationship to error-related brain activity. *Journal of Psychophysiology*, 2013.

[290] E. K. Ritzl, A. M. Wohlschlaeger, N. Crone, A. Wohlschlaeger, L. Gingis, C. Bowers, and D. F. Boatman. Transforming electrocortical mapping data into standardized common space. *Clinical EEG and neuroscience*, 38(3):132–136, 2007.

[291] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.

[292] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141, 1996.

[293] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[294] F. Rosenow and K. Menzler. Invasive eeg studies in tumor-related epilepsy: When are they indicated and with what kind of electrodes? *Epilepsia*, 54:61–65, 2013.

[295] C. Rottschy, S. B. Eickhoff, A. Schleicher, H. Mohlberg, M. Kujovic, K. Zilles, and K. Amunts. Ventral visual cortex in humans: cytoarchitectonic mapping of two extrastriate areas. *Human brain mapping*, 28(10):1045–1059, 2007.

[296] J. Ruescher, O. Iljina, D.-M. Altenmüller, A. Aertsen, A. Schulze-Bonhage, and T. Ball. Somatotopic mapping of natural upper-and lower-extremity movements and speech production with high gamma electrocorticography. *Neuroimage*, 81: 164–177, 2013.

[297] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[298] M. F. Rushworth. Intention, choice, and the medial frontal cortex. *Annals of the New York Academy of Sciences*, 1124(1):181–207, 2008.

[299] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus. Correcting robot mistakes in real time using eeg signals. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 6570–6577. IEEE, 2017.

[300] D. Sammler, S. Koelsch, T. Ball, A. Brandt, C. E. Elger, A. D. Friederici, M. Grigutsch, H.-J. Huppertz, T. R. Knösche, J. Wellmer, et al. Overlap of musical and linguistic syntax processing: intracranial erp evidence. *Annals of the New York Academy of Sciences*, 1169(1):494–498, 2009.

[301] G. Schalk, J. Kubanek, K. Miller, N. Anderson, E. Leuthardt, J. Ojemann, D. Limbrick, D. Moran, L. Gerhardt, and J. Wolpaw. Decoding two-dimensional movement trajectories using electrocorticographic signals in humans. *Journal of neural engineering*, 4(3):264, 2007.

[302] R. Schirrmeister, L. Gemein, K. Eggensperger, F. Hutter, and T. Ball. Deep learning with convolutional neural networks for decoding and visualization of eeg pathology. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2017 IEEE*, pages 1–7. IEEE, 2017.

[303] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[304] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[305] S. Schröer, I. Killmann, B. Frank, M. Völker, L. Fiederer, T. Ball, and W. Burgard. An autonomous robotic assistant for drinking. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 6482–6487. IEEE, 2015.

[306] R. Schubert, S. Haufe, F. Blankenburg, A. Villringer, and G. Curio. Now you'll feel it, now you won't: Eeg rhythms predict the effectiveness of perceptual masking. *Journal of cognitive neuroscience*, 21(12):2407–2419, 2009.

[307] A. Schulze-Bonhage, H. J. Huppertz, R. M. Comeau, J. B. Honegger, J. M. Spreer, and J. K. Zentner. Visualization of subdural strip and grid electrodes using curvilinear reformatting of 3d mr imaging data sets. *American journal of neuroradiology*, 23(3):400–403, 2002.

[308] M. L. Seghier, A. Ramlackhansingh, J. Crinion, A. P. Leff, and C. J. Price. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *Neuroimage*, 41(4):1253–1266, 2008.

[309] F. Sepulveda. Brain-actuated control of robot navigation. In *Advances in Robot Navigation*. InTech, 2011.

[310] P. Sermanet and Y. LeCun. Traffic sign recognition with multi-scale convolutional networks. In *IJCNN*, pages 2809–2813, 2011.

[311] A. Shenhav, M. M. Botvinick, and J. D. Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

[312] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244, 2000.

[313] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[314] A. Sinai, C. W. Bowers, C. M. Crainiceanu, D. Boatman, B. Gordon, R. P. Lesser, F. A. Lenz, and N. E. Crone. Electrocorticographic high gamma activity versus electrical cortical stimulation mapping of naming. *Brain*, 128(7):1556–1570, 2005.

[315] D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty—i. *Bell System Technical Journal*, 40(1):43–63, 1961.

[316] R. Spehlmann. The averaged electrical responses to diffuse and to patterned light in the human. *Electroencephalography and clinical neurophysiology*, 19(6):560–569, 1965.

[317] W. Speier, C. Arnold, and N. Pouratian. Integrating language models into classifiers for bci communication: a review. *Journal of neural engineering*, 13(3):031002, 2016.

[318] M. Spüler, M. Bensch, S. Kleih, W. Rosenstiel, M. Bogdan, and A. Kübler. Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a p300-bci. *Clinical Neurophysiology*, 123(7): 1328–1337, 2012.

[319] D. Sridharan, D. J. Levitin, and V. Menon. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, 105(34):12569–12574, 2008.

[320] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[321] M. Staudt, L. F. Ticini, W. Grodd, I. Krägeloh-Mann, and H.-O. Karnath. Functional topography of early periventricular brain lesions in relation to cytoarchitectonic probabilistic maps. *Brain and language*, 106(3):177–183, 2008.

[322] M. Stoeckel, A. Kleinschmidt, A. Ebner, O. Witte, and R. Seitz. Reorganization of motor representation in a patient with epilepsia partialis continua as shown by [o15]-labeled butanol positron emission tomography and functional magnetic resonance imaging. *Journal of Neuroimaging*, 12(3):276–281, 2002.

[323] K. Stubbs, P. J. Hinds, and D. Wettergreen. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems*, 22(2), 2007.

[324] C. Studholme, E. Novotny, I. Zubal, and J. S. Duncan. Estimating tissue deformation between functional images induced by intracranial electrode implantation using anatomical mri. *Neuroimage*, 13(4):561–576, 2001.

[325] Z. Y. Sun, S. Klöppel, D. Rivière, M. Perrot, R. Frackowiak, H. Siebner, and J.-F. Mangin. The effect of handedness on the shape of the central sulcus. *Neuroimage*, 60(1):332–339, 2012.

[326] J. Talairach. Approche nouvelle de la neurochirurgie de l'épilepsie: méthodologie stéréotaxique et résultats thérapeutiques. In *Congres An nuel de la Societe de Neurochirurgie de Langue Francaise. Marseille: Masson, 25-28 Juin*, 1974.

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

[328] Z. Tang, C. Li, and S. Sun. Single-trial eeg classification of motor imagery using deep convolutional neural networks. *Optik-International Journal for Light and Electron Optics*, 130:11–18, 2017.

[329] M. Tangermann. *Feature Selection for Brain-Computer Interfaces*. PhD thesis, Eberhard-Karls-Universität Tübingen, 2007.

[330] J. R. Themanson, P. J. Rosen, M. B. Pontifex, C. H. Hillman, and E. McAuley. Alterations in error-related brain activity and post-error behavior over time. *Brain and cognition*, 80(2):257–265, 2012.

[331] F. Thinnes-Elker, O. Iljina, J. K. Apostolides, F. Kraemer, A. Schulze-Bonhage, A. Aertsen, and T. Ball. Intention concepts and brain-machine interfacing. *Frontiers in psychology*, 3:455, 2012.

[332] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.

[333] G. Thut, A. Nietzel, S. A. Brandt, and A. Pascual-Leone. α-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. *Journal of Neuroscience*, 26(37):9494–9502, 2006.

[334] J. Timmer. What can be inferred from surrogate data testing? *Physical review letters*, 85(12):2647, 2000.

[335] C. Toro, G. Deuschl, R. Thatcher, S. Sato, C. Kufta, and M. Hallett. Event-related desynchronization and movement-related cortical potentials on the ecog and eeg. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 93(5):380–389, 1994.

[336] V. L. Towle, J. D. Hunter, J. C. Edgar, S. A. Chkhenkeli, M. C. Castelle, D. M. Frim, M. Kohrman, and K. E. Hecox. Frequency domain analysis of human subdural recordings. *Journal of Clinical Neurophysiology*, 24(2):205–213, 2007.

[337] L. T. Trujillo and J. J. Allen. Theta eeg dynamics of the error-related negativity. *Clinical Neurophysiology*, 118(3):645–668, 2007.

[338] A. U. Turken and D. Swick. The effect of orbitofrontal lesions on the error-related negativity. *Neuroscience letters*, 441(1):7–10, 2008.

[339] L. K. Tyler, W. D. Marslen-Wilson, B. Randall, P. Wright, B. J. Devereux, J. Zhuang, M. Papoutsi, and E. A. Stamatakis. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*, 134(2): 415–431, 2011.

[340] M. Uematsu, N. Matsuzaki, E. C. Brown, K. Kojima, and E. Asano. Human occipital cortices differentially exert saccadic suppression: intracranial recording in children. *Neuroimage*, 83:224–236, 2013.

[341] M. Ullsperger, H. A. Harsay, J. R. Wessel, and K. R. Ridderinkhof. Conscious perception of errors and its relation to the anterior insula. *Brain Structure and Function*, 214(5-6):629–643, 2010.

[342] M. Ullsperger, C. Danielmeier, and G. Jocham. Neurophysiology of performance monitoring and adaptive behavior. *Physiological reviews*, 94(1):35–79, 2014.

[343] M. Ullsperger, A. G. Fischer, R. Nigbur, and T. Endrass. Neural mechanisms and temporal dynamics of performance monitoring. *Trends in cognitive sciences*, 18 (5):259–267, 2014.

[344] H. T. van Schie, R. B. Mars, M. G. Coles, and H. Bekkering. Modulation of activity in medial frontal and motor cortices during error observation. *Nature neuroscience*, 7(5):549, 2004.

[345] J. D. Velleman. How to share an intention. *Philosophy and phenomenological research*, 57(1):29–50, 1997.

[346] C. Vesper, S. Butterfill, G. Knoblich, and N. Sebanz. A minimal architecture for joint action. *Neural Networks*, 23(8-9):998–1003, 2010.

[347] F. Vidal, B. Burle, M. Bonnet, J. Grapperon, and T. Hasbroucq. Error negativity on correct trials: a reexamination of available data. *Biological psychology*, 64(3): 265–282, 2003.

[348] J. J. Vidal. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1):157–180, 1973.

[349] J. J. Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65 (5):633–641, 1977.

[350] M. Völker. *Error-related brain responses in high-density EEG*. University of Freiburg, Freiburg, Germany, 2015.

[351] M. Völker, L. D. Fiederer, S. Berberich, J. Hammer, J. Behncke, P. Kršek, M. Tomášek, P. Marusič, P. C. Reinacher, V. A. Coenen, et al. The dynamics of error processing in the human brain as reflected by high-gamma activity in noninvasive and intracranial eeg. *NeuroImage*, 173:564–579, 2018.

[352] M. Völker, J. Hammer, R. T. Schirrmeister, J. Behncke, L. D. Fiederer, A. Schulze-Bonhage, P. Marusič, W. Burgard, and T. Ball. Intracranial error detection via deep learning. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.

[353] M. Völker, R. T. Schirrmeister, L. D. Fiederer, W. Burgard, and T. Ball. Deep transfer learning for error decoding from non-invasive eeg. In *Brain-Computer Interface (BCI), 2018 6th International Conference on*, pages 1–6. IEEE, 2018.

[354] W. G. Walter. Contingent negative variation: an electric sign of sensori-motor association and expectancy in the human brain. *Nature*, 230:380–384, 1964.

[355] C. Wang, I. Ulbert, D. L. Schomer, K. Marinkovic, and E. Halgren. Responses of human anterior cingulate cortex microdomains to error detection, conflict monitoring, stimulus-response mapping, familiarity, and orienting. *Journal of Neuroscience*, 25(3):604–613, 2005.

[356] Y.-T. Wang, K.-C. Huang, C.-S. Wei, T.-Y. Huang, L.-W. Ko, C.-T. Lin, C.-K. Cheng, and T.-P. Jung. Developing an eeg-based on-line closed-loop lapse detection and mitigation system. *Frontiers in neuroscience*, 8:321, 2014.

[357] D. Weber et al. Bioelectromagnetism matlab toolbox. *Google Scholar*, 2005.

[358] C. Weiß. *Basiswissen medizinische statistik*. Springer-Verlag, 2013.

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

[360] D. Welke*, J. Behncke*, M. Hader, R. T. Schirrmeister, A. Schönau, B. Eßmann, O. Müller, W. Burgard, and T. Ball. Brain responses during robot-error observation. *Kognitive Systeme*, 2017. *These authors contributed equally.

[361] J. Wellmer, J. Von Oertzen, C. Schaller, H. Urbach, R. König, G. Widman, D. Van Roost, and C. E. Elger. Digital photography and 3d mri–based multimodal imaging for individualized planning of resective neocortical epilepsy surgery. *Epilepsia*, 43(12):1543–1550, 2002.

[362] D. Wenke, S. Atmaca, A. Holländer, R. Liepelt, P. Baess, and W. Prinz. What is shared in joint action? issues of co-representation, response conflict, and agent identification. *Review of Philosophy and Psychology*, 2(2):147–172, 2011.

[363] P. Werbos. Beyond regression:" new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*, 1974.

[364] J. R. Wessel, C. Danielmeier, and M. Ullsperger. Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of cognitive neuroscience*, 23(10):3021–3036, 2011.

[365] J. R. Wessel, C. Danielmeier, J. B. Morton, and M. Ullsperger. Surprise and error: common neuronal architecture for the processing of errors and novelty. *Journal of Neuroscience*, 32(22):7528–7537, 2012.

[366] L. White, T. Andrews, C. Hulette, A. Richards, M. Groelle, J. Paydarfar, and D. Purves. Structure of the human sensorimotor system. i: Morphology and cytoarchitecture of the central sulcus. *Cerebral cortex (New York, NY: 1991)*, 7(1): 18–30, 1997.

[367] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1 (6):80–83, 1945.

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

[369] J. R. Wolpaw and D. J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the national academy of sciences*, 101(51):17849–17854, 2004.

[370] A. I. Yang, X. Wang, W. K. Doyle, E. Halgren, C. Carlson, T. L. Belcher, S. S. Cash, O. Devinsky, and T. Thesen. Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage*, 63(1):157–165, 2012.

[371] N. Yeung, M. M. Botvinick, and J. D. Cohen. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4): 931, 2004.

[372] J. Yordanova, M. Falkenstein, J. Hohnsbein, and V. Kolev. Parallel systems of error processing in the brain. *Neuroimage*, 22(2):590–602, 2004.

[373] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[374] T. Zander, C. Kothe, S. Jatzev, M. Luz, A. Mann, S. Welke, R. Dashuber, and M. Rötting. Das phypa-bci: Ein brain-computer-interface als kognitive schnittstelle in der mensch-maschine-interaktion. *Prospektive Gestaltung von Mensch-Technik-Interaktion*, 13:183–185, 2007.

[375] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[376] A. Zell. *Simulation neuronaler netze*, volume 1. Addison-Wesley Bonn, 1994.

[377] H. Zhang, R. Chavarriaga, Z. Khaliliardali, L. Gheorghe, I. Iturrate, and J. d R Millán. Eeg-based decoding of error-related brain activity in a real-world driving task. *Journal of neural engineering*, 12(6):066028, 2015.

[378] K. Zilles, A. Schleicher, C. Langemann, K. Amunts, P. Morosan, N. Palomero-Gallagher, T. Schormann, H. Mohlberg, U. Bürgel, H. Steinmetz, et al. Quantitative analysis of sulci in the human cerebral cortex: development, regional heterogeneity, gender difference, asymmetry, intersubject variability and cortical architecture. *Human brain mapping*, 5(4):218–221, 1997.