Budapest University of Technology and Economics

Faculty of Natural Sciences - Természettudományi Kar

Department of Cognitive Science - Kognitív Tudományok Tanszék

THESIS

DIPLOMAMUNKA

Towards Naturalistic BCI: Leveraging Deep Learning to Decode Brain Activity During Passive Listening

Természetes BCI előkészítése: Mély tanulás alkalmazása az agyi aktivitás dekódolására a passzív beszédhallgatás során

Számítógépes és Kognitív Idegtudomány MSc

Computational and Cognitive Neuroscience MSc

Author - Szerző

Milán András Fodor

Advisor - Konzulens

Dr. Péter Márton Rácz

Dr. Tamás Gábor Csapó

February 4, 2024

betétlap

”Diplomamunka feladat a Számítógépes és Kognitív Idegtudomány mesterképzési (MSc) szak hallgatói számára”

ANNOTATION

The aim of the study is to investigate the complex mechanisms of speech perception and ultimately decode the electrical changes in the brain to produce speech.

I attempt to decode heard speech from intracranial electroencephalographic (iEEG) data using deep learning methods.

I set out with the goal to aid in the advancement of brain-computer interface (BCI) technology for speech synthesis, and I also hoped to provide an additional perspective on the cognitive processes of speech perception.

The approach I used in this research was novel, as it diverged from the conventional focus on speech production and instead chose to investigate neural representations of perceived speech. This angle opened up a complex perspective, potentially allowing for the study of more sophisticated neural patterns. Leveraging the power of deep learning models, my research aimed to establish a connection between these intricate neural activities and the corresponding speech sounds.

Despite my approach not having achieved a breakthrough yet, promising results have been obtained, particularly in identifying neural patterns related to speech perception. While precise speech synthesis from neural activity remains a challenge, correlations observed between the original and decoded signals reveal a strong link between the cognitive imprints of speaking and listening.

Therefore, my research indicates progress in the field of BCIs by shedding light on the vast potential of decoding neural activity during speech perception, which could be crucial in creating a more naturalistic, communication-focused BCI system.

ANNOTÁCIÓ

A diplomamunka célja a beszédészlelés összetett mechanizmusainak vizsgálata és végső soron az agyban történő elektromos változások dekódolása, ez által beszéd előállítása.

A hallott beszédet mélytanulásos módszerekkel kísérlem meg dekódolni intrakraniális elektroenkefalográfiai (iEEG) adatokból.

A célomnak tekintettem, hogy eredményeimmel elősegítsem a beszédszintézisre szolgáló agy-számítógép interfészek (BCI) technológiájának fejlődését, továbbá a beszédészelelés kognitív folyamatainak megértéséhez is szerettem volna egy extra támpontot nyújtani.

Az általam alkalmazott megközelítés újszerű volt, mivel eltért a beszédprodukcióra fókuszáló hagyományos iránytól, és ehelyett az észlelt beszéd neurális reprezentációinak vizsgálatát választottam. Ez a szemszög egy olyan összetett perspektívát nyit meg, amely lehetővé teheti a potenciálisan magasabb szintű neurális minták tanulmányozását. A kutatásom a mélytanulási modellek erejét kihasználva arra törekedett, hogy kapcsolatot hozzon létre e bonyolult neurális tevékenységek és a megfelelő beszédhangok között.

Annak ellenére, hogy áttörést még nem ért el a megoldásom, ígéretes eredmények születtek, különösen a beszédészleléssel összefüggő neurális minták azonosítása terén. Bár a precíz beszédszintézis idegi aktivitásból továbbra is egy kihívás, az eredeti és a dekódolt jelek között megfigyelt összefüggések mentén megfigyelhető a szoros kapcsolat a beszélés és a hallgatás kognitív lenyomata között.

A kutatásom tehát előrelépést jelez a BCI-k terén, mert rávilágít a beszédészlelés közbeni neurális aktivitás dekódolásának hatalmas lehetőségeire, amelyek kulcsfontosságúak lehetnek egy természetesebb feltételek mellett működő, kommunikációra orientált BCI rendszer megalkotásához.

Contents

1. INTRODUCTION

The brain-computer interfaces (BCIs) field is a fascinating amalgamation of neuroscience, computer science, and engineering. Over the past several decades, it has emerged as a focal point for researchers worldwide, fueled by the tantalizing possibility of enabling direct communication between the human brain and digital systems. Such a feat could have far-reaching implications, including transcending conventional interaction paradigms and opening up new avenues of accessibility for individuals with disabilities.

This intriguing field caught my attention during the 'Deep Learning in Practice' class I attended, where a group project focusing on synthesizing spoken words from EEG data using deep learning was assigned. The opportunity to work on this project opened my eyes to the potential this unique convergence of fields could offer. This experience deeply resonated with me, compelling me to delve further into this fascinating area of research.

The remarkable progress in BCI research has been largely propelled by several synergistic factors. First, significant advancements in neuroscience have exponentially deepened our understanding of the brain's complexities. Second, the explosive evolution of technology has led to quantum leaps in computational power, allowing us to handle the vast and complex neural data with increasing finesse. Lastly, a pressing societal need for novel assistive technologies, especially those that can significantly enhance the communication abilities of individuals with disabilities, has further accelerated this research field.

This Master's Thesis situates itself in the midst of this dynamic field of BCI research, with a specific focus on an exciting subset - speech synthesis. The aim of speech synthesis in BCIs is to decode and reconstruct speech from neural activity, thereby bridging the gap between thoughts and vocal communication. This capability could have a profound impact on individuals who have lost their ability to speak due to stroke, neurodegenerative diseases, or traumatic brain injuries, providing them with a powerful tool for communication.

In the pursuit of creating a more effective speech synthesis BCI, this thesis emphasizes the importance of developing a naturalistic BCI. The term "naturalistic" refers to a BCI that closely emulates real-world conditions, aiming to create a more intuitive and user-friendly system. To this end, the research undertaken in this thesis leverages cutting-edge deep learning techniques to decode brain activity recorded during passive listening.

Conventionally, BCIs for speech synthesis have been developed based on the paradigm of active speech production, where individuals try to speak or imagine speaking while their neural activity is recorded. While these approaches have yielded noteworthy insights, they might not entirely capture the nuances and complexity of cognitive processes inherent in speech perception and production. Therefore, this research argues that studying neural activity during passive listening could offer a fresh perspective and a more comprehensive understanding of these cognitive processes. This knowledge could potentially enhance the development of more naturalistic and user-friendly BCI systems for speech synthesis.

The thesis is methodically structured into seven designed chapters, each contributing towards a comprehensive elucidation of the research process.

Chapter 2 forms the theoretical bedrock of this thesis. It unravels the fundamental principles and complex mechanisms related to speech synthesis and the operation of BCIs. This chapter provides an overview of the cognitive underpinnings of both listened and spoken speech, summarizing the most recent advancements in the field of speech synthesis. A significant part of this chapter delves into deep learning, discussing its underlying principles, the various methodologies used, and its transformative role in decoding brain activity. This theoretical grounding serves as the scaffold upon which the research question is formulated and the subsequent methodologies are developed.

Chapter 3 introduces the central research question, which explores the potential of applying deep learning techniques to intracranial electroencephalography (iEEG) data recorded during passive listening. This research question sets the stage for a unique investigation that could advance BCI-based speech synthesis and provide novel cognitive insights.

Chapter 4 expounds upon the various research methodologies deployed in this study. This chapter details the dataset used, data preprocessing techniques employed, and the comprehensive procedure for deep learning training. The two primary deep learning models utilized in this study are the fully connected deep neural networks (FC-DNN) and 2D convolutional neural networks (2D-CNN). The rationale behind employing these models is their inherent capacity to adeptly handle high-dimensional data, a characteristic property of brain activity during passive listening.

Chapter 5 presents an exhaustive analysis of the empirical results derived from the application of the Fc-DNN and 2D-CNN models. It methodically evaluates the performance of each model, leading to a comparative assessment of different deep learning approaches in the specific context of BCI-based speech synthesis. This chapter goes beyond just the raw metrics to interpret what these results signify for our understanding of naturalistic BCI systems.

Chapter 6 broadens the discussion by situating the findings within the larger context of existing BCI and speech synthesis literature. This chapter is dedicated to dissecting the implications of the findings, identifying potential limitations of the study, and carving out avenues for future research. It provides a thoughtful exploration of the interplay between our results, existing knowledge, and the wider horizons of this field.

The final chapter, Chapter 7, offers a comprehensive conclusion to the thesis. It encapsulates the key findings, reflects on their theoretical and practical implications, and ponders upon the future trajectory of this research. This chapter is about the preceding discussions and points towards promising applications and areas of future investigation, thereby paving the way for ongoing exploration in the realm of naturalistic BCI for speech synthesis.

This Master's Thesis is my attempt to contribute to the evolving field of BCI, specifically focusing on speech synthesis. By investigating the potential of deep learning techniques to decode brain activity during passive listening, I aspire to add a new shade to our understanding of the cognitive processes involved in speech perception and production. I hope that my research will bring us a step closer to developing more naturalistic BCIs, opening up new avenues of communication for those who are currently voiceless. The ultimate ambition of this research extends beyond the confines of academia—it envisions a future where individuals with disabilities could engage with their environment using communication tools that are more closely aligned with the process of natural speech, potentially revolutionizing their quality of life and societal integration.

2. THEORETICAL OVERVIEW

This chapter is designed to offer a comprehensive theoretical overview, structured to provide an in-depth understanding of the relevant concepts and advancements in the field. It is organized into five main sections, each addressing a key aspect of the research topic:

'Towards Naturalistic BCI:Leveraging Deep Learning to Decode Brain Activity During Passive Listening'

The first section, **'Speech Synthesis'**, delves into the fundamental mechanisms of speech synthesis, providing a detailed look at traditional Text-to-Speech technology before transitioning to the innovative Brain-to-Speech methods. The significant role and importance of speech synthesis in the contemporary world is also discussed.

The second section, **'Deep Learning'**, delves into the key concepts and methodologies of deep learning, a subset of machine learning. It starts with an introduction to the fundamental principles, covering aspects like neural network architectures, layers, neurons, and activation functions. Essential training procedures, such as gradient descent and backpropagation, are also discussed. This deep dive into deep learning serves as a primer for understanding the mechanisms and methodologies used in decoding brain activity for our research.

The third section, **'Brain-Computer Interfaces (BCIs)'**, focuses on the intersection of neurology and technology. This section begins with a historical exploration of BCIs, followed by a detailed explanation of the basic principles that govern their operation. The different types of brain signals utilized by BCIs, the various paradigms and applications of BCIs, and the ongoing efforts towards creating a more naturalistic BCI are explored.

The fourth section, **'The Cognitive Background of Listened and Spoken Speech'**, delves into the cognitive processes underlying speech. This section explores the neural basis of speech perception, the cognitive framework of spoken speech, and the interaction between speech perception and production, providing a connection between the cognitive processes and the physical act of speech.

Finally, the fifth section, **'Advances in Speech Synthesis'**, offers a comprehensive review of the state-of-the-art in speech synthesis. This section highlights how technological advancements have improved our ability to generate synthetic speech that is increasingly naturalistic and intelligible.

Through this theoretical framework, the aim is to establish a robust foundation to then pose the

research question.

2.1 Speech synthesis

Speech synthesis is a broad area of research that encompasses various techniques for generating human-like speech from different input sources, such as text or brain activity [155]. This technology plays a crucial role in various applications, including assistive communication devices for individuals with speech and language impairments [60], natural language processing systems for machine translation and text summarization [82], and human-computer interaction in smart devices and virtual assistants [42]. As the demand for more natural and intelligible synthetic speech grows, researchers continue to explore and develop innovative approaches to improve the quality and flexibility of speech synthesis systems. Speech synthesis has come a long way since the early days of text-to-speech systems, which convert written text into spoken words. Today, with advances in brain-computer interfaces (BCIs), we are moving closer to a new milestone in speech synthesis: the ability to synthesize speech with help of data recorded in the brain. This emerging technology has the potential to revolutionize communication for people with speech impairments and to enable new forms of human-computer interaction. In this overview, we will explore the evolution of speech synthesis from text-to-speech to brain-to-speech, and we will discuss the importance of this technology in modern society.

2.1.1 Text-to-Speech

Text-to-speech (TTS) is a well-established area of speech synthesis research that focuses on converting written text into audible speech [83]. These systems have evolved over time, starting with rule-based methods [12] and transitioning to data-driven approaches using machine learning algorithms [67]. Deep learning techniques, such as recurrent neural networks (RNNs) and transformer-based models, have significantly improved TTS quality in recent years [121, 147].

TTS systems have a wide range of applications, including accessibility tools for visually impaired individuals [110], language learning software [59], and conversational AI systems [133].

2.1.2 Brain-to-Speech

Another exciting area of speech synthesis research involves the use of brain-computer interfaces (BCIs), in recent years, the field BCIs has begun to intersect with speech synthesis, sparking fascinating advancements in a sector known as neural speech synthesis. This emerging area of research

focuses on directly decoding and generating speech from neural activity, a concept that has vast implications for individuals suffering from severe speech and motor impairments [35]. Particularly, it offers new hope to those afflicted with conditions like amyotrophic lateral sclerosis (ALS), locked-in syndrome, or post-stroke paralysis, which severely hamper conventional means of communication [33, 24]. This forward-looking research represents a beacon of hope, potentially paving the way for these individuals to regain the ability to communicate.

The technology underpinning neural speech synthesis hinges on capturing the neural correlates of intended speech and then translating this information into synthetic speech. This is achieved through a complex interplay of data acquisition, preprocessing, feature extraction, and the application of various machine learning algorithms [67, 11].

Recent strides in BCIs have illustrated promising results in decoding and synthesizing speech from different forms of neural data, such as electrocorticography (ECoG) and intracortical recordings [157, 10]. Furthermore, several studies have ventured into the fusion of neural data with articulatory movement data, like tongue and lip movements, or the activity of the vocal cords, in an attempt to enhance the accuracy and intelligibility of the synthesized speech [30, 93].

This multimodal approach embraces a more comprehensive representation of the speech production process. It acknowledges that speech production is a complex orchestration, one that encompasses not only the brain but also the elaborate choreography of the articulatory system [53, 11].

The methodologies adopted in neural speech synthesis largely depend on the type and location of brain recordings. ECoG and intracortical recordings, although offering high-resolution data, are invasive techniques and entail risks linked with surgery [27]. Conversely, non-invasive techniques, such as EEG or fMRI, offer a safer alternative but trade-off spatial resolution and temporal precision [30, 38].

Besides the mode of data acquisition, different signal processing and machine learning techniques are deployed to decode the recorded data, which significantly influence the performance and robustness of the speech synthesis system [100]. In recent years, deep learning has been increasingly used in neural decoding due to its ability to model complex, non-linear relationships in high-dimensional data [67].

Despite the challenges faced, the potential implications of neural speech synthesis are immense. It has the potential to bypass traditional communication barriers and open new channels of interaction for those unable to speak due to physiological constraints. As the technology matures and as our understanding of the neural correlates of speech deepens, neural speech synthesis could be-

come a revolutionary communication tool [45]. Consequently, it remains an incredibly important, albeit challenging, area of ongoing research that necessitates interdisciplinary collaboration and a sustained research effort.

2.1.3 The Essential Role of Advancing Brain-Computer Interface and Deep Learning for Speech Synthesis

Speech synthesis is undoubtedly a crucial element in our contemporary society, providing potential benefits in multiple sectors ranging from health care and education to the technology and entertainment industry. However, a particularly transformative application of speech synthesis is found in the conjunction of neuroscience and advanced technology, specifically within the domain of Brain-Computer Interfaces (BCIs). The importance of propelling this technology forward is immense. BCIs are capable of decoding neural activity associated with speech, thereby creating alternative avenues of communication for individuals who are unable to speak through traditional means. This advancement in technology can drastically revolutionize communication, particularly for those suffering from debilitating motor or speech impairments [33, 24].

BCIs symbolize the epitome of synergy between the human brain and technology, leveraging electrical signals produced by the brain to manipulate external devices. Advancing this technology is critically important as it enables a new form of communication. This form of speech synthesis bypasses physical articulation, making it an extraordinary lifeline for individuals who cannot verbally communicate due to physical or neurological limitations.

Consider those living with conditions such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome, or those who have suffered from severe strokes. For these individuals, the ability to communicate has been greatly diminished or entirely eradicated. Presently, these individuals frequently depend on alternative and augmentative communication (AAC) systems. However, these systems can be slow, burdensome, and insufficient for conveying complex messages or emotions. The application of BCI technology in these circumstances could provide a more intuitive, fluid, and natural form of expression, thus restoring the individuals' ability to communicate. As such, continued progress and refinement in BCI technology is not merely an intriguing scientific pursuit but a pressing humanitarian imperative [33, 24].

BCI technology's significance extends beyond the healthcare sector. The development of systems capable of transmuting thought into speech could revolutionize how we interact with technology. Such a shift would significantly impact human-computer interaction, leading to more efficient, intuitive, and personalized interactions. In an era dominated by virtual assistants and AI-based

applications, BCIs could provide a seamless link between our thoughts and technology, making these systems feel more like extensions of our cognition than separate entities [115].

Deep learning plays an integral role in propelling these advancements in BCI technology. Deep learning, a subset of machine learning, uses artificial neural networks with various abstraction layers to 'learn' from large amounts of data. While traditional machine learning algorithms require manual feature extraction, deep learning automates this process, making it an instrumental tool for decoding brain signals [90].

In the context of BCIs, deep learning models can be trained to decode neural signals associated with speech, transforming these electrical impulses into intelligible words and sentences. The success of deep learning is largely attributable to the hierarchical representation of data, which mirrors the functioning of the human brain. This ability to model complex, non-linear relationships makes deep learning uniquely suitable for the challenge of decoding and synthesizing speech from brain activity [100]. As a result, advancements in deep learning directly fuel the progress in the field of BCI-driven speech synthesis.

Moreover, the advancement of BCI technology is integral to the development of inclusive technology, a key component in fostering digital accessibility. The commitment to create a world where technology can serve all individuals equitably, irrespective of their physical or cognitive abilities, is a cornerstone of modern society. In this endeavour, BCIs represent a significant breakthrough. By enabling intuitive, seamless interaction with digital platforms, BCIs not only cater to individuals with severe disabilities but benefit all users [30, 108].

In conclusion, the necessity to advance BCI technology and deep learning for speech synthesis is profound. It holds the potential to dramatically enhance the lives of those with severe speech and motor impairments, redefine the landscape of human-computer interactions, and propel us towards a more inclusive and accessible society. The boundless potential of BCIs, powered by deep learning, symbolizes a beacon of hope for a future where communication is a universal right, unimpeded by physical or neurological barriers [30, 108].

2.2 Deep Learning

2.2.1 Introduction to Deep Learning

Deep learning, a specialized subset of machine learning, concentrates on building and applying artificial neural networks with numerous layers, thereby introducing 'depth' into the network. Artificial neural networks are computational models inspired by the human brain's functioning and

structure. They endeavor to mimic the brain's ability to learn from vast quantities of data. A single-layered neural network can make rudimentary predictions. However, with the inclusion of multiple hidden layers, the network can significantly improve the accuracy of these predictions by modeling more complex representations and relationships within the data [52].

Over the past few years, deep learning has been instrumental in propelling the field of artificial intelligence (AI). Its efficacy has been established across a variety of applications and domains. For instance, deep learning algorithms have demonstrated profound success in image and speech recognition, enabling computers to interpret and categorize visual and auditory data with remarkable accuracy. In the realm of natural language processing, deep learning models can understand, interpret, generate, and translate human languages effectively. They've also been utilized for bioinformatics, facilitating the analysis of biological data and contributing to advancements in personalized medicine [90, 52, 143].

2.2.2 Exploring Neural Networks and Architecture

At its core, deep learning is underpinned by the concept of the artificial neural network. Such a network consists of numerous simple processing nodes, or 'neurons,' systematically organized into layers. The architecture of a typical neural network comprises an input layer, several hidden layers, and an output layer. The input layer receives raw data, mirroring the role of sensory input in biological neurons. The hidden layers process the input data, each adding a level of abstraction and complexity, while the output layer generates the final result of the network [52].

Different problems require different neural network architectures. Some of the most commonly used include the Feedforward Neural Network (FNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN). The FNN is a basic type of network where information moves unidirectionally—from the input layer to the output layer—without looping back. On the other hand, more complex networks such as CNNs excel at tasks involving image and video data processing, owing to their proficiency in preserving spatial information. RNNs and LSTMs are especially suited to dealing with sequential data like time series or natural language, thanks to their ability to remember previous inputs in the sequence [90, 52].

2.2.3 The Training Process of Neural Networks

Training a neural network involves adjusting the network's weights to minimize the difference between its predicted output and the actual output. This process is typically achieved using a

combination of the backpropagation algorithm and gradient descent optimization. The network's learning procedure is guided by its mistakes: it adjusts the weights to minimize the error between the predicted output and the actual one. This error is propagated backward through the network, hence the term 'backpropagation' [139].

Training deep learning models is not a trivial task. It requires substantial computational power and voluminous amounts of data. Deep learning models, while celebrated for their ability to learn complex representations from data, are also notorious for overfitting, particularly when the amount of training data is small. Overfitting refers to a situation where the model learns the noise and outliers in the training data to the point where it negatively impacts the model's performance on new data. Several regularization techniques, such as dropout, weight decay, and early stopping, are often employed to mitigate overfitting [152].

2.2.4 Deep Learning in Speech and Language Processing

Deep learning has instigated a paradigm shift in the field of speech and language processing. It has substantially enhanced automatic speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS) systems. ASR systems, for instance, have benefited greatly from deep learning, showing a marked reduction in word error rates. RNNs, with their capacity to handle temporal sequences, are commonly used in language modeling, allowing the model to account for the influence of past words on the present context [153].

LSTMs and GRUs, both variations of RNNs, have been especially effective in NLP tasks such as machine translation, due to their ability to mitigate the vanishing gradient problem and deal with long-range dependencies in the text. In the area of speech synthesis, deep learning approaches have contributed to the development of end-to-end systems that convert text to speech in a single step, circumventing the need for multiple handcrafted components. These models have produced synthesized speech of unprecedented naturalness and intelligibility, blurring the lines between human and synthesized speech [147, 153].

2.2.5 Deep Learning: A Transformational Force

Deep learning, with its aptitude for high-level abstraction and dealing with massive datasets, has revolutionized machine learning and AI. Its capabilities have been employed to make significant advances in many areas, including speech synthesis, natural language processing, image recognition, and more. Its potential continues to be explored in numerous novel applications, including the

burgeoning field of brain-computer interfaces, which present an exciting and challenging frontier for the application of deep learning methodologies [171].

However, as deep learning models become increasingly prevalent, concerns about their impact and implications are also growing. In particular, the ”black box” problem, which refers to the lack of interpretability and transparency of these models, is a significant issue. The complexities of deep learning models often make it difficult to understand why they make specific decisions or predictions, raising concerns, especially when these models are used in high-stakes decisions [32].

Moreover, the data-driven nature of deep learning models means they may perpetuate and amplify existing biases in the data they are trained on. This raises significant ethical and fairness considerations that must be addressed, particularly when these models are deployed in sensitive domains such as hiring, lending, or judicial decision-making [120].

2.2.6 Future Directions and Challenges

The future of deep learning research includes addressing the aforementioned challenges and expanding the frontiers of its application. More efficient training algorithms, new architectures that can learn from fewer examples, improved interpretability and transparency of models, and addressing fairness and ethical considerations are some of the critical areas of focus. Furthermore, as AI becomes increasingly integrated with society, it is crucial to consider societal and ethical aspects when developing and deploying these models [32, 120].

2.3 Brain-Computer Interfaces

Brain-Computer Interfaces (BCIs) are systems that enable direct communication between the human brain and external devices by decoding neural signals into actionable commands [106]. BCIs have the potential to revolutionize various fields, such as assistive technology for individuals with severe motor disabilities [21], neurorehabilitation for stroke patients [132], gaming and immersive experiences [91], military applications for enhancing cognitive performance [113], and many more. This literature review aims to provide an extensive overview of BCIs, including their history, underlying principles, and the types of brain signals used. Additionally, the review will cover different BCI paradigms, applications, ethical considerations, and the challenges faced in the field, such as signal acquisition, processing, and user training.

2.3.1 History and Evolution of BCIs

The concept of Brain-Computer Interfaces (BCIs) has evolved significantly since its inception. Hans Berger, a German psychiatrist, first laid the groundwork in the early 20th century by recording human brain activity using electroencephalography (EEG). Berger's discovery of the alpha rhythm, a prominent oscillatory activity in the brain, marked the beginning of modern EEG research [18].

In the 1960s and 1970s, the possibility of utilizing brain signals to control external devices began to take shape. Grey Walter in the 1960s harnessed EEG signals to control a slide projector, and by 1973, Jacques Vidal demonstrated the first successful BCI, in which humans could control a cursor on a computer screen using their brain signals [160, 161].

Following Vidal's groundbreaking research, various BCI systems were developed, each harnessing different types of brain signals, such as event-related potentials (ERPs), steady-state visual evoked potentials (SSVEPs), and motor imagery-related EEG patterns. One remarkable breakthrough was the BCI developed by Lawrence Farwell and Emanuel Donchin in the 1980s. It was based on the P300 component of the ERP and allowed users to select letters on a computer screen by focusing their attention on specific stimuli [47]. This innovation demonstrated the potential for communication and control by individuals with severe motor impairments.

The 1990s brought more advancements. Jonathan Wolpaw and colleagues developed a BCI system that enabled users to move a computer cursor in two dimensions using slow cortical potentials (SCPs) [167]. This ability to achieve continuous control of a computer cursor, rather than simply selecting from a set of predefined options, marked a significant milestone in BCI research. Around the same time, researchers also explored motor imagery-related EEG patterns for BCI control. Pfurtscheller and colleagues demonstrated that imagining specific motor actions, such as left or right-hand movement, resulted in distinct patterns of event-related desynchronization (ERD) and event-related synchronization (ERS) in the EEG. These patterns could then be used to control a cursor on a computer screen [124].

The development of BCI systems for various applications continued to progress, driven by advances in technology and neuroscience. In 2006, a significant milestone was reached when Hochberg and colleagues reported the successful control of a robotic arm by a human with tetraplegia using an intracortical BCI [68]. This demonstrated the potential of BCIs for neuroprosthetic control, providing a proof-of-concept for the restoration of lost motor function through direct brain control.

BCIs also found their place in neurorehabilitation, particularly for the rehabilitation of motor functions after a stroke. Daly and Wolpaw reviewed the use of BCIs in this context, highlighting their

potential to facilitate functional recovery by promoting cortical reorganization and neural plasticity [39]. They emphasized the importance of integrating BCI systems with conventional rehabilitation techniques to maximize their therapeutic benefits.

Furthering the reach of BCIs, their application expanded into the entertainment industry, such as gaming and virtual reality. In 2009, the first consumer-grade BCI headset, the Emotiv EPOC, was released, enabling users to control video games and virtual environments using their brain activity [13]. This development marked the beginning of a new era for BCIs, as it brought the technology out of the lab and into the hands of the general public.

More recently, the focus of BCI research has been on improving the performance and usability of the systems, while also exploring new applications and paradigms. One such development is the advent of hybrid BCIs, which combine multiple BCI paradigms or integrate BCIs with other physiological signals, such as electromyography (EMG) or electrooculography (EOG). This combination can improve the performance and robustness of the systems [126].

Another noteworthy development is the concept of passive BCIs, which monitor the user's cognitive state without requiring explicit tasks, enabling the development of adaptive systems that can respond to the user's needs and preferences [173].

The field has also benefited from the integration of machine learning techniques, such as deep learning and reinforcement learning. These techniques have the potential to enhance BCI performance by facilitating the development of more sophisticated models and algorithms [142].

While invasive BCIs, such as intracortical recordings, offer high spatial and temporal resolution, they also carry significant risks and ethical concerns. Therefore, the development of high-performance non-invasive BCIs remains a priority in the field [117].

As the field of BCI research continues to grow, it is anticipated that future developments will focus on addressing the remaining challenges, such as signal quality, user training, inter-subject variability, and ethical considerations. Additionally, the exploration of novel applications and paradigms will continue to expand the potential impact of BCIs in various domains.

2.3.2 Principles of BCIs

Overview of the BCI process A typical BCI system involves several steps, including:

- Signal acquisition
- Preprocessing

- Feature extraction
- Control/command execution

The main goal of a BCI is to identify specific patterns in the brain signals and translate them into meaningful actions or commands for controlling external devices or applications.

Various signal acquisition methods are used in BCI research, including electroencephalography , magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and intracortical recordings [117]. Each method has its advantages and disadvantages in terms of spatial and temporal resolution, invasiveness, cost, and signal quality.

- EEG: Non-invasive, high temporal resolution, relatively low cost, but low spatial resolution [118].
- MEG: Non-invasive, high temporal and spatial resolution, but expensive and requires a shielded environment [58].
- fMRI: Non-invasive, high spatial resolution, but low temporal resolution and high cost [99].
- Intracortical recordings: Invasive, high spatial and temporal resolution, but carry significant risks and ethical concerns [68].
- Control/command execution

Signal processing techniques are used to enhance the quality of the acquired brain signals and remove artifacts or noise. Common preprocessing techniques include filtering (e.g., bandpass or notch filtering) and artifact removal algorithms, such as independent component analysis (ICA) or principal component analysis (PCA) [41].

Feature extraction is the process of identifying relevant features or patterns from the preprocessed brain signals that can be used for classification. Features can include time-domain features (e.g., amplitude, latency), frequency-domain features (e.g., spectral power, coherence), and time-frequency features (e.g., wavelet coefficients) [102]. Feature selection methods aim to identify the most discriminative features for classification, reducing the feature dimensionality and computational complexity. Common techniques include recursive feature elimination, mutual information, and Fisher's discriminant ratio [101].

Classification algorithms are used to identify the specific brain patterns associated with different mental tasks or cognitive states. Various machine learning algorithms have been employed in BCI

research, such as linear discriminant analysis (LDA), support vector machines (SVM), artificial neural networks (ANN), and deep learning [142]. The choice of classifier depends on the specific BCI application, the nature of the brain signals, and the desired performance characteristics.

The final step in the BCI process involves translating the classified brain patterns into meaningful commands or actions for controlling external devices or applications. Control strategies can include discrete or continuous control, depending on the specific application and user needs [112]. For example, P300-based BCIs typically use discrete control (e.g., selecting letters or icons), while motor imagery-based BCIs can support continuous control (e.g., moving a cursor or robotic arm).

2.3.3 Types of Brain Signals

In the realm of BCI, different types of brain signals each have unique characteristics:

- Event-related potentials (ERPs)
- Motor imagery and associated EEG patterns
- Slow cortical potentials (SCPs)
- Other brain signals and their potential for BCI development

Event-related potentials (ERPs) are time-locked changes in the EEG signal that occur in response to specific sensory, cognitive, or motor events. ERPs are characterized by their polarity, latency, and amplitude. One of the most well-known ERP components used in BCIs is the P300, which is a positive deflection occurring approximately 300 milliseconds after the presentation of a rare or unexpected stimulus [154]. P300-based BCIs are widely used for communication and control, especially for individuals with severe motor impairments [47].

Steady-state visual evoked potentials (SSVEPs) are oscillatory brain responses elicited by repetitive visual stimulation at a constant frequency. SSVEPs are widely used in BCIs due to their high signal-to-noise ratio, minimal user training, and easy implementation [159]. SSVEP-based BCIs often use flickering visual stimuli, and the frequency of the flickering stimulus corresponds to the frequency of the SSVEP response in the user's brain activity [66].

Motor imagery is the mental rehearsal of a motor action without overt movement. Motor imagery-based BCIs decode users' intentions by analyzing changes in their EEG patterns, such as event-related desynchronization (ERD) and event-related synchronization (ERS) [123]. These changes

occur during motor planning, execution, and imagery, making them suitable for BCI applications [116].

Slow cortical potentials (SCPs) are slow shifts in the EEG signal, lasting from several hundred milliseconds to several seconds, and have been utilized in BCI applications. SCPs are thought to be related to cortical excitability, and can be either positive or negative depending on the underlying neural processes [23]. SCP-based BCIs have been developed for communication and control, particularly for individuals with severe motor disabilities such as those with locked-in syndrome or amyotrophic lateral sclerosis (ALS) [87].

Several other brain signals have been explored for BCI applications, including gamma-band activity (GBA), which is associated with various cognitive processes such as attention, memory, and perception [78]. GBA-based BCIs have shown promise in applications such as attention monitoring and cognitive workload assessment [122]. Another potential BCI signal is the readiness potential (RP), which is a slow negative potential that precedes self-initiated movements [97]. RP-based BCIs have been proposed for movement prediction and control applications [6].

2.3.4 Signal Acquisition Methods

EEG is a non-invasive method for recording electrical activity in the brain using electrodes placed on the scalp. EEG has high temporal resolution and has been widely used for BCI systems due to its portability, affordability, and ease of use [104]. However, EEG signals have low spatial resolution and are susceptible to artifacts from muscle and eye movements, as well as electrical noise [51].

Magnetoencephalography (MEG) measures the magnetic fields produced by neuronal activity in the brain. MEG provides better spatial resolution than EEG and is less affected by artifacts, but it requires a magnetically shielded room and is more expensive [58]. Functional magnetic resonance imaging (fMRI) measures changes in blood oxygenation levels in response to neural activity. fMRI has high spatial resolution but has lower temporal resolution compared to EEG and MEG. Despite its potential, fMRI-based BCIs are limited by the bulky and expensive equipment and the need for users to remain still during scanning [151].

Near-infrared spectroscopy (NIRS), a possibly cheaper and simpler alterntaive to fMRI, uses near-infrared light to measure changes in blood oxygenation levels, which reflect brain activity. NIRS has moderate spatial and temporal resolution and is less sensitive to artifacts than EEG. NIRS-based BCIs have been explored for various applications, but they are not as common as EEG-based BCIs [114].

Transcranial magnetic stimFBCIulation (TMS) is a non-invasive technique that induces electrical activity in the brain by applying a magnetic field to the scalp. While TMS is primarily used as a tool for studying brain function, recent studies have explored its potential as a signal acquisition method for BCIs [140]. However, the practical use of TMS for BCIs is currently limited due to its high cost and the technical challenges associated with integrating TMS with other signal acquisition methods.

Intracortical recordings involve the implantation of microelectrodes directly into the cortex of the brain, allowing for the measurement of neural activity with high spatial and temporal resolution. These invasive techniques have been used in the development of BCIs for motor control and sensory feedback, with promising results in both animal and human studies [144, 69]. However, intracortical recordings raise concerns about the long-term stability and safety of the implanted electrodes, as well as ethical considerations related to invasive procedures [138].

Intracranial electroencephalography (iEEG), like ECoG or Stereoelectroencephalography (sEEG), involves placing electrodes directly on the surface of the brain, typically during surgical procedures. iEEG provides a compromise between spatial resolution and invasiveness, offering higher spatial and temporal resolution compared to EEG and is less susceptible to artifacts. This makes iEEG an attractive option for BCI applications, especially those involving speech synthesis, as it can accurately capture the high-resolution neural signals associated with speech production [62]. Despite its invasive nature, the use of iEEG in BCIs is primarily limited to patients undergoing neurosurgical procedures for conditions such as epilepsy. However, iEEG has been successfully utilized in a variety of BCI applications, including motor control, communication, and even music performance,indicating its high potential for future BCI development [162]. For example, iEEG has been used to decode imagined speech, allowing users to control a speech synthesizer with their thoughts. This technology could provide a new communication method for individuals who are unable to speak due to injury or disease [7]. However, more research is needed to improve the performance and usability of iEEG-based BCIs, and to address the ethical and practical issues associated with their use.

2.3.5 BCI Paradigms and Applications

Brain-Computer Interfaces (BCIs) can be categorized into different paradigms based on the type of neural activity they utilize and the methods used to decode these activities. Each paradigm has unique characteristics, which makes it suitable for certain applications. Here, I discuss some of the major BCI paradigms and their applications.

P300-based BCIs rely on event-related potentials (ERPs) in EEG signals, specifically the P300 wave, a positive ERP component that occurs around 300 ms after a rare or unexpected stimulus [129]. The P300 wave is employed in a speller paradigm, a communication interface for people with severe motor impairments, initially developed by Farwell and Donchin [47]. It has since found applications in neurorehabilitation, helping patients recover from stroke and other neurological conditions by providing a novel means of communication and control [89]. Further applications include the control of wheelchairs [134] and gaming interfaces [146].

Steady-state visual evoked potentials (SSVEPs) are oscillatory responses to visual stimuli at a specific frequency [128]. SSVEP-based BCIs have been utilized in a variety of applications due to their high information transfer rate and minimal training requirements. These include cursor control [36], virtual reality environments [149], neuroprosthetics [76], and gaming [98].

Motor imagery (MI)-based BCIs work by decoding EEG patterns associated with the mental imagination of movement, such as the mu and beta rhythms [125]. These BCIs have been particularly impactful in neurorehabilitation, assisting stroke and spinal cord injury patients in regaining motor function [5]. MI-based BCIs have also been used to control robotic arms [70] and for navigating virtual environments [92].

Slow cortical potentials (SCPs) are slow voltage shifts in the EEG that can be controlled voluntarily [20]. SCP-based BCIs have been employed in communication interfaces for individuals with severe motor impairments, including those in the locked-in state [88]. They have also shown promise in neurorehabilitation [158] and in controlling neuroprosthetics [22].

Passive BCIs monitor brain activity without requiring the user's active control, hence they are often used to assess cognitive or emotional states [131]. They have found applications in workload and fatigue assessment [79], emotion recognition [85], and adaptive automation systems that adjust their operation according to the user's state [130].

Hybrid BCIs combine two or more BCI paradigms or incorporate non-BCI input modalities to improve performance and usability [126]. By leveraging the strengths of multiple paradigms, hybrid BCIs can overcome some of the limitations inherent to individual BCI systems. They have been applied in various domains, such as communication [170], neurorehabilitation [132], and robotic control [107].

One of the most profound applications of BCIs is in restoring communication for individuals with severe motor impairments, including those in the locked-in state. Various BCI paradigms, such as P300 and SCP-based BCIs, have been used to create spelling devices that allow these individuals

to communicate [88, 47].

In recent years, the field has made significant strides in using BCIs for speech synthesis. This is a particularly exciting development for those unable to speak due to conditions like amyotrophic lateral sclerosis (ALS) or paralysis. BCIs for speech synthesis aim to decode neural activity related to imagined or attempted speech and convert it into spoken words. This has been achieved using intracranial EEG (iEEG) signals, which offer high spatial and temporal resolution [62, 8].

For instance, Angrick et al. [8] demonstrated a BCI that could decode continuously spoken numbers from iEEG signals with high accuracy. Similarly, Herff et al. [62] developed a BCI that could synthesize speech from iEEG signals recorded while subjects vocalized text. While these developments are promising, challenges remain in terms of improving the accuracy and robustness of speech decoding, reducing the invasiveness of the recording techniques, and making the technology more accessible and user-friendly [29].

Despite these challenges, BCIs for speech synthesis represent a significant step forward in augmenting human communication and offer a promising avenue for restoring the ability to speak in individuals with severe speech and motor impairments.

2.3.6 The Path Towards a More Naturalistic BCI

Brain-computer interfaces (BCIs) have evolved significantly over the years, moving from rudimentary signal detection to complex communication systems. However, achieving seamless, naturalistic communication between humans and computers through BCIs still presents a number of challenges. These challenges are particularly pronounced in the realm of speech synthesis based on passively perceived auditory signals.

The quality of recorded brain signals is one of the most critical aspects of BCI research. Noise and artifacts from muscle activity, eye movements, and environmental interference can significantly degrade the signal quality, hindering effective decoding [48]. The development and application of advanced signal processing techniques and artifact rejection algorithms are needed to enhance the quality and reliability of these recorded signals [156]. Moreover, the application of machine learning methods can also help improve the classification and interpretation of these signals [103].

Decoding passively perceived auditory signals, such as those generated during listening to speech or music, is a complex task. It necessitates the design of sophisticated machine learning algorithms that can effectively identify and extract relevant features from intricate and noisy brain signals [37]. Recent advances in deep learning and artificial intelligence may offer robust tools for tackling this

challenge [169]. The integration of these technologies with BCIs could bring us closer to realizing more naturalistic communication capabilities.

Inter-individual variability in BCI performance is another major challenge. Differences in brain anatomy, cognitive abilities, and other factors can cause significant variations in BCI performance between individuals [26]. Addressing this issue requires the development of personalized BCI models that account for this inter-subject variability, which could enhance the performance and usability of BCIs for a broader range of individuals [77].

As we move towards a future where BCIs may become a part of everyday life, ethical considerations also gain paramount importance. Issues concerning privacy, consent, and potential misuse of the technology need to be addressed [172]. Establishing transparent and responsible guidelines for BCI research and applications is imperative to ensure that the technology is used in a manner that respects individual autonomy and promotes societal benefit.

The potential impact of these advancements in BCI technology is vast. For individuals with severe motor impairments or communication disorders, BCIs could offer a transformative means of expression and interaction with the world. For the broader public, they could redefine the way we interact with technology, opening up new avenues in entertainment, education, and more.

In conclusion, while significant challenges remain in the development of a more naturalistic BCI for speech synthesis, the potential benefits make it a worthy pursuit. Overcoming these hurdles will require an interdisciplinary approach that leverages advances in neuroscience, computer science, engineering, and ethics. The prospect of AI-based BCI systems capable of decoding passively perceived auditory signals holds immense promise for the future of BCIs and human-computer interaction.

2.4 The Cognitive Background of Listened and Spoken Speech

Understanding the cognitive background of listened and spoken speech is fundamental to the field of brain-computer interfaces (BCIs) and speech synthesis. The process of speech, both in terms of perception and production, is a complex interplay of neural networks and cognitive processes. The human brain's capacity to interpret and generate speech is a testament to its intricate functioning and flexibility. In this section, I delve into the neural and cognitive foundations of speech, from how we perceive spoken language to how we produce it. This exploration is crucial to the thesis as it sets the groundwork for understanding how BCIs can leverage deep learning to decode brain activity during passive listening, leading to more naturalistic speech synthesis.

2.4.1 Speech Production

The intricate task of spoken speech production engages a wide spectrum of cognitive and neural mechanisms, ranging from the initiation of thoughts to the muscular movements articulating those thoughts as speech. This elaborate process is facilitated by a network of brain regions, including but not limited to, Broca's area, the motor cortex, the cerebellum, Wernicke's area, and the superior temporal gyrus [56, 64]. This section aims to delve into the cognitive and neural underpinnings of spoken speech, shedding light on the stages of speech production, the specific roles various brain regions play, and the neural representation of speech sounds.

The journey of spoken speech production can be conceptualized as a sequence of stages, each contributing a critical component to the process. These stages are as follows: conceptualization, linguistic planning, motor planning, and motor execution [95, 75].

Conceptualization is the birthplace of speech, where the intent or message to be conveyed is formulated. This cognitive process is more than just the generation of an idea; it involves selecting the most effective way to communicate the intended message, considering the context, the audience, and the speaker's goals.

Following conceptualization is linguistic planning, a stage where the abstract ideas are translated into linguistic structures. This stage is characterized by the selection of appropriate words, organizing them into grammatically correct structures, and encoding the sentence's semantic and syntactic information. Linguistic planning also involves phonological encoding, where selected words are transformed into a string of phonemes, the most basic units of sound in a language [94].

Motor planning is the next stage, serving as the bridge between the cognitive and physical realms of speech production. Here, the string of phonemes from the linguistic planning stage is converted into a series of motor commands. These commands are instructions for the articulatory muscles, dictating the precise movements required to produce the speech sounds [55]. This stage is particularly complex due to the high degree of coordination required among various speech muscles, including those in the larynx, pharynx, tongue, and lips.

Finally, motor execution is where thought transforms into audible speech. This stage involves the actual movement of the articulatory muscles as per the motor commands, resulting in the production of speech sounds. The process is closely monitored by the auditory and somatosensory systems, ensuring the produced speech aligns with the intended message [56].

While these stages provide a simplified view of the speech production process, it's important to

note that in reality, these processes are likely overlapping and recursive, with constant feedback and adjustments taking place at each stage [63]. Furthermore, the brain regions involved in speech production do not work in isolation; instead, they interact dynamically as part of a broader network, with each region contributing to various aspects of the process.

2.4.2 Speech Perception

Speech perception is a marvel of human cognitive ability, requiring the integration of complex auditory pathways and specialized brain regions to decode the rich tapestry of spoken language. The process begins with the conversion of acoustic signals into neural signals in the auditory system, but this transformation is only the first step in a fascinating journey.

Sound waves initially reach the ear, where they are converted into electrical signals by the cochlea in the inner ear. Specialized hair cells within the cochlea are sensitive to different frequencies, allowing them to break down complex sounds into distinct frequency components. This tonotopic organization, where specific cells respond preferentially to different frequencies, provides a foundation for the complex spectral analysis necessary for speech perception [109].

The converted signals then travel along the auditory nerve to subcortical structures, passing through several key stations, including the cochlear nucleus, superior olivary complex, and the inferior colliculus. These structures are essential in preprocessing the auditory signal, extracting critical features such as pitch, loudness, and localization cues, and further refining the signal as it travels through the pathway [31].

The next waypoint is the thalamus, specifically the medial geniculate body. The thalamus acts as a primary relay station, receiving sensory input from multiple pathways and dispatching it to the appropriate cortical areas. From there, the signals are transmitted to the primary auditory cortex (A1) in the temporal lobe. This marks the signal's entry into the cortical realm where higher-order processing begins [150].

Within the cortical domain, the A1 plays a vital role. Its highly organized structure, characterized by a tonotopic map, is fundamental in processing the complex spectral content of speech. Yet, A1 is only one piece of the puzzle. Comprehension of speech necessitates the involvement of a network of higher-order auditory cortical areas, each specialized in dissecting specific facets of the speech signal.

One of these areas is the superior temporal gyrus (STG). The STG displays sensitivity to both spectral and temporal features of speech, highlighting its critical role in parsing the rich auditory

information contained in spoken language. Neuroimaging studies have pointed to the active involvement of STG in processing phonemes, syllables, and words [19, 74].

Another key player in the perception of spoken language is the left posterior superior temporal sulcus (pSTS). This region integrates auditory and visual speech information to generate a unified percept, crucial for understanding spoken language in a real-world context where both auditory and visual cues are present. The pSTS also contributes significantly to processing the syntactic and semantic aspects of speech [61, 40].

Not to be overlooked is the left inferior frontal gyrus (IFG), home to Broca's area. While traditionally associated with speech production, Broca's area, and its surrounding region, is now recognized for its involvement in various aspects of speech comprehension, such as processing syntactic and semantic structures [49, 57].

Beyond the individual brain regions involved, speech perception is fundamentally underpinned by the integration of both temporal and spectral information. Temporal processing involves analyzing changes in sound amplitude over time, vital for perceiving rhythm, stress, and intonation in speech. In contrast, spectral processing involves analyzing the distribution of energy across different frequencies, which is crucial for distinguishing different phonemes. Research has shown that these two types of information are processed in parallel in the auditory cortex, with different neuronal populations responding preferentially to one type of information [174]. The anterior regions of the STG have been found to be more sensitive to spectral information, while the posterior regions are more involved in processing temporal information [15, 175].

Furthermore, studies have suggested that the human brain utilizes a combination of auditory object analysis and speech-specific processing mechanisms to understand spoken language. Auditory object analysis refers to the brain's ability to organize the complex acoustic input into meaningful auditory objects, like the sounds of individual speech sources in a noisy environment. This requires the extraction and integration of various acoustic features, such as pitch, timbre, and spatial location [54]. On the other hand, speech-specific processing mechanisms are specialized for processing the unique acoustic properties of speech, such as formant transitions and voice-onset time, crucial for distinguishing different phonemes [44].

Moreover, the processing of speech also involves top-down influences from higher cognitive functions, including attention, memory, and expectation. For instance, our ability to understand speech in noisy conditions, known as the cocktail party phenomenon, is partly attributed to our capacity to focus attention on a particular sound source while ignoring others [28]. Likewise, our previous

knowledge and expectations about the content and structure of speech also shape our perception, as evidenced by the influence of lexical and semantic context on the perception of ambiguous phonemes [141].

By understanding these intricate processes and the complex interplay between different brain regions during speech perception, we gain valuable insights that can be leveraged in decoding brain activity associated with speech and language tasks. The current research findings, showing that deep learning models can identify and decode shared patterns and features in neural activity during passive listening and spoken speech, represent a significant advancement in our understanding of this fascinating and complex cognitive process.

2.4.3 The Interaction between Speech Perception and Production

Understanding the intricacies of spoken language processing requires an examination of the complex interplay between speech perception and production. Most of the above mentioned important brain regions are shown on Figure 2.1. This dynamic interaction forms the core of the perception-production loop, a fundamental aspect of language processing that enables individuals to adapt their speech output based on the auditory input they receive. [65].

A key player in this interaction is the motor system. Once thought to be only involved in speech production, recent studies have highlighted its role in speech perception as well. The listener's motor system may contribute to the perception of speech by activating the articulatory gestures associated with the perceived speech sounds. This suggests that our understanding of speech might involve 'simulating' the motor actions of speech production in our minds, a concept often referred to as 'motor resonance' or 'embodied cognition' [166, 105, 145].

Complementing the motor system's role is the mirror neuron system, a network of neurons that activate both when an individual performs an action and when they observe the same action performed by someone else. Initially discovered in the premotor cortex of monkeys, the human mirror neuron system has been implicated in various aspects of language processing, including both speech perception and production [43, 135, 46, 50].

Auditory feedback is another critical component in the interaction between speech perception and production. This feedback allows speakers to monitor and adjust their speech output based on the acoustic input they receive. Notably, disruptions to auditory feedback, such as through delayed or altered auditory feedback, can cause significant disturbances in speech production, highlighting the importance of the perception-production loop in maintaining fluent and accurate speech [72, 80,

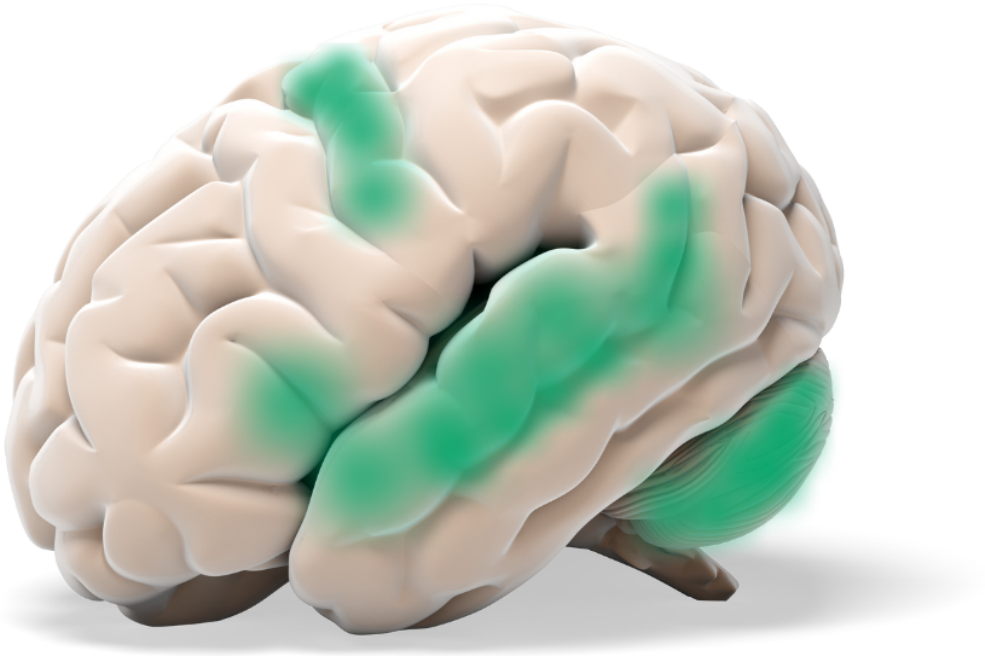

Figure 2.1: Broca's area, the motor cortex, the cerebellum, Wernicke's area, and the superior temporal gyrus, posterior superior temporal sulcus highlighted as important areas of the brain regarding speech based on [56, 64, 55, 61] .

14].

Finally, the dual-stream model of speech processing provides a broader context for understanding these interactions. The model proposes two main pathways in the brain for processing speech and language: a ventral stream for mapping sound to meaning, and a dorsal stream for mapping sound to motor representations. This model has been influential in recent research, suggesting that the interaction between speech perception and production involves a complex network of pathways, rather than a single linear process [63].

Understanding these neural mechanisms of speech perception and their interaction with speech production not only contributes to our understanding of spoken language processing, but it also has significant implications for the field of speech synthesis.

Traditionally, speech synthesis models have focused primarily on speech production mechanisms. These models often take linguistic input and generate speech sound by simulating the physical

and acoustic processes of speech production [83]. However, the complexity of these mechanisms, which involves intricate coordination of numerous muscles and precise control of airflow, makes them challenging to accurately model and replicate.

In contrast, using speech perception as a basis for speech synthesis could offer several advantages. One major advantage is the potential to create more natural-sounding and intelligible synthetic speech. Given that speech perception mechanisms are fine-tuned to extract the most linguistically relevant features from the speech signal, a synthesis model based on these mechanisms could potentially focus on generating these critical features, leading to improved intelligibility [127]. Furthermore, as speech perception mechanisms are constantly adapting to the variability in the speech signal, such a model could potentially be more robust to changes in speaking style, accent, or environmental noise.

Another advantage is that a perception-based model could potentially make the synthesis process more computationally efficient. As speech perception involves extracting meaningful linguistic units from the complex acoustic signal, a synthesis model based on these processes could potentially generate speech by assembling these units, rather than simulating the entire production process [137].

In conclusion, the interaction between speech perception and production is a complex and dynamic process that involves the interplay of multiple systems and pathways in the brain. Understanding these interactions is critical for developing a comprehensive model of listened and spoken speech processing. Furthermore, considering the potential benefits, speech perception mechanisms could provide a promising alternative approach for the development of future speech synthesis systems.

2.5 Advances in Speech Synthesis

2.5.1 Traditional Approaches to Speech Synthesis

Rule-based speech synthesis systems rely on manually crafted rules to generate speech sounds from text input [3]. These rules describe the relationships between the phonemes, prosody, and other acoustic properties of speech. The formant synthesis method is a classic example of rule-based synthesis, where the formant frequencies are adjusted according to predefined rules to produce the desired speech sounds [84]. Despite their relatively low computational requirements, rule-based systems often produce less natural-sounding speech compared to data-driven approaches [25].

Concatenative synthesis uses a large database of pre-recorded speech segments, which are then concatenated to generate the output speech [73]. This approach often produces high-quality and

natural-sounding speech, as it relies on actual human speech recordings. However, the size of the speech database and the complex search algorithms required for optimal segment selection can lead to computational challenges [177]. Parametric synthesis generates speech by modeling the underlying acoustic properties of the human vocal tract [119]. Linear predictive coding (LPC) is a parametric synthesis technique, where the speech signal is approximated by an all-pole filter based on a linear prediction model [111]. While parametric synthesis allows for greater control over the speech output and requires fewer computational resources, the generated speech can sound less natural compared to concatenative synthesis [86].

2.5.2 Brain-Computer Interfaces for Speech Synthesis

Decoding speech from neural activity is a critical component of BCI-based speech synthesis. Studies have shown that it is possible to reconstruct intelligible speech from electrocorticography (ECoG) signals [10] and intracortical recordings [67]. These studies have utilized machine learning techniques, including deep learning algorithms, to map neural activity to speech features and generate intelligible speech.

Recent advancements in deep learning have opened up new possibilities in decoding speech from neural signals. For instance, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to extract temporal and spatial features from EEG and iEEG signals and map these features to speech content. This has resulted in substantial improvements in the quality and intelligibility of the synthesized speech.

The ultimate goal of BCI-based speech synthesis is generating speech directly from brain activity. A key development in this field has been the use of deep learning models that map brain activity to vocal tract movements, which are then used to synthesize speech [10]. This approach has demonstrated the potential to generate intelligible and natural-sounding speech directly from neural signals. However, significant challenges remain in terms of improving the quality and robustness of the synthesized speech and making these systems more practical for real-world applications.

2.5.3 Modern Approaches to Speech Synthesis

Statistical parametric synthesis uses statistical models, particularly machine learning algorithms, to map text input to acoustic features, which are then used to generate speech waveforms [163]. This approach typically employs hidden Markov models HMMs) [164] or deep neural networks (DNNs) [176] for acoustic modeling. Statistical parametric synthesis allows for greater flexibility

and smaller memory requirements compared to concatenative synthesis. However, the generated speech can sometimes exhibit unnatural artifacts due to the limitations of the underlying models and vocoders [163].

Deep learning techniques have revolutionized speech synthesis in recent years. One popular approach is the WaveNet model. WaveNet is a deep generative model for raw audio waveforms that has significantly advanced the state-of-the-art in text-to-speech (TTS) synthesis [168]. It is based on a deep convolutional neural network (CNN) that directly models the conditional probability distribution of the waveform given a linguistic input. WaveNet has achieved unprecedented naturalness in TTS, outperforming traditional parametric and concatenative TTS techniques. [121]

Another influential approach is the Tacotron model, another end-to-end speech synthesis system that converts text directly to speech using a sequence-to-sequence neural network architecture. [2, 147] A key feature of Tacotron is its use of attention mechanisms, which allow the model to focus on different parts of the input sequence at different times during the synthesis process [165]. It combines the strengths of both WaveNet and traditional TTS systems, producing high-quality synthesized speech with a simpler and more efficient training process. Tacotron has been followed by several improved versions, such as Tacotron 2 [148], further enhancing the naturalness and quality of synthesized speech.

2.5.4 Towards a More Naturalistic BCI for Speech Synthesis

Developing a more naturalistic BCI for speech synthesis involves overcoming several significant challenges, the most critical ones being improving the quality of decoded speech, reducing the invasiveness of neural recording techniques, and handling inter-subject variability [16]. Deep learning, with its unrivaled ability to model complex systems, is particularly suited to tackle these hurdles. This, coupled with novel signal processing methods, holds promise for significant advancements in the realm of BCI-based speech synthesis systems [9].

Understanding the cognitive processes underpinning both listened and spoken speech is of paramount importance for refining BCI-based speech synthesis systems. Insights from research into the neural basis of speech perception [63] and the neural encoding of listened speech [71] can provide valuable information that can be used to improve the decoding and generation of speech from neural signals. This requires identifying common patterns of neural activity during both processes and leveraging this information to develop algorithms capable of reconstructing intelligible speech.

Moreover, a promising avenue for development lies in the fusion of artificial intelligence (AI) and

BCI technologies. The convergence of these two fields has the potential to revolutionize the realm of speech synthesis [16]. The power of AI algorithms can be harnessed to decode and generate speech from neural signals, potentially leading to the development of BCIs that are more naturalistic and efficient. With further refinement, it is conceivable that BCIs could handle noisy signals and passive perception, including those derived from individuals participating in real-world conversations, thereby closing the gap between BCI-based speech and naturally produced speech.

Yet, as promising as AI and BCI convergence is, it's worth noting that there are unique challenges associated with this integration. Firstly, developing algorithms that can accurately decode neural signals into speech is a complex task, as it requires a deep understanding of the brain's speech mechanisms and the ability to handle high-dimensional, noisy data. Secondly, the diversity of human brains and the nature of individual variability present significant obstacles for creating universally applicable models. Overcoming these challenges will require continued advancements in AI, a more nuanced understanding of brain function, and the development of personalized models that can adapt to the specific neural activity patterns of each individual.

Furthermore, the invasiveness of current neural recording techniques is a considerable obstacle to the widespread adoption of BCIs. To facilitate long-term, daily use of BCIs, it is crucial to develop non-invasive or minimally invasive neural recording techniques that do not compromise the quality of recorded signals. Such advancements may require innovative approaches in the fields of neuroscience and biomedical engineering, such as developing high-resolution imaging technologies, creating more sensitive sensor arrays, or devising techniques for external modulation of neural activity.

Lastly, ethical considerations will become increasingly important as BCIs evolve and become more widespread. Protecting the privacy and autonomy of BCI users is paramount, as these systems have the potential to access highly sensitive personal information contained within an individual's neural signals. The development of comprehensive ethical guidelines and safeguards will therefore be a crucial component of future BCI research and development.

In conclusion, developing a more naturalistic BCI for speech synthesis is a complex and multidisciplinary endeavor that requires advancements in multiple fields, including neuroscience, artificial intelligence, biomedical engineering, and ethics. Despite the challenges, the potential benefits of such systems, such as restoring communication abilities to individuals with speech impairments, make this a worthwhile and exciting area of research. This thesis contributes to this ambitious goal by providing insights into the cognitive processes underlying speech and offering a foundation for the application of AI techniques in BCI-based speech synthesis.

3. RESEARCH QUESTION AND OBJECTIVES

The intricate and multifaceted nature of speech processing in the human brain has been explored in the preceding chapters, focusing on the mechanisms of both speech perception and production. Building on this foundation of cognitive neuroscience, attention is now turned towards the interplay between these processes and the potential for harnessing this understanding to enhance speech synthesis methodologies. The focus lies within the realm of brain-computer interfaces (BCIs) and deep learning, two technological arenas that have exhibited considerable promise for advancing our ability to decode and generate human speech.

The overarching research question is:

Can the application of deep learning methodologies to intracranial electroencephalography (iEEG) data, recorded during passive listening of speech, enhance the development of BCIs for speech synthesis, and in doing so, provide novel insights into cognitive speech processing and previous BCI approaches?

This research question is predicated on the unique potential of BCIs to bridge the gap between the neural basis of speech and practical applications in speech synthesis. By utilizing rich, high-resolution iEEG data, nuanced insights into the neural underpinnings of listened speech can be gleaned and these insights can be translated into improved BCI models.

To address this research question, the rationale for the choice of BCIs and deep learning for speech synthesis must first be made clear. BCIs are poised to harness the intricacies of brain activity, providing a direct conduit between the neural basis of speech and its synthetic reproduction. Deep learning, with its capacity for handling high-dimensional data and identifying complex patterns, presents an optimal toolset for modeling the intricate, multi-faceted processes of speech perception and production.

Furthermore, the rationale for framing this research within the context of passive listening of speech is articulated. While much of the previous research in the field of BCIs for speech has focused on speech production, the realm of listened speech remains relatively uncharted. It is proposed that this perspective holds untapped potential, offering a fresh vantage point for understanding speech processing and creating more accurate and naturalistic synthetic speech.

The ultimate objective in this study is to leverage a dataset of iEEG recordings collected during passive listening of speech, applying deep learning algorithms to create a model capable of syn-

thesizing speech from this neural data. It is anticipated that this approach will yield two primary outcomes: first, the advancement of BCI technology for more naturalistic speech synthesis, which holds significant implications for aiding individuals with speech and communication disorders. Second, this research is expected to enrich the theoretical understanding of cognitive speech processing, shedding new light on the complex interplay between speech perception and production, and how these processes are mirrored in the neural structures of the brain.

In summary, the research question guiding this study is rooted in the intersection of cognitive neuroscience, BCIs, and deep learning. By fusing these disciplines, uncharted territories in speech synthesis are to be explored, ultimately contributing to both practical applications and theoretical understanding of speech processing.

4. METHODS

4.1 Dataset

This study employs the 'Open multimodal iEEG-fMRI dataset' [17], which is a publicly available dataset that allows for a comprehensive investigation of the neural correlates of speech and language processing in humans. The dataset is unique in that it combines intracranial electroencephalography (iEEG) data with functional magnetic resonance imaging (fMRI) data, enabling the analysis of brain activity at both high spatial and temporal resolutions, making it a great fit for the study.

4.1.1 Participants

The dataset for this study was acquired from fifty-one patients with medication-resistant epilepsy who were admitted to the University Medical Center Utrecht for diagnostic procedures. These procedures involved intracranial electrode implantation for the purpose of identifying the source of the patients' seizures and considering the surgical removal of the corresponding brain tissue. Of these patients, sixteen granted written permission for the use of their clinical data for research, and forty-seven participated in the scientific research conducted by our group by giving their written informed consent. The patients' ages ranged widely, with an average age of 25 and a standard deviation of 15, including 32 females. For the patients under 18 years old, informed consent was obtained from their parents or legal guardian. The Medical Ethical Committee of the University Medical Center Utrecht approved the study, adhering to the Declaration of Helsinki (2013).

4.1.2 Experimental Procedures

The patients participated in two main types of experiments: movie-watching and resting state. The movie-watching experiment, which involved the patient watching a short film, was part of the standard battery of clinical tasks for presurgical functional language mapping. The resting state experiment, which required the patients to rest for three minutes, was conducted for research purposes. For those patients who did not participate in a separate resting state task, a 3-minute 'natural rest' period was selected from their 24/7 clinical iEEG recordings.

4.1.3 Electrode Implantation

The patients were implanted with a variety of electrode types based on clinical needs. Forty-six patients had subdural electrocorticography (ECoG) grids with 2.3 mm exposed diameter, inter-electrode distance of 10 mm, and between 48 and 128 contact points. Six patients had high-density (HD) ECoG grids with 1.3 mm exposed diameter, inter-electrode distance 3–4 mm, and either 32, 64 or 128 contact points. Furthermore, sixteen patients had stereoelectroencephalography (sEEG) electrodes with between 4 and 173 contact points. Most patients had electrode coverage over peri-sylvian areas and the frontal and motor cortices.

4.1.4 Data Acquisition

Intracranial EEG (iEEG) data were acquired using a 128-channel recording system (Micromed, Treviso, Italy) during the experimental tasks. The majority of patients' data were sampled at 512 Hz and filtered at 0.15–134.4 Hz, while in some cases, the data were sampled at 2048 Hz and filtered at 0.3–500 Hz. An external reference electrode was used for signal referencing, typically placed on the mastoid part of the temporal bone. Besides, six patients had their HD ECoG data recorded either simultaneously with the clinical channels or in separate sessions.

4.1.5 Stimuli

The stimulus for the movie-watching experiment was a 6.5-minute short movie composed of fragments from "Pippi on the Run" (Pårymmen med Pippi Långstrump, 1970). The movie was edited to form a coherent plot and consisted of 13 interleaved blocks of speech and music, each 30 seconds long. The movie was originally in Swedish but dubbed into Dutch. Detailed annotations of the audio and video content of the movie stimulus were provided.

4.1.6 Structural Data Acquisition

For the majority of patients, structural T1 images were obtained using a Philips Achieva 3 T MRI scanner. However, one patient had a 7T structural scan, and twenty had a 3 T scan with varying parameters.

4.1.7 De-identification and Data Sharing

All patient data included in this dataset have been thoroughly de-identified, and all patients gave consent to share this data publicly. As a result, this dataset can be shared with the research com-

munity while still ensuring patient confidentiality.

4.2 Data Availability

The dataset utilized in this study is openly accessible at:
https://openneuro.org/datasets/ds003688. To protect patient confidentiality, all personal identifiable information has been removed, and individual MRI scans have been defaced. Furthermore, the order in which subjects are presented in the dataset has been randomized, ensuring no identifiable pattern exists (such as alphabetical order or order by date of the experiment). This allows for data sharing with the research community while maintaining patient confidentiality.

4.3 Data Validation

In this study, the specific focus was on utilizing the iEEG data from the 'Open multimodal iEEG-fMRI dataset'. This was chosen due to its high temporal resolution and the direct neural recordings it provides. The iEEG data offers a unique opportunity to examine the fine-grained temporal dynamics of neural activity during speech and language processing, which is crucial for understanding the neural mechanisms underlying speech synthesis. By analyzing the iEEG data, it is possible to gain insights into the neural representations of speech, the temporal patterns of brain activity associated with different linguistic features, and the interactions between brain regions involved in the processing of speech and language. Leveraging these insights, the aim was to develop deep learning models that can accurately decode listened speech from iEEG data and generate synthesized speech based on the neural representations. Ultimately, the findings can contribute to the advancement of brain-computer interface technology, enabling the development of more naturalistic and efficient speech synthesis systems that are directly controlled by brain activity. To prepare, validate and clean the iEEG data for training and analysis, several preprocessing steps were performed using MNE-Python (https://mne.tools).

First, channels of type "ECoG" and "sEEG" were selected, and previously identified bad channels were discarded. To account for line noise at 50 Hz and its harmonics, a notch filter was applied to the data. The data was then re-referenced using the common average signal, and band-specific neural signals were extracted using the Hilbert transform for the following frequency range (60-120 Hz).

To ensure data quality, the subjects' neural activity during speech and music blocks was compared by the team behind the dataset [17]. They used an ordinary least squares fit to the HFB

envelopes with the block design boxcar function. The fit and statistical analysis were performed using the Python package statsmodels (https://www.statsmodels.org). Given the potential delay in the brain's response to auditory input and the possibility of varying delays across electrodes, the fit was calculated per electrode at all time lags within 1 second after the sound onset. The best fit across the lags was recorded, along with the lag value. The significance of the fit was assessed parametrically based on the t-statistic for the block design regression weight, with only positive t-statistics reported, corresponding to higher responses to speech and lower responses to music (for the block design predictor with zeros in music blocks and ones in speech blocks) that are significant at $p < 0.01$, Bonferroni corrected for the number of electrodes and lags.

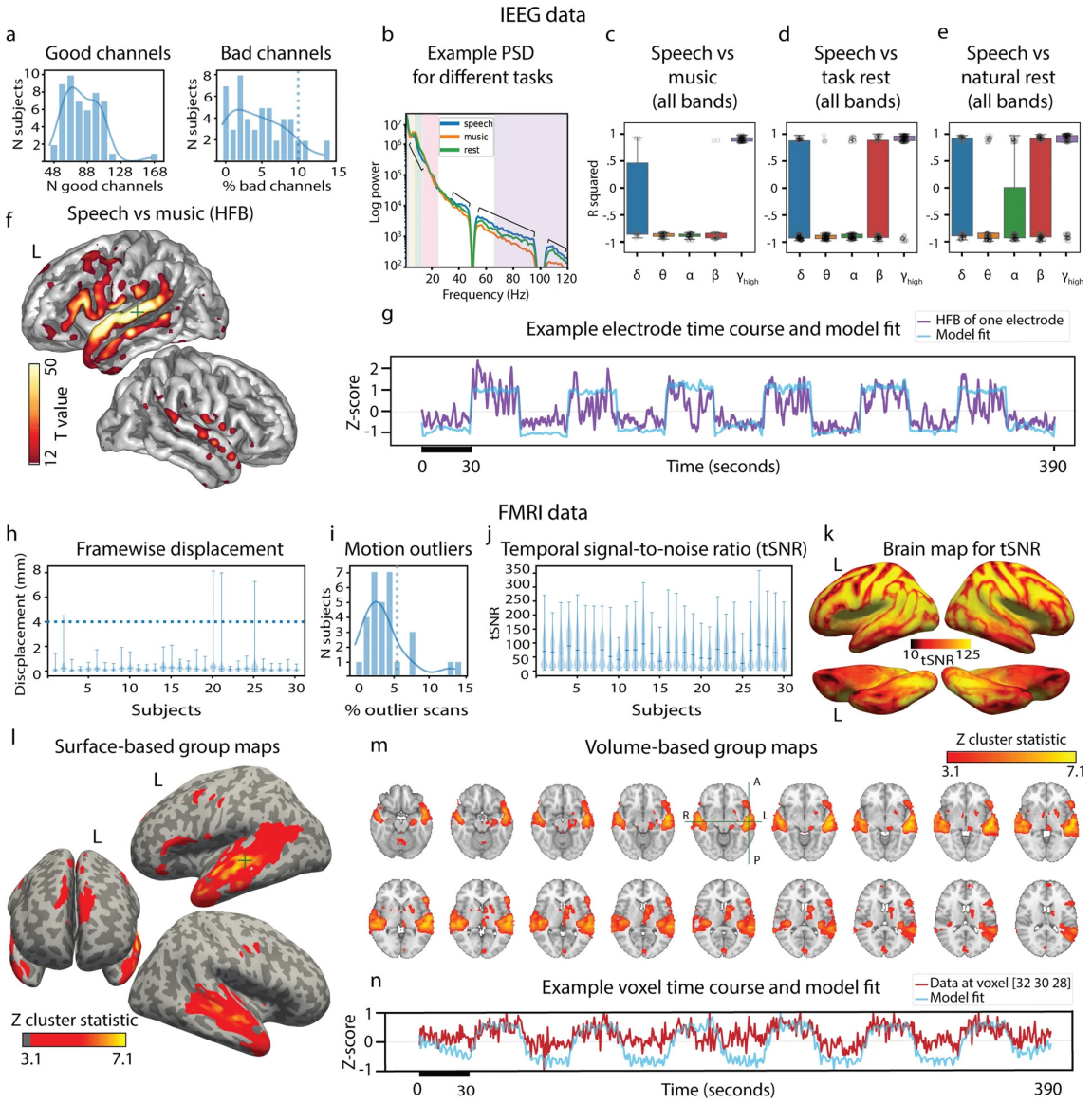

Figure 4.1: Dataset overview [17]

An overview of the dataset can be seen in fig. 4.1. Technical data validation in iEEG (a–g) and

fMRI (h-n). (a) Histograms of good and bad channels over all subjects with estimated distribution density (solid line) and cut-off at 10% of bad channels (dotted line). (b) Example of power spectral density (PSD) plot for one electrode in one subject per condition: speech, music and rest. Different frequency bands are highlighted (δ, θ, α, β and high frequency band, HFB). (c-e) Boxplots of signed r-squared values (significant at p<0.01) for three comparisons in the neural data: speech vs music (c), speech vs task rest (d) and speech vs natural rest (e), separately per frequency band. Boxes outline 25th and 75th quantiles, round markers show individual electrodes. (f). Brain map (all patients) for results of the linear regression of HFB data to the block design in the movie-watching task (speech and music blocks). Only positive t-statistic values significant at p<0.01 are shown. For visualization, a 2d Gaussian kernel (10mm in width) was applied to each electrode's central coordinate. Example time course fitted by the model (block design+audio envelope) and observed HFB in one electrode of one subject. (h). Framewise displacement in fMRI data. Violin plots show entire data range, horizontal lines show medians. Dotted line shows voxel size (4mm). (i) Histogram of motion outliers with estimated distribution density (solid line) and cut-off at 5% of bad volumes. (j) Violin plots of the temporal signal-to-noise ratio (tSNR), same display setup as in h. (k) Brain map of tSNR averaged over all subjects and projected onto the average surface. (l) Group-level statistics based on results of the first-level analysis fitting a general linear model on fMRI data using the block design (with default FSL parameters). (m) Same group-level brain map on volume slices. (n) Example time course fitted by the model (block design+motion parameters) and observed fMRI in one voxel of one subject.

Additionally, signed r-squared values (calculated as a Pearson correlation coefficient squared, preserving the sign of correlation) between speech and music blocks, speech blocks and task rest, and speech blocks and natural rest were computed. To reduce the number of multiple comparisons (number of electrodes × frequency bands), the analysis was performed only on the electrodes with a significant linear fit to the block design (see the analysis above). The three comparisons (speech vs. music, speech vs. task rest, and speech vs. natural rest) were made separately for all extracted iEEG frequency bands. The significance of reported r-squared values was determined parametrically, with reported values significant at $p < 0.05$, Bonferroni corrected for the number of electrodes and frequency bands.

4.3.1 Prime Subjects

In order to facilitate the most effective and meaningful analysis for this study, I used a rigorous selection process for the subjects. This process was primarily oriented around a key determinant -

the level of correlation that each subject demonstrated with the speech envelope during the movie, which was done by the team who compiled the dataset [17]. The significance of this criterion lies in the crucial interplay between neural activity and the speech envelope, a fundamental aspect when considering brain-computer interfaces designed for speech synthesis.

This subject selection methodology originated from the hypothesis that individuals whose neural activity closely mirrored the dynamic ebb and flow of the speech envelope would be ideal candidates for this study. Such a correlation implies a substantial degree of neural engagement and synchronization with the auditory stimuli, vital characteristics for the research we aimed to conduct. From the pool of potential subjects, four individuals were eventually selected as shown in fig. 4.2.

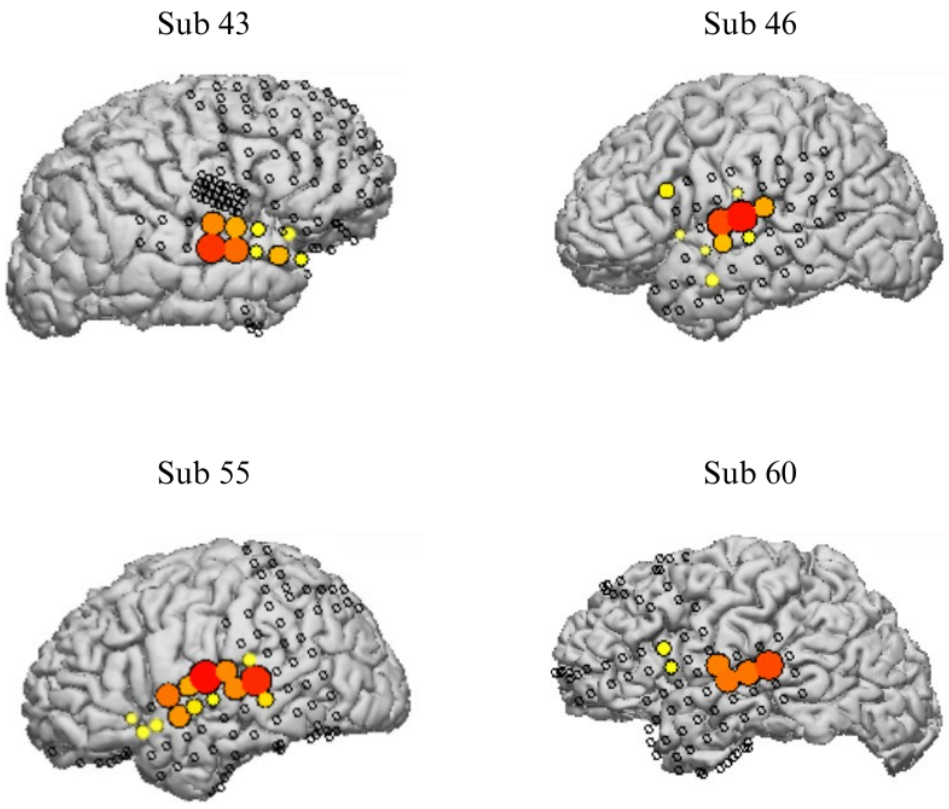

Figure 4.2: The 4 subjects with highest correlation with speech envelope. [17]

The selected participants displayed notably high correlation values, suggesting a robust connection between their neural activity and the speech envelope presented in the movie. It's pertinent to note that the pronounced correlation likely results from the placement of the intracranial electrodes covering key areas associated with speech perception and production, as outlined in the theoretical

overview. These areas include the Broca's area, the motor cortex, the cerebellum, Wernicke's area, and the superior temporal gyrus.

This selection process was meticulously calibrated to ensure that we recruited subjects whose neural responses would most likely yield the richest and most insightful data for the ambitious endeavor of decoding and reconstructing speech from neural signals.

Alongside the data-driven selection, another layer of manual selection was integrated to ensure the coverage of essential regions of the brain critical to the study. Subject 38 was particularly chosen based on their exceptional coverage of electrodes over the motor cortex, the Broca's area and the superior temporal gyrus . As these are instrumental in the generation of neural signals related to speech production, a comprehensive representation from this area significantly enhances the richness of data collected.

This unique electrode placement in Subject 38 might provide a great opportunity for more accurate and nuanced speech reconstructions. By employing both quantitative and qualitative selection

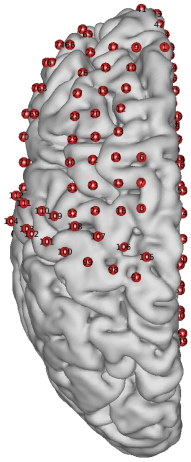

Figure 4.3: The electrode positions for Subject 38. extracted from the Open multimodal iEEG-fMRI dataset [17]

criteria, the selection process has been meticulously designed to identify the subjects who are most likely to contribute valuable and insightful data to the study. This holistic approach leverages statistical metrics and the critical understanding of the brain's speech production mechanisms. By doing so, it not only maximizes the potential for successful speech decoding and reconstruction from neural signals, but also fuels the advancement in our understanding of how the brain produces speech. This refined understanding will be a substantial contribution to the development of highly effective brain-computer interfaces for speech synthesis, marking a significant stride in neuroscience and technology.

4.4 Preparing Data for training

The first significant step in the data preparation phase involves cropping the EEG data to match the duration of the stimulus presented during the experiment. This was done based on the annotations

embedded in the EEG data, which marked the start and end points of the stimulus presentation. The stimulus, a 6.5-minute long movie, served as the temporal benchmark. This approach ensures that the EEG data corresponds strictly to the period when the subjects were actively engaged with the stimulus, thereby focusing the analysis on task-relevant neural activity.

The trimmed EEG data, now aligned with the movie duration, provides a more accurate and meaningful basis for subsequent steps in the training process.

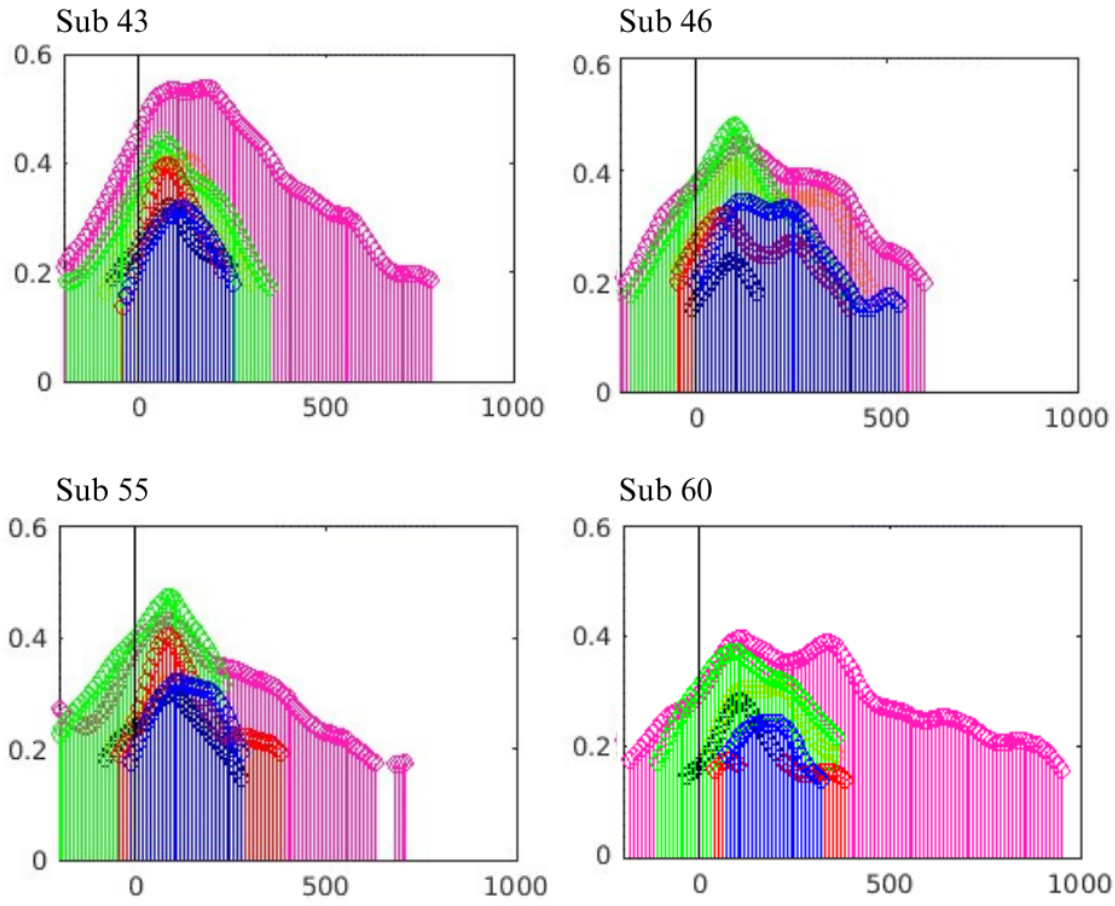

Figure 4.4: Lagplots of the cross-correlation of the electrode's high frequency band signal and the sound envelope [17].

Figure 4.4 provides a visual representation of the cross-correlation between the electrode's high-frequency band signal and the sound envelope. On the y-axis, the correlation magnitude can be seen, and on the x-axis, the positive time lag in milliseconds. The positive time lag signifies that the brain data lags behind the audio. This lag is illustrated through different colors, each representing a separate 30-second speech block. The average delay observed is approximately 150 milliseconds. This delay between the neural response and the audio stimulus is a crucial consideration in our study, as it influences the timing of the speech synthesis process. To account for this observed lag,

the audio was shifted backwards by 150 milliseconds. This adjustment ensures that the decoded speech aligns more accurately with the original auditory stimulus, thus enhancing the naturalness and intelligibility of the synthesized speech.

When further preparing for training, several steps are taken to extract features from the iEEG and mel spectrogram data using the Hilbert transform. The process consists of the following steps:

Linear detrending: The iEEG data is linearly detrended using scipy.signal.detrend to eliminate any linear trends or biases present in the data.

Calculation of the number of windows: Based on the window length and frame shift parameters, the number of windows is determined. This indicates the number of feature vectors to be extracted.

High-Gamma bandpass filtering: The iEEG data undergoes bandpass filtering in the high-gamma frequency range using an IIR filter designed with scipy.signal.iirfilter. The frequency range is defined by the bandpass_min and bandpass_max parameters.

Attenuation of line noise harmonics: Optionally, bandstop filters can be applied to attenuate specific harmonics of line noise, such as the first and second harmonics. However, these lines are currently commented out in the code.

Creation of the feature space: The absolute values of the Hilbert transform of the filtered iEEG data are computed using the hilbert3 function. The resulting data represents the amplitude envelope of the high-gamma frequency band.

Window-based feature extraction: The feature space is divided into overlapping windows, and for each window, the mean amplitude values across channels are calculated. This results in a feature vector for each window.

Stacking of features: The extracted feature vectors are stacked together using a function with a specified model order and step size. The function responsible for this is not shown in the provided code snippet.

Lastly, the lengths of the iEEG and mel spectrogram data are compared, and the minimum length is used to truncate the data. This ensures that the extracted features and the input data have the same length for further processing.

In summary, this part of the process involves detrending, filtering, and feature extraction using the Hilbert transform to extract high-gamma band envelope features from the iEEG data. The extracted features are then stacked and prepared for further analysis.

In the later phase of preparing data for training, a crucial step involved the careful extraction of

EEG data corresponding to the segments where speech was present in the movie. The aim was to align the EEG data with the speech-specific segments of the movie, thereby focusing the analysis on the brain's response to auditory speech stimuli.

To achieve this, the raw_car data, representing the preprocessed EEG data, and the mel_data, representing the Mel spectrogram data, were selectively sliced based on the provided annotations. These annotations were contained within the events array, indicating the start and end times of each speech segment.

The process was executed through a loop that iterated over each speech segment in the movie. The start and end times of each speech segment were converted into corresponding indices in the raw_car data using the time_as_index method, taking into account the sampling frequency of the EEG data. A similar calculation was made to obtain corresponding indices in the mel_data, considering the frame shift parameter.

Once the start and end indices were determined for each speech segment, the EEG data and Mel spectrogram data were cut and appended into raw_car_cut and mel_data_cut arrays, respectively. These arrays represented the EEG data and Mel spectrogram data, now specifically aligned with the movie's speech segments.

The result of this procedure was a refined set of data (raw_car_cut and mel_data_cut), precisely encapsulating the EEG responses to speech stimuli, thus enhancing the relevance and accuracy of the subsequent deep learning model training.

After meticulously preparing the iEEG and mel spectrogram data, it's worth emphasizing the importance of this phase. The careful alignment of the EEG data with the speech segments of the movie, the consideration of the time lag between the brain data and the audio stimulus, and the detailed feature extraction were all vital in ensuring the quality and relevance of our data for training. As I now transition to explaining the structure and function of the deep learning models, these data preparation steps serve as the sturdy foundation on which our models operate and generate meaningful results.

4.5 Deep Learning Training

Deep learning, a subset of machine learning, excels in various domains, such as computer vision, natural language processing, and bioinformatics, owing to its capability to extract intricate and abstract patterns from vast, high-dimensional data sets, such as images, audio, and in this study, intracranial electroencephalogram (iEEG) data. The primary motivation for choosing deep learn-

ing for this investigation is its unrivaled competence in handling high-dimensional, intricate data and its well-established efficacy in related tasks.

In this context, Fully Connected Deep Neural Networks (Fc-DNNs) and 2D Convolutional Neural Networks (2D-CNNs) were chosen. The selection of these two network architectures was dictated by their inherent properties and their suitability for the task of predicting mel-spectrograms from iEEG data.

It's crucial to underscore that these models were not selected arbitrarily. Instead, it was the result of a comprehensive and iterative process of evaluating various hyperparameters and network configurations. Numerous model architectures, training strategies, and optimization techniques were assessed, with each iteration refining the approach based on the performance results. The configurations that consistently performed optimally were chosen for the final models, which will be detailed in the following sections. This rigorous process of model selection and optimization underscores the robustness and thoroughness of the methodology adopted in this study.

To promote transparency and repeatability of the research conducted in this study, all code, files, and associated scripts used in data preprocessing, model training, and result analysis have been publicly shared. These resources can be accessed in the GitHub repository at:
https://github.com/MILANIUSZ/speech2brain2speech.

The training process was carried out using an RTX 3070 graphics card and an AMD Ryzen 5 3600 processor. The training environment was set up using Docker, specifically utilizing the public image "thegeeksdiary/tensorflow-jupyter-gpu." Instructions for setting up a similar environment can be found at https://thegeeksdiary.com/2023/01/29/how-to-setup-tensorflow-with-gpu-support-using-docker/. This setup ensured optimal utilization of the available hardware resources for efficient model training and evaluation.

4.5.1 Fc-DNN

Fully Connected Deep Neural Networks (Fc-DNNs), also known as Multilayer Perceptrons (MLPs), are among the simplest types of neural networks. They comprise several layers of neurons, with each neuron connected to all neurons in the preceding layer, hence "fully connected". Fc-DNNs are versatile and robust, capable of approximating any function given a sufficient number of neurons and layers. Their success in a variety of tasks, including regression tasks similar to the one in this study, is well-documented.

In this study, a fully connected Deep Neural Network (Fc-DNN) model was constructed with a

single hidden layer, containing 3000 neurons. The selection of this architecture was not arbitrary but emerged from a series of systematic experimentation with various configurations of layers and neurons. The guiding principle behind the choice was not just to optimize performance metrics, such as accuracy, but also to ensure a robust model that generalizes well and avoids overfitting.

When the complexity of the model was escalated by increasing the number of layers and neurons, while there was no substantial degradation in accuracy, noticeable issues with overfitting surfaced. This overfitting was marked by the model's outputs - the generated melspectrograms exhibited aberrant characteristics significantly deviating from what one would anticipate from a typical mel-spectrogram. These peculiarities suggested an over-complex model that had potentially learnt the training data too well, capturing even the noise, hence producing outputs that were rather unrealistic.

Therefore, in an effort to strike an effective balance between model complexity and performance, and to avoid the undesirable effects of overfitting, the optimal configuration was determined to be a Fc-DNN model with one hidden layer comprising of 3000 neurons. This architecture resulted in a model that not only performed admirably in terms of accuracy, but also demonstrated a better capacity for generalization, evident from the melspectrograms that it produced, aligning closely with expectations.

The Rectified Linear Unit (ReLU) activation function was used for the input layer due to its ability to create sparse representations and mitigate the vanishing gradient problem. A linear activation function was selected for the output layer since this is a regression task, and the output can be any real number. The Adam optimizer was chosen due to its adaptive learning rate capabilities, leading to quicker and more stable convergence.

To prevent overfitting, a common problem in deep learning where the model memorizes the training data too well and performs poorly on unseen data, early stopping was used, and the best model weights were saved during training.

Below is the detailed process of preparing data and building the Fc-DNN model:

Firstly, the data was split into training, validation, and test sets, with 80% of the data allocated for training, 10% for validation, and the remaining 10% for testing.

The EEG data was scaled to the range [0,1] using the MinMaxScaler. It is crucial to normalize the data to ensure that all inputs have the same scale, which contributes to more stable training and improved performance.

The output mel-spectrogram data was also scaled, but using the StandardScaler to transform the data to have zero mean and unit variance.

The Fc-DNN model was constructed using the Sequential API from Keras. The model comprised an input layer with 3000 neurons and a ReLU activation function, and an output layer with 80 neurons and a linear activation function. The choice of 3000 neurons for the input layer corresponds to the dimensionality of the input data, and the 80 neurons in the output layer correspond to the dimensionality of the mel-spectrogram data.

The model was compiled with the Mean Squared Error (MSE) as the loss function and Adam as the optimizer. MSE is a common choice for regression tasks as it penalizes large errors more than small ones, leading to more accurate models.

To avoid overfitting, an EarlyStopping callback was used, which stopped the training when the validation MSE did not improve by at least 0.0001 for two consecutive epochs. Furthermore, the weights of the model with the smallest validation loss were saved using the ModelCheckpoint callback.

The model was then trained for a maximum of 50 epochs with a batch size of 32. The validation data was used to tune the hyperparameters and monitor the model's performance during training. The training process was logged using the CSVLogger callback, and the scalers used to normalize the data were saved for future use.

The process described above constitutes the model's training phase, during which the model learns to map the EEG data to the mel-spectrogram data. The performance of this model was subsequently evaluated on the test set, which had not been seen by the model during training. The results of this evaluation will be discussed in the following sections.

The entirety of the method is detailed in the Jupyter notebook "speech2brain2speech_FC-DNN"

4.5.2 2D-CNN

Two-Dimensional Convolutional Neural Networks (2D-CNNs) are a specialized type of neural network designed to process data with a grid-like topology, such as an image composed of pixels. They have shown tremendous success in image and audio processing tasks due to their ability to capture both local and global patterns in the data.

In this study, the 2D-CNN was applied to process the spectrogram data derived from EEG recordings. The data was split into training and test sets, with the training set comprising 80% of the total

data. Both the input and output data were then normalized using the mean and standard deviation calculated from the training set. The input data was reshaped to match the required 3D structure for 2D-CNN, i.e., (9, 127). A sliding window approach was used to convert the data into 3D blocks.

The architecture of the 2D-CNN model consisted of three convolutional layers with kernel sizes of (13, 13) and activation function as 'swish'. The convolutional layers were interspersed with dropout layers to prevent overfitting. The architecture also included a max pooling layer to reduce the dimensionality of the input and highlight the most salient features.

The model then passed through a flatten layer, transforming its input into a one-dimensional array. A densely connected layer followed the flatten layer, with an activation function 'swish'. Another dropout layer was added before the final output layer. This output layer was a dense layer with a linear activation function, matching the shape of the training spectrogram.

The model was compiled with the 'Adam' optimizer and the 'mean squared error' as the loss function. To control overfitting and fine-tune the learning rate, an early stopping mechanism was applied, and a learning rate reduction on plateau scheme was implemented. The model was trained over 100 epochs with a batch size of 128. The best weights obtained during training were saved and used to predict the test data.

Upon completion of training, the model's predicted spectrogram was saved. The predicted data was then inverse transformed, which involved scaling it back by the standard deviation and adding the mean to regain the original scale of the spectrogram data. The reconstructed spectrogram was then saved for further evaluation.

The entirety of the method is detailed in the Jupyter notebook "speech2brain2speech_2D-CNN"

4.5.3 Evaluation Methods

The performance of the deep learning models, namely the fully-connected deep neural network (Fc-DNN) and the 2D Convolutional Neural Network (2D-CNN), was assessed using a combination of quantitative measures and qualitative visualizations.

Quantitatively, the mean squared error (MSE) was used as the primary metric for evaluating the models' ability to predict the Melspectrogram data from EEG signals. The MSE measures the average squared difference between the predicted and actual Melspectrogram values, with lower values indicating better performance.

Qualitatively, the models' outputs were visualized in the form of reconstructed Melspectrograms.

These visualizations allowed for a more intuitive and direct understanding of the models' performance. The reconstructed Melspectrograms were generated by feeding the predicted Melspectrogram data back into an inverse Mel-spectrogram transformation, which essentially converts the data back into the time domain.

The reconstructed Melspectrogram plots provide a powerful visual tool for comparing the original and reconstructed signals. Any significant discrepancies between the original and the reconstructed Melspectrograms can indicate areas where the models are underperforming. For example, if certain frequency components are consistently missing in the reconstructed Melspectrograms, it may suggest that the models are struggling to capture those particular components.

Lastly, the evaluation also involved an auditory analysis, whereby the reconstructed Melspectrograms were converted back into audio signals. This was accomplished using an inverse short-time Fourier transform (iSTFT), which transforms the frequency domain data back into the time domain. The resulting synthesized audio signals provided an additional layer of evaluation, allowing for an auditory assessment of the models' performance. Any audible differences between the original and synthesized audio signals can help to further identify areas where the models may be underperforming.

In summary, the evaluation of the models incorporated a combination of quantitative metrics, visual spectrogram analysis, and auditory analysis. These methods together provided a comprehensive assessment of the models' performance in predicting Melspectrogram data from EEG signals.

5. RESULTS

In this chapter, the results obtained from the deep learning models, specifically Fc-DNN and 2D-CNN (A.0.1,A.0.2, will be presented. These models were trained on the collected brain activity data during passive listening and spoken speech tasks. The performance of each model is discussed, emphasizing the potential applications in brain-computer interfaces for speech synthesis and communication.

5.1 Fully-connected Deep Neural Network (Fc-DNN)

The Fc-DNN was trained on the data from four different subjects, and the performance of the model for each subject is summarized in Table 5.1. The table presents the best training loss and validation mean squared error (MSE) achieved for each subject.

Subject	Best Training Loss	Best Validation MSE
38	0.021	0.6982
43	0.0336	0.7381
46	0.2643	0.7923
55	0.2015	0.721
60	0.39	0.6520

Table 5.1: Performance of the Fc-DNN for each subject.

The training loss values represent how well the model is able to predict the Melspectrogram data from the EEG signals during training. Lower training loss indicates a better fit of the model to the training data. The validation MSE, on the other hand, provides a measure of the model's performance on unseen data, with lower MSE values representing better generalization performance.

From Table 5.1, it can be observed that the model achieved the lowest training loss with subject 43, indicating the model was able to fit the training data most effectively for this subject. On the other hand, the model demonstrated the best generalization performance on unseen data with subject 60, as indicated by the lowest validation MSE.

5.2 Two-Dimensional Convolutional Neural Network (2D-CNN)

Just like the Fc-DNN, the 2D-CNN model was trained on the data from the four different subjects. The performance metrics for the 2D-CNN, specifically the best training loss and the validation mean squared error (MSE) for each subject, are outlined in Table 5.2.

Subject	Best Training Loss	Best Validation MSE
38	0.4121	0.7023
43	0.5321	0.7326
46	0.8043	0.6920
55	0.9605	0.7879
60	0.9039	0.7922

Table 5.2: Performance of the 2D-CNN for each subject.

The 2D-CNN model performance is evaluated using the same metrics as the Fc-DNN model: the training loss, the validation MSE and listening to the synthesized audio. The training loss values demonstrate the accuracy of the model in predicting the Melspectrogram data from the EEG signals during the training phase. A smaller training loss indicates a better fit of the model to the training data. The validation MSE offers a measure of the model's performance on unseen data, with smaller MSE values indicating better generalization performance.

5.3 Audio synthesis

The synthesized audio resulting from both models was subjected to human evaluation to gauge its perceptual quality and intelligibility. While the speech was not comprehensible, the audio was not a complete blur either. The model was successful in capturing some significant auditory features such as the silences between the speech segments, which is a critical aspect in determining the rhythm and pace of spoken language. However, the primary challenge remains in the precise reconstruction of speech content, implying that the models in this form has yet to effectively decode the complex relationship between brain activity and speech production. Further research and model optimization could lead to more intelligible and natural-sounding synthesized speech.

6. DISCUSSION

6.1 Speech Decoding

The investigations outlined in this thesis explore the capacity of deep learning models for decoding brain activity during passive listening and spoken speech. The selected deep learning architectures, Fully Connected Deep Neural Network (Fc-DNN) and the 2D Convolutional Neural Network (2D-CNN), have demonstrated promising potential especially in light of the limited training data), successfully reducing test loss in a consistent manner. Intriguingly, in certain instances, the implemented models also showed a decrease in validation loss, limited solely by the early stopping mechanism enforced to avoid overfitting.

Despite the relative success of these models in certain facets of the decoding process, there remain considerable obstacles to overcome. A notable challenge in the current study has been the difficulty in reaching a satisfactory level of accuracy for both validation and test sets concurrently, despite the application of numerous training iterations. This issue manifests in the predicted Melspectrograms, which, although indicating some recognition and learning of patterns within the data (as depicted in Figure 6.1) the 2D-CNN captures the right areas in terms of intensity, while the FC-DNN can capture more detail, they both fall short of providing a realistic spectrogram. Consequently, the audibility and clarity of the synthesized speech generated from these Melspectrograms were below anticipated levels.

This divergence between expectations and outcomes invites a discussion regarding existing research within the field of speech synthesis and decoding. It is noteworthy that the present study diverges from a significant proportion of prior research by focusing on decoding during passive listening scenarios, which have been less explored and present their own unique challenges. Unlike active tasks such as reading or speaking, which elicit overt motor and auditory brain responses, passive listening does not rely on such explicit signals. This factor considerably complicates the decoding process, making the development of effective decoding algorithms a challenging task.

In the pursuit of comparable research, a prominent study by Anumanchipalli et al. [10] merits attention. This study, which entailed the synthesis of intelligible speech from neural activity recorded while participants were engaged in reading aloud, reported successful outcomes. A crucial difference to note is that, unlike our study, Anumanchipalli et al. relied on active tasks involving overt speech, which would be expected to facilitate the decoding process through the generation of both

motor and auditory brain responses. Consequently, the contrast between the results obtained in their study and those obtained in ours can be largely attributed to the differing nature of the tasks involved.

Simultaneously, the present study finds closer alignment with investigations such as the one conducted by Akbari et al. [1]. Here, the researchers aimed to decode spectrograms from brain activity recorded during passive listening tasks. Their findings, which also reported challenges in generating realistic spectrograms and clear, synthesized speech, resonate with the challenges encountered in the present study. However, it is crucial to exercise caution when drawing comparisons across

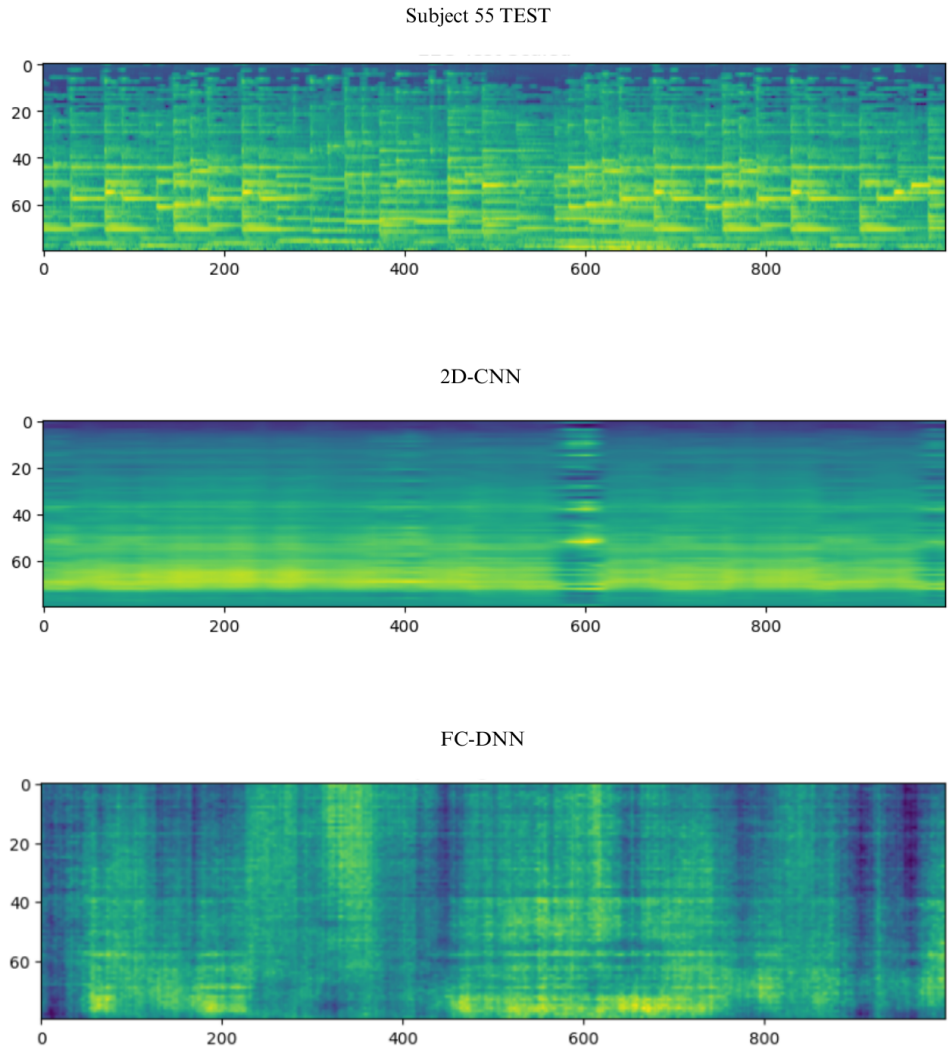

Figure 6.1: Melspectograms for sub 55

studies. Methodological variations, including differences in data collection techniques, preprocessing steps, model architectures, and evaluation metrics, can significantly influence the outcomes

and interpretability of results. For instance, some studies might employ invasive electrocorticography (ECoG) for data collection, resulting in high-resolution data, while others might utilize non-invasive methods such as EEG or fMRI.

Despite these disparities, it is imperative to recognize the broader trends within the field. These trends underscore the nascent state of speech decoding, particularly in the context of passive listening scenarios. They highlight the multi-faceted complexities of the task and indicate the significant advancements that remain to be accomplished before the reliable synthesis of clear and audible speech from passive listening brain activity becomes feasible.

The challenges and lessons drawn from this study contribute invaluably to the broader understanding of thhe field. They underscore both the potential and limitations inherent in the application of deep learning models for speech decoding, providing key insights that are likely to inform and shape future investigations in this rapidly evolving field. Moreover, these findings underscore the necessity for continual development and refinement of the methodologies and approaches employed in this domain, catalyzing further progress in our understanding and manipulation of the intricate relationship between brain activity and speech.

6.1.1 Limitations and shortcomings

This research, while significant in terms of its findings and the potential it exhibits, does bear certain limitations that need to be acknowledged and addressed in future iterations of this work.

Firstly, one of the major challenges encountered in the course of this research pertained to the synchronization of EEG and audio data. The process of achieving precise alignment of these two inherently different types of data is critical to the effective training of the neural network models. Given potential latency issues and the different natures of these signals, aligning them accurately is a complex task. Misalignment in the data could compromise the quality of the training set and, consequently, the performance of the models. Therefore, there is a crucial need for future research to focus on devising robust methods that can facilitate the accurate synchronization of data, thereby improving the quality of the training dataset.

Secondly, the quantity and diversity of the dataset play a significant role in the performance of deep learning models. In this study, the dataset size was relatively limited, which likely imposed constraints on the models' ability to learn, generalize, and maintain robustness. Deep learning models typically thrive on large and diverse datasets, so the scope of the data available here might have put a damper on the performance of the models. Therefore, future research needs to prioritize the

gathering of larger, more diverse datasets, encompassing data from a wider range of subjects under varying experimental conditions and different brain signal modalities. By doing so, the training process can be significantly enhanced, leading to improved performance and generalizability of the models.

Thirdly, the study, while it did gather data from regions of the brain known to play a role in speech perception and production, could benefit from a more targeted approach in terms of data collection in future iterations. A more nuanced understanding of the roles played by different areas in speech processing could inform the selection of brain regions for data collection in future studies, leading to a richer, more informative dataset. This, in turn, could serve to improve both the performance of the models and our overall understanding of speech-related neural activity.

Fourthly, the duration of the stimuli used in the current study could stand to be extended in future studies. The six-minute film, while it did provide a reasonable dataset, falls somewhat short in terms of duration. Longer stimuli could potentially allow for the detection of more nuanced patterns in brain activity. By providing a richer dataset, they could contribute significantly to the development of more robust and accurate models.

Fifthly, the choice of neural network architecture and hyperparameters is a significant factor in determining the performance of the models. While the Fully Connected Deep Neural Network (Fc-DNN) and the 2D Convolutional Neural Network (2D-CNN) used in this study showed promising results, there remains considerable room for further exploration and improvement. More advanced deep learning architectures, such as recurrent neural networks (RNNs), transformers, (or with more available data) attention-based models, could potentially offer improved performance when it comes to decoding brain signals related to speech.

Sixthly, incorporating a scenario where subjects audibly reproduce the speech they hear could provide a valuable reference for future studies. This approach could serve to offer an additional layer of data for comparison and a more comprehensive understanding of the differences and similarities between perceived and spoken speech at the neural level.

Seventhly, the issue of interpretability remains a significant challenge in the world of deep learning. Despite their high performance, these models often act as ”black boxes,” providing little to no insight into the features or patterns they have learned. As such, there is a pressing need for the development of techniques that can enhance the interpretability of neural networks. Doing so could help bridge the gap between machine learning and neuroscience, offering valuable insights into the cognitive processes underlying speech perception and production.

Lastly, while this study focused on EEG data, other brain signal acquisition methods, such as Magnetoencephalography (MEG) or functional Magnetic Resonance Imaging (fMRI), could be utilized in future studies. These modalities could potentially offer richer and more detailed data, possibly leading to improved model performance and a deeper understanding of the neural mechanisms involved in speech processing.

In conclusion, while this research has shed light on the potential of deep learning in decoding neural signals for speech synthesis, it also underscores the need for addressing several limitations and shortcomings. By doing so, future work can continue to enhance our understanding of speech perception and production at the neural level and contribute to the development of more effective and naturalistic BCIs.

6.2 Cognitive Conclusions

The results presented in this thesis are testament to the power of deep learning models in decoding brain activity during passive listening and spoken speech. In this journey of discovery, the exploration of intricate cognitive and neural processes during speech perception and production has led to insights with profound implications for cognitive science, neuroscience.

A compelling revelation from this research is the identification of shared features in the neural activity associated with both passive listening and spoken speech. This finding does more than just confirm the existence of these shared patterns; it shows that they can be successfully identified and decoded by deep learning models. This paves the way for further exploration of the cognitive frameworks and neural substrates that underpin these processes [63]. It also offers potential evidence in support of the 'motor theory of speech perception' which suggests that the cognitive processes involved in understanding and producing speech may be fundamentally intertwined [96].

Such revelations have the potential to expand our understanding of auditory cognition, a critical factor in tasks such as music perception, voice recognition, and auditory scene analysis. The discovery that meaningful patterns in brain activity can be extracted even during passive listening could drastically change our perception of these cognitive processes. There might be more subtleties and complexities involved than we have so far assumed [34].

Additionally, the shared patterns in brain activity identified during passive listening and spoken speech could provide evidence supporting 'neural reuse'. This theory postulates that specific neural circuits can be reused across a multitude of cognitive functions [4]. Evidence for neural reuse could significantly advance our understanding of brain organization and function, leading to revolutionary

approaches in fields such as neuropsychology and neurorehabilitation.

The results also hold promise in shedding light on 'mirroring' processes in the brain, a concept underpinned by the discovery of 'mirror neurons'. These neurons were first discovered in the context of motor actions but have since been suggested to play a role in cognitive functions such as speech [136]. If the existence of 'speech mirror neurons' can be confirmed, this could open up new research avenues to explore their role in empathy, language learning, and communication.

Furthermore, the successful application of deep learning models in this research underscores their potential in the development of more advanced brain-computer interfaces (BCIs). With the ability to decode brain activity during speech-related tasks, we could potentially develop BCIs that offer more seamless and intuitive communication between the human brain and external devices [81].

Beyond BCIs, the application of deep learning models could significantly impact the wider field of neuroscience. These models' ability to learn from high-dimensional data and capture intricate patterns makes them ideally suited to explore the complexity of brain activity. As our research demonstrates, deep learning models can help uncover intricate insights into the workings of the brain, expanding their potential for even more innovative applications in neuroscience.

It is also important to highlight the interdisciplinary nature of this research. Our work stands at the intersection of machine learning, cognitive science, and neuroscience, thus exemplifying the power of interdisciplinary research. By combining insights from these different fields, we can gain a more comprehensive and nuanced understanding of complex phenomena like speech, which might not be possible within the confines of a single discipline.

Lastly, while our research has yielded promising results, it also opens up a multitude of questions and avenues for future research. It invites us to delve deeper into the complex workings of the human brain, pushing the boundaries of our understanding and sparking new ideas and innovations.

In conclusion, this thesis underscores the power of deep learning models in decoding brain activity during speech-related tasks and offers invaluable insights into the cognitive and neural processes involved in speech perception and production. These findings not only pave the way for more advanced research but also lay a solid foundation for further explorations into the fascinating world of human cognition.

7. CLOSURE AND WAY FORWARD

7.1 Leveraging Deep Learning to Decode Brain Activity During Passive Listening

In the quest to decode the intricacies of speech perception and production, this study ventured into a relatively uncharted territory: the use of deep learning methods applied to intracranial electroencephalography (iEEG) data recorded during passive listening of speech. The goal was not only to advance the field of brain-computer interfaces (BCIs) for speech synthesis but also to gain insights into cognitive speech processing.

Deep learning, with its ability to model complex, non-linear relationships, has shown promise in decoding high-dimensional iEEG data. When applied to data recorded during passive listening, these models successfully identified patterns of neural activity associated with the perception of speech sounds. Even though the current stage of the research did not achieve accurate speech synthesis from the neural activity, the strong correlations found between the original and decoded signals suggest that this approach is viable and holds potential for future investigations.

The use of passive listening iEEG data introduced a novel angle to the existing methodologies. Unlike the traditional focus on speech production, this approach allows for the examination of neural representations of listened speech, which arguably could be more nuanced and intricate. The deep learning models, in this case, were trained to map the relationship between these complex neural patterns and the corresponding speech sounds, providing a powerful tool for decoding brain activity during speech perception.

In addition to this, the employment of deep learning algorithms enabled the processing of vast amounts of iEEG data in ways that were not previously possible. Deep learning models, with their sophisticated architectures, can capture higher-order interactions and dependencies within the data. This makes them particularly suited for tasks like speech decoding, where the input data can be multi-dimensional and highly complex.

The findings of this research provide valuable insights into cognitive speech processing. They demonstrate that the neural activity in the brain during passive listening of speech contains enough information, similar to during speaking, to differentiate between various speech sounds. This reinforces the relevance of speech perception in future speech synthesis research, suggesting that a comprehensive understanding of speech processing should encompass both production and perception aspects.

In conclusion, although the direct synthesis of speech from brain data recorded during passive listening remains a challenge for the future, the present study has shown that deep learning might help to decode brain activity during speech perception. This represents a significant advancement in the field of BCIs and provides a promising direction for further research in speech synthesis.

7.2 Towards Naturalistic BCI

As we stand on the threshold of the next era in brain-computer interfaces (BCIs), the aspiration to realize naturalistic BCIs for speech synthesis remains a driving force behind research across numerous domains. Progress made thus far, including the ability to decode neural activity during passive listening of speech via deep learning methods, has edged us closer to this goal. However, a significant gap still exists that needs to be bridged.

Achieving naturalistic BCIs for speech synthesis necessitates advancements in several areas. Foremost, the enhancement of neural recording resolution is essential. This implies not only the refinement of recording techniques but also the improvement in processing and analyzing the recorded data. Current methodologies, such as intracranial electroencephalography (iEEG), offer high temporal resolution but are invasive and hence, impractical for widespread use. Non-invasive techniques, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), come with their own set of limitations. Therefore, the inception of innovative and efficient neural recording techniques is a pressing need.

Next, the decoding algorithms require further refinement. Despite the beneficial application of deep learning, these models necessitate extensive datasets for training and substantial computational resources. There is a need for more research to engineer efficient algorithms capable of accurately decoding speech-related neural signals.

Moreover, the realization of naturalistic BCIs for speech synthesis might entail the incorporation of multi-modal data amassed through various means and different locations. This could include neural activity from different brain regions involved in speech perception and production, as well as the activity of muscles involved in articulation, such as those in the tongue and larynx. Supplementary data from tools like cameras or other sensors capturing facial movements and visual cues could also be invaluable. This integration of diverse data sources would present a holistic view of the speech process and could potentially lead to more accurate and naturalistic speech synthesis.

An essential component of the future of BCIs for speech synthesis also lies in advancements in our neurobiological and cognitive understanding of speech processes. Gaining deeper insights into

how the brain functions during speech perperception and production, understanding the intricate neural networks, the role of various brain areas, the cognitive processes involved, and how these elements harmoniously facilitate fluent speech, are all crucial. These advancements in our neuroscientific understanding could provide valuable insights that propel the development of more effective, accurate, and naturalistic BCIs.

This study reinforces the importance of exploring the role of perceived speech in the development of BCIs for speech synthesis. While the task of synthesizing audible speech from neural activity recorded during passive listening is formidable, the results obtained so far are promising. The statistical correlation between the original and decoded signals indicates that the information encapsulated in the brain's neural activity during speech perception holds substantial potential for enhancing our understanding of speech processing and consequently, improving the development of BCIs for speech synthesis.

In conclusion, the pathway towards naturalistic BCIs for speech synthesis is intricate, requiring advancements across multiple domains. Despite the length of the road and the multitude of challenges, each step forward brings us closer to the goal of creating BCIs that can genuinely mimic natural speech and are usable in natural settings. This opens a realm of possibilities for individuals with speech impairments and deepens our understanding of the human brain's remarkable capabilities.

Acknowledgements

The author would like to express his gratitude to the many researchers and institutions who have contributed to the field of brain-computer interfaces and related technology. The efforts of those individuals have paved the way for advancements in the understanding of human brain function and communication.

I would like to express my deepest gratitude and sincere appreciation to Frigyes Viktor Arthur, whose guidance, support, and profound knowledge greatly contributed to the completion of this work.

In addition, my sincere thanks go to Julia Berezutskaya, whose invaluable insights and unique dataset were instrumental to the realization of this project.

Bibliography

[1] Hassan Akbari, Yiyuan Gao, Mikhail Belkin, and Alejandro Ribeiro. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1):874, 2019.

[2] Antti Alastalo et al. Finnish end-to-end speech synthesis with tacotron 2 and wavenet. 2021.

[3] Jonathan Allen, M. Sharon Hunnicutt, and Dennis Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.

[4] Michael L Anderson. Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4):245–266, 2010.

[5] Kai Keng Ang and Cuntai Guan. Transcranial direct current stimulation and eeg-based motor imagery bci for upper limb stroke rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(3):534–541, 2012.

[6] Kai Keng Ang, Cuntai Guan, Karen Sui Geok Chua, Beng Ti Ang, Christopher Wee Keong Kuah, Chuanchu Wang, Kok Soon Phua, Zheng Yang Chin, and Haihong Zhang. A large clinical study on the ability of stroke patients to use an eeg-based motor imagery brain-computer interface. *Clinical EEG and Neuroscience*, 42(4):253–258, 2011.

[7] M. Angrick, C. Herff, E. Mugler, M.C. Tate, M.W. Slutzky, D.J. Krusienski, T. Schultz, and J.S. Brumberg. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of Neural Engineering*, 20(2):026018, 2023.

[8] Marco Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

[9] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

[10] Gopala K. Anumanchipalli, Josh Chartier, and Edward F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[11] Frigyes Viktor Arthur and Tamás Gábor Csapó. Towards a practical lip-to-speech conversion system using deep neural networks and mobile application frontend. *CoRR*, abs/2104.14467, 2021. URL https://arxiv.org/abs/2104.14467.

[12] Joan L.G. Baart and Vincent J. van Heuven. From text to speech; the mitalk system: Jonathan allen, m. sharon hunnicutt and dennis klatt (with robert c. armstrong and david pisoni): Cambridge university press, cambridge, 1987. xii+216 pp. £25.00. *Lingua*, 81(2):265–270, 1990. ISSN 0024-3841. doi: https://doi.org/10.1016/0024-3841(90)90014-C. URL https://www.sciencedirect.com/science/article/pii/002438419090014C.

[13] Nicholas A Badcock, Petroula Mousikou, Yatin Mahajan, Peter De Lissa, Johnson Thie, and Genevieve McArthur. Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps. *PeerJ*, 1:e38, 2013.

[14] William Roger Balfrey. *Delayed auditory feedback as a function of levels of anxiety*. PhD thesis, Oklahoma State University, 1965.

[15] Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. Voice-selective areas in human auditory cortex. *Nature*, 403(6767):309–312, 2000.

[16] Julia Berezutskaya, Luca Ambrogioni, Nick F Ramsey, and Marcel AJ van Gerven. Towards naturalistic speech decoding from intracranial brain data. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3100–3104. IEEE, 2022.

[17] Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9 (1), March 2022. doi: 10.1038/s41597-022-01173-0. URL https://doi.org/10.1038/s41597-022-01173-0.

[18] Hans Berger. Über das elektrenkephalogramm des menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, 87(1):527–570, 1929.

[19] Jeffrey R Binder, James A Frost, Thomas A Hammeke, Robin W Cox, Stephen M Rao, and Thomas Prieto. Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10(5):512–528, 2000.

[20] Niels Birbaumer. Brain-computer-interface research: coming of age. 2006.

[21] Niels Birbaumer, Nasser Ghanayim, Thilo Hinterberger, Ingrid Iversen, Boris Kotchoubey, Andrea K"ubler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[22] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[23] Niels Birbaumer, Thilo Hinterberger, Andrea Kubler, and Nicola Neumann. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[24] Niels Birbaumer, N Ghanayim, T Hinterberger, I Iversen, B Kotchoubey, A Kübler, J Perelmouter, E Taub, and H Flor. The thought translation device (ttd) for completely paralyzed patients. *IEEE transactions on rehabilitation engineering*, 8(2):190–193, 2000.

[25] Alan W Black and Paul Taylor. Statistical parametric speech synthesis. *Proceedings of the IEEE*, 88(8):1239–1254, 2000.

[26] Benjamin Blankertz, Laura Acqualagna, Sven Dähne, Stefan Haufe, Matthias Schultze-Kraft, Irene Sturm, Marija Uščumlic, Markus A Wenzel, Gabriel Curio, and Klaus-Robert Müller. The berlin brain-computer interface: progress beyond communication and control. *Frontiers in neuroscience*, 10:530, 2016.

[27] J D Breshears, J L Roland, M Sharma, C M Gaona, Z V Freudenburg, R Tempelhoff, M S Avidan, and E C Leuthardt. Neural decoding of cursor motion using a kalman filter. *Advances and applications in statistical sciences*, 4(2):141–155, 2011.

[28] A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

[29] Jonathan S Brumberg, Kevin M Pitt, and Jeremy D Burnison. Brain–computer interfaces for communication and control. *Communicative disorders review*, 3(2):1–15, 2011.

[30] J.S. Brumberg, A. Nieto-Castanon, P.R. Kennedy, and F.H. Guenther. Brain-computer interfaces for communication in paralysis: a clinical experimental approach. *Journal of Speech, Language, and Hearing Research*, 54(1):361–382, 2011.

[31] Nell B Cant and Christina G Benson. Parallel auditory pathways: projection patterns of the different neuronal populations in the dorsal and ventral cochlear nuclei. *Brain research bulletin*, 60(5-6):457–474, 2003.

[32] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20–23, 2016.

[33] H. Cecotti and N. Birbaumer. Brain-computer interface for communication and control. *Clinical Neurophysiology*, pages 106–119, 2011.

[34] Maria Chait and David Poeppel. Neural and behavioral correlates of auditory cognition: Speech and music perception. 1:265–290, 2018.

[35] Shreya Chakrabarti, Hilary M Sandberg, Jonathan S Brumberg, and Dean J Krusienski. Progress in speech decoding from the electrocorticogram. *Biomedical Engineering Letters*, 5:10–21, 2015.

[36] M Cheng, X Gao, S Gao, and D Xu. Design and implementation of a brain-computer interface with high transfer rates. In *IEEE Transactions on Biomedical Engineering*, volume 49, pages 1181–1186. IEEE, 2002.

[37] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor. The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10:604, 2016.

[38] Anders M Dale. Optimal experimental design for event-related fmri. *Human brain mapping*, 8(2-3):109–114, 1999.

[39] Janis J Daly and Jonathan R Wolpaw. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, 2008.

[40] Matthew H Davis and Ingrid S Johnsrude. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research*, 229(1-2):132–147, 2007.

[41] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[42] Nina Dethlefs, Maarten Milders, Heriberto Cuayáhuitl, Turkey Al-Salkini, and Lorraine Douglas. A natural language-based presentation of cognitive stimulation to people with dementia in assistive technology: A pilot study. *Informatics for Health and Social Care*, 42 (4):349–360, January 2017. doi: 10.1080/17538157.2016.1255627. URL `https://doi.org/10.1080/17538157.2016.1255627`.

[43] G di Pellegrino, L Fadiga, L Fogassi, V Gallese, and G Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental brain research*, 91(1):176–180, 1992.

[44] R. L. Diehl, A. J. Lotto, and L. L. Holt. Speech perception. *Annual Review of Psychology*, 55(1):149–179, 2004. doi: 10.1146/annurev.psych.55.090902.142028.

[45] John P Donoghue. Bridging the brain to the world: a perspective on neural interface systems. *Neuron*, 60(3):511–521, 2008.

[46] L Fadiga, L Craighero, G Buccino, and G Rizzolatti. Speech listening specifically modulates the excitability of tongue muscles: a tms study. *European Journal of Neuroscience*, 15(2): 399–402, 2002.

[47] Lawrence A Farwell and Emanuel 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.

[48] Mehrdad Fatourechi, Ali Bashashati, Rabab K Ward, and Gary E Birch. Emg and eog artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118(3): 480–494, 2007.

[49] Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological Reviews*, 91(4):1357–1392, 2011.

[50] Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996.

[51] Irina Goncharova, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. Emg and eog artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 114:480–494, 2003.

[52] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[53] Gábor Gosztolya, Ádám Pintér, László Tóth, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. Autoencoder-based articulatory-to-acoustic mapping for ultrasound silent speech interfaces. *CoRR*, abs/1904.05259, 2019. URL `http://arxiv.org/abs/1904.05259`.

[54] T. D. Griffiths and J. D. Warren. What is an auditory object? *Nature reviews. Neuroscience*, 5(11):887–892, 2004. doi: 10.1038/nrn1538.

[55] F. H. Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102(3):594–621, 1995.

[56] Frank H Guenther. Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, 39(5):350–365, 2006.

[57] Peter Hagoort. On broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, 9(9):416–423, 2005.

[58] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413, 1993.

[59] Zöe Handley. Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10):906–919, 2009. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2008.12.004. URL `https://www.sciencedirect.com/science/article/pii/S0167639308001842`. Spoken Language Technology for Education.

[60] Mark S Hawley, Pam Enderby, and Phil Green. Assistive technology for speech and language disorders. *Disability and Rehabilitation: Assistive Technology*, 2(1):17–28, 2007.

[61] Grit Hein and Robert T Knight. The superior temporal sulcus is crucial for social communication. *The Superior Temporal Sulcus is Crucial for Social Communication*, 5(1):721–727, 2008.

[62] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 10:217, 2016.

[63] G. Hickok and D. Poeppel. The functional neuroanatomy of language. *Physics of life reviews*, 4(3):255–266, 2007.

[64] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

[65] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2011.

[66] Thilo Hinterberger, Andrea Kübler, Jochen Kaiser, Nicola Neumann, and Niels Birbaumer. A brain–computer interface (bci) for the locked-in: comparison of different eeg classifications for the thought translation device. *Clinical neurophysiology*, 114(3):416–425, 2003.

[67] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[68] Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehs, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442 (7099):164–171, 2006.

[69] Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick van der Smagt, and John P Donoghue. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375, 2012.

[70] Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick van der Smagt, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485 (7398):372–375, 2012.

[71] C.R. Holdgraf, W. de Heer, B. Pasley, J. Rieger, N. Crone, J.J. Lin, R.T. Knight, and F.E. Theunissen. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, 7:13654, 2016.

[72] John F Houde and Michael I Jordan. Sensorimotor adaptation in speech production. *Science*, 279(5354):1213–1216, 1998.

[73] Andrew Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the Acoustics, Speech, and Signal Processing Conference*, pages 373–376, 1996.

[74] Tetsuya Iidaka, Norihiro Sadato, Hiroki Yamada, and Yoshiharu Yonekura. Functional asymmetry of human prefrontal cortex in verbal and non-verbal episodic memory as revealed by fmri. *Cognitive Brain Research*, 9(1):73–83, 2000.

[75] Peter Indefrey and Willem JM Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, 2004.

[76] Zafer İşcan and Vadim V Nikulin. Steady state visual evoked potential (ssvep) based brain-computer interface (bci) performance under different perturbations. *PloS one*, 13(1): e0191673, 2018.

[77] Vinay Jayaram, Mohammad Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Transfer learning with intelligence: A data-driven approach for enhancing bci performance. *Neural Computation*, 28(12):2499–2522, 2016.

[78] Jochen Kaiser and Werner Lutzenberger. Human gamma-band activity: a window to cognitive processing. *Neuroreport*, 16(3):207–211, 2005.

[79] Ivo Käthner, Selina C Wriessnegger, Gernot R Müller-Putz, Andrea Kübler, and Sebastian Halder. Effects of mental workload and fatigue on the p300, alpha and theta band power during operation of an erp (p300) brain–computer interface. *Biological psychology*, 102: 118–129, 2014.

[80] Shira Katseff, John Houde, and Keith Johnson. Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Language and speech*, 55(2):295–308, 2012.

[81] Spencer Kellis, Kip Miller, Kyle Thomson, Richard Brown, Paul House, and Bradley Greger. Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus. *Journal of Neural Engineering*, 17(3):036023, 2020.

[82] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744, July 2022. doi: 10.1007/s11042-022-13428-4. URL `https://doi.org/10.1007/s11042-022-13428-4`.

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

[84] Dennis H Klatt. Software for a cascade/parallel formant synthesizer. In *Journal of the Acoustical Society of America*, 1980.

[85] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. In *IEEE Transactions on Affective Computing*, number 1, pages 18–31. IEEE, 2012.

[86] Paul Konstantin Krug, Simon Stone, and Peter Birkholz. Intelligibility and naturalness of articulatory synthesis with vocaltractlab compared to established speech synthesis technologies. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 102–107, 2021.

[87] Andrea Kübler, Nicola Neumann, Barbara Wilhelm, Thilo Hinterberger, and Niels Birbaumer. Brain-computer communication: self-regulation of slow cortical potentials for verbal communication. *Archives of physical medicine and rehabilitation*, 82(11):1533–1539, 2001.

[88] Andrea Kübler, Femke Nijboer, Jürgen Mellinger, Theresa M Vaughan, Helmut Pawelzik, Gerwin Schalk, Dennis J McFarland, Niels Birbaumer, and Jonathan R Wolpaw. Patients with als can use sensorimotor rhythms to operate a brain-computer interface. *Neurology*, 64 (10):1775–1777, 2005.

[89] Andrea Kübler, Femke Nijboer, and Niels Birbaumer. Brain-computer interfaces for communication and motor control-perspectives on clinical applications. *Toward Brain-Computer Interfacing*, pages 373–391, 2007.

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

[91] Anatole Lecuyer, Fabien Lotte, Richard B Reilly, Robert Leeb, Michitaka Hirose, and Mel Slater. Brain-computer interfaces, virtual reality, and videogames. In *Computer*, pages 66–72. IEEE, 2008.

[92] Robert Leeb, Doron Friedman, Gernot R Müller-Putz, Reinhold Scherer, Mel Slater, and Gert Pfurtscheller. Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):473–482, 2007.

[93] Eric C Leuthardt, Charles Gaona, Mohit Sharma, Nicholas Szrama, Jarod Roland, Zac Freudenberg, Jamie Solis, Jonathan Breshears, and Gerwin Schalk. Using the electrocorticographic speech network to control a brain–computer interface in humans. *Journal of neural engineering*, 8(3):036004, 2011.

[94] W. J. M. Levelt, A. Roelofs, and A. S. Meyer. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38, 1999.

[95] Willem JM Levelt. *Speaking: From intention to articulation*. MIT press, 1993.

[96] Alvin M Liberman and Ignatius G Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

[97] Benjamin Libet, Benjamin Libet, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) the unconscious initiation of a freely voluntary act. *Neurophysiology of consciousness*, pages 249–268, 1993.

[98] Yijun Liu, Zongtan Zhou, and Dewen Hu. A frequency recognition method based on canonical correlation analysis for ssvep-based bcis. *IEEE Transactions on Biomedical Engineering*, 58(6):2009–2017, 2011.

[99] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fmri signal. *Nature*, 412(6843):150–157, 2001.

[100] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger. A review of classification algorithms for eeg-based brain-computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2015.

[101] Fabien Lotte and Cuntai Guan. A review of recent advances in feature selection for brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(5):512–528, 2011.

[102] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain-computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

[103] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, A Rakotomamonjy, and Florian Yger. A tutorial on eeg signal-processing techniques for mental-state recognition in brain–computer interfaces. *Guide to EEG signal processing for BCI*, pages 133–161, 2018.

[104] Scott Makeig, Christian Kothe, Tim Mullen, Nima Bigdely-Shamlo, Zeynep Zhang, and Kenneth Kreutz-Delgado. Evolving signal processing for brain–computer interfaces. *Proceedings of the IEEE*, 100:1567–1584, 2012.

[105] William Matchin and Gregory Hickok. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498, 2020.

[106] Dennis J McFarland and Jonathan R Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011.

[107] José del R Millán, Rüdiger Rupp, Gernot R Müller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kübler, Robert Leeb, et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, 4, 2010.

[108] G. Mirabella. Overcoming the limitations of the traditional paradigms of brain-computer interfaces. *Journal of Neuroscience Methods*, 325:108346, 2019.

[109] Brian CJ Moore. *An Introduction to the Psychology of Hearing*. Brill, 2012.

[110] Jack Mostow, Greg Aist, Paul Burkhead, Albert Corbett, Andrew Cuneo, Susan Eitelman, Cathy Huang, Brian Junker, Mary Beth Sklar, and Brian Tobin. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29(1):61–117, July 2003. doi: 10.2190/06ax-qw99-eq5g-rdcf. URL `https://doi.org/10.2190/06ax-qw99-eq5g-rdcf`.

[111] Firra M Mukhneri, Inung Wijayanto, and Sugondo Hadiyoso. Voice conversion for dubbing using linear predictive coding and hidden markov model. *Journal of Southwest Jiaotong University*, 55(4), 2020.

[112] Gernot R Müller-Putz, Reinhold Scherer, Christian Brauneis, and Gert Pfurtscheller. Towards noninvasive hybrid brain-computer interfaces: framework, practice, clinical application, and beyond. *Proceedings of the IEEE*, 103(6):926–943, 2015.

[113] Charles N Munyon. Neuroethics of non-primary brain computer interface: focus on potential military applications. *Frontiers in Neuroscience*, 12:696, 2018.

[114] Noman Naseer and Keum-Shik Hong. fnirs-based brain-computer interfaces: a review. *Frontiers in Human Neuroscience*, 9:3, 2015.

[115] C. Nass and S. Brave. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press, 2005.

[116] Christa Neuper and Gert Pfurtscheller. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International journal of psychophysiology*, 43(1):41–58, 2001.

[117] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation. *Methods in Molecular Biology*, 671:449–470, 2012.

[118] Ernst Niedermeyer and Fernando H Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[119] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050, 2019.

[120] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[121] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[122] Ramaswamy Palaniappan. Utilizing gamma band to improve mental task based brain-computer interface design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(3):299–303, 2006.

[123] Gert Pfurtscheller. Event-related synchronization (ers): an electrophysiological correlate of cortical areas at rest. *Electroencephalography and clinical neurophysiology*, 83(1):62–69, 1992.

[124] Gert Pfurtscheller and Agustin Aranibar. On the existence of different types of central beta rhythms below 30 hz. *Electroencephalography and clinical Neurophysiology*, 86(3):218–226, 1993.

[125] Gert Pfurtscheller and Christa Neuper. Functional brain imaging based on erd/ers. *Vision research*, 41(10-11):1257–1260, 2001.

[126] Gert Pfurtscheller, Brendan Z Allison, Clemens Brunner, Gunther Bauernfeind, Teodoro Solis-Escalante, Reinhold Scherer, Thorsten O Zander, Gernot Mueller-Putz, Christa Neuper, and Niels Birbaumer. The hybrid bci. *Frontiers in Neuroscience*, 4:42, 2010.

[127] Colin Phillips, Alec Marantz, Martha McGinnis, David Pesetsky, K Wexler, A Yellin, David Poeppel, T Roberts, and H Rowley. Brain mechanisms of speech perception: A preliminary report. *MIT Working Papers in Linguistics*, 26:125–163, 1995.

[128] Terence W Picton, S Bentin, P Berg, Emanuel Donchin, SA Hillyard, R Johnson Jr, GA Miller, W Ritter, DS Ruchkin, MD Rugg, et al. Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria, 2000.

[129] John Polich. Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10):2128–2148, 2007.

[130] Alexander T Pope, Edward H Bogart, and David S Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(1-2): 187–195, 1995.

[131] Janna Protzak, Klas Ihme, and Thorsten Oliver Zander. A passive brain-computer interface for supporting gaze-based human-machine interaction. In Constantine Stephanidis and Margherita Antona, editors, *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, pages 662–671, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39188-0.

[132] Ander Ramos-Murguialday, Doris Broetz, Massimiliano Rea, Leonhard L"aer, Ozge Yilmaz, Fabricio Lima Brasil, Giulia Liberati, Marco Rocha Curado, Eliana Garcia-Cossio, Alkis Vyziotis, et al. Brain–machine interface in chronic stroke rehabilitation: A controlled study. *Annals of neurology*, 74(1):100–108, 2013.

[133] Antoine Raux and Maxine Eskenazi. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 1–10, 2008.

[134] Brice Rebsamen, Cuntai Guan, Haihong Zhang, Chuanchu Wang, Cheeleong Teo, Marcelo H Ang, and Etienne Burdet. A brain controlled wheelchair to navigate in familiar environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18 (6):590–598, 2010.

[135] 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.

[136] Giacomo Rizzolatti and Corrado Sinigaglia. Mirrors in the brain: How our minds share actions and emotions. 2008.

[137] Andrew Rosenberg, Raul Fernandez, and Bhuvana Ramabhadran. Measuring the effect of linguistic resources on prosody modeling for speech synthesis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5114–5118. IEEE, 2018.

[138] Simone Rossi, Mark Hallett, Paolo M Rossini, Alvaro Pascual-Leone, Safety of TMS Consensus Group, et al. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical neurophysiology*, 120(12):2008–2039, 2009.

[139] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[140] J. Ruohonen and J. Karhu. Transcranial magnetic stimulation: language function. *Journal of Neurolinguistics*, 23(3):355–371, 2010.

[141] A. G. Samuel. Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, 110(4):474–494, 1981. doi: 10.1037/0096-3445.110.4.474.

[142] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[143] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, 2015.

[144] Andrew B Schwartz. Cortical neural prosthetics. *Annual Review of Neuroscience*, 27:487–507, 2006.

[145] Sophie K Scott. Perception and production of speech: Connected, but how? In *Speech Perception and Spoken Word Recognition*, pages 33–46. Psychology Press, 2016.

[146] Eric W Sellers and Emanuel Donchin. A brain-computer interface for long-term independent home use. *Amyotrophic Lateral Sclerosis*, 9(5):279–293, 2008.

[147] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *arXiv preprint arXiv:1712.05884*, 2018.

[148] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *ArXiv*, abs/1712.05884, 2018.

[149] Hakim Si-Mohammed, Ferran Argelaguet Sanz, Géry Casiez, Nicolas Roussel, and Anatole Lécuyer. Brain-computer interfaces and augmented reality: A state of the art. In *Graz Brain-Computer Interface Conference*, 2017.

[150] John F Smiley, Troy A Hackett, Todd M Preuss, Cynthia Bleiwas, Khadija Figarsky, J John Mann, Gorazd Rosoklija, Daniel C Javitt, and Andrew J Dwork. Hemispheric asymmetry of primary auditory cortex and heschl's gyrus in schizophrenia and nonpsychiatric brains. *Psychiatry Research: Neuroimaging*, 214(3):435–443, 2013.

[151] Bettina Sorger, Joel Reithler, Brigitte Dahmen, and Rainer Goebel. A real-time fmri-based spelling device immediately enabling robust motor-independent communication. *Current Biology*, 22(14):1333–1338, 2012.

[152] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[153] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112, 2014.

[154] Samuel Sutton, Max Braren, Joseph Zubin, and E. Roy John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188, 1965.

[155] Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[156] Jose A Uriguen and Begonya Garcia-Zapirain. Eeg artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering*, 12(3):031001, 2015.

[157] Mariska J Vansteensel and Beata Jarosiewicz. Brain-computer interfaces for communication. *Handbook of clinical neurology*, 168:67–85, 2020.

[158] Theresa M Vaughan, William J Heetderks, Leonard J Trejo, William Z Rymer, Michael Weinrich, Melody M Moore, Andrea Kübler, Bruce H Dobkin, Niels Birbaumer, Emanuel Donchin, et al. Brain-computer interface technology: a review of the second international meeting. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 11(2):94–109, 2003.

[159] Francois-Benoit Vialatte, Mathilde Maurice, Justin Dauwels, and Andrzej Cichocki. Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. *Progress in neurobiology*, 90(4):418–438, 2010.

[160] Jacques J Vidal. Toward direct brain-computer communication. In *Annual review of biophysics and bioengineering*, volume 2, pages 157–180. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 1973.

[161] William Grey Walter. Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. *Nature*, 203(4943):380–384, 1964.

[162] S. Wang, M. Parsons, J. Stone-McLean, P. Rogers, S. Boyd, K. Hoover, O. Meruvia-Pastor, M. Gong, and A. Smith. A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery. *Expert review of medical devices*, 13(12):1063–1072, 2016.

[163] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415, 2019.

[164] Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda. A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical

parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:157–170, 2019.

[165] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V Le, Yannis Agiomyrgiannakis, Rob Clark, et al. Tacotron: Towards end-to-end speech synthesis. *ArXiv*, abs/1703.10135, 2017.

[166] Stephen M Wilson, Ayse Pinar Saygin, Martin I Sereno, and Marco Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, 7(7): 701–702, 2004.

[167] Jonathan R Wolpaw and Dennis J McFarland. An eeg-based brain-computer interface for cursor control. *Electroencephalography and clinical Neurophysiology*, 78(3):252–259, 1991.

[168] Yi-Chiao Wu, Tomoki Hayashi, Patrick Lumban Tobing, Kazuhiro Kobayashi, and Tomoki Toda. Quasi-periodic wavenet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1134–1148, 2021.

[169] Minda Yang, Sameer A Sheth, Catherine A Schevon, Guy M Mckhann Ii, and Nima Mesgarani. Speech reconstruction from human auditory cortex with deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[170] Xuntian Yin, Bo Xu, Changyang Jiang, Yong Fu, Zhiguo Wang, Hongyi Li, and Guangming Shi. A hybrid bci based on eeg and fnirs signals improves the performance of decoding motor imagery of both force and speed of hand clenching. *Journal of Neural Engineering*, 10(2): 026–014, 2013.

[171] Han Yuan and Bin He. Brain-computer interfaces: a review. *IEEE transactions on neural systems and rehabilitation engineering*, 27(9):1721–1731, 2019.

[172] Rafael Yuste, Sara Goering, Gary Bi, Jose M Carmena, Adrian Carter, Joseph J Fins, Phoebe Friesen, Jack Gallant, Jane E Huggins, Judy Illes, Philipp Kellmeyer, Eran Klein, Adam Marblestone, Christine Mitchell, Erik Parens, M Pham, Alan Rubel, Norihiro Sadato, Laura Specker Sullivan, Meredith Teicher, David Wasserman, Anna Wexler, Meredith Whittaker, and Jonathan Wolpaw. Four ethical priorities for neurotechnologies and ai. *Nature News*, 551(7679):159, 2017.

[173] Thorsten O Zander and Sabine Jatzev. A passive brain-computer interface for supporting gaze-based human-computer interaction. *International Journal of Human-Computer Interaction*, 27(1):69–84, 2011.

[174] Robert J Zatorre and Pascal Belin. Spectral and temporal processing in human auditory cortex. *Cerebral cortex*, 11(10):946–953, 2001.

[175] Robert J Zatorre, Pascal Belin, and Virginia B Penhune. Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, 6(1):37–46, 2002.

[176] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2, 2023.

[177] Xiao Zhou, Zhen-Hua Ling, and Li-Rong Dai. Unitnet: A sequence-to-sequence acoustic model for concatenative speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2643–2655, 2021.

Appendix

Listing A.0.1: FC-DNN

```
min_len = np.min((len(eeg), len(mel_data)))
eeg = eeg[0:min_len]
mel_data = mel_data[0:min_len]

train_index = np.arange(0, int(0.8 * eeg.shape[0]))
test_index = np.arange(int(0.8 * eeg.shape[0]), eeg.shape[0])

# train-validation-test split
eeg_train = eeg[0 : int(len(eeg) * 0.8)]
eeg_valid = eeg[int(len(eeg) * 0.8) : int(len(eeg) * 0.9)]
eeg_test =  eeg[int(len(eeg) * 0.9) : ]

melspec_train = mel_data[0 : int(len(mel_data) * 0.8)]
melspec_valid = mel_data[int(len(mel_data) * 0.8) : int(len(
   mel_data) * 0.9)]
melspec_test =  mel_data[int(len(mel_data) * 0.9) : ]

# scale input to [0-1]
eeg_scaler = MinMaxScaler()
# eeg_scaler = StandardScaler(with_mean=True, with_std=True)
eeg_train_scaled = eeg_scaler.fit_transform(eeg_train)
eeg_valid_scaled = eeg_scaler.transform(eeg_valid)
eeg_test_scaled  = eeg_scaler.transform(eeg_test)

# scale outpit mel-spectrogram data to zero mean, unit variances
melspec_scaler = StandardScaler(with_mean=True, with_std=True)
melspec_train_scaled = melspec_scaler.fit_transform(
   melspec_train)
```

```
melspec_valid_scaled = melspec_scaler.transform(melspec_valid)
melspec_test_scaled  = melspec_scaler.transform(melspec_test)


model = Sequential()
model.add(
    Dense(
        3000,
         input_dim=eeg_train_scaled.shape[1],
         kernel_initializer='normal',
         activation='relu'))

model.add(
    Dense(
         80,
         kernel_initializer='normal',
         activation='linear'))
```

Listing A.0.2: 2D-CNN

```
method = '2D-CNN'
result_path = os.path.join(os.getcwd(), f"results_{method}")
winLength = 0.05
frameshift = 0.01
audiosr = 16000

spectrogram = mel_data
data = eeg
pt=subject

    # Create a train and test split from data, test is 20% of
       the data
    train_index = np.arange(0, int(0.8 * data.shape[0]))
    test_index = np.arange(int(0.8 * data.shape[0]), data.shape
       [0])
```

```
# Initialize an empty spectrogram to save the reconstruction
    to
rec_spec = np.zeros(spectrogram.shape)

# Z-Normalize with mean and std from the training data
mu = np.mean(data[train_index, :], axis=0)
std = np.std(data[train_index, :], axis=0)
trainData = (data[train_index, :] - mu) / std
testData = (data[test_index, :] - mu) / std

# Z-Normalize with mean and std from the training data --
    output
mu = np.mean(spectrogram[train_index, :], axis=0)
std = np.std(spectrogram[train_index, :], axis=0)
trainSpectrogram = (spectrogram[train_index, :] - mu) / std
testSpectrogram = (spectrogram[test_index, :] - mu) / std

print('Input␣shape:␣', trainData.shape)
print('Input␣shape:␣', testData.shape)

# Find the right shape for the input, as it should be 3D,
    like 1143 is 9*127
new_shape = int(trainData.shape[1] / 9)

# reshape input from 1143 to 9*127
trainData = trainData.reshape(-1, 9, new_shape)
testData = testData.reshape(-1, 9, new_shape)
print('Input␣shape:␣', trainData.shape)

sts = 6
window_size = sts * 4 + 1
n_to_skip = np.floor(window_size // 2).astype(np.int64)
```

```
print('Input␣shape:␣', trainData.shape)

#conversion to 3D blocks
trainData = strided_app(trainData, window_size, 1)
trainSpectrogram = trainSpectrogram[n_to_skip:(
    trainSpectrogram.shape[0] - n_to_skip)]

testData = strided_app(testData, window_size, 1)
testSpectrogram = testSpectrogram[n_to_skip:(testSpectrogram
    .shape[0] - n_to_skip)]

print('Input␣shape:␣', trainData.shape)
print('Input/validation␣shape:␣', testData.shape)
print('Output␣shape:␣', trainSpectrogram.shape)

model = Sequential()
model.add(InputLayer(input_shape=trainData.shape[1:]))
model.add(Conv2D(filters=40,
                 kernel_size=(13, 13),
                 strides=(sts, 2),
                 activation=tensorflow.nn.swish,
                 padding="same",
                 kernel_initializer=keras.initializers.
                     he_uniform(seed=None),
                 kernel_regularizer=regularizers.l1(0.00001)
                     ,
                 input_shape=trainData.shape[1:]))

model.add(Dropout(0.1))
model.add(Conv2D(filters=400, kernel_size=(13, 13), strides
    =(2, 2), activation=tensorflow.nn.swish,
                 padding="same", kernel_initializer=keras.
                     initializers.he_uniform(seed=None),
                 kernel_regularizer=regularizers.l1(0.00001)
```

```
                        ))
model.add(Dropout(0.1))
model.add(Flatten())
model.add(
    Dense(1000, activation=tensorflow.nn.swish,
        kernel_initializer=keras.initializers.he_uniform(seed=
        None),
            bias_initializer=keras.initializers.he_uniform(
                seed=None),
            kernel_regularizer=regularizers.l1(0.000005)))
model.add(Dropout(0.1))
model.add(Dense(trainSpectrogram.shape[1], activation='
    linear'))

plot_model(model, to_file=f"model_{method}.png", show_shapes
    =True, show_layer_names=True)

model.compile(
        loss='mean_squared_error',
        metrics=['mean_squared_error'],
        optimizer='adam')
earlystopper = EarlyStopping(monitor='val_mean_squared_error
    ', min_delta=0.001, patience=3, verbose=1,
                                 mode='auto')
lrr = ReduceLROnPlateau(monitor='val_mean_squared_error',
    patience=2, verbose=1, factor=0.5, min_lr=0.0001)

print(model.summary())

if not (os.path.isdir('models/')):
    os.mkdir('models/')

# early stopping to avoid over-training
model_name = 'models/iEEG_to_melspec_2D-CNN_sp-' + pt
```

```
# csapot: temporarily disabled
checkp = ModelCheckpoint(
    model_name +
    '_weights_best.h5',
    monitor='val_loss',
    verbose=1,
    save_best_only=True,
    mode='min')

# Run training
history = model.fit(trainData, trainSpectrogram,
                    epochs=100, batch_size=64, shuffle=False
                        , verbose=1,
                    callbacks=[earlystopper, checkp, lrr],
                    validation_data=(testData,
                        testSpectrogram),
                    )



# load back best weights
model.load_weights(model_name + '_weights_best.h5')

rec_spec = model.predict(testData)

# inverse transform
# testSpectrogram=(spectrogram[test,:]-mu)/std
rec_spec = rec_spec * std + mu

print('start saving wav')

# Save reconstructed spectrogram
os.makedirs(os.path.join(result_path), exist_ok=True)
np.save(os.path.join(result_path, f'{pt}_predicted_spec.npy'
```

```
    ), rec_spec)


# remove model file
os.remove(model_name + '_weights_best.h5')
del model
# Run garbage collection
gc.collect()
```