Master's thesis

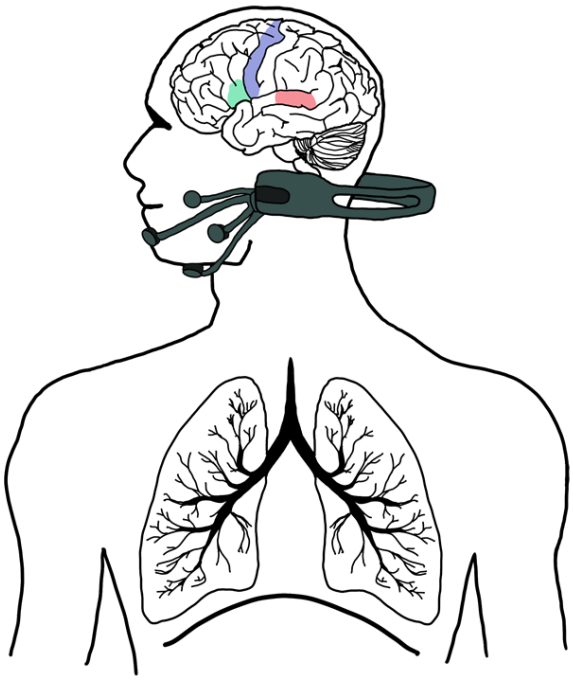
NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

Mathias Gullikstad Backsæther

Silent Speech Communication Using Facial Electromyography

The development of two silent speech interfaces using EMG-to-text, EMG-to-speech, and an Emotiv Epoc+ sensor

Master's thesis in MTNANO
Supervisor: Prof. Giampiero Salvi
June 2021



NTNU
Kunnskap for en bedre verden

Mathias Gullikstad Backsæther

Silent Speech Communication Using Facial Electromyography

The development of two silent speech interfaces using EMG-to-text, EMG-to-speech, and an Emotiv Epoc+ sensor

Master's thesis in MTNANO
Supervisor: Prof. Giampiero Salvi
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

NTNU

Kunnskap for en bedre verden

"[L]ife is like an extremely difficult, horribly unbalanced videogame. When you're born, you're given a randomly generated character, with a randomly determined name, race, face, and social class. Sometimes the game might seem easy. Even fun. Other times it might be so difficult you want to give up and quit. But unfortunately, in this game you only get one life. Some people play the game for a hundred years without ever figuring out that it's a game, or that there is a way to win it. To win the videogame of life you just have to try to make the experience of being forced to play it as pleasant as possible, for yourself, and for all of the other players you encounter in your travels."

- Ernest Cline, Ready Player Two [5]

ABSTRACT

Speech is of immense importance to human society and is the natural enabler for cooperation between humans. Unfortunately, there are situations where vocalized speech is not an option. Interest in the possibility of silent speech devices has continued to increase with the technological revolution of the last couple of decades. One possible modality for a silent speech interface is facial electromyography (EMG): electrical signals generated from muscle activation when moving the articulators without any vocalization. The aim of this project is to contribute to this field of research by showing that a standardized headset originally meant for recording brain waves can be used for EMG-based silent speech recognition. The Emotiv Epoc+ EEG headset was used to collect five corpora. Two of the corpora included time-synced audio recordings. Five different neural network architectures, as well as a Hidden Markov Model (HMM) classifier, were used for single word classification. An average recognition rate of 93.3% over four speakers was achieved on a vocabulary of three words using a recurrent neural network (RNN). The remaining corpora were collected by one speaker with session-independent word recognition of 85.4% accuracy on a vocabulary of 10 words. A convolutional neural network (CNN) was used for this, and the same architecture resulted in 63.2% word accuracy on the joint vocabulary of the NATO phonetic alphabet and digits. Two functional silent speech interfaces were furthermore created. One was based on EMG-to-text spelling out sentences. This system correctly classified an average of 82.7% of the characters in six test sentences. The other system utilized EMG-to-speech and was able to synthesize digits with the voice of the author. 20 synthesized digits were correctly classified 73.5% of the time by human listeners. This thesis shows that the Emotiv Epoc+ sensor can indeed be used for an EMG-based silent speech interface, and this sensor is proposed as a standardized platform for future silent speech research.

SAMMENDRAG

Språk er uvurderlig for mennesket som art, og tale som kommunikasjonsmiddel muliggjør samarbeid mellom mennesker hver dag. Allikevel finnes det ulike situasjoner der vokalisert tale ikke er et alternativ. Interessen for et fungerende system som muliggjør lydløs tale har økt de siste årene, i takt med teknologiske nyvinninger innen elektronikk og programmering. En mulig modalitet for slik lydløs tale kan være signaler fra muskelbevegelser i ansiktet, såkalt elektromyografi (EMG). Disse muskelbevegelsene er tett tilknyttet produksjonen av tale og kan dermed oversettes til lydbølger eller tekst ved hjelp av maskinlæring. Målet med denne masteroppgaven er å vise at et standardisert apparat for måling av hjernebølger, en Emotiv Epoc+ EEG-sensor, kan brukes til EMG-basert lydløs tale. Emotiv-sensoren ble brukt til å samle 5 datasett som i et par tilfeller også inkluderte mikrofonopptak. Fem ulike former for nevrale nettverk ble brukt, i tillegg til en Skjult Markov Modell (HMM), til å klassifiser enkeltord. En gjennomsnittlig nøyaktighet på 93.3% over 4 ulike talere på et 3-ords vokabulær ble oppnådd ved å bruke et nevralt nettverk med tilbakekoblinger (RNN). De resterende datasettene ble samlet inn av forfatteren selv, og 85.4% nøyaktighet i ord-gjenkjenning ble oppnådd på 10 ulike ord ved hjelp av et konvolusjonalt nevralt nettverk (CNN). Tilsvarende resultat var 63.2% på 39 ulike ord. Videre ble to ulike systemer for lydløs tale utviklet. Den ene baserte seg på å stave hvert tegn, noe som førte til 82.7% korrekt plassering av tegnene i seks testsetninger. Det andre systemet ble utviklet ved hjelp av EMG-til-tale og genererte lyd tilsvarende tallene 0 til 9 med forfatteren sin egen stemme. Fra 20 slike genererte lyder ble 73.5% av dem korrekt gjenkjent av personer som lyttet til lydklippene. Disse resultatene viser for aller første gang at en Emotiv Epoc+ sensor kan bli brukt til lydløs EMG-basert talegjenkjenning, og denne sensoren blir foreslått som en standardisert løsning for framtidig forskning på EMG-basert lydløs tale.

PREFACE

This master thesis was written as a part of a Master of Technology degree in nanotechnology at the Norwegian University of Science and Technology (NTNU). The work presented was carried out during the autumn, winter, and spring of 2020/2021 under the supervision of Prof. Giampiero Salvi at the Department of Electronic Systems. Everything presented in this master thesis, unless specified otherwise, is original and unpublished work conducted by the author, including all results, figures, and graphs. Note that parts of this report were previously presented in the project thesis written for the course TFE4570 and are reused here with the approval from the project supervisor Prof. Giampiero Salvi. This is most true for the Introduction and Theory chapters, as well as parts related to corpora 1 and 2.

I want to thank my supervisor for taking this project on, even though it was not listed as a choosable topic and included many potential pitfalls. With his advice, ideas, and thorough knowledge of speech recognition, it was possible to take this project further than what had been possible otherwise. I would further like to extend my gratitude to three of my fellow students: Isak, Nicolai, and Simen, for their input and discussions throughout the project and the inspiration they brought towards starting this thesis. Lastly, I want to thank Susanne Rosvoll, who with invaluable moral support helped me get through this project and all of the motivation it brought knowing that completing the thesis would give me more time to spend with her.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ARCHITECTURES

ACRONYMS

ADC	analog-to-digital converter
AI	artificial intelligence
API	application programming interface
ASR	automatic speech recognition
CNN	convolutional neural network
DFT	Discrete Fourier Transform
DNN	deep neural network
EEG	electroencephalography
EMG	electromyography
GMM	Gaussian mixture model
GRU	gated recurrent unit
GUI	graphical user interface
HMM	hidden Markov model
LSTM	long short-term memory
MFCC	Mel-frequency cepstral coefficient
MRI	magnetic resonance imaging
NAM	non-audible murmur
NN	neural network
NTNU	Norwegian University of Science and Technology
PLP	perceptual linear prediction
RNN	recurrent neural network
SPS	samples per second
SSI	silent speech interface
STFT	Short-Time Fourier Transform
TTS	text-to-speech
WER	word error rate

1 INTRODUCTION

As far as we know, humans are the only species to have developed communication at such an advanced level. Speech enables not only the sharing of thoughts and intentions but also a way of sharing knowledge. As a result of immense efforts by researchers in the area of automatic speech recognition (ASR), speech is also increasingly being used as a mode of interaction with consumer technologies. However, there are some scenarios where communication by speech is unpractical. Loud background noise or reluctance to disturb nearby listeners are examples where other means of communication could be more effective. Additionally, communication by speech is often impossible for those with significant speech impediments. This motivates the need for a functional silent speech interface (SSI), a device that can record a speakers silent intention and translate it either to text or a speech waveform.

When producing silent speech, one still uses the spoken language rather than having to learn a new form of communication. Different kinds of SSIs have been imagined in both science fiction and research projects of various forms in the last 60 years. The way the artificial intelligence (AI) HAL9000 in *2001 - A Space Odyssey* [43] uses lip-reading of a camera feed to understand that the two astronauts on board are conspiring against it is one example. Another is the DARPA Advanced Speech Encoding program from the early 2000s that aimed to enable silent and noise-prone communication for the American military forces.

There are different expressions that convey more or less the same meaning as *silent speech*. These include covert, subvocal, sub-auditory, non-acoustic, sub-acoustic, imagined, and inner speech, as well as the term subvocalization. Note that some sources use these expressions as the process of *thinking out loud inside your head*, while others include everything from thinking to mouthing words without vocalizing. In this thesis, the term silent speech is used as *the process of moving one's mouth as if speaking but without making any noticeable sound*.

Scientific research on silent speech can be traced all the way back to the 1950s, but it was not until the 1960s that electromyography (EMG) electrodes were used to record activation of facial muscles[20]. At that time, it was done mostly to research the role of unconscious facial micro-movements in memory and problem-solving. Some scientists regarded this kind og silent speech as the principal mechanism of thought [63], others described it as essential for establishing and maintaining speech code representations in short-term memory. EMG was during those years limited as a research tool by the electronic equipment and computational power available, exemplified by this excerpt of the data analysis part of the 1977 Garrity [20] paper *Electromyography: A review of the current status of subvocal speech research*:

> "Two principal techniques have been used in studies of subvocal speech to date to analyze EMG data: measurement of the amplitude of the single highest (or several highest) polygraph pen deflection(s) per trial segment (e.g. stimulus presentation or delay periods), and analog computer routines for squaring and integrating voltage values over trial segments."

With only polygraph pen EMG recordings and analog data processing available, the notion of detecting and classifying silent speech probably seemed unattainable at the time. However, by 1985 two Japanese scientists had made the first EMG-based silent speech system using three sensors and recognizing 5 vowels with 71% accuracy [61]. In 2003, researchers from the NASA Ames Research Center published results showing they were able to classify silent speech. Their vocabulary consisted of 6 words related to controlling a Mars rover using one pair of EMG sensors, and they achieved 92% accuracy using hidden Markov models (HMMs) and simple neural networks (NNs) [34]. The current state-of-the-art in EMG-based silent speech recognition is the solution by a group of researchers from Massachusetts, USA, where Meltzner et al. [50] achieved a 91.1% recognition rate on a 2200-word data set. Unfortunately, there is currently no EMG equipment available that is precise, mobile, and available at a low cost. This project adopts an electroencephalography (EEG) headset originally designed for reading brainwaves, the Emotiv Epoc+ [13], to work as a facial EMG sensor.

1.2 OBJECTIVES

The aim of this master thesis is to show that an existing EEG hardware solution more readily available than medical-grade EMG sensors can be used to detect facial muscle signals and that those signals can be used for silent speech communication. There are multiple potential use cases for an SSI, but none are yet available outside the scope of a few research groups. This thesis gives insight into the opportunities of an EMG-based silent speech solution and the current achieved results, with the following main objectives:

(1) Analyze whether the chosen sensor can be used to recognize silently spoken single words from a small vocabulary.

(2) Enable a method for efficient collection of both EMG and audio datasets while using the Emotiv Epoc+.

(3) Discover types of classification and feature extraction methods that work well with the available data.

(4) Work on understanding the challenges connected to session independence and the potential for a direct EMG-to-speech solution.

1.3 LITERATURE STUDY

The literature study for this thesis was based primarily on three text books (Freitas et al. [16], Huang et al. [27], and Yu and Deng [69]) describing spoken language processing, automatic speech recognition, and silent speech interfaces, respectively. Additionally, articles and theses from two scientific milieus were of the utmost importance. One in Massachusetts, USA (Kapur [37], Kapur, Kapur, and Maes [38], Kapur et al. [39], Meltzner et al. [50, 51], and Wadkins [64]) and the other in Germany (Denby et al. [9], Denby et al. [10], Diener and Schultz [11], Janke and Diener [32], Maier-Hein et al. [48], and Wand [65]).

The aforementioned resources provide a sound basis for a review of the current state of EMG-based silent speech. Additionally, they are essential for the discussions regarding the validity and significance of this project.

1.4 THESIS STRUCTURE

The structure of this thesis mostly follows the standard IMRaD (Introduction, Methods, Results, and Discussion) format, with the addition of Theory and Conclusion chapters. In the Theory chapter, the reader is presented with the most critical topics of this work. Note that a decision was made to go wide rather than deep as many different topics

are included. The Methods chapter primarily describes the topics that are common for all the work conducted, as well as a concise rundown of each of the five corpora. All results are sorted based on the corresponding corpus as well as the two functional SSIs, for the most part following the same timeline as each experiment was conducted. Then the Discussion chapter summarizes and compares the results based on speaker- and session-dependence with previous studies. It further goes into more in-depth analyses of sensor placement, electrode subsets, and signal artifacts. Finally, the Conclusion chapter rounds off the report and looks ahead with points on how this work can be continued in the future.

1.5 A NOTE ON REPRODUCIBILITY

Be aware that implicit bias may unintentionally be present in the used methods, collected training data, or how the results are interpreted. Four out of the five corpora were collected on only one speaker, the author of this thesis. This avoided the challenges of speaker dependence and was the most practical solution given the objective of social distancing as a result of the Covid19-pandemic. However, note that Corpus 1 was collected with four different speakers as a preliminary effort to create a venture based on the concept of EMG-based silent speech. All of these speakers were male of approximately the same age and with similar backgrounds. As inherent bias in training data is an important topic, a much more diverse group of subjects should be included if more experiments on speaker dependence are conducted. The sensor used for this work was bought by a startup founded by the four subjects used in this study, with money awarded from *Trønderenergibidraget*, a fund for early startup ideas at Norwegian University of Science and Technology (NTNU). This thesis will refer to source code multiple times, used both to collect data from the Emotiv sensor and for processing, visualizing, and training algorithms using this data. This code and the five corpora collected are not available online but can be requested from the author at mathias.backsaether@gmail.com.

2 THEORETICAL BACKGROUND

This chapter covers the theory relevant to describing the experiments that have been conducted and how to interpret the results. Speech production is first explained briefly, then speech recognition, both by classical probabilistic approaches and modern machine learning. A deep dive is done into different ways silent speech might become obtainable follows before the topic of electromyography is explained.

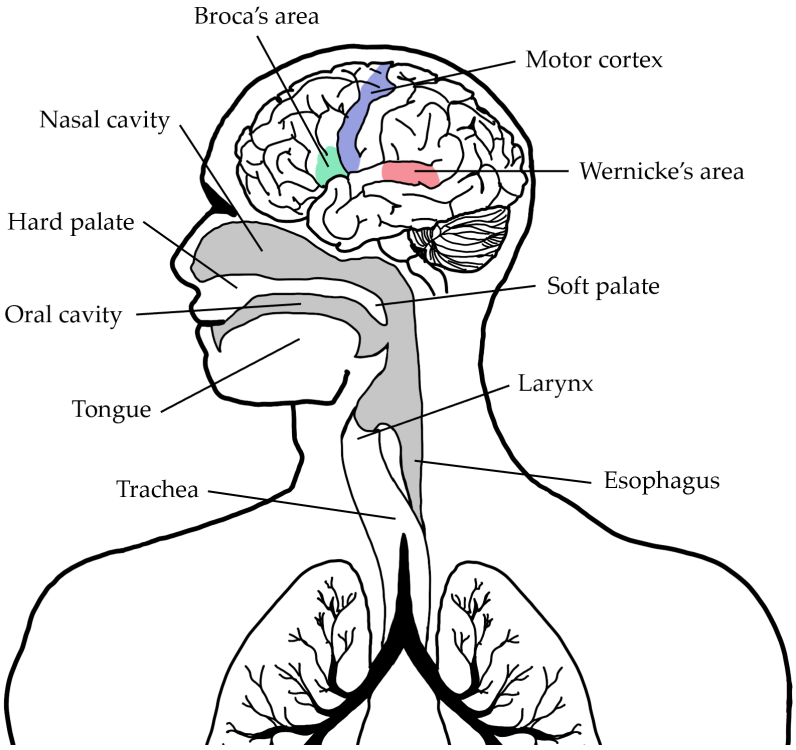


Figure 1.: A visualization of the most essential parts of the speech production pipeline.

2.1 SPEECH PRODUCTION

Speech production is a very complex process. It starts in the brain, specifically with language understanding in Wernicke's area. Signals travel from Wernicke's area to Broca's area, which is essential to speech formulation and articulation. Broca's area is located close to the motor cortex, the part of the brain that controls movement. Action potentials activated in the motor cortex travel through the cranial nerves to the peripheral nervous system, where efferent nerve

cells activate specific muscle fibers in a highly coordinated manner to produce the movements needed for speech [36, 65]. This process is visualized in Figure 1, which will be the guideline for the coming chapter.

Sound is longitudinal pressure waves moving through air. The different human speech sounds are therefore a result of the amplitude and frequency of these pressure waves. To be able to control these, human speech production includes both phonation and articulation. Phonation is the production of sound by moving air from the lungs through the larynx in a periodic manner, while articulation is the fine-tuned movement of the articulators to move the air pressure from the lungs through the vocal tract. The articulators include the tongue, lips, jaw, and soft and hard palate. For every word one utters, all of the needed muscles have to be coordinated and precisely controlled [27].

2.2 SPEECH RECOGNITION

For a machine to detect and recognize speech, an interface will have to record and classify signals somewhere along the path from thought to vocalized speech. For ASR systems based on sound, this requires a microphone that records the pressure waves in the air. The sampling rate for speech is often set to 16 kHz, slightly higher than double that of the highest relevant frequencies. The analog input is then transformed into a digital signal that can be analyzed by looking at the frequencies of the pressure waves. One way to analyze this digital, and therefore discrete, signal is to run it through a Discrete Fourier Transform (DFT). The resulting spectrum will give information about the number of frequencies present in the signal, up to half that of the sampling rate. If we are looking at an audio-recording one second long, sampled at 16 kHz, a DFT will only give information about frequencies present up to 8 kHz, and for the one-second recording as a whole. Because some of the most important aspects of speech are based on the order of speech sounds present for only short windows in time, a DFT on the whole recording rarely gives the needed information for speech recognition. Short-Time Fourier Transforms (STFTs) are therefore used to make spectra of smaller windows in time, usually 25 milliseconds long. Using information from these spectra directly or looking at the complete information from a series of windows composing a spectrogram, it is possible to extract very useful features from the speech signal [27].

2.2.1 *Feature extraction of speech signals*

The naïve approach to use a speech signal would be to input the raw signal into one's classification model of choice. Historically, this would not lead anywhere, as the meaning of speech depends heavily on the frequencies in airwaves. Three values are especially interesting from a spectrum, the fundamental frequency f_0, and the two formant frequencies F1 and F2. f_0 corresponds to the pitch of a voice, i.e. the rate of vocal-fold cycling, while F1 and F2 indicate which vowel is spoken. All three are marked in the example spectrum of the vowel /iy/ in Figure 2. Note that the amplitude of the different frequencies in the spectrum is not linear but logarithmic, shifting it to decibels (dB), the unit used for the loudness of sounds. The spectral envelope, drawn in red in Figure 2, corresponds to how the shape of the vocal tract filters the glottal pulse during speech. Values for the fundamental and formant frequencies, as well as the spectrum itself, constitute some of the most important features used for ASR [27].

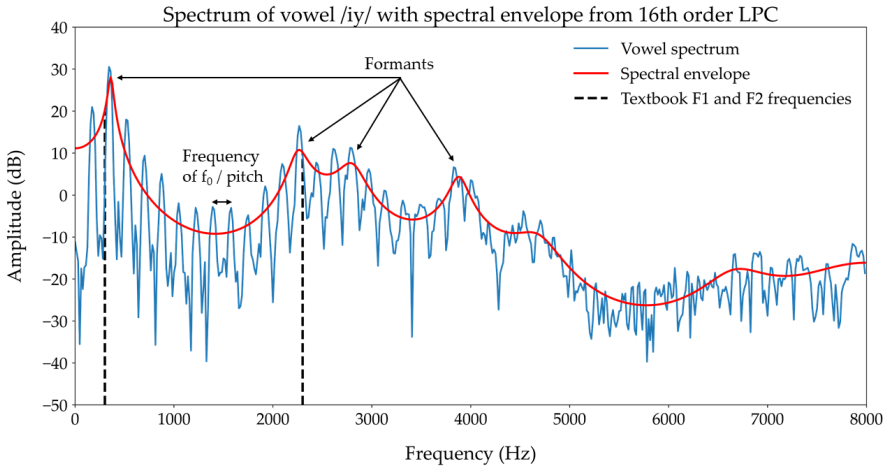


Figure 2.: A spectrum of the vowel /iy/ with linear predictive coding (LPC) coefficients of order 16 used for the spectral envelope. The value for f_0 is found by looking at the distance between tops, while the formants are seen as tops in the spectral envelope. The textbook values for F1 and F2 (300 and 2300 Hz) [27, Table 2.5] were added as stippled black lines.

Mel-Frequency Cepstral Coefficients

Probably the most popular features for speech applications are the Mel-frequency cepstral coefficients (MFCCs) [1], invented in 1980 by Davis and Mermelstein [8]. They are based on two crucial insights. One is the fact that taking the inverse Fourier transform of a logarithmic spectrum returns valuable information about the periodic structures in frequency spectra [27]. This changes the data to the *que-*

frency-domain, and a spectrum is renamed *cepstrum*. In this cepstrum, information about the low-frequency formats and the fundamental frequency, as seen in the log power spectrum of Figure 2, will be discernible. The second insight is that the human auditory system is based on a non-linear frequency scale. For us humans, the perceptual difference between frequencies is much more significant with lower frequencies, easily recognizable by comparing the perceived difference between 100 and 200 Hz, and 1000 and 1100 Hz. Both the Mel- and Bark-scales are empirically developed scales where the same distance on each of the scales corresponds to the same change in pitch perception. The following formula is the basis for calculating the *mel* value from a frequency,

$$Mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right).$$

Using this Mel-scale and overlapping triangular windows to create mel-filterbanks, then applying a discrete cosine transform[1], the result is the MFCC features. Correspondingly, features named perceptual linear prediction (PLP) can be derived using the Bark-scale [27]. Later developments have led to another feature extraction algorithm called power-normalized cepstral coefficients, which suppresses background noises and improves recognition accuracy compared to MFCCs and PLPs [40]. However, there are examples of research groups that use the raw audio signal for speech recognition and synthesis by using modern deep learning methods [56, 70]. Note that for many applications, independently of the selected feature extraction method, first and second time-derivatives of the features are used as additional inputs to the learning model.

Even though the principle of MFCCs is based on the human cochlea, the easy availability of MFCC libraries and their usefulness in speech recognition has led to MFCCs being used for EMG-based silent speech as well [38, 50]. The typical frequencies of the Mel-scale are not suited for the frequency range of an EMG signal but can be re-scaled using the frequency range of the input signal and the same building principles as the original MFCCs.

2.2.2 *Degrees of recognition*

Speech recognition can be sorted into four different degrees of recognition: isolated words, a few connected words, continuous speech, and spontaneous speech. The recognition rate of isolated words is easily calculated by looking at how many of the words were correctly

1 The discrete cosine transform is used as it is a simpler version of the inverse Fourier transform that returns only real values, and at the same time, decorrelates energy in the overlapping mel-filterbanks.

classified. However, it is not as easy to evaluate the performance of a speech recognition system for spontaneous speech with multiple sentences. When comparing the output-sequence of words from the recognition system with the original reference transcription, errors will continually lead to the need for re-alignment of the two sequences. There are typically three types of such word recognition errors in speech recognition; substitution (*subs*), deletion (*dels*), and insertion (*ins*). The word error rate (WER) was therefore defined in the following way [27, Equation 9.3]:

$$\text{Word Error Rate} = 100\% \times \frac{subs + dels + ins}{\text{No. of words in the correct sentence}}.$$

The WER is frequently used throughout this report to describe the performance of a speech recognition system. Performance of single word classification is usually kept to the recognition rate for simplicity, which is the same as 100% minus WER for single word classification.

2.3 A PROBABILISTIC APPROACH TO SPEECH RECOGNITION

In the very beginning of ASR, only isolated words were recognized. By the late 1990s, real-time language dictation systems with large enough vocabularies for more widespread use became available [28]. One of the reasons for these advancements was the representation of speech as a hidden Markov process, which builds on the concept of a Markov chain; A stochastic model of random processes based on the probabilities of an initial state-distribution and state transitions [69]. In an ordinary Markov chain, each state corresponds to a deterministically observable event. However, when introducing a non-deterministic process that includes hidden states and another set of observable states depending on the hidden Markov process, we have the hidden Markov model (HMM) [27]. As a layer between the observations and the hidden Markov process, the probabilistic Gaussian mixture model (GMM) is often used to fit the real-world data, such as the relevant speech features, into probability distributions. Figure 3 shows an overview of how the GMMs and hidden Markov processes are combined. This GMM-HMM method resulted in very effective speech recognition systems from the 1980s onward [28]. HMMs are still widely used for ASR today. Typically in combination with the MFCC features mentioned previously.

Phonemes are the building blocks of spoken language, making up the words we speak. In a typical ASR system using HMMs, a model is built hierarchically from ground states, via phonemes and words, to sen-

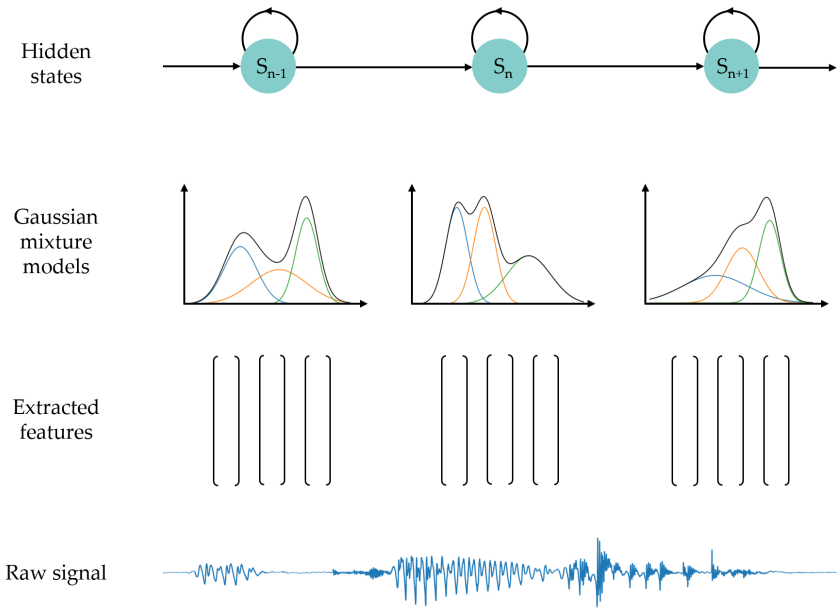


Figure 3.: A simplified overview of the GMM-HMM process where the raw audio is split into windows, relevant features extracted, then fitted by the GMMs and finally connected with the hidden states of the HMM.

tences. This is done by maximizing the probability for a sequence of words based on a combination of the acoustic, lexical, and language models. Despite how well GMM-HMMs work for speech recognition, GMMs cannot optimally fit the non-linear properties of speech [28]. Since the introduction of GMM-HMMs, more advanced methods have been used to minimize the WER further and generalize speech recognition systems to become more speaker-independent.

2.4 MACHINE LEARNING USED FOR SPEECH RECOGNITION

To solve the issue of fitting non-linear data, feed-forward deep neural networks (DNNs) were introduced to ASR during the late 1980s [25]. DNNs work by iteratively training the parameters of each node. A set error between the proposed solution by the network and the correct answer is backpropagated through the network, and each parameter is updated according to the learning rate. Combinations of DNNs and HMMs resulted in a significant reduction of the WER for the best-performing ASR systems at the time. The performance of these new machine learning methods increased proportionally with more training data, going hand-in-hand with the growing availability of processing power through the 2000s [28]. As the DNN does not hold any temporal information, it was a very good match with the HMM.

By itself, a well-working NN method was not possible until the advancement of recurrent neural networks (RNNs) that keep a *memory* structure expressed as internal states between the different nodes. It does not only feed information forward but also loops it back recurrently. One issue with this is the vanishing/exploding gradient, the fact that information from long before in a time series will have a disproportionate impact on the result as the gradient of the early information will be repeated for every timestep during training. The long short-term memory (LSTM) and gated recurrent unit (GRU), two more advanced versions of the RNN, were developed by researchers to solve this problem. Figure 4 shows a simplified version of the difference between the DNN and RNN, while Figure 5 visualizes the differences between the simpleRNN, LSTM, and GRU units [69].

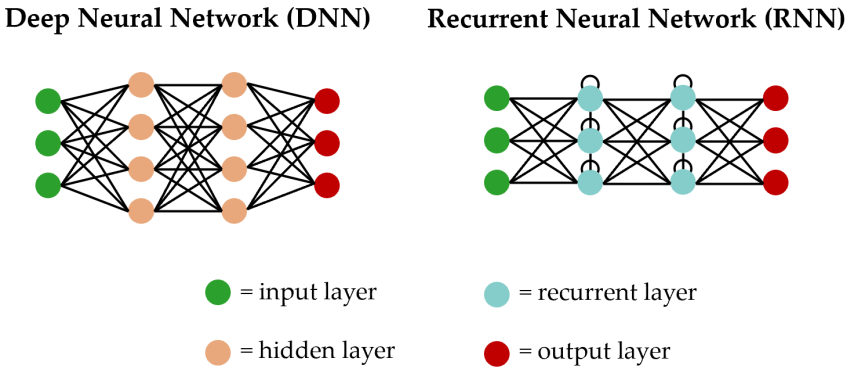


Figure 4.: Visualizations of the network architectures of deep neural networks (DNNs) and recurrent neural networks (RNNs). Note that the connections that loop back in the recurrent layers have a time-delay compared to the rest of the connections.

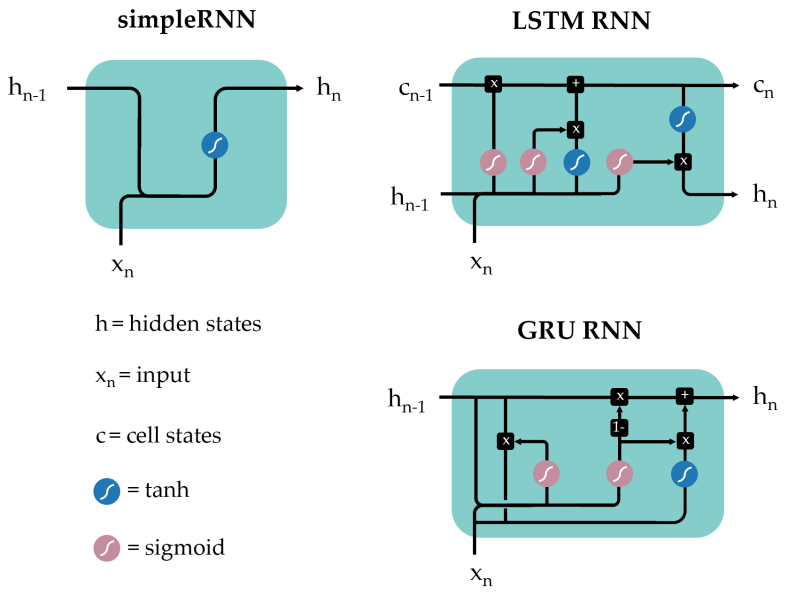


Figure 5.: The inner workings of the simpleRNN, LSTM and GRU versions of an RNN node. *tanh* and *sigmoid* are two different activation functions transforming a signal to either between -1 and 1 or 0 and 1, respectively.

The convolutional neural network

Another appealing NN architecture is the convolutional neural network (CNN). It is structured in a way that makes it possible to extract certain shapes for each convolution layer by a filter that moves around the input data. These shapes can later be combined to the representation of more complex structures. Figure 6 shows a simplification of this process, including two fully connected hidden layers in the end as that is often used. CNNs are most famous for revolutionizing image classification but have also been used on audio data where the convolution layers look for features along the time-axis of the input data. CNNs have to a limited degree been used for vocalized speech recognition, but more so for EMG-based silent speech [38, 69].

Performance optimization

To get neural networks (NNs) to perform well, hyperparameter optimization is of great importance. This involves tuning all of the parameters that the designer of a NN can decide. The most obvious choices include what type of NN is best fitted for the problem at hand, how large the NN should be, i.e. the number of layers and hidden nodes, what an appropriate size of the input is, and how to extract features from the original data. Then, there are numerous decisions to be made within the NN structure itself, such as dropout layers that randomly remove a certain percentage of nodes for each round of

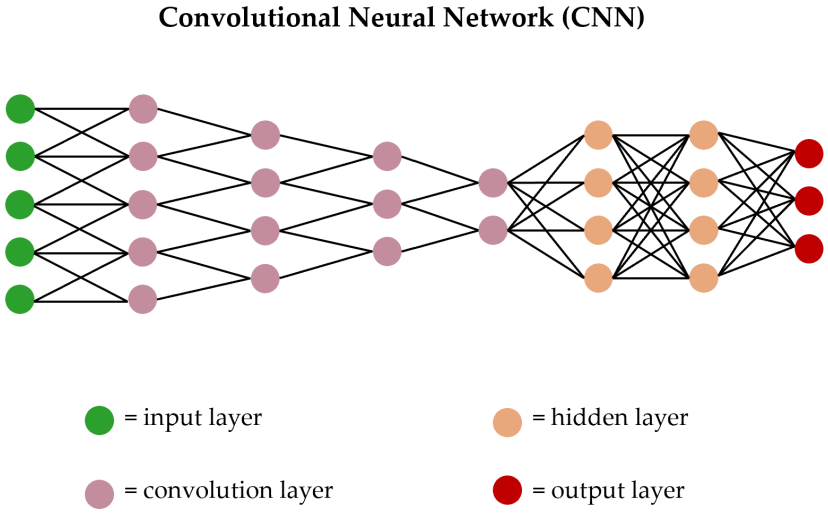


Figure 6.: Visualization of a typical network architecture of a convolutional neural network (CNN). Note that the hidden layers are often called dense layers as well.

backpropagation training or pooling layers that blurs the signal by taking the average or max value from a certain number of nodes and feeding that forward to the next layer in the NN. There also exist different activation functions that can be chosen for each layer of the NN. These determine how the input into the node is processed before being sent to the next layer of nodes. Moreover, different loss functions describe precisely how to calculate the loss during the training of the NN. Lastly, the learning rate and the number of epochs decide how much the weights of each node will change for each training step and how many rounds of training will be performed, respectively. Both can have a significant impact on performance.

2.5 SILENT SPEECH

2.5.1 *Modalities to detect silent speech*

To enable silent speech it is necessary to read signals earlier in the speech production process than audible vocalization. These signals can be anything between neuron activity in the brain and inaudible whispers. Revisit Figure 1 for a recap of this process and how it connects.

Silent speech based on the central nervous system

Because all the information necessary for speech is present in the brain, it should theoretically be possible to achieve silent speech based on signals recorded from the brain. As even our most complex thoughts and emotions form in the brain, future brain interfaces might even surpass silent speech and enable communication on a much more abstract level. An important distinction when it comes to silent speech based on brain signal is whether or not the interface is invasive, i.e. if an operation is needed to install the device. For speech-impaired patients with disorders that target the brain or the connection between the brain and muscles, only direct brain recordings will be able to restore speech [22]. Facebook announced during their F8 conference in 2017 that they aimed to "creat[e] a silent speech system capable of typing 100 words per minute straight from your brain" [14]. One of their collaborating research groups published a paper in Nature in 2020 reporting a WER of only 3% on a vocabulary of 250 words using electrocorticography (ECoG), electrodes placed on top of the brain, beneath the skull [49]. Another American research group showed in 2020 that a patient with two brain implants was able to type with his mind at a rate of 90 characters a minute on an unlimited vocabulary. The initial WER of 25.1% was decreased to 1.5% using an offline bidirectional decoder and a language model [68]. These results show that functional SSIs are closer than ever, but so far only with invasive brain surgery.

Non-invasive solutions based on the brain include the inherent challenge that signals from the brain are distorted by the skull. They are therefore limited to recording either brain waves or activation of brain areas, not singular or a small group of neurons. EEG is likely the non-invasive brain interface that has been proposed most seriously as a modality for silent speech as it can be used on freely behaving humans. One result using EEG for silent speech shows high rates of classification accuracy only when distinguishing two different vowels [30]. Another report a WERs of around 75% on a vocabulary of 29 words [42]. Contrarily, one study from 2016, using the same Emotiv Epoc+ sensor used for this project (as an EEG headset the way it was intended), achieved an average recognition rate of 67.03% on 30 different classes envisioned by the subjects [44]. This shows that EEG as an SSI is indeed possible.

Detection of movement

Subsequent to the brain activity related to silent speech, signals are detectable in the peripheral nervous system between the brain stem and muscles. Kapur, Kapur, and Maes [38] place weight on the fact that their system, although based on EMG-sensors, can detect acti-

vation of the peripheral efferent neurons without any noticeable facial movement. Most EMG-based systems, however, also use signals from muscle activation interlinked with movement. To date, it seems an EMG-based SSI is the most promising non-invasive method for enabling functional silent communication with both the speed and accuracy of vocalized speech. The most striking evidence of this is the results from Meltzner et al. [50], describing an EMG-based SSI that only has an 8.9% WER on a 2200 vocabulary.

Another option is to use magnets to detect the movement of the tongue and lips, which has achieved good recognition of smaller vocabularies. However, this option introduces the obvious disadvantage that the magnets have to be operated into the user's mouth for long-term use to be achieved [15, 26]. A third option is to capture the movement of the lips by using cameras in the same way as lip-reading is an attainable skill for humans. This method is limited by the need for a camera in front of the user's mouth but may still be useful as this is already present when using modern-day personal computers or smartphones. Taking advantage of the fact that there is an almost unlimited amount of video with corresponding text available online, Google DeepMind has made a system trained on almost 3600 hours of training data that was able to read lips much more precisely than human professional lipreaders (WER of 40.9% compared to 86.4 – 92.9%) [59].

Almost-silent glottal activity

The final category of possible SSIs is dependant on some form of glottal activity, but it can be so low that it is still not discernible. First out of these systems was the non-audible murmur (NAM) microphone, proposed in 2003 by Nakajima et al. [54]. It is based on a stethoscopic microphone placed behind the ear that captures vibrations from the vocal tract through the skin. These vibrations are present with a lower-than-whisper murmur of words. Even though there are some challenges with noise from clothing, hair, and respiration, the NAM microphone is currently used to a small extent in Japan.

Other methods using the vibration of bone or skin are in use today, but primarily as ways to reduce background noise in settings of vocalized speech. One instance is loud military combat environments, with the solutions of the Danish company Invisio [6] already existing. Another alternative is the use of ingressive speech, where a special microphone is placed less than 2 mm from the mouth, and non-detectable speech commands can be recognized. Fukumoto [18] reported a WER of 1.8% on a speaker-dependent system with a vocabulary of 85 command sentences.

Multimodal silent speech interfaces

Various SSI modalities concerning the different stages of speech production have been covered in this section so far. Additionally, it is possible to combine multiple modalities as well. Multimodal interfaces introduce the possibility of richer input data that can give silent speech recognition with lower WERs. Nevertheless, they also introduce challenges, like how to synchronize data from the different sensors optimally. One example is how Freitas et al. [17] used a combination of EMG and real-time magnetic resonance imaging (MRI) to detect nasal vowels inherent to the Portuguese language during silent speech. A complete summary of multimodal SSIs can be found in Freitas et al. [16, Chapter 4].

2.6 ELECTROMYOGRAPHY

2.6.1 *Recording muscle activation*

Muscular activation is always preceded by an electrochemical current through the nervous system and into muscle fibers. The corresponding voltage potential propagates through tissue from the activation site and will eventually reach the surface of the skin. As the signal gets attenuated along the way, the most precise measurement of this signal is invasive electromyography (EMG) using conductive needles, which is the preferred method for some medical diagnostic tests of muscle response and to detect neuromuscular abnormalities. A more practical application of EMG detection outside of the hospital is surface electromyography,[2] where electrodes are placed on the skin. Therefore, the recorded signal from such sensors is an attenuated signal from the surrounding muscles with a stronger signal from muscle fibers closer to the sensor [52]. Most uses of surface EMG are with a bipolar configuration where two electrodes are placed along the muscle of interest with approximately 2 cm in between. A third reference electrode is placed on a place with little to no muscle activity so that the resulting signal is the difference between the skin potential of the two electrodes relative to the reference. The most important property of a surface electrode is to minimize the electrode-skin impedance to reduce noise. Different kinds of dry and wet electrodes exist for different purposes. The widely accepted standard for EMG electrodes is of the type silver (Ag)/silver chloride (AgCl) with an added conductive gel between the electrode and skin [4, 65].

There are additionally other ways of obtaining EMG measurements from multiple sensors. For instance, with an electrode array that has

2 Surface electromyography is usually abbreviated sEMG, but will after this section be denoted as EMG.

one common reference and the possibility of looking at the difference between all the electrodes in the array. Breakthroughs in material science have also made it possible to produce flexible and super-thin EMG arrays for facial recordings of surface EMG. Exemplified here with the Nature article: *A Wearable High-Resolution Facial Electromyography for Long Term Recordings in Freely Behaving Humans* [29].

2.6.2 *Data processing of EMG-signals*

The interesting frequency range for EMG signals is heavily dependent on the sensors and what the signals are used for. A wide range can be put at 0.5 to 2000 Hz, while $20 - 2000$ Hz is often used for medical purposes [52]. However, if one were to include the full bandwidth, unnecessary noise would be included due to both biological and technical artifacts. These artifacts include amplification noise, the possibly high impedance between the sensors and the skin, powerline interference at either 50 or 60 Hz, motion noise typically at $1 - 10$ Hz, heart activity, and cross-talk between muscles [4, 65]. The EMG signal is therefore often filtered extensively, rectified to avoid a mean of zero,[3] and smoothed. Different features are then extracted based on the specific application. Frequency ranges chosen for previous EMG-based SSIs vary extensively from as low as $0.5 - 8$ Hz [64] to even broader than the textbook range, $0 - 2.5$ kHz [50].

Comparing EEG and EMG

Electroencephalography (EEG) is a non-invasive measurement technique to read brain waves, with electrodes placed on top of the head. The electrodes used for most measures of biopotentials (usually denoted ExG, including EEG and EMG) are usually more or less similar. Placement and size range, as well as the frequency bandwidths and recorded amplitudes (in mV), on the other hand, are somewhat different for each application. EEG usually operates with smaller voltage amplitudes and a lower frequency range than EMG, but there is still an overlap in the lower end of the frequency range. Chapter 2.1 of *Neural Engineering* gives a complete rundown of the different types of biopotential measurements and electrodes [2].

2.7 EMG-TO-TEXT

As seen in the Introduction, there has been research on the topic of facial EMG measurements related to silent speech since the 1960s. Since then, research on EMG-based SSIs has mainly focused on EMG-to-text, and usually session- and speaker-dependent single word classifica-

3 Rectification of the EMG signal is conducted to identify the overall strength of the neural signal, and thus the total muscle activation in an area.

tion in English. Some papers have looked at session and speaker-independent models, but that makes classification much more difficult. Table 1 shows an overview of most previous studies on single-word classification using facial EMG sensors. The columns show the original publication for the results, the publication year, vocabulary size for the main corpus in the study, the recognition rate for that corpus using the optimal method in each paper, and lastly, whether or not that method was session-independent. A couple of papers are listed twice as they present both session-dependent and -independent results. All results used for the literature review for this thesis were speaker-dependent.

Table 1.: Single word classification results from previously published work in EMG-to-text.

Source	Year	Vocabulary	Accuracy [%]	Session-independent
[61]	1985	5	64	No
[53]	1991	10	60	No
[3]	2001	10	93	No
[33]	2005	10	73	No
[48]	2005	10	97.4	No
[48]	2005	10	76.2	Yes
[35]	2006	108	68	No
[65]	2014	108	85	No
[65]	2014	108	73	Yes
[60]	2017	5	64.7	No
[38]	2018	10	92	No
[50]	2018	65	90.4	No
[47]	2019	10	72	No
[71]	2020	10	79.5	No
[72]	2020	10	93	No

Based on the accuracies presented in Table 1, it is evident that EMG-to-text is far from a solved scientific problem. Even though some papers presented classification accuracies above 90% on 10 words before 2005, several more recent studies still achieve accuracies between 70 and 80 percent. These differences are usually a result of the focus of the study, different amounts of available training data, how much effort is put into optimizing the classification methods, and whether the publication came from a research group with much previous experience in the field of EMG based silent speech or not. In the last few years, more focus has furthermore been put into EMG-based silent speech in different languages, e.g. Soon et al. [60] using Malay and

Ma et al. [47] using Chinese.

The truly state-of-the-art in EMG-to-text has been mentioned a couple of times already in this thesis but is not listed in the table above. This is because it is a model based on mapping EMG data to phonemes, not words. Meltzner et al. [50] collected multiple corpora, and their smallest corpus had a 65-word vocabulary that they used for word classification with an average accuracy of 90.4%, as presented in Table 1. The largest of their corpora, however, included a vocabulary of 2200 words. Instead of having a 2200-class classification model, the authors used MFCC features and GMM-HMMs to map 50ms windows of EMG data to a tri-phone model. This final speaker-dependent tri-phone model was then evaluated on 1200 continuous phrases from the 2200-word vocabulary of their final corpus with an average WER of 8.9%. The main disadvantage of the results from Meltzner et al. [50] is the fact that they used limited, highly expensive medical-grade EMG sensors that are difficult to acquire. This might be solved with a custom EMG headset made for silent speech sometime in the future, as something like this does not exist yet. This thesis therefore proposes that the Emotiv Epoc+ sensor might be a good option in the meantime.

2.8 EMG-to-speech

The EMG-based SSIs discussed so far have all focused on translating facial muscle activation to text, either directly through single-word classification or by mapping EMG data to phonemes. This translation from EMG to text has also been the main focus area of this Master's thesis. However, a subcategory of research on EMG-based silent speech focuses on generating speech waveforms from EMG signals. This approach results in potentially no restrictions on vocabulary in the corpus or even language, as certain muscle movements link to certain sounds made by the speaker. Additionally, a functional EMG-to-speech solution could preserve the voice of the speaker and possibly facets such as pronunciation, dialect, tone, and tempo as well [31].

2.8.1 *Speech synthesis*

With a high-level view of EMG-to-speech, there are two main approaches. One is to map the input EMG data *directly* to the waveform of a speech signal that can be played on a speaker. The other is to make use of existing research on speech synthesis and build the EMG-to-speech system out of two blocks; The first block takes EMG data as input and outputs some intermediate values that can then be combined with an already fine-tuned framework for synthesizing a waveform in the second block. With enough training data and a sufficiently advanced

learning model, a direct EMG-to-speech approach should be possible but is yet to be implemented. Therefore, a short introduction to the field of speech synthesis is in order.

Traditionally, speech synthesis was divided into two approaches, concatenative and parametric. With a concatenative speech synthesizer, previously recorded segments of speech, phones or words, are played back in a new order. This approach often results in high naturalness of the voice but is dependant on large datasets of prerecorded recordings and is limited to the voice of the person in the recordings. A parametric approach, on the other hand, will synthesize speech from parameters such as formants and fundamental frequency. Therefore, a parametric approach gives a broader range of possible voices and is not limited by previous recordings but has struggled with lower naturalness [24]. Even so, most early operating systems for personal computers came with a form of text-to-speech (TTS) based on formant or articulatory synthesis with high intelligibility, although with a robot-sounding voice.

Furthermore, a general challenge with TTS systems is that the underlying text needs to be translated into the basis of the speech synthesizer, either phonemes, formants, or the articulatory parameters. In 2016, DeepMind published a paper presenting WaveNet, a modern approach to speech synthesis using a form of CNNs named dilated causal convolutions to generate raw audio waveforms [55]. It achieved much higher scores of naturalness than the current best concatenative and parametric approaches used by Google at the time. Since then, a multitude of different deep learning methods has been used to achieve very human-like TTS systems. For instance, the combination of Tacotron2 [58] and WaveGlow [56] where Tacotron2 transforms text to mel-spectrograms, while WaveGlow generates waveforms from those mel-spectrograms realtime.[4] In the space of TTS, complete end-to-end methods without any vocoder are also starting to become available, e.g. the newly published Wave-Tacotron that removes the intermediate step of mel-spectrograms or MFCCs [67].

2.8.2 *Previous work in EMG-to-speech*

The first published article on the topic of EMG-to-speech was from Lam, Leong, and Mak [45] in 2006. There, the authors used two electrodes and a simple NN with two layers to map EMG data to 7 different sounds. From 2009 onward, researchers from the German silent speech community have published at least four papers on the

4 A simple-to-use example of how to use pre-trained versions of the Tacotron2 and WaveGlow models is available online: https://pytorch.org/hub/nvidia_deeplearningexamples_waveglow/.

topic [11, 12, 32, 62]. In 2009, they presented a GMM-based synthesis technique with a limited vocabulary where 84.3% out of 108 words were recognized correctly by humans listening to the synthesized audio. However, only 20.2% of the words were correctly recognized when the input used for audio synthesis was from EMG data recorded *silently*. This inconsistency highlights a general challenge with EMG-to-speech: that the EMG data recorded for training usually is from vocalized speech (because that is what is recorded during training to have matching audio data), while the EMG data for a practical SSI is from silent speech. The research group's later papers include improvements in selected features, the usage of different NN-based methods [32], the introduction of an unlimited vocabulary [11], and more realistic speech synthesis by improving the mapping from EMG to the fundamental frequency [12].

The current state-of-the-art in EMG-based speech synthesis is nonetheless from a group in Berkeley, US, with their 2020 paper *Digital Voicing of Silent Speech* [19]. Gaddy and Klein [19] achieved an impressive 3.6% WER using human evaluators on the digital voicing of silent speech from sentences built with a limited vocabulary of 67 words. Using their unlimited vocabulary, the WER from listening tests was 68%, but with a very natural-sounding voice[5] and on EMG data from silent speech, not vocalized speech. To achieve this, they recorded all sentences twice. Once by vocalized speech recording audio and EMG, then a separate time recording only EMG data during silent speech. The silent speech EMG could then be matched with the audio output targets by using a target-transfer approach. They further used a large LSTM-network with three bidirectional layers of 1024 units to transfer silent speech EMG data to MFCCs, which were finally used as inputs for a version of the DeepMind WaveNet vocoder [55] trained with their own data connecting MFCCs and speech waveforms.

5 The dataset and samples of synthesized speech from the Gaddy and Klein [19] study are available online: https://doi.org/10.5281/zenodo.4064408.

3 MATERIALS AND METHODS

For this project, five different corpora were collected; see Table 2 for an overview. Corpus 1 included three words in Norwegian: *stein*, *saks*, and *papir*, corresponding to rock, paper, scissors - the hand game usually played between two players. Four subjects, all male and aged 25 ± 1 years with Norwegian as their mother tongue voluntarily participated in collecting data for this corpus. For the remainding corpora, there was only one speaker, the author of this report. Corpus 2 included ten words, the digits 0 to 9 in English, and was collected using silent speech. For corpora 3, 4, and 5, a custom software program written for this project in Python was used to streamline data collection and enable recording of both EMG and audio data. Corpus 3 also had a vocabulary of the digits 0 to 9 and was recorded during vocalized speech simultaneously recording from the Emotiv sensor and a Blue Yeti microphone. Corpus 4 further increased the number of words in a single vocabulary to 29. Its vocabulary was the Norwegian extension of the NATO phonetic alphabet, where each word corresponds to a letter.[1] Corpus 5 was collected to include the possibility of a EMG-to-speech approach unlimited by vocabulary. It consists of EMG- and audio-recordings in an audiobook format that totals 7 hours of recordings from reading the novel Neuromancer [21].

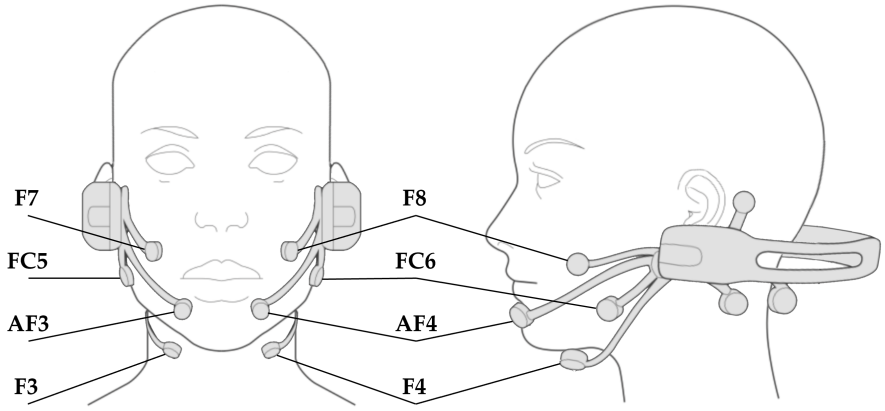


Figure 7.: The 8 chosen channels of the Emotiv Epoc+ and where they are located on the face when the sensor is used 'upside-down' as in this project. Adapted figure from Emotiv [13] with permission.

1 Alpha, Bravo, Charlie, Delta, etc.

Table 2.: An overview of the 5 collected corpora.

Corpus	Vocabulary	Sessions	Speakers	Samples
1	3 words	1	4	600
2	10 digits	7	1	4120
3	10 digits	15	1	6430
4	29 word NATO alphabet	12	1	5481
5	Unlimited[2]	9	1	7.0 h[3]

3.1 EMOTIV EPOC+ SENSOR

For all the experiments covered in this report, an Emotiv Epoc+ 14 channel EEG headset was used to collect data [13]. The electrodes are of the type Ag/AgCl with an additional felt pad soaked in a saline solution to achieve good skin contact. Out of the 14 electrodes, 8 were deemed relevant as they covered the face when the Emotiv sensor was turned upside down. Table 3 lists the relevant sensors and the muscles they cover, while Figure 7 shows their placement on a face. Note that since the Emotiv sensor is symmetrical, each sensor is placed as pairs, covering the same muscles on each side of the face. As a result, the sensor gives potentially redundant measurements. Still, the data from all sensors was used due to the possibility of inconsistent sensor placement from session to session. Furthermore, there is the possibility of users using muscles slightly asymmetric during silent speech, something that might influence the session and speaker independence. Other sensors such as a magnetometer and accelerometer are included in the Emotiv Epoc+ sensor but were not used for any experiments. A buildup of particles from the saline solution might occur from prolonged use, which can degrade the sensor signal and worsen session dependence in the subsequent data processing. To mitigate this effect, all the felt pads and electrodes were thoroughly cleaned semi-regularly, in accordance with the Emotiv Epoc+ guideline documents [13].

2 The vocabulary in *Neuromancer* is not unlimited, it is probably somewhere between 1000 and 10000 words. However, as Corpus 5 was used for speech synthesis, which include the possibility of synthesising words that were not in the original vocabulary, it's vocabulary is described as unlimited.

3 For Corpus 5, the value in the Samples column is the total number of hours recorded. When comparing this with the other corpora, 6430 samples in Corpus 3 corresponds to about 1.8 hours of training data given the fact that each sample is 1 second long.

Table 3.: Numbers, names and corresponding facial muscles for the 8 relevant sensors out of the 14 sensors on the Emotiv Epoc+ sensor. Note that the sensor names originally describe different brain regions used for EEG measurements.

Sensor #	Sensor name	Corresponding muscles
1	F3	Right sternohyoid & sternothyroid
2	F4	Left sternohyoid & sternothyroid
3	FC5	Right risorius
4	FC6	Left risorius
5	AF3	Orbicularis oris (lower lip - right side)
6	AF4	Orbicularis oris (lower lip - left side)
7	F7	Right zygomatic major
8	F8	Left zygomatic major

3.2 SIGNAL PROCESSING

As seen in Section 2.6.2, processing of the raw EMG signal is crucial for good detection. Most of this is done internally on the Emotiv sensor, which samples sequentially through a single analog-to-digital converter (ADC) at a rate of 2048 samples per second (SPS), later downsampled to 256 SPS. Digital notch filters at 50 Hz and 60 Hz are present to remove interference from the electrical power supply, independently of location. A built-in digital fifth-order Sinc filter then gives the resulting bandwidth of 0.16–43 Hz. Examples of normalized data from the Emotiv sensor are seen in Figure 8, where selected instances of the three different words in Corpus 1 are visualized.

Data from the Emotiv sensor was collected using WebSocket through the internal Emotiv Cortex service, where one and one data package had to be requested through their application programming interface (API). Initially, during work for the project report, this was done by using single Python scripts, running linearly. When using these scripts, the actual samples per second (SPS) of the Emotiv sensor turned out to be 280, not the expected 256 SPS. However, it slowed down to 125 SPS when simultaneously recording from the built-in computer microphone as a result of running everything linearly. A choice was then made to only use data from the Emotiv headset for corpora 1 and 2.

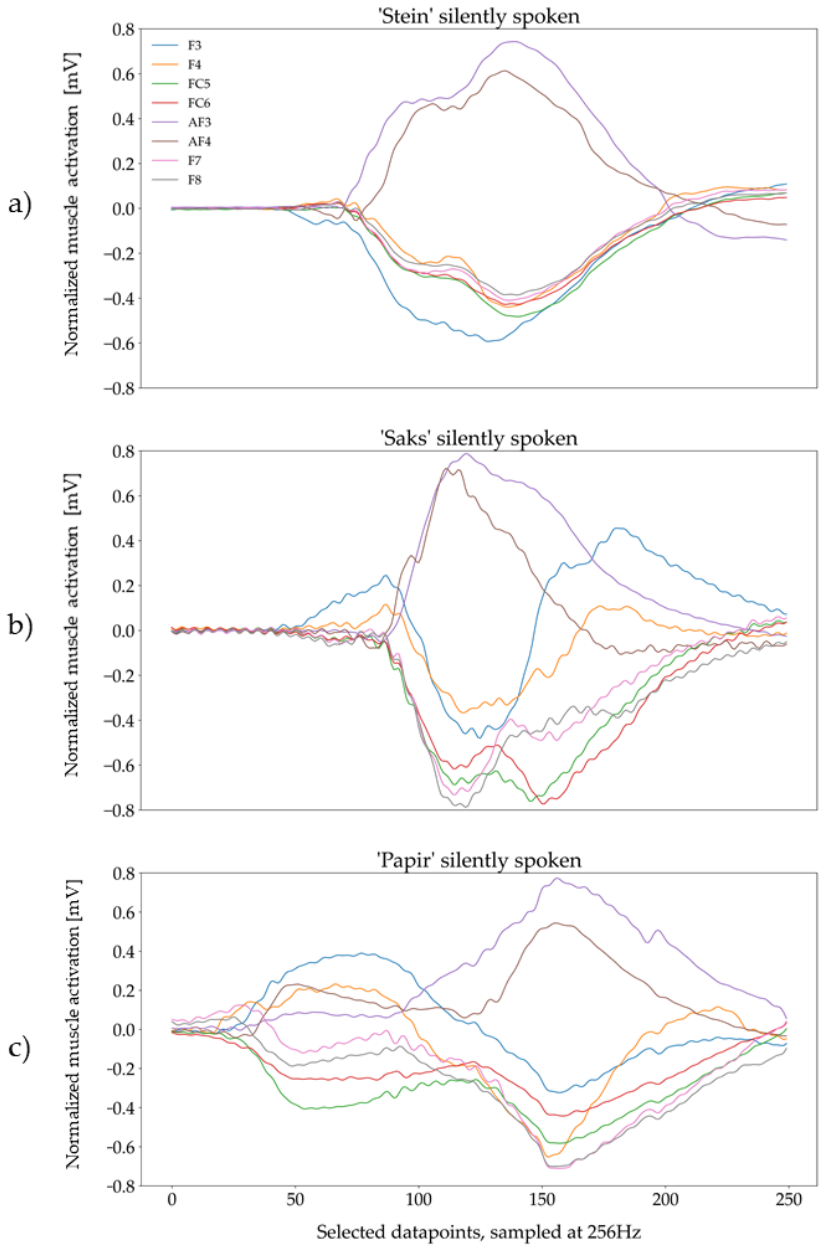


Figure 8.: Normalized data from Corpus 1, visualizing the difference between the tree words in the corpus. Legend and x-axis are common for all three plots. a) Shows the 8 different sensor values for a selected instance of 'stein' silently spoken. Likewise for b) with 'saks' and c) with 'papir'.

3.2.1 *Custom Python GUI*

It was decided to spend part of the master thesis to build a custom Python graphical user interface (GUI) software program from scratch to solve the issues with the data collection pipeline used for the project report. The two main reasons for this decision were to enable simultaneous recording of EMG and audio signals, and to make the process of recording large quantities of data much more efficient. The resulting software enabled long-duration recording sessions, the possibility of playing back earlier recordings of EMG data with the classification of words using previously trained models, and a mode for live visualization and speech recognition of silent speech. To solve the issue of simultaneously recording EMG and audio, parallel processing using *threads* was implemented. With this new program, about four times the amount of previously collected data was collected for corpora 3, 4, and 5. A screenshot of the program can be viewed in Figure 9, and screen recordings from using the program are available online.[4] Note that with this setup, the SPS for the Emotiv sensor was stable at 256 Hz for all recordings, while all audio recordings were conducted at a rate of 16 kHz.

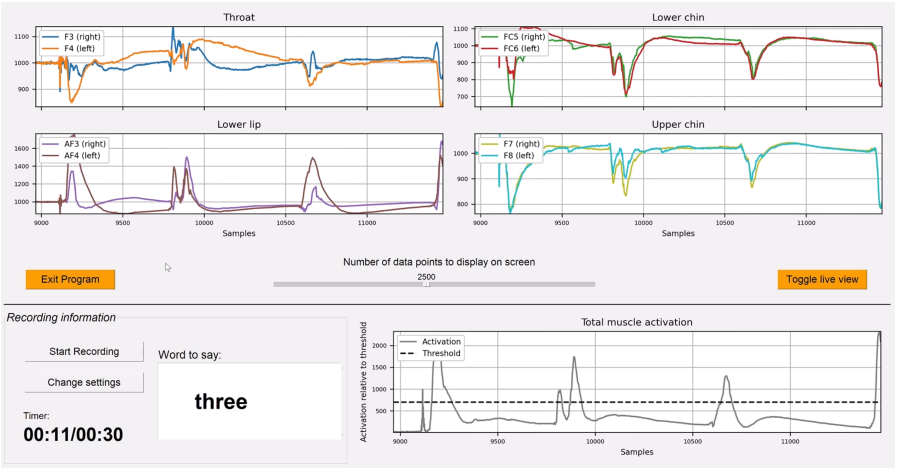


Figure 9.: A screenshot of the *subVocal* program made by the author. The program was running in the recording mode, where live data can be viewed while words are prompted, when this screenshot was taken.

4 https://drive.google.com/drive/folders/159X_Owdh6JoadVWggexwSA-oGckDrmNe?usp=sharing

Subjects for Corpus 1 participated in recording sessions in which they were asked to silently move their mouth as if speaking, reciting words displayed on a computer screen. They could simultaneously observe the live data from the Emotiv sensor to minimize the amount of noise resulting from movement between each word. In the recording session for Corpus 1B, where three of the subjects participated, two different setups were used. One setup where each subject recited the words displayed silently, and another where each word was spoken with vocalization.

Data from the different recordings were processed and correctly labeled before it was used for single-word classification. All code used for the recording sessions, processing of data, visualization and recognition was written in Python version 3.7 by the author. Both Jupyter [41] and Google Colab [23] notebooks, in addition to pure Python scripts, were used for these purposes.

3.3.1 *Each of the five corpora*

Corpus 1

Corpus 1 was intended as an initial data set to test whether the silent speech recognition principle worked with the Emotiv sensor. It included three words collected by four subjects, with a total of 600 samples (50 per subject per word), comparable to the first studies on EMG-based silent speech classification [34, 61]. In Corpus 1B, three of the speakers from Corpus 1 collected 20 instances of each word in Corpus 1 silently, as well as 20 vocalized, totaling 120 samples per speaker. The additional Corpus 1B was collected to perform initial testing on session independence and the difference in EMG recognition between silent and vocalized speech. It should be noted that the recording session for Corpus 1B was conducted several months after the main body of Corpus 1. Different kinds of feature extraction and classification methods were effectively tested on Corpus 1 before they were used with Corpus 2.

Corpus 2

The aim for Corpus 2 was to study the effects of more words in the selected vocabulary, more training data, and the effect of multiple recording sessions.[5] Seven recording sessions were conducted, resulting in 4120 samples equally distributed between the ten digits. Considering that the resulting data for each word corresponds to 1 second

5 Note that a *session* is defined as all recordings conducted within a time frame without removing the Emotiv sensor between recordings.

of the recording, this gives about 1.1 hours of training data. Table 4 shows the relevant information on the different recording sessions.

Table 4.: Information on the 7 recording sessions included in Corpus 2.

Session	Date of recording	Samples recorded
$2-1$	10.11.2020	580
$2-2$	12.11.2020	740
$2-3$	13.11.2020	360
$2-4$	16.11.2020	160
$2-5$	16.11.2020	560
$2-6$	17.11.2020	860
$2-7$	20.11.2020	860

Corpus 3

Corpus 3 was collected by the author vocalizing the digits 0 to 9 in a series of recording sessions. Corpus 3 enabled the possibility of comparing single word recognition when using EMG and audio data. Because corpora 2 and 3 had the same vocabulary, it was initially planned to compare models trained on silent and vocalized speech between the two corpora. However, because Corpus 3 was collected after the completion of the custom Python GUI, it was recorded at a rate of 256 SPS. As Corpus 2 had an SPS of 280 Hz, models were not compatible between the two corpora. Corpus 3 included 15 sessions with a total of 643 recordings, each consisting of the ten words in the vocabulary in a randomized order, see Table 5. This gives a total of 6430 samples, 50% more than in Corpus 2.

Corpus 4

To see the effect of more words in the vocabulary, Corpus 4 was collected. The Corpus 4 vocabulary consisted of 29 words, the NATO phonetic alphabet, designed such that each word is as distinct from the others as possible when spoken out loud. When combining the vocabularies for Corpus 3 and Corpus 4, it is possible to perform complete spoken communication independently of language by spelling each word. A total of 12 sessions were recorded for Corpus 4, with 189 recordings of 29 utterances, totaling 5481 samples. This gives 3.4 times fewer utterances per word when compared with Corpus 3. Table 6 lists all the different sessions of Corpus 4.

Table 5.: Information on the 15 recording sessions included in Corpus 3.

Session	Date of recording	Samples recorded
3 – 1	7.03.2021	480
3 – 2	9.03.2021	600
3 – 3	9.03.2021	220
3 – 4	11.03.2021	550
3 – 5	18.03.2021	300
3 – 6	18.03.2021	200
3 – 7	19.03.2021	300
3 – 8	22.03.2021	400
3 – 9	23.03.2021	520
3 – 10	24.03.2021	570
3 – 11	25.03.2021	700
3 – 12	26.03.2021	200
3 – 13	26.03.2021	400
3 – 14	29.03.2021	490
3 – 15	29.03.2021	500

Table 6.: Information on the 12 recording sessions included in Corpus 4.

Session	Date of recording	Samples recorded
4 – 1	8.03.2021	464
4 – 2	10.03.2021	667
4 – 3	11.03.2021	145
4 – 4	11.03.2021	435
4 – 5	17.03.2021	580
4 – 6	18.03.2021	290
4 – 7	18.03.2021	580
4 – 8	18.03.2021	580
4 – 9	19.03.2021	435
4 – 10	23.03.2021	725
4 – 11	29.03.2021	290
4 – 12	29.03.2021	290

Corpus 5

Corpus 5 was collected as a testbed for direct EMG-to-speech applications. By having a dataset that included time-synced facial EMG data and recorded audio, it would be possible to train models that take EMG as the input and returns an audio waveform. As described in Section 2.8.1, generating a raw speech signal is anything but trivial. However, if successful, it would enable a functional SSI unlimited by vocabulary. With this goal in mind, it was decided to collect synchronized EMG and audio data by reading a science fiction novel loud while wearing the Emotiv sensor and using a the Blue Yeti microphone, similar to how audiobooks are recorded. The science fiction novel Neuromancer [21] was chosen as it includes a rich vocabulary, and 7 hours were recorded across nine recording sessions.

3.4 FEATURE EXTRACTION

Before inputting the collected data into classification or regression models, it is usually sensible to process the data in one of several possible ways. Some methods are general, e.g. normalizing the data between 0 and 1, while others are domain-specific. When extracting features from the audio signal, either MFCCs or mel-spectra were chosen as those have been the preferred features for speech for a long time. Feature extraction for facial surface-EMG, on the other hand, was a more challenging choice. As the research area of EMG-based SSIs is relatively young and based on only a few scientific milieus, no superior feature extraction method for the EMG signal has been found. This choice of features also heavily depends on what type of EMG sensor is used and how the signal is pre-processed. To study the effect of different possible feature extraction methods, seven methods were selected and tested on Corpus 3 to see what gave the best results. These methods were:

1. The **raw signal** of each of the eight electrodes with standard values of 2000 to 8000.
2. Data **normalized** on the basis of each electrode within a recording. Calculated from dividing each channel by its mean value. All electrodes centered around 1, removing any effects of inter-electrode shifting. Values typically between 0.5 and 2.
3. Using the ***minmax***_*scale* function from the *sklearn.preprocessing* library, data for each word was scaled between 0 and 1 for each of the electrodes separately.
4. Like the one before, but adding the first order time-**deltas** (derivatives/the difference between consecutive values) as 8 additional features.

5. **Mel-frequency cepstral coefficients (MFCCs)** calculated using the *python_speech_features* library on the raw data for each of the words in the dataset. They were altered to fit the frequency range of the EMG signal and a list of the parameters used is presented in Table 7.

6. The same MFCC features as above, but with added time-**deltas**.

7. An **activation** signal combining the absolute values of all electrodes after centering the normalized values around 0.

Visualizations of all the different feature extraction methods are presented in Figure 10 and the parameters chosen for the MFCC features for EMG data is listed in Table 7. About 200 different combinations of the MFCC features were tested before the final parameters were found.

Table 7.: A list of the parameters that were used in the Mel-frequency cepstral coefficients (MFCCs) function from the *python_speech_features* library to genereate features for EMG-based silent speech. Note that the cepstral coefficients for each electrode were appended to the same axis, then the axes of the returned output features were reorganized to better match the input layer of the NNs.

Parameter	Value
No. of cepstrum for each window	5
Length of analysis window in seconds	0.12s
Step between successive windows	0.01s
No. of filters in filterbank	26
Size of the FFT	512
Append energy-sum to 0th cepstral	False

A common challenge when classifying single words is the different lengths of each utterance. This challenge was solved by setting a fixed length of 250 samples per utterance, ~1 second of recording. The eight different channels were handled as different features when inputting the data to neural networks, making the training data of shape $N \times T \times D$ when experiments of single word classification were conducted. Here, N = the number of training samples in a session, $T = 250$, and D = the number of features. When using the raw, normalized, or minmax-scaled signal from each of the channels $D = 8$, and when including their first derivatives $D = 16$. For MFCCs, T is dependant on the lengths of the windows and steps while D = the number of cepstrum times the number of electrodes.

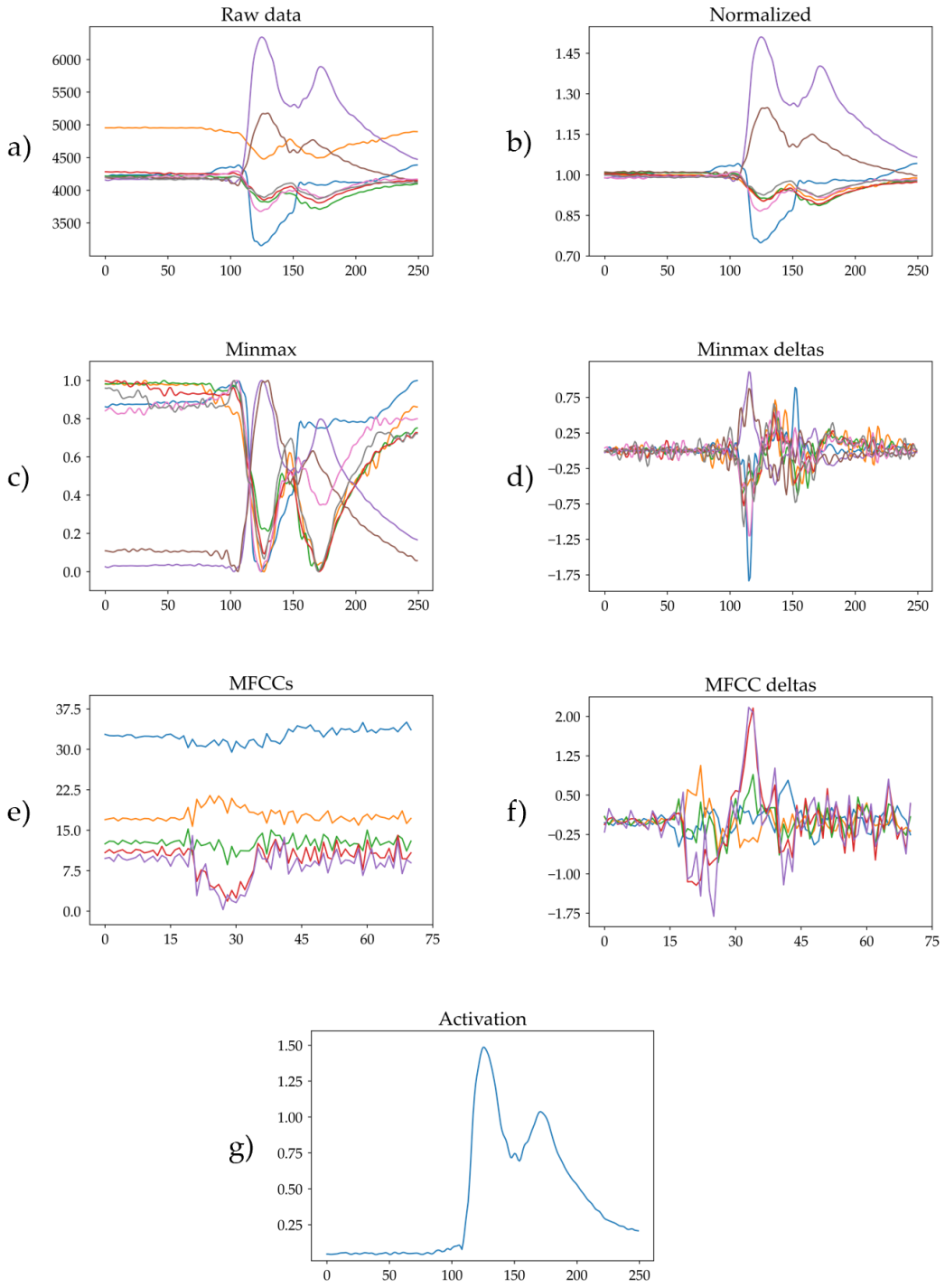


Figure 10.: Data from all 7 different feature extraction methods on one single utterance from session $4-12$. **a)** The plotted raw data for all 8 electrodes as it is received from the sensor. **b)** Data normalized for each electrode. **c)** Scaled data using the *minmax_scale* function from from the *sklearn.preprocessing* library. **d)** The deltas for *c)*. **e)** Calculated MFCCs for electrode 1. Note that MFCCs are calculated and used for the other seven electrodes as well. **f)** Deltas for *e)*. **g)** The total muscle activation signal from all 8 electrodes.

When using a fixed-length method that is shift-dependent, it is essential to synchronize the data from the Emotiv sensor correctly with the sample indices for when each word was silently spoken. The words prompted to the subject during a recording came from a randomized list of words containing an even distribution of each word in the corpus vocabulary. Words from this list were prompted to the test subject at specific intervals, and the timestamp for each word was used as initial indexing. Then the *find_peaks* function in the *scipy.signal* library [7] was used to find the maximum muscle activation related to each timestamp to set the final index for each utterance. The data used for the *find_peaks* function was the sum of the absolute values of the eight normalized sensor values (the activation signal), see Figure 11. Using these final indices, half of the word length of 250 samples was selected in each direction, such that the peak always appeared as the time-midway point in any given sample. The same method was used for single word classification of the audiofiles in Corpus 3, where the word length in samples was set to 16000, one second of recording.

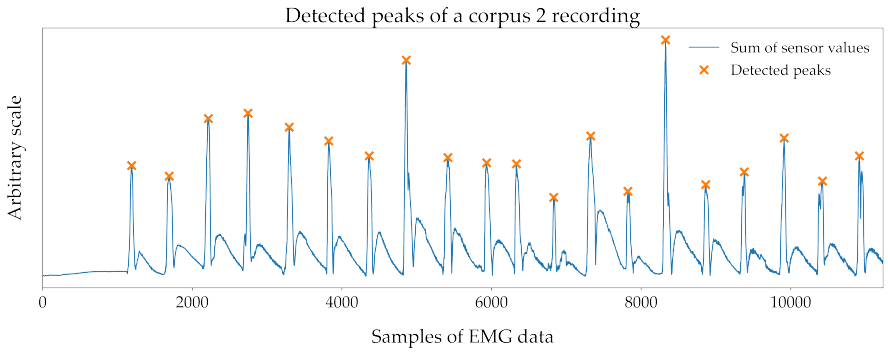


Figure 11.: To select the EMG data for each silently spoken word correctly, the peaks of the total muscle activation as the sum of the absolute values for the 8 different sensors were detected.

3.5 CLASSIFICATION ALGORITHMS

Because EMG-to-text is a classification task relevant classification methods were evaluated. This section is divided into two parts focusing on NNs and the GMM-HMM method respectively.

3.5.1 *Neural networks*

Classification was first conducted on Corpus 1 as a testbed to find algorithms that worked on the Emotiv EMG data. Based on results from similar studies (e.g. [38, 64]), both RNNs and CNNs were considered suitable options for classification.

Corpora 1 and 2

Because the initial objective of this project was to show a working principle of EMG-based silent speech rather than to optimize the classification perfectly, little effort went into hyperparameter optimization for corpus 1 or 2. By using data from Corpus 1 and inspiration from an Udemy course on TensorFlow 2 [57], parameters for semi-optimal versions of a CNN model and three different versions of the RNN: a simple RNN, a GRU, and an LSTM, were selected. These four different NN-based classification algorithms were then used for continued testing with corpora 1 through 4.

Architecture 1: Layers of the CNN that was used for classification.

```
x_train, x_test, y_train, y_test = train_test_split(X, Y, stratify=Y,
    test_size=0.2)

# Build the CNN model
K = 10   # number of categories for classification

i = Input(shape=x_train[0].shape)
x = Conv1D(32, 3, strides=1, activation='relu')(i)
x = MaxPooling1D(2)(x)
x = Dropout(0.5)(x)
x = Conv1D(64, 3, strides=1, activation='relu')(x)
x = MaxPooling1D(2)(x)
x = Dropout(0.5)(x)
x = Conv1D(128, 3, strides=1, activation='relu')(x)
x = MaxPooling1D(2)(x)
x = Dropout(0.5)(x)
x = Dense(512, activation='relu')(x)
x = Flatten()(x)
x = Dropout(0.5)(x)
x = Dense(K, activation='softmax')(x)

model = Model(i, x)
model.compile(optimizer=Adam(lr=0.001), loss='
    sparse_categorical_crossentropy')
```

The initial CNN architecture as written in Python using TensorFlow 2 can be seen in Architecture 1, and the same for the simple RNN in Architecture 2.[6] Some different activation functions, optimizers, and loss functions were tested, and the best-performing setup for Corpus 2 is presented below. It was decided to keep the same train/test split at 20% for all tests and to have the same distribution of the different classes in the test set as the train set, an equal distribution between all classes for both corpora. Initial results showed that the CNN drastically outperformed all three RNN models. However, after the introduction of an iterative decreasing learning rate and using only every eight value of the T-direction as an effort to solve the issue of the vanishing gradient, all three RNNs performed on par

6 The architectures for the LSTM and GRU RNN-types can be found in Appendix A.1.

with the CNN on Corpus 1. The CNN network was trained for 100 epochs with a learning rate of 0.001, while the RNNs were trained for 1000 epochs. Their learning rate decreased from 0.005, via 0.001 and 0.0005, to 0.00025, each for 250 epochs. Neither a dynamic learning rate nor increasing the number of epochs made the CNN perform any differently. The *sparse categorical crossentropy* was used as the loss function for all four NN architectures.

Architecture 2: Layers of the simple RNN that was used for classification.

```
# Remove most of the data (leave every 8th timestep)
X = X[:,0:250:8,:]

x_train, x_test, y_train, y_test = train_test_split(X, Y, stratify=Y,
    test_size=0.2)

# Build the simple RNN model
K = 10    # number of categories for classification
M = 100   # number of RNN nodes

i = Input(shape=x_train[0].shape)
x = SimpleRNN(M, activation='relu')(i)
x = Dense(K, activation='softmax')(x)

model = Model(i, x)
```

After Corpus 2 was collected, it was decided to continue the use of all four NNs to get a broad range of results for the silent speech classification of 10 digits. The main experiments were the multiple train/test split classification tests conducted on each of the seven recording sessions. Additionally, tests were conducted where one of the sessions was set as the test set, while an increasing number of the remaining sessions were used as the training data. Only the CNN method was used for these final tests, and no more than one test/train split was used. Both to decrease the needed time for neural network model training. This experiment was conducted to investigate whether six sessions were enough to classify a completely unseen set of data, as well as to see the effects of session independence as more and more sessions were used as training data.

Corpora 3 and 4

When continuing work with single word classification on corpora 3 and 4, all the available feature extraction methods and classification architectures were cross-tested to find the match with the most potential. Then, more effort was put into further increasing the accuracy of the selected feature-architecture match by systematically optimizing the model hyperparameters. Several thousand neural networks were trained with varying features as input and network parameters to find the optimal combination. A list of the parameters

that were tested and the resulting optimal combination is presented in Table 8. Multiple rounds of hyperparameter optimization were carried out, starting very wide and with only five training epochs for each model. After each round, the most promising value ranges were chosen, and new hyperparameters were tested in a more narrow search with an increased number of training epochs. After more than 2500 different configurations of the CNN were trained on corpus 3, a new and more optimal version was selected. This new CNN model architecture named CNN2 is presented in Architecture 3 and was used for most of the remaining experiments.

Table 8.: A list of the parameters that were optimized for the CNN model. Note that any change that increased complexity/training time without increasing performance was seen as less optimal. The resulting optimal model structure and parameters as seen in the rightmost column were named CNN2.

Parameter	Search window	Optimal combination
No. of conv. layers	[1, 4]	4
Size of 1st conv. layer	[8, 120]	110
Conv. size multiplier	[1.0, 2.0]	2.0
Conv. kernel	[1, 4]	4
Conv. stride	[1, 3]	3
Dropout	[0.0, 0.5]	0.0
No. of dense layers	[1, 3]	1
Size of dense layer(s)	[25, 500]	250
Learning rate	[0.005, 0.0001]	0.00075
Epochs	[1, 100]	50

Using Corpus 3, efforts were then made to study whether it would be possible to cluster the 15 recorded sessions by training the final CNN model on one session and subsequently testing it on another session. When presenting the resulting classification accuracies in a matrix, the diagonal would represent training and testing on the data from the same session, but with a $50-50$ split between the training and testing datasets. Next, the effects of selecting different subsets of electrodes were tested. First, subsets of single electrodes were selected, and the CNN2 model was trained and tested. Then all of the possible combinations of 2 electrodes were tested, and finally, the two different sides of the face were compared. For these experiments, the CNN2 model was trained for 50 epochs and using one train/test split to save time.

Subsequently, the audio recordings from Corpus 3 were processed and classification conducted on the 10 digits using MFCC features and both the CNN2 and a series of HMM trained on each of the digits. Finally, classification was tested on Corpus 4 by itself and the joint vocabulary of corpora 3 and 4.

Architecture 3: The hyperparameter optimized version of a CNN for single word classification, named CNN2.

```
x_train, x_test, y_train, y_test = train_test_split(X, Y, stratify=Y,
    test_size=0.2)

# Build the CNN2 model
K = 10   # number of categories for classification

i = Input(shape=x_train[0].shape)
x = Conv1D(110, 4, strides=3, activation='relu')(i)
x = MaxPooling1D(2, strides=1)(x)
x = Conv1D(220, 4, strides=3, activation='relu')(x)
x = MaxPooling1D(2, strides=1)(x)
x = Conv1D(440, 4, strides=3, activation='relu')(x)
x = MaxPooling1D(2, strides=1)(x)
x = Conv1D(880, 4, strides=3, activation='relu')(x)
x = MaxPooling1D(2, strides=1)(x)
x = Dense(250, activation='relu')(x)
x = Flatten()(x)
x = Dense(K, activation='softmax')(x)

model = Model(i, x)
model.compile(optimizer=Adam(lr=0.00075), loss='
    sparse_categorical_crossentropy')
```

3.5.2 *Hidden Markov Models*

Based on the results from the state-of-the-art EMG-to-text, GMM-HMMs were seen as a potential classification method in addition to the NNs [50]. Using the *hmmlearn* package for Python, several versions of HMMs were implemented. For the audio waveforms of Corpus 3, one GMM-HMM was trained for each of the 10 digits in the vocabulary. Then the log probability was computed for each of the GMM-HMMs iterating through each sample in the test set. Each sample was classified into the category with the corresponding max log probability, and the model accuracy was calculated. This process was then conducted for 10 different train-test splits. Standard parameters were used unless for the number of states in the GMM, which was set to 5. The same method was used on the EMG part of Corpus 3 for the cross-testing of features and classification methods.

3.6 FUNCTIONAL SILENT SPEECH INTERFACE

The optimal result from this work would be a functional SSI based on the Emotiv Epoc+ sensor that could be used in realtime. Two options were viewed as within reach of the scope of this thesis:

1. Silent speech spelling by achieving session-independent word detection combining the vocabularies of corpora 3 and 4.
2. Translating EMG signals to speech waveforms independent of vocabulary by using Corpus 5, either directly or with some intermediate feature-space.

To test the functionality of these two methods, six sentences were created so that they included at least one occurrence of every letter of the English alphabet as well as the digits 0 to 9. The six sentences are presented in Table 9, and the character distribution is presented in Figure 12. Each of the sentences was recorded four times. Both vocalized and silent speech was used, and the sentences were recorded spoken directly and with each letter spelled out.

Table 9.: The six sentences used for testing the viability of a functional silent speech interface (SSI) system.

No.	Sentence
1	78 knights rode up the steep hill.
2	The answer to life and the universe is 42.
3	On May 5th 2021 SpaceX successfully landed SN15.
4	Pi equals approximately 3.1415.
5	Trondheim has its own jazz festival.
6	Roughly 20,600 lines of code were written for this thesis.

3.6.1 *EMG-to-text by spelling*

By using a CNN2 model trained on the joint data of corpora 3 and 4, it should be possible with a complete session-independent SSI with an unlimited vocabulary, as long as every word is spelled out. Because Corpus 3 was recorded using vocalized speech, while silent speech was used for Corpus 4, it was expected to see the CNN2 recognize words from the two vocabularies differently between the silent and vocalized versions of the spelled out sentences. As no words for *space*, *comma*, or the *dot/period* were part of the vocabularies, words for the three Norwegian-specific letters *ærlig*, *østen* and *åse* were used for this purpose. 10 different versions of the CNN2 architecture were trained on both corpora 3 and 4 using different train/test splits. They were

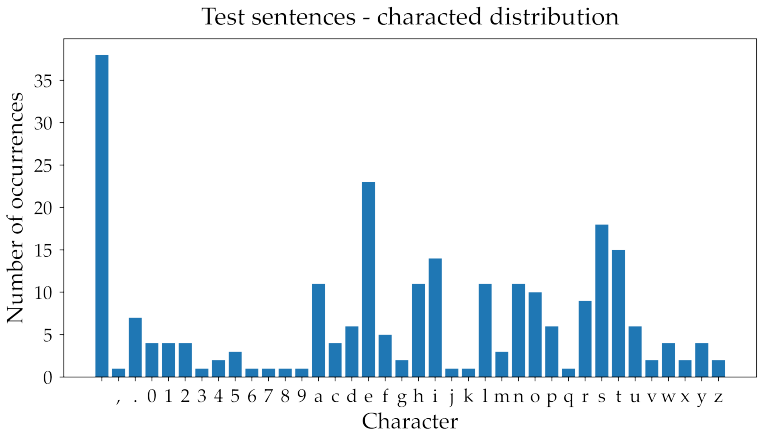


Figure 12.: The character distribution for the sum of all 6 test sentences is presented as a bar plot. Note that the leftmost character is the *space* between words.

then used to predict the sequence of characters, letters of the alphabet and numbers, for the six test sentences. The per-character accuracy was calculated for each of the models, then a system drawing the most often selected character over all 10 models, as well as the final results after using a standard grammar correction tool.

3.6.2 *EMG-to-speech*

To investigate whether an EMG-to-speech approach would be possible with data from the Emotiv sensor and the available feature extraction methods and machine learning models, pairs of speech waveforms and EMG data from Corpus 3 were studied. Numerous different model architectures, intermediate features, and speech synthesis methods were tested. Until the very last weeks of this project, all combinations only resulted in unintelligible sounds. Then, by using a version of the Nvidia WaveGlow model [56], mel-spectrograms as intermediate features, and a custom neural network architecture combining CNN- and LSTM-layers, intelligibly was achieved. The final architecture was named EMG-Net and is presented in Architecture 4, which takes the minmax and delta features from the EMG data as inputs and returns the corresponding mel-spectra for a 1-second window. Waveforms were synthesized from 20 EMG samples from Corpus 3 that the EMG-Net model had not been trained on. Ten human listeners were then asked to classify each of the digits based on the synthesized audio. Each of the listeners was sent the link to a Google Form with audio embedded and a questionnaire where they selected which number they thought each of the samples sounded like. A separate choice was available for the cases where the listeners

were uncertain or thought the sample did not sound like a digit.[7]

Architecture 4: EMG-Net, the model created for this thesis to transform the EMG minmax and delta features to mel-spectrograms calculated from the corresponding audio samples.

```
# Build the EMG-Net model

i = Input(shape=x_train[0].shape)
x = Conv1D(110, 4, strides=1, activation='relu', padding='causal',
    dilation_rate=1)(i)
x = MaxPooling1D(2, strides=2)(x)
x = LSTM(32, return_sequences=True, dropout=0.5)(x)
x = Conv1D(220, 4, strides=1, activation='relu', padding='causal',
    dilation_rate=2)(x)
x = MaxPooling1D(2, strides=2)(x)
x = LSTM(64, return_sequences=True, dropout=0.5, )(x)
x = Conv1D(440, 4, strides=1, activation='relu', padding='causal',
    dilation_rate=4)(x)
x = LSTM(128, return_sequences=True, dropout=0.5)(x)
x = Conv1D(880, 4, strides=1, activation='relu', padding='causal',
    dilation_rate=8)(x)
x = LSTM(254, return_sequences=True, dropout=0.5)(x)
x = MaxPooling1D(2, strides=2)(x)
x = Flatten()(x)
x = Dense(K, activation='linear')(x)

model = Model(i, x)
model.compile(optimizer=Adam(lr=0.00075), loss='mse')
```

As a result of the master thesis deadline approaching quickly, there was only minimal time available for experiments regarding Corpus 5 and the possibility of a functional SSI taking EMG to speech with an unlimited vocabulary. Two EMG-Net versions were separately trained on 1-second and 130-millisecond windows of Corpus 5 before they were used to generate speech from the EMG data recorded from the six test sentences. The plan was to use human listeners once again to judge the performance of the model and compare their input with the actual test sentences to calculate the resulting WER per listener. However, the resulting synthesized audio clips for both window sizes were unintelligible, sounding only like background noise.

7 The Google Form can be found here: https://forms.gle/98v4mWtv2bAgwRzT8

4 RESULTS

This chapter presents the achieved results on all corpora and the two SSIs. Results from every corpus are presented in order. The session- and speaker-dependent results are presented first within each corpus, before results regarding the session or speaker independence. For most experiments, 10 different trials using a different random split between training and testing data were conducted. As a result of recording both EMG- and audio-data for corpora 3 and 5, it was possible to compare the results using the Emotiv sensor with similar methods conducted on the audio signal. Results regarding the two functional SSIs are placed at the very end of this chapter.

4.1 CORPUS 1

4.1.1 *Recognition rate*

Classifying the three different words of Corpus 1 turned out to be a relatively simple task for the four selected NNs. In practice, each algorithm only had 120 training examples to learn from as each speaker was tested independently, and 20% of the total samples were used for testing the accuracy. Table 10 shows the average accuracy after 10 different train/test splits for each classification method on each speaker and the total average for each method. Notice the consistently high accuracy overall, the higher recognition rate on Speaker 3, and the fact that the simple RNN method was the one with the highest classification, even though both LSTM and GRU are more advanced versions of the same NN. To get a sense of the variance in the recognition data, the corresponding boxplots for the 10 train/test splits of both the CNN and simple RNN methods are presented in Figure 13. Similar boxplots for the LSTM and GRU results are found in Appendix A.2.1.

4.1.2 *Speaker independence*

The CNN was also used to test speaker independence by training on data from 3 speakers and testing on the 4th. This gave results no better than random chance at a 33% recognition rate and shows obvious overfitting. Table 11 lists the results after one trial per speaker, which indicates as expected that much more training data and on many more speakers is necessary to overcome the challenge of speaker independence.

Table 10.: Speaker dependent recognition rate [in %, average (± standard deviation)] for the different speakers in Corpus 1 using the four different classification methods.

Speaker	CNN	simpleRNN	LSTM	GRU
1	81.1 (±9.6)	85.0 (±6.0)	84.3 (±9.2)	83.7 (±7.4)
2	93.2 (±6.1)	94.3 (±7.5)	88.3 (±10.6)	88.7 (±7.8)
3	97.1 (±3.6)	96.7 (±3.3)	97.0 (±3.5)	95.3 (±6.0)
4	90.5 (±8.1)	97.3 (±3.3)	89.7 (±7.2)	89.3 (±8.0)
Average	90.5 (±6.8)	93.3 (±5.7)	89.8 (±5.2)	89.3 (±4.8)

Table 11.: Train and test accuracy [in %] of the CNN architecture trained on 3 speakers and tested on the 4th. As the results were no better than chance after one trial, only that trial was conducted. Values are therefore presented without any standard deviation.

Test data speaker	Train accuracy	Test accuracy
1	95.3	36.9
2	99.3	28.7
3	100	35.1
4	100	29.3

4.1.3 *Session independence & effect of vocalization*

Using Corpus 1B, session independence and the effect of vocalization were researched to a limited degree. Data from Corpus 1 on each of the respective speakers was used as training data for the CNN, while data from Corpus 1B was used as the testing data. Silent and vocalized data were tested separately. As the results show in Table 12, some recognition is present. The vocalized samples gave notably higher recognition than the silent ones. 10 train/test splits were conducted, and both the average values and standard deviation are presented.

Table 12.: Recognition rate [in %, average (± standard deviation)] of silently spoken and vocalized data from Corpus 1B, where the CNN was trained using data from Corpus 1, is presented for three of the original four speakers.

Speaker	Silent speech recognition	Vocalized recognition
1	58.0 (±9.7)	74.0 (±8.7)
3	44.3 (±13.5)	62.0 (±6.7)
4	38.3 (±7.1)	41.0 (±8.9)

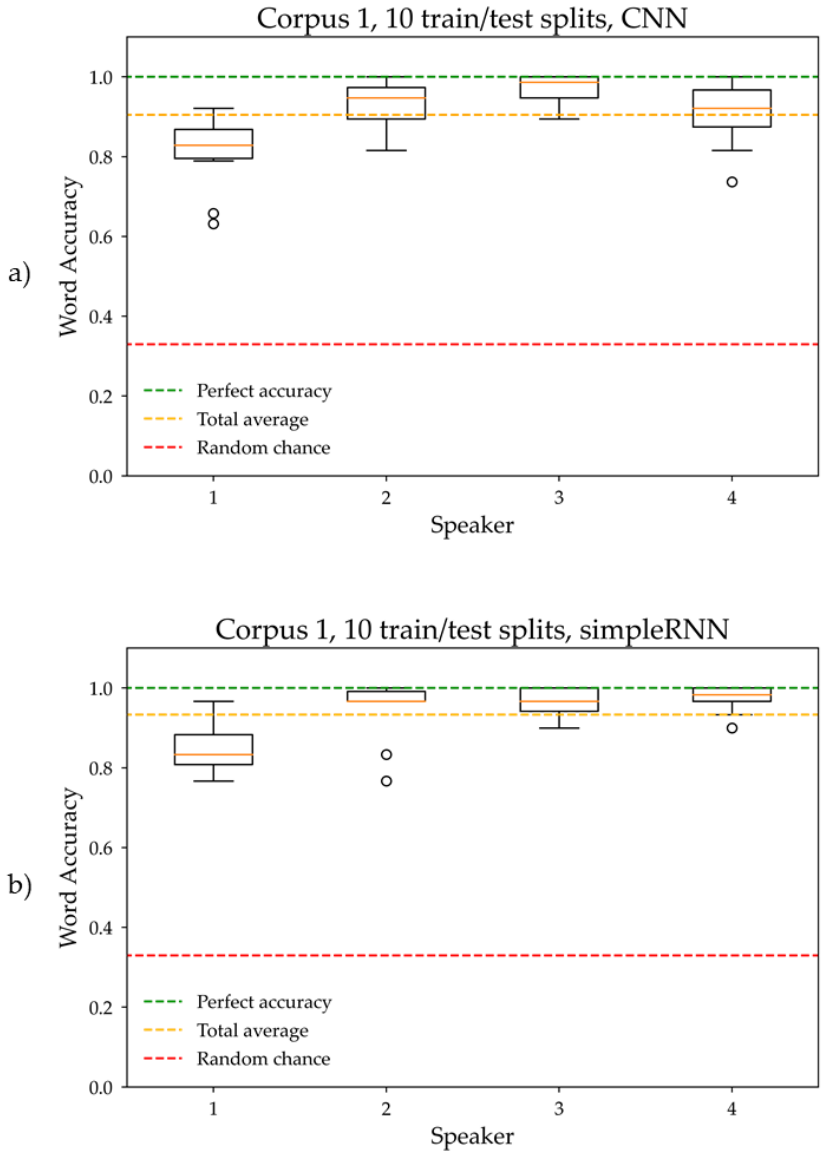


Figure 13.: Boxplots of the recognition rate for Corpus 1 using the **a)** CNN and **b)** simpleRNN architectures. 10 different train/test splits were performed with the median for each speaker as the solid orange line in each box. The box extends to lower and upper quartile, while the whiskers extend from the box to show the range of the data. Circles are regarded as outliers, but still calculated into the total average.

4.2.1 *Recognition rate*

More data is generally needed with a 10-word classification task compared with a 3-word classification to achieve the same recognition. In this case, between 128 and 688 training samples were available for each session. All 4 of the classification algorithms were tested for Corpus 2, and average recognition rates per NN varied from 42.3% to 64.6%. The single-highest recognition rate, 84.7%, was achieved using the CNN. All of the average values per session and NN is presented in Table 13, while Figure 14 shows the boxplots for each of the train/test splits on the different sessions using the CNN and simple RNN methods. Similar boxplots for the LSTM and GRU NNs are found in Appendix A.2.2. The recognition rate is much lower than for Corpus 1, with some sessions being easier to recognize than others. Independently of the classification method, sessions 2-2, 2-3, 2-6, and 2-7 gave better results. The CNN resulted in a higher classification than the other NNs and achieved recognition rates between 70 and 80% for sessions 2-2, 2-6, and 2-7.

Table 13.: Recognition rate [in %] for the different sessions in Corpus 2 using the four different classification methods.

Session	CNN	simpleRNN	LSTM	GRU
$2-1$	53.3 ($\pm$3.3)	36.6 ($\pm$2.6)	40.6 ($\pm$3.0)	32.2 ($\pm$1.5)
$2-2$	74.1 ($\pm$2.3)	56.0 ($\pm$7.0)	58.1 ($\pm$3.5)	45.8 ($\pm$3.2)
$2-3$	68.6 ($\pm$4.6)	64.0 ($\pm$4.7)	56.8 ($\pm$5.2)	47.6 ($\pm$5.6)
$2-4$	56.2 ($\pm$3.9)	40.3 ($\pm$8.0)	36.6 ($\pm$7.7)	33.8 ($\pm$4.6)
$2-5$	47.6 ($\pm$3.6)	39.8 ($\pm$2.5)	44.6 ($\pm$3.5)	37.6 ($\pm$3.3)
$2-6$	78.8 ($\pm$3.8)	60.1 ($\pm$5.7)	62.0 ($\pm$4.2)	51.5 ($\pm$4.3)
$2-7$	73.5 ($\pm$3.9)	62.6 ($\pm$6.4)	52.1 ($\pm$6.6)	46.9 ($\pm$3.8)
Average	64.6 ($\pm$11.2)	51.4 ($\pm$11.1)	50.1 ($\pm$8.9)	42.3 ($\pm$6.8)

The confusion matrix for one of the train/test splits using the CNN on session 2-7 is presented in Figure 15. Notice that digits with similar phonemes at the beginning of the word, and thus muscle movement when pronounced, are more often misclassified as each other — *zero*, *six*, and *seven*, as well as *four* and *five*, are examples of this.

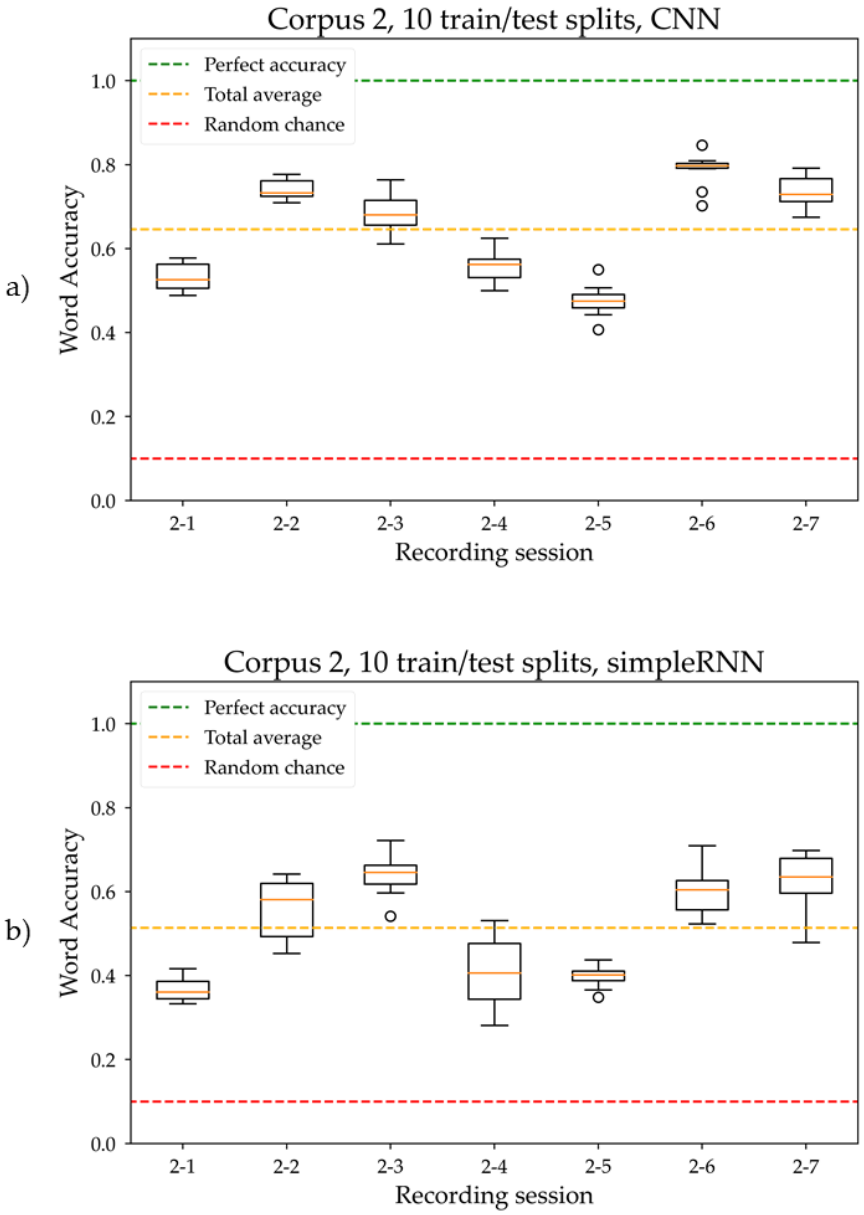


Figure 14.: Boxplots of the classification results for Corpus 2 using the **a)** CNN and **b)** simpleRNN architectures. Total average accuracies of 64.6% for the CNN and 51.4% for the simpleRNN are marked with the orange stippled lines.

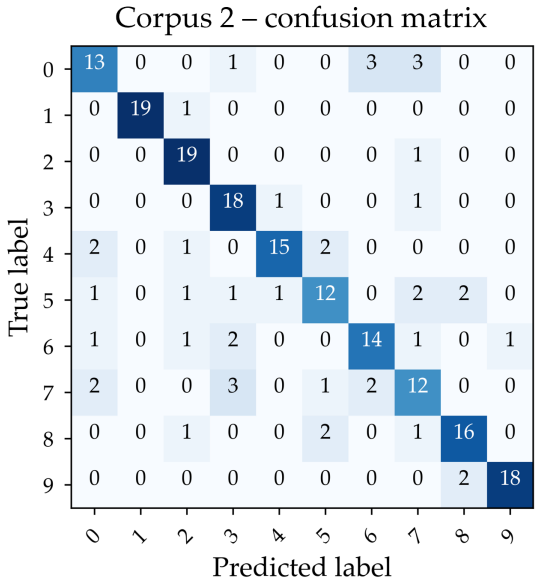


Figure 15.: Confusion matrix of session 2 – 7 using the CNN. A perfect classifier would have all values on the diagonal, as true and predicted labels are on the y- and x-axis respectively. Note that the test set was made with an equal distribution between the 10 digits and such each of the rows sum to 20.

4.2.2 *Session independence and generalization*

One final experiment was conducted on the topic of session independence for Corpus 2. As seven sessions were collected as a part of Corpus 2, it was hypothesized that training on an increasing number of sessions could give higher recognition rates on an unseen session. For every session, one to six of the other sessions were used as the training data, and the classification accuracy for each instance of the CNN was used as an indicator of how well it was able to generalize learning from multiple sessions to data from an unseen session. Because the number of samples in each session is different, and some sessions are easier to recognize than others, the following results should be used as an indication only. Only one train/test split was conducted due to time constraints, and all results for this experiment are shown in Figure 16. Even though only one of the plots shows a steady increase in recognition rate for each added session to the training data (session 3), the average clearly shows an increasing trend. Furthermore, six out of the seven had a higher recognition after training on six different sessions, compared with training on only one.

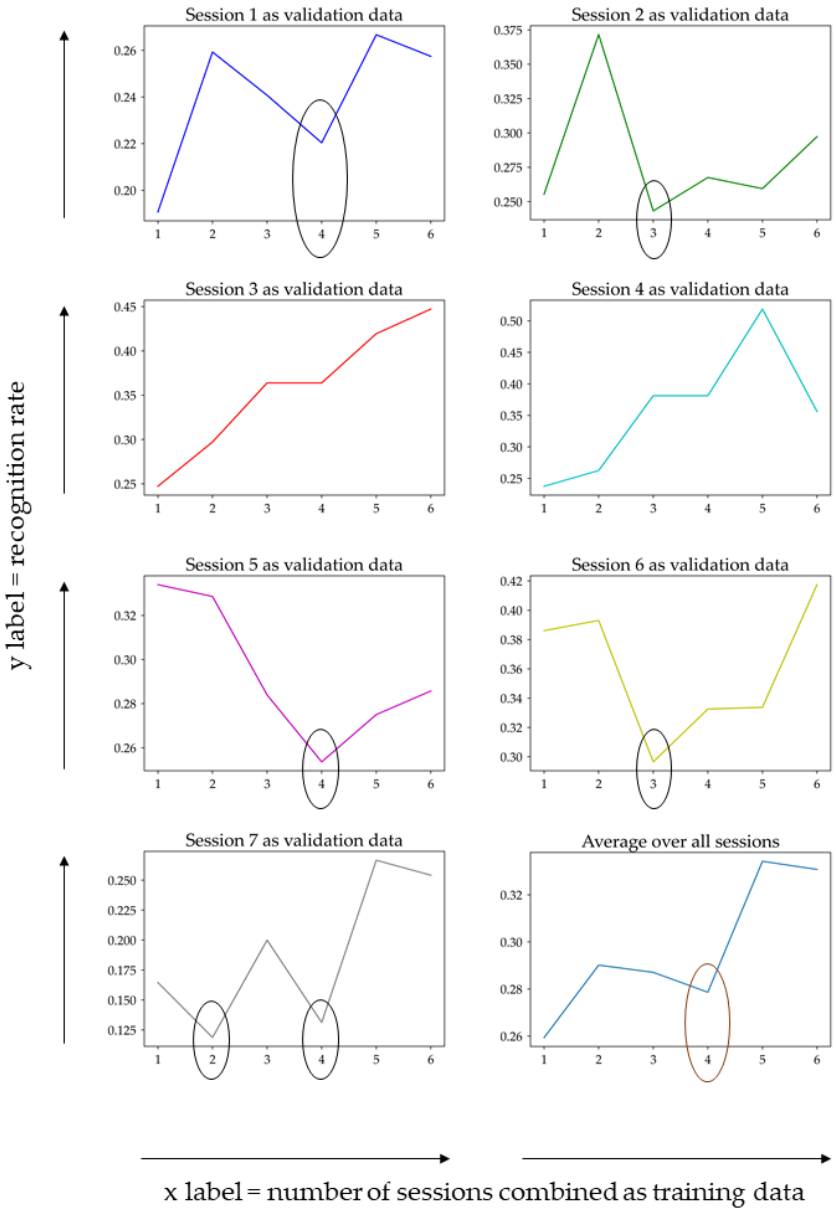


Figure 16.: Plots of session independence for the 7 different sessions in Corpus 2 and their average (the lowest graph to the right). Observe that there are significant drops after adding some of the sessions, marked with black ellipses. This is also visible in the average, seen with the red ellipse. For a discussion on this phenomena, look to Section 5.2.

4.3 CORPUS 3

Corpus 3 included both EMG and audio data on a vocabulary of 10 words where each of the digits 0 to 9 were spoken with vocalization over 15 different recording sessions. This corpus could therefore be used for EMG and audio based recognition separately as well as EMG-to-speech. Because Corpus 3 contained the most samples, experiments related to cross-session results and the effect of selecting subsets of electrodes were conducted using this corpus.

4.3.1 *EMG-based recognition*

With more than double the amount of sessions in Corpus 3 as compared with Corpus 2, higher accuracy on a session independent model is expected. However, as the number of samples in each session lies in the same area for both corpora, session-dependent accuracy would be expected to be more or less equal between the two corpora when only accounting for the number of samples.

Session dependent results

All six different classification methods were used on a per session basis on prepared datasets using the *minmax + deltas* features. Every experiment was run with 10 trials, and the results are presented in boxplots as with corpora 1 and 2. Word classification accuracies for the CNN and CNN2 architectures are presented in Figure 17, while Figure 18 presents the corresponding results for the simpleRNN and GMM-HMM. Results for the LSTM and GRU NNs can be found in Appendix A.2.3.

Observe that even using the same CNN architecture, the average accuracy for Corpus 3 is more than 13 percentage points higher than for Corpus 2. The average for CNN2 is even 6 percentage points higher and scored as high as 93.4% accuracy averaging the results for Session 3-2. The highest single trial was 97.1% showing the potential for classification accuracies closing in on 100%.

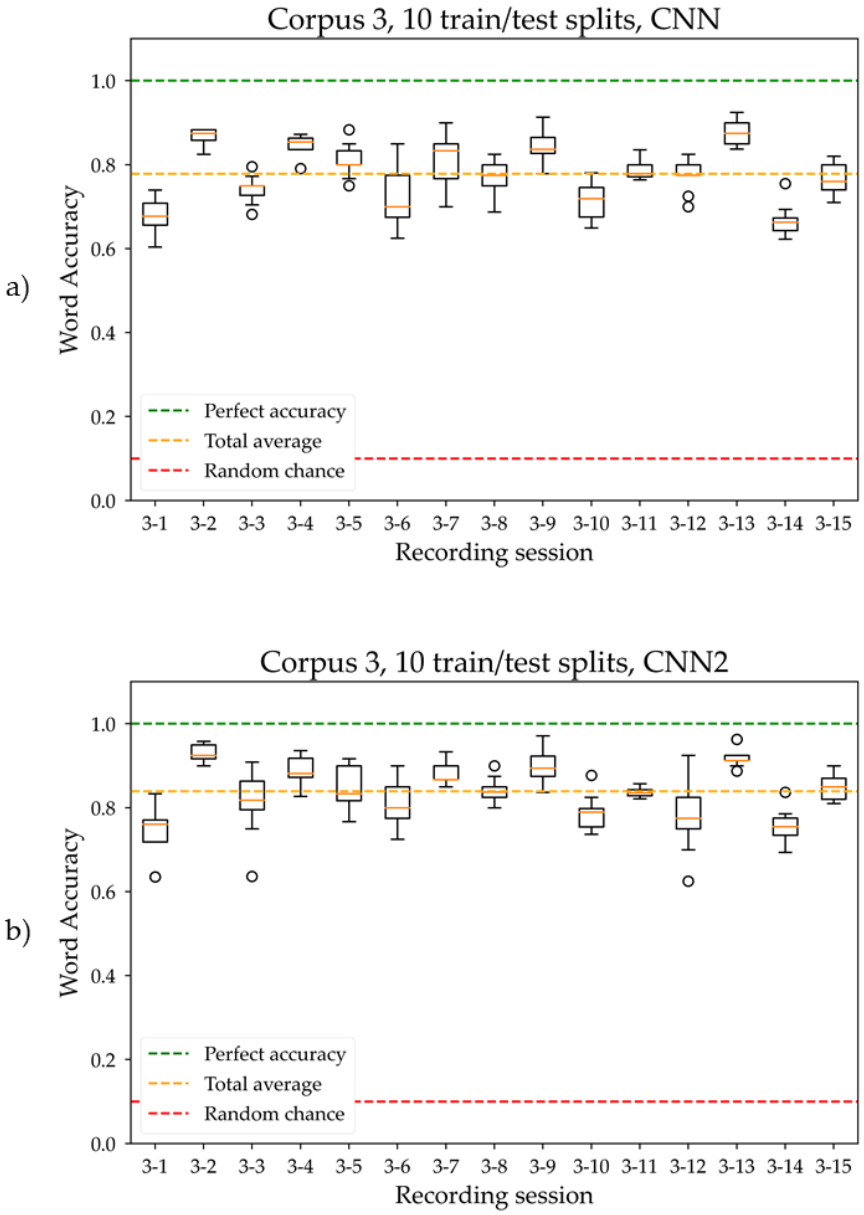


Figure 17.: Boxplots of the classification results for Corpus 3 using the **a)** CNN and **b)** CNN2 architectures. Total average accuracies of 77.8% for the CNN and 83.9% for the CNN2 are marked with the orange stippled lines.

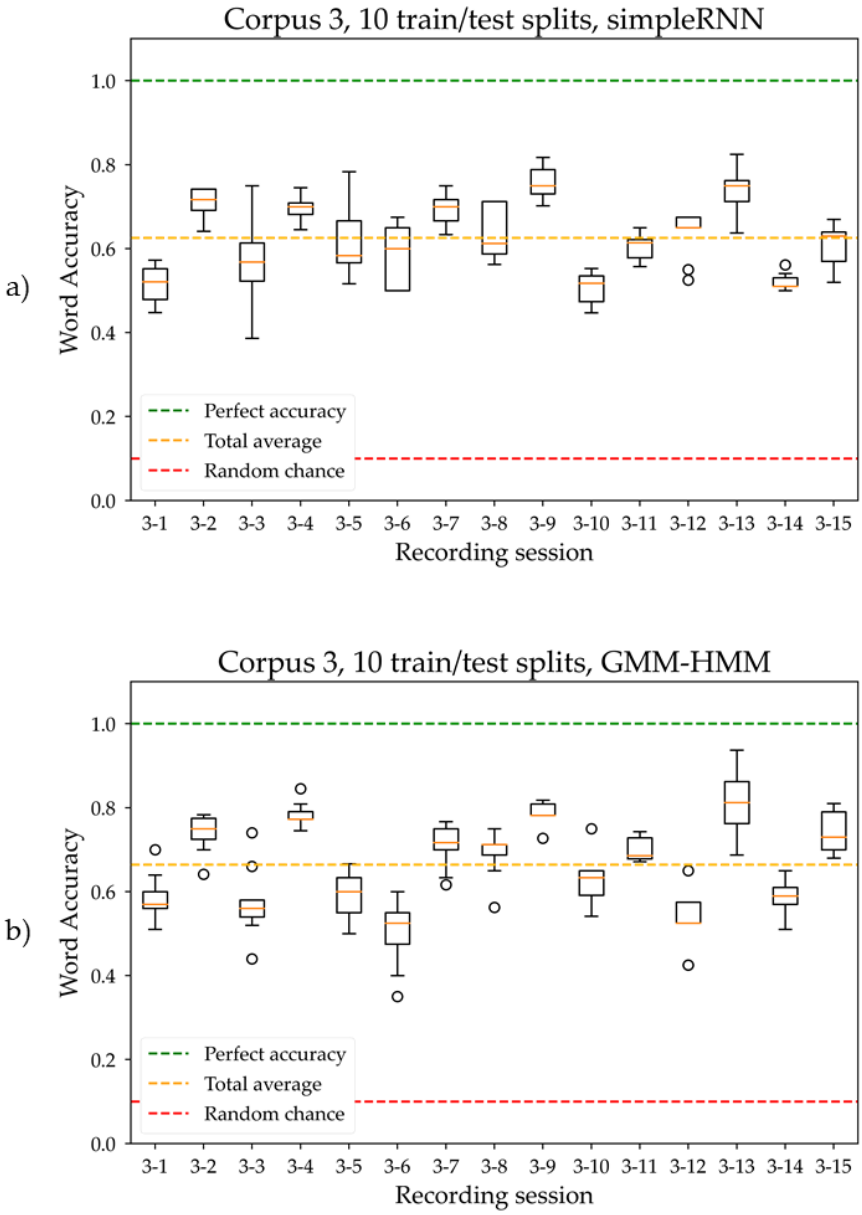


Figure 18.: Boxplots of the classification results for Corpus 3 using the **a)** simpleRNN and **b)** GMM-HMM architectures. Total average accuracies of 62.6% for the simpleRNN and 66.4% for the GMM-HMM are marked with the orange stippled lines.

Session independent results

For a functional SSI, session independence is essential. Therefore, it is important to find the optimal combination of EMG features and classification method on all the sessions in a corpus combined. Corpus 3 was selected for this, and all pairs of features and classification methods were tested.[1] The results are presented in Table 14. For the CNNs and GMM-HMM, 10 train/test splits were performed and the mean accuracy over all trials are presented. Because the training time for the different RNNs was an order of magnitude higher than the other classification methods, only 2 train/test splits were performed for them.

Table 14.: Mean recognition rate [in %] comparing the different feature extraction techniques and classification methods. Train and test data was randomly selected from the whole dataset of Corpus 3 and 10 train/test splits were performed or the CNNs and HMM, while only 2 train/test splits were used for the RNNs.

Features	CNN	sRNN	LSTM	GRU	HMM	CNN2
Raw data	15.3	10.0	10.0	10.0	52.9	74.4
Normalized	40.7	43.0	50.8	40.1	48.9	78.8
Minmax	67.6	52.1	55.5	40.9	56.9	81.3
Minmax + deltas	70.0	52.2	53.9	43.9	60.6	**85.4**
MFCCs	27.3	35.9	20.5	14.0	55.3	68.7
MFCCs + deltas	48.3	19.1	28.3	29.5	54.7	65.9
Activation	34.0	24.4	38.2	24.4	33.3	52.5

From these results, it was evident that some combinations of features and classification algorithms fare better than others. Because the top score was initially found to be the combination of minmax + deltas and the first CNN, this combination was chosen for hyperparameter optimization, giving the total highest score of 85.4% using the same features and the new CNN2 architecture. It is important to note that out of all the training, the GMM-HMM method was much faster than the other methods. Calculating the accuracy over 10 trials for all the different feature classes took about 45 minutes using the GMM-HMMs, compared to 5 hours calculating only two trials for each of the RNNs on the same features.

Furthermore, there are several additional interesting observations to be found in Table 14. For instance, notice that the GMM-HMM and CNN2 solutions both scored well on all feature extraction methods, even on the raw data. When it comes to adding additional features

1 A table including the standard deviation for all values can be found in Appendix A.3, as those were omitted in this section because of limited space.

in the form of time-deltas, this increases the recognition rate in most cases, but notably not in all. The simpleRNN, for example, achieved a much lower score when adding the MFCC deltas. The minmax + deltas features gave the best results overall, while the raw data and activation signal performed the worst, but these results show that this effect can be mitigated by the choice of classification method.

Cross-session results

A cross-session accuracy matrix was calculated to investigate the difference between different sessions further. Every session was used both as training and test data for the CNN2 architecture, and the resulting accuracies are presented in Figure 19. As expected, the diagonal representing the inter-session results shows the highest accuracies, but there are also some signs of clusters between certain sessions. If two sessions result in high accuracies, it is visible in the form of stronger blue colors as a mirror image on both sides of the diagonal, possibly indicating that the sensor placement was close to similar between those sessions. This is most visible for sessions 3-2 and 3-4, as well as 3-12 and 3-13. As both of these examples are close in time, other factors than sensor placement might very well be important as well.

Selecting a subset of electrodes

The results from experiments on subsets of electrodes are presented in three sections. First, Figure 20 shows the accuracy of one and one electrode when used on the Corpus 3 dataset with 10 trials each. Secondly, a diagonal heatmap is shown in Figure 21, with pairs of two and two electrodes used. Notice that the diagonal shows the average accuracy for all combinations for each electrode. Information above the diagonal would be the same as below, as using electrode 1 and 3 is the same as using 3 and 1, and is therefore not shown. Lastly, the accuracy of classification based on the left vs. right side of the face is presented for each of the 15 sessions if Corpus 3 in Figure 22.

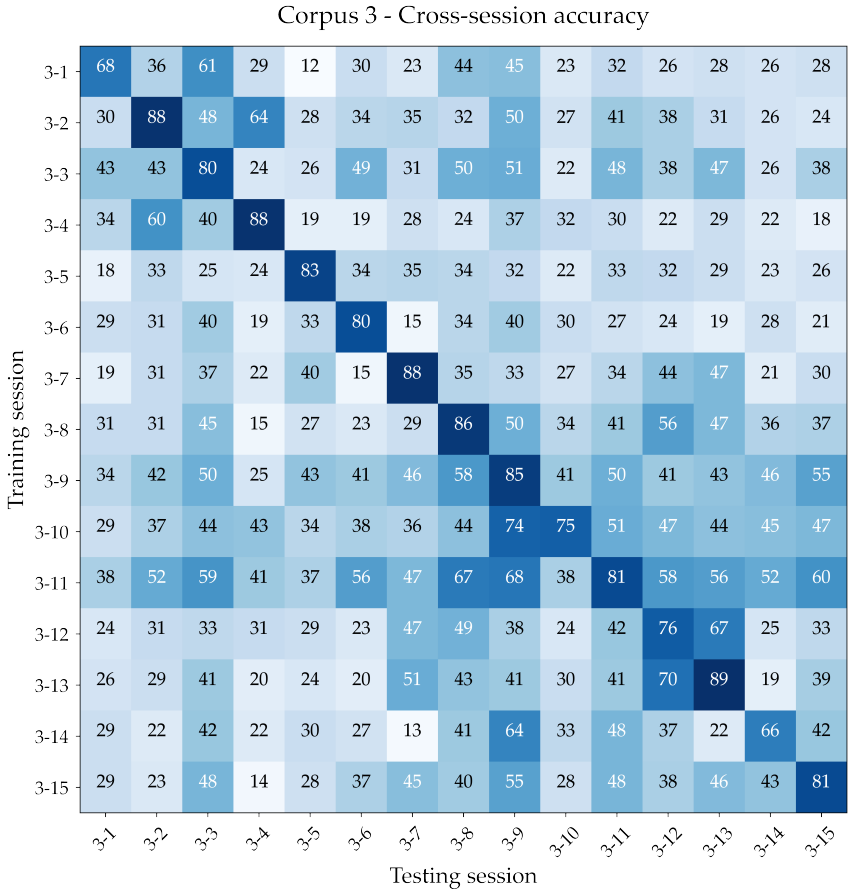


Figure 19.: The calculated classification accuracy [in %] for the 15 different sessions of Corpus 3. Each row represents the session the CNN2 model was trained on, while the columns corresponds to which session was used as the test set.

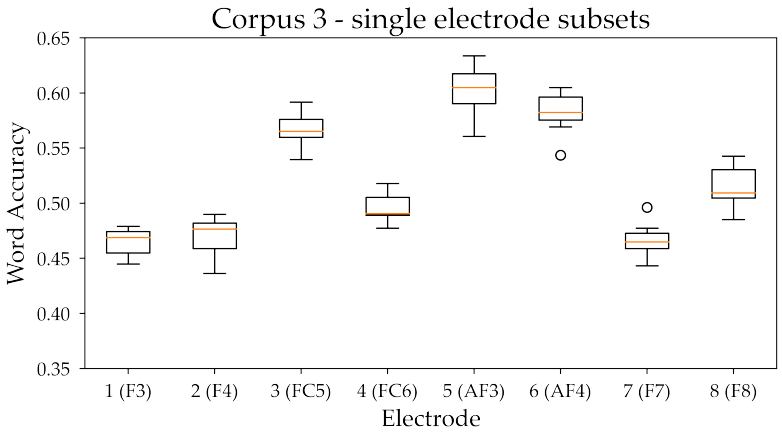


Figure 20.: Boxplot of the classification results for Corpus 3 using CNN2 model and only data from one and one electrode.

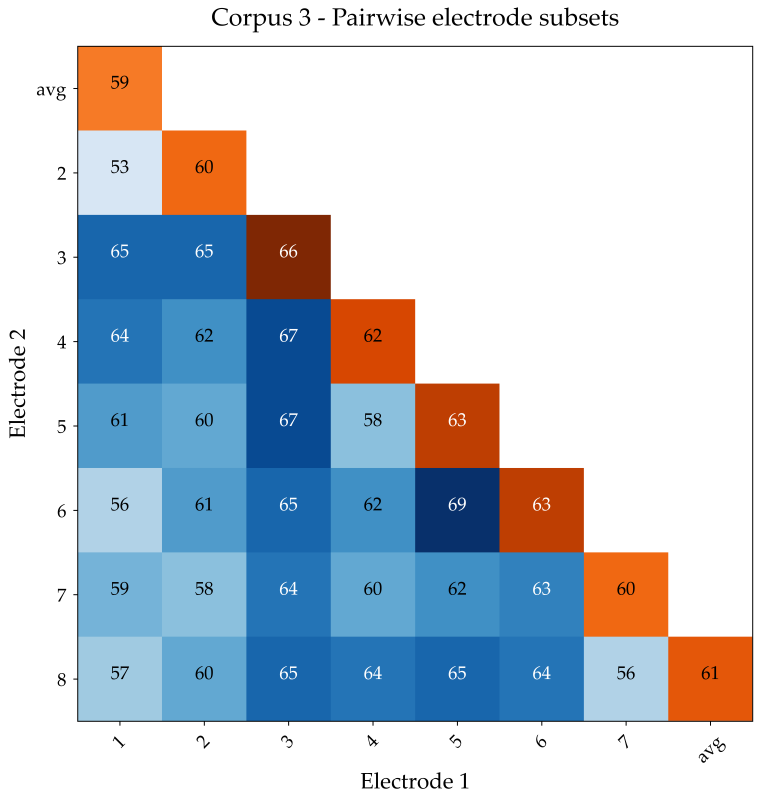


Figure 21.: The calculated classification accuracy using the CNN2 model using pairs of electrodes. The diagonal (in hues of orange) shows the average values for every pair that includes that electrode.

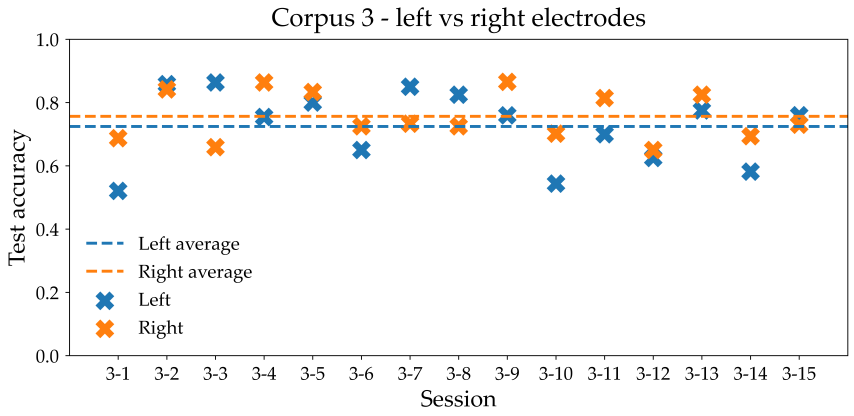


Figure 22.: The calculated classification accuracy using the CNN2 model on either the left (electrodes 2, 4, 6 and 8) or right (electrodes 1, 3, 5 and 7) side of the face. Calculated for each of the 15 sessions in Corpus 3. Average values for each of the left (72.5%) and right (75.7%) is included as stippled lines.

Observe that out of the single electrodes, AF3 and AF4 are the two with the highest word accuracy and thus the most important for EMG-based speech recognition out of the electrodes available using the Emotiv Epoc+. These two sensors are the ones placed on the lower lip, which makes sense as a lot of speech movement is present in this area. F3 and F4, which are placed on the throat, are the two that contributes the least to a reliable word classification. These results are backed by the pairwise electrode subsets, where the overall highest score is with the combination of AF3 and AF4, while the lowest score comes from combining F3 and F4. Electrode FC5, located on the lower right jaw, looks to be of great importance as well, with a high single electrode word accuracy and high accuracies overall when combined with other electrodes as viewed in Figure 21. No definitive difference between the left and right side is observed as this is highly dependent on each of the single sessions, see Figure 22. Nevertheless, based on Figure 20 it seems electrodes FC5, FC6, F7, and F8 show signs of some asymmetry in facial muscle activation during speech for the author.

4.3.2 *Audio-based recognition*

As a comparative measure, the audio signals collected as part of Corpus 3 were classified in the same way as the EMG data. By using MFCC features for each of the uttered digits, both the original CNN and the GMM-HMM methods were tested with 10 train/test splits. The CNN model achieved a 99.4($\pm$0.2)% accuracy on session independent single word recognition of the 10 digits, while the corresponding number was 98.9($\pm$0.2)% for the GMM-HMM. These results show that by using features designed for speech recognition, outstanding accuracy can be achieved on session independent single word recognition of vocalized speech with 6430 samples.

4.4 CORPUS 4

4.4.1 *Recognition rate*

Corpus 4 has more words in its vocabulary and fewer samples per word than the other corpora. Thus the resulting accuracies are expected to be lower. As with Corpus 3, all six different classification methods were used on a per session basis on prepared datasets using the *minmax + deltas* features. Every experiment was run with 10 trials, and the results are presented in boxplots. Word classification accuracies for the CNN and CNN2 architectures are presented in Figure 23, while Figure 24 presents the corresponding results for the simpleRNN and GMM-HMM. Results for the LSTM and GRU NNs can be found in Appendix A.2.4.

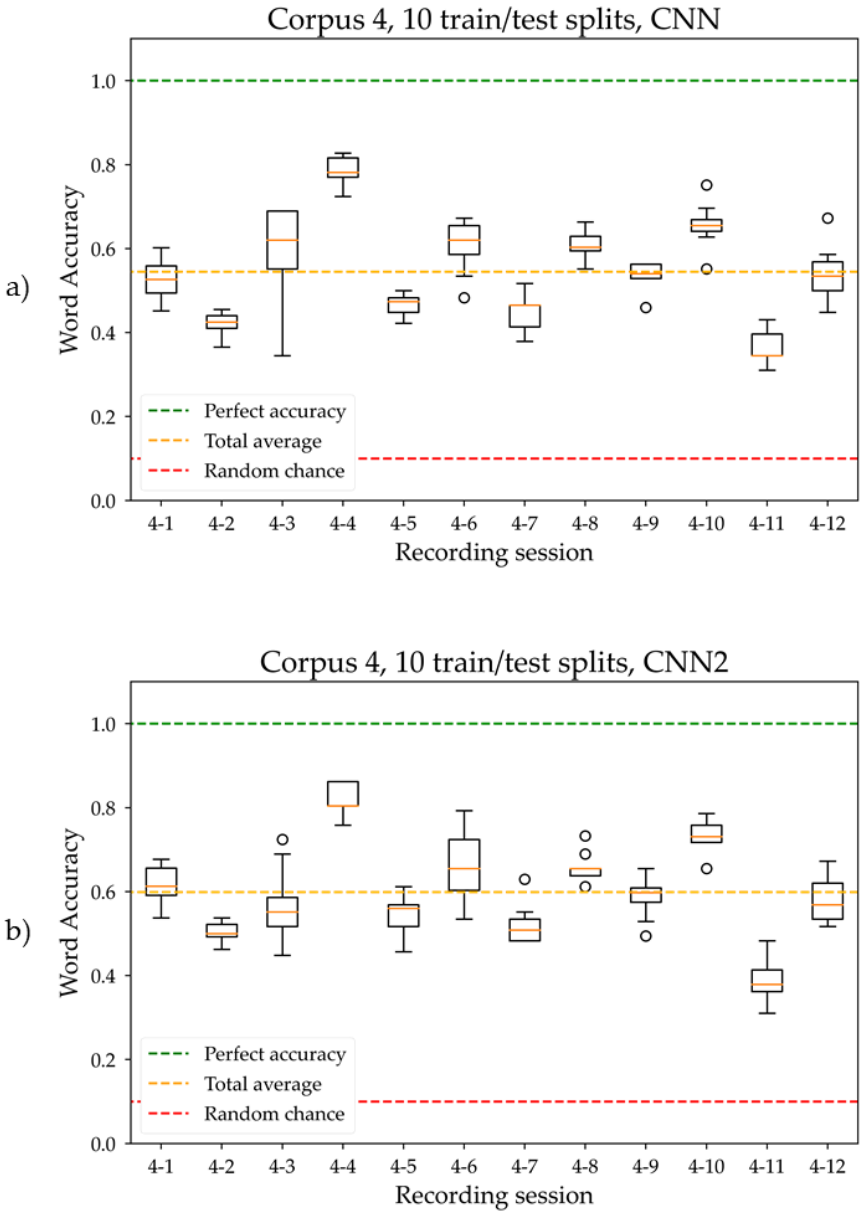


Figure 23.: Boxplots of the classification results for Corpus 4 using the **a)** CNN and **b)** CNN2 architectures. Total average accuracies of 54.5% for the CNN and 59.9% for the CNN2 are marked with the orange stippled lines.

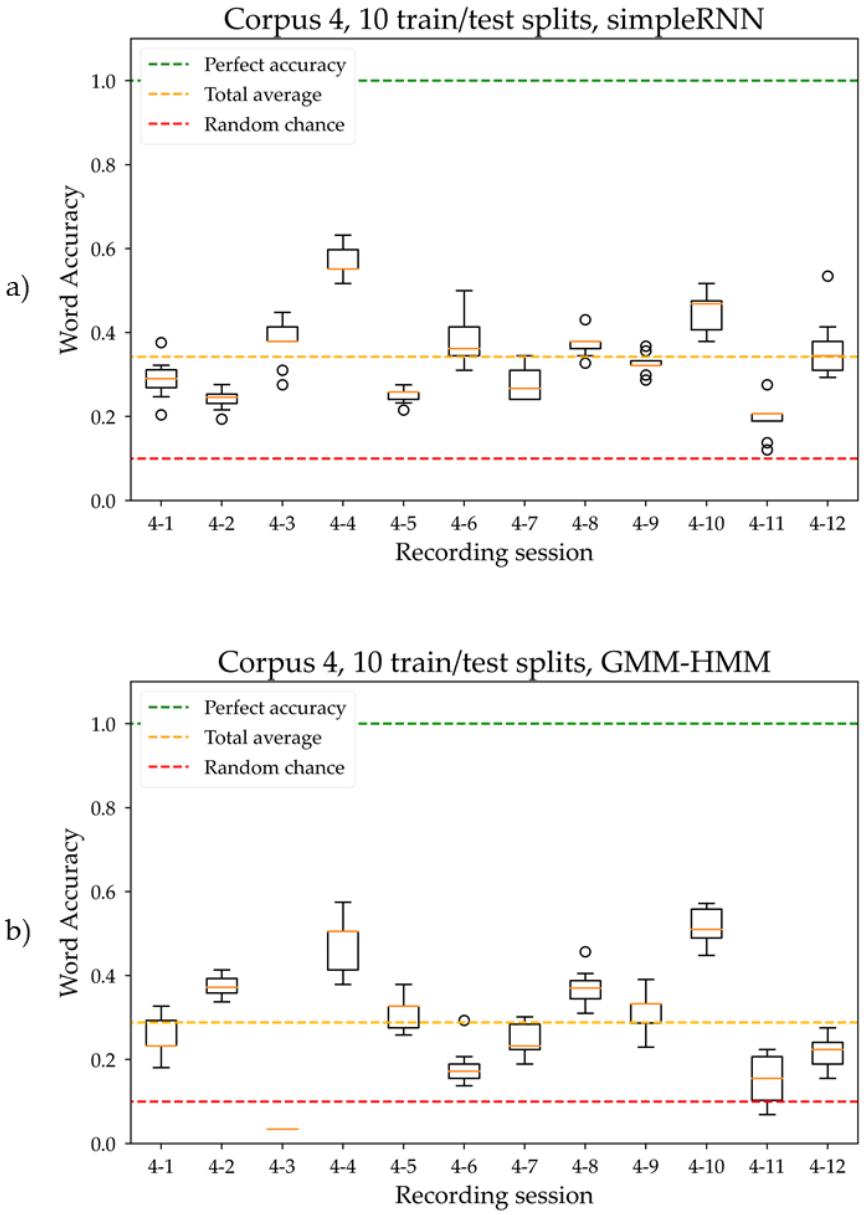


Figure 24.: Boxplots of the classification results for Corpus 4 using the **a)** simpleRNN and **b)** GMM-HMM architectures. Total average accuracies of 34.2% for the simpleRNN and 28.8% for the GMM-HMM are marked with the orange stippled lines.

As described in Section 3.6 in the Materials and Methods chapter, two different angles were approached to communicate six test sentences using the Emotiv sensor. One using the combined vocabulary of corpora 3 and 4, totaling 39 words to spell each character of the six sentences out. The other by utilizing a pre-trained speech synthesizer and training the EMG-Net model to transform EMG data to mel-spectra.

4.5.1 *EMG-to-text by spelling*

Combining the vocabularies of corpora 3 and 4 gives 39 different classes for classification. It was therefore relatively likely that a model trained on both corpora would have somewhat lower accuracy than one trained on Corpus 4 alone, which had an average accuracy of 59.9% over all 12 sessions. Ten different CNN2 models were trained on different train/test splits of the combined data of corpora 3 and 4. All resulted in accuracies between 58.7 and 64.8%, with the average being 63.2%, higher than expected. A confusion matrix of the first of those models is presented in Figure 25.

The ten models already trained on corpora 3 and 4 were then used to predict the sequence of characters in the six test sentences listed in Table 9. 120 sentences were thus generated (6 sentences spoken both with vocalization and silently, times 10 models). Even though the average accuracy over all 120 sentences was 53.8%, the resulting sentences were close to impossible to interpret. Notably, the accuracy was higher for the silently spoken sentences compared to the vocalized ones (58.4 to 49.1%). The three sentences with the highest accuracy are presented in Table 15, and with accuracies above 70%, it is barely possible to understand the sentences if the six original sentences are known beforehand. This showed that much higher accuracies or further treatment of the model predictions were needed for a functional EMG-based spelling-SSI.

Table 15.: The three predicted sentences with the highest word classification accuracy out of the 120 sentences generated.

Accuracy	Predicted sentence
73.5%	f8 kn8ghtk ro8e8up the87tcep h,ll.
73.5%	98 ynyoht7 3ode up the sttep jdll.
72.3%	on hay 5th 2621 5paced 9ue3e982ully 3anded an158

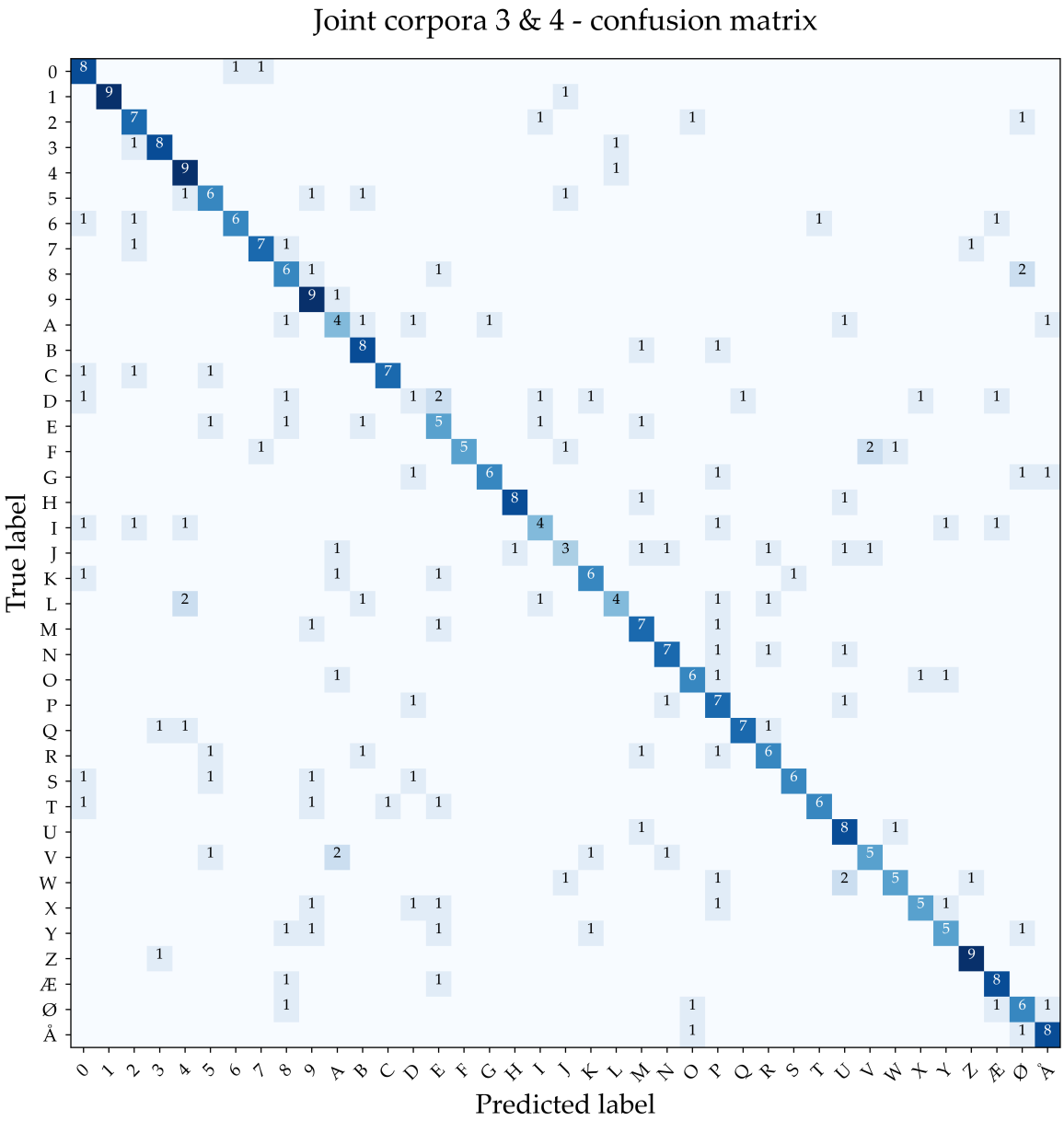


Figure 25.: Confusion matrix of the joint vocabularies of corpora 3 and 4. The CNN2 model was trained on data from both corpora and then tested on 10 words for each class of previously unseen data. This confusion matrix corresponds to an accuracy of 63.3%.

To increase the character-based accuracy, a simple selector was constructed that for each character in the six test sentences chose the character predicted by most of the ten models. This resulted in an average accuracy of 80.5%. With the addition of a simple spell checker from the Python *spellchecker* library [46], a couple of spelling mistakes were corrected, and the average accuracy rose to 82.7% over all six sentences, ranging from 75 to 100%. Table 16 presents the final predictions of the six test sentences and the per-character accuracy for each of them.

Table 16.: Predictions for each of the six test sentences and the accuracy of correctly classified characters.

No.	Accuracy	Predicted sentence
1	85.3%	78 knights rode8up tjtesteep jill.
2	100%	the answer to life and the universe is 42.
3	76.6%	on hay 5th 2621 spacex 9uc3e98full81landed 9n15.
4	75.0%	pa equals alpcoeim2tel8 321415t.
5	83.3%	orond1eim has its own j2zzofest9val8
6	75.9%	roughly 2so2t.ol9nesoof code were written foroth.5 fjedis.

4.5.2 *EMG-to-speech*

Both corpora 3 and 5 were used in the effort to develop a functional EMG-to-speech SSI, as those were the two datasets with matching EMG and audio data.

Digits

Audio files for two recordings of 10 digits each were synthesized using the EMG-Net and WaveGlow architectures.[2] The resulting recognition rate after 10 volunteers had listened to the audio files were an average of 73.5% (±2.4%). Observe that the 10 last digits were much more intelligible than the 10 first, indicating that one of the EMG recordings was easier for the model to recognize based on the available training data. As both recordings were selected at random and not used for training, this might result from which recording session they were a part of. Notice further that a few of the samples were recognized as digits by the listeners but not the digits that they were supposed to sound like. These digits were also more likely to be recognized differently between listeners. For instance was sample 5 (the 5th sound in the linked audio file) recognized as

2 The 20 synthesized digits are available as a single audio file online: https://drive.google.com/file/d/1MxNySKPeRXpOIJx2d4PcHRVWhgNRw34l/view?usp=sharing

1, 5, 7, 9, and 'other' by different listeners, while the actual digit was supposed to be 6. This shows that the model sometimes generates sounds between different digits, making it difficult to correctly classify what was said.

To represent the synthesized audio visually, mel-spectra for each of the 10 digits in the second recording are compared with the corresponding mel-spectra from the actual audio recording in Figure 26. From these mel-spectra it is clear that the general shape and pattern of the mel-spectra are well represented in the synthesized digits, but they lack some of the resolution. However, this was close enough because all 10 digits shown in Figure 26 were correctly classified by every one of the listeners except for one who classified 9 of them correctly.

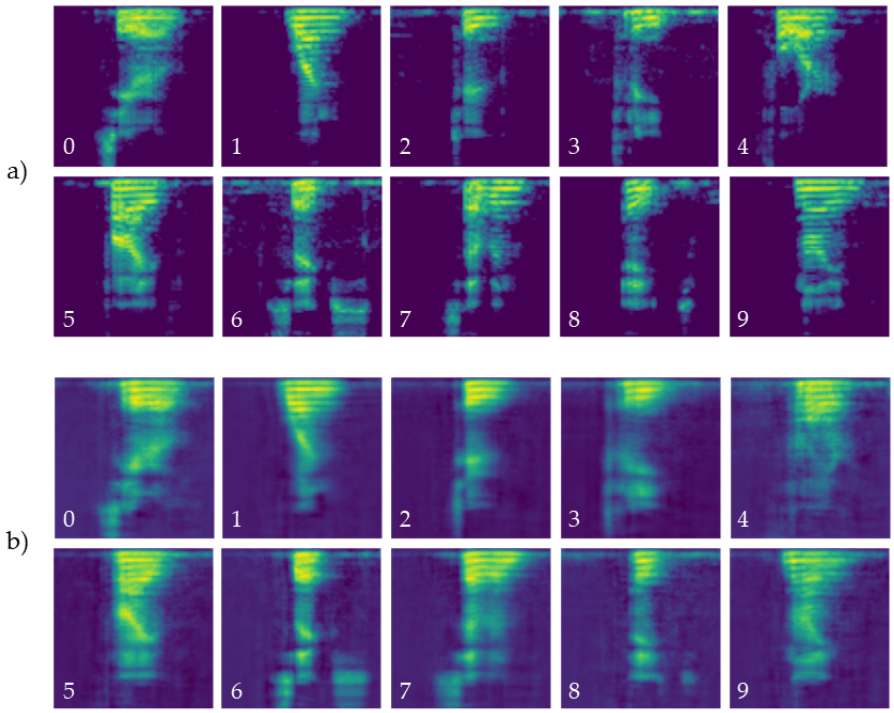


Figure 26.: A comparison of **a)** mel-spectra from audible speech waveforms and **b)** mel-spectra generated from EMG data using EMG-Net. The spectra are sorted based on what digit they represent and the generated mel-spectra are from the last 10 digits from listening test. 80 mel filterbanks were used (y-axis) and the 1 second recording resulted in 79 time windows along the x-axis.

Unlimited vocabulary - test sentences

As mentioned in Section 3.6.2, there was unfortunately not enough time before the deadline to synthesize any intelligible speech from

the six test sentences using Corpus 5 as the training data. The model predicted almost the same mel-spectra for both window lengths independently of the input EMG data, resulting in more or less static background noise.

5 DISCUSSION

Many different results have been presented in this thesis. To keep track of the most important results, this chapter starts with a short summary.

Single-word silent speech recognition has been proven with vocabularies of 3, 10, 29, and 39 words. The classification accuracy was naturally highest for the smallest vocabulary of 3 words, with 93.3% accuracy averaging all four speakers. The average accuracy on the vocabulary of 10 words was 85.4% with a session-independent model and 63.2% for the session-independent model with a vocabulary of 39 words. These results were only achieved after significant efforts were put into selecting the best features and models for an EMG-to-text SSIs. This effort resulted in a functional SSI that, with an average precision of 82.7%, translated facial muscle activation during the spelling of six test sentences into text. One of the sentences was even perfectly predicted, showing real potential for such a system.

Functional EMG-to-speech should also be possible, as shown by Gaddy and Klein [19]. For this thesis, a unique model architecture has been created named EMG-Net that can synthesize digits with high intelligibility when combined with a pre-trained WaveGlow model. EMG-Net was further trained on data from Corpus 5, but no intelligibility was achieved on the test sentences showing that there is still a long way to go for a practical SSI based on this EMG-to-speech system. This chapter compares the achieved results with previous studies on EMG-based silent speech, comments interesting observations from throughout the project, and takes a turn into the topic of possible signal artifacts.

5.1 SESSION DEPENDENT RESULTS

Using Corpus 1, the four initial classification algorithms gave recognition rates above 80% for all speakers, with a recognition rate as high as 97.3% on Speaker 4. The difference between speakers might result from naturally more precise articulation leading to more recognizable silent speech, less shifting of sensors during recording, or other, unknown factors. Overfitting was observed for most training sessions as the training accuracy almost always reached $95 - 100\%$. Additional variance between each train/test split was observed, and changing from a static learning rate to one that decreased four

times during training for all three RNNs definitively helped in this regard. The plotted train and test loss before and after introducing a dynamic learning rate is seen in Figure 27, which shows a drastic change in the smoothness of the loss. The same change from a static to dynamic learning rate did not significantly impact the CNN.

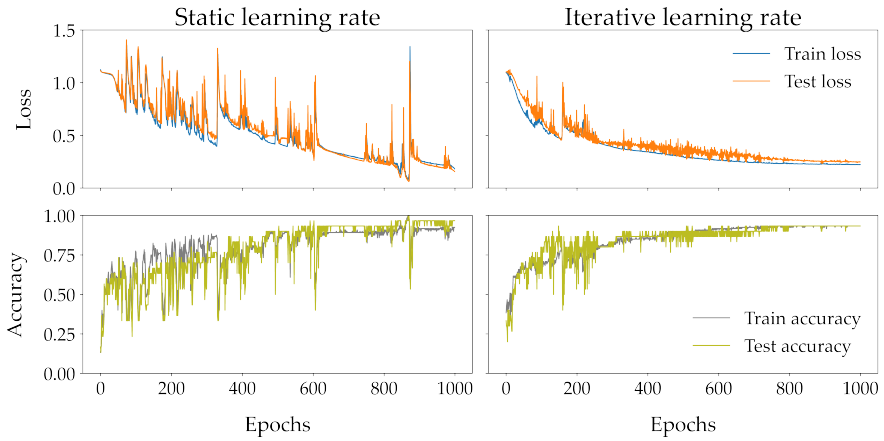


Figure 27.: Plots of the loss and accuracy of both the train and test sets for the static learning rate that was initially used (left) and the dynamic learning rate (left). The data is from the LSTM algorithm on Corpus 1.

However, with vocabularies larger than three words, the CNN architectures and the GMM-HMM model gave overall better results than the RNN models. The reason for this is unknown, but one possibility might be that more time was spent optimizing the initial CNN method, and extensive hyperparameter optimization was only conducted for the CNN2 architecture. Another possibility could be the large window sizes of 1 second used for recognition. Most other speech recognition systems, be it for audible or silent speech, use shorter windows of time and then combine data from multiple windows to predict the correct phonemes or words. Because RNNs are notorious for the exploding/vanishing gradient problem, using only every eight samples in time is probably not the best solution. Interestingly, both the LSTM and GRU models performed worse than the simple RNN for corpora 2 through 4, even though these models were constructed to fare better than the simple RNN against the exploding/vanishing gradient problem. The reason for this is likely the fact that many more nodes were used for the simple RNN. One might then ask why not more nodes were added to the LSTM and GRU models. This decision was made because these two models were much slower to train than the simple RNN when using the same number of nodes.

Interestingly, the relative accuracies between different recording sessions were quite consistent independently of the classification method, despite the differences in overall performance. For Corpus 2, four of the sessions achieved 20 percentage points higher recognition rates when averaging over all classification methods than the remaining three. The same tendency is present for corpora 3 and 4, where some sessions consistently result in higher accuracy. This phenomenon was observed when calculating the cross-session accuracy presented in Figure 19 in the Results chapter. There, some sessions gave much higher accuracies, and signs of clusters between individual sessions were seen. Figure 28 visualizes this session inconsistency by showing the average accuracy over 10 trials for each of the 15 sessions in Corpus 3 using all the developed classification methods. The same trend is visible independently of the classification method, showing that some sessions are doubtlessly more suited for single-word classification than others.

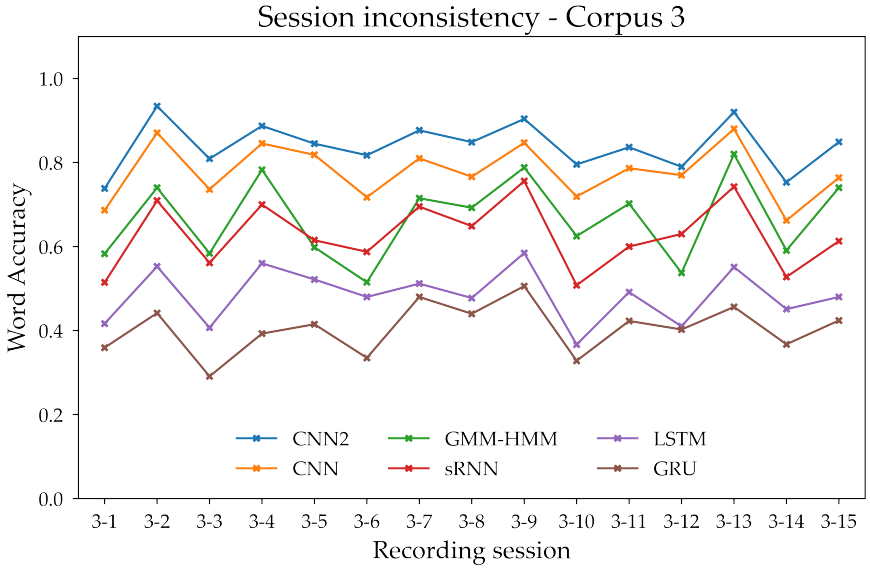


Figure 28.: Average accuracies for the six different classification methods on each of the 15 recording sessions of Corpus 3. A clear trend between accuracy and session is seen independently of the method used.

This inconsistency between sessions brings the question of why this has happened and whether it is possible to make all future recording sessions better by changing the method slightly. The length of each recording session was decided by pragmatism rather than a set sample goal. This decision resulted in a different number of samples available for each session, an apparent possible explanation, as more training data is known to give better recognition. However, as seen in Figure 29, the trend between the number of samples and the average

accuracy for each session is not very obvious. The data used is from the average accuracies from the initial CNN model on sessions from corpora 2 through 4. Regression lines were added for each corpus, with R-values of 0.58, 0.17, and −0.02, indicating a slightly positive correlations for Corpus 2 but not for the others.[1]

Other factors must therefore be considered as well. Possible explanations include irregular cleaning of the sensor, sensor movement during a session, and how well the original sensor placement of each session matches muscle activation features easily recognized. Another possibility is the prospect that the subject might be getting more used to performing silent speech over time, as has been previously reported by Wand [65] and Meltzner et al. [50].

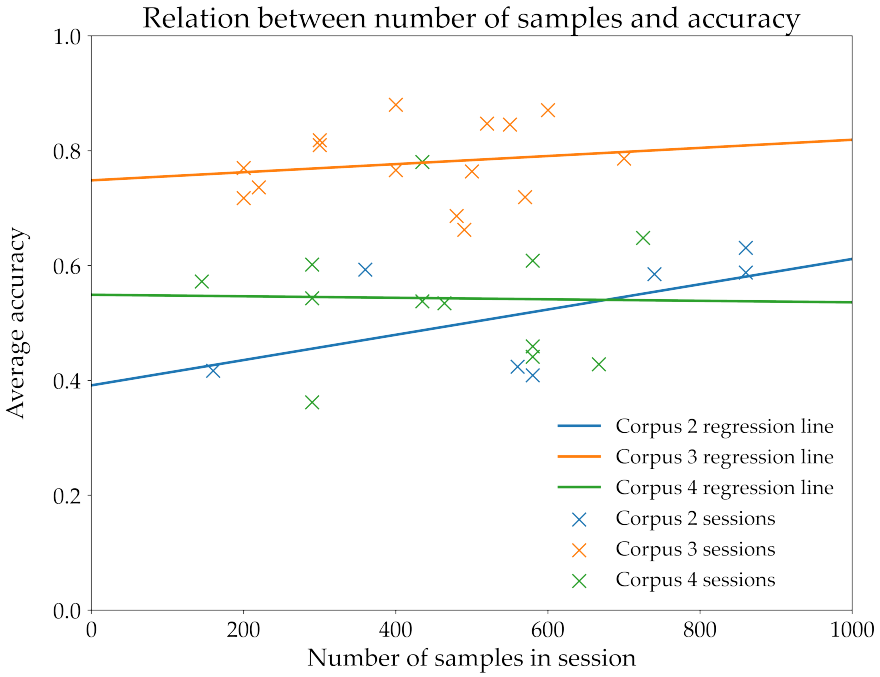


Figure 29.: Scatter plot of the number of samples and average word accuracy for the all sessions of corpora 2 through 4. Regression lines are added for each corpus.

5.2 SESSION AND SPEAKER INDEPENDENT RESULTS

Challenges with session- and speaker independence are regarded as two of the most critical barriers for functional EMG-based silent speech interfaces [16, Chapter 1.5]. The most in-depth analysis of session independence found during the literature review for this

1 A straight line will not be accurate for this data when extended further to the sides but might indicate a trend nonetheless. Zero samples would always lead to no recognition, and one would never get above a recognition rate of 100%.

thesis was Wand [65, Chapter 8.1], where the use of 7 and 15 sessions as training data were compared. There, the author found that the average accuracy on single word classification using a vocabulary of 108 words increased from 65.9% using 7 sessions to 73.0% using 15 sessions. This was compared to 85.0% accuracy using a session-dependent classification system.

Initial results using Corpus 1 regarding speaker independence showed no recognition above random chance (Table 11). Using Corpus 1B, at least some recognition independently of sessions were found (Table 12), notably with very little available data. Using Corpus 2, more results were collected on session independence using 7 sessions and training on an iteratively increasing number of sessions. Those results showed an average of 25.9% accuracy using 1 session, 33.1% using 6, and 64.6% on a session-dependent system, indicating the same tendency as in Wand [65, Chapter 8.1]. However, there were obvious drops in accuracy after adding certain sessions to the pool of training data, seen in Figure 16. It is believed that this might be the result of the session inconsistency effect discussed in Section 5.1, where data from some sessions are better suited for EMG-to-text classification, and training with data from the "bad" sessions therefore throws the learning algorithm off. This conclusion is nonetheless speculative, and the accuracy highly depends on what data are used for testing. The best approach is likely to use as many sessions for training data as possible to generalize the recognition to future sessions, as needed for a session independent system.

After the implementation of a completely new, more streamlined data collection program, it was possible to collect more data for the remaining datasets. Therefore, it was believed that corpora 3 and 4, with 15 and 12 sessions respectively, would result in higher degrees of session independence. This was indeed the effect as the average accuracy using the CNN2 architecture was, remarkably, higher when training on all the sessions of Corpus 3 with 85.4% than when averaging the accuracies over single sessions, which was 83.9%. It should be noted that the highest average accuracy when looking only at the best-performing single session within Corpus 3 was as high as 93.4%, as seen for Session 3-2 in Figure 28. This high accuracy on unseen sessions enabled the functional EMG-to-text spelling SSI to be completely session-independent, working well on data recorded independently of the main corpora.

With that being said, it is important to acknowledge the fact that from Corpus 2 to 3, the accuracy increased significantly even with the same vocabulary and sessions of approximately the same number of samples, clearly seen in Figure 29. One possible reason for this

might be the fact that Corpus 2 consisted of silently spoken digits while Corpus 3 was vocalized. This effect is previously seen in other studies such as Maier-Hein et al. [48], where the authors argue it might be linked to a stronger muscle activation for vocalized than silent speech. This difference between muscle activation during silent and vocalized speech is also highly problematized by Gaddy and Klein [19], who used dynamic time warping to align EMG data from silent and vocalized speech. However, other factors must play a role, for Corpus 4 is silently spoken and still achieves higher accuracies than Corpus 2, even with almost triple the number of words in its vocabulary. A combination of the consistent sampling rate of 256 Hz, the author becoming more used to silent speech, and the fact that the NATO phonetic alphabet is to distinguish the different words easily, is believed to be the reason.

Results from this project on speaker and session independence are clearly in line with previous research where session independence on one speaker seems to be an easier challenge to solve than speaker independence. Both Maier-Hein et al. [48] and Wand [65] have reported some session independence by having more data or using specific methods to increase the recognition between sessions, while no work has been found that claims to have solved the challenge of speaker dependence. As described in 2009 by Denby et al. [9], the situation of speaker independence using EMG might be pretty different from audio-based speech recognition, as the features depend directly on the speaker's anatomy and the exact synaptic coding inherent of her/his articulatory muscles. It might, however, be solvable with more data as today's audio-based speech recognition systems are based on many orders of magnitude more training data than even the most extensive EMG-based system. In Meltzner et al. [50], the authors write that they believe:

> "... training effective subject-independent models will require an additional data-set of subvocal speech recorded from a large and diverse population representative of typical end-users. Once obtained, these data could be combined with recent Deep Learning algorithms that have advanced the state of acoustic ASR to human recognition levels (and is the basis for recognition capabilities of the commercial virtual assistants such as Siri, Alexa, and Cortana). Using a deep-learning approach for each desired user, a subject-specific model could be created by using a small amount of that subjects' data to adapt the network weights of the much larger average deep neural network baseline model."

To be able to collect such a large and diverse data set as they describe, it is first necessary to find a common practice between research com-

munities on how to collect and process EMG data for silent speech. The Emotiv Epoc+, or preferably a sensor actually made for facial EMG measurements, might be a potential candidate for such a common platform.

5.3 DIRECT COMPARISON WITH OTHER STUDIES

5.3.1 *Single word classification*

Section 2.7 in the Theory chapter includes a table listing the single-word classification results from previously published work in EMG-to-text. Table 17 shows the same table, including the results from this thesis. This place results in this work on par with many of the previous results, even from more recent studies. The accuracy achieved for this thesis is lower than others have achieved on larger vocabularies but include session-independency, an important factor for practical SSIs.

Table 17.: Comparison of single word classification results from this project with similar previous work.

Source	Year	Vocabulary	Accuracy [%]	Session independent
[61]	1985	5	64	No
[53]	1991	10	60	No
[3]	2001	10	93	No
[33]	2005	10	73	No
[48]	2005	10	97.4	No
[48]	2005	10	76.2	Yes
[35]	2006	108	68	No
[65]	2014	108	85	No
[65]	2014	108	73	Yes
[60]	2017	5	64.7	No
[38]	2018	10	92	No
[50]	2018	65	90.4	No
[47]	2019	10	72	No
[71]	2020	10	79.5	No
[72]	2020	10	93	No
This project	2020	3	93.3	No
This project	2021	10	85.4	Yes
This project	2021	39	63.2	Yes

5.3.2 *Electrode subsets and optimal placement*

This thesis is not the only study in electrode subsets and optimal sensor placement. Most notably, a group from the Shenzhen Institutes of Advanced Technology ran experiments using 120 different electrodes resulting in two papers describing the effects of electrode placement [72] and the number of electrodes used [66]. They found that when divided into three groups, the electrodes on the whole neck resulted in the highest accuracy compared with electrodes on each side of the face or only in the center of the neck. Unfortunately, they did not place any electrodes below the lower lip's sides, which was the most important electrode placement in this thesis. A more easily comparable electrode setup was that of Wadkins [64], who reported that out of their 8 electrodes, the one directly below the chin was the most important. The most closely matched electrodes on the Emotiv Epoc sensor to that are F3 and F4 (electrodes 1 and 2), which were deemed least important in the experiments related to subsets of electrodes, as seen in Section 4.3.1. These inconsistencies show that there is no current best practice on electrode placement and that it might be dependent on the speaker and the types of electrodes used.

Wang et al. [66] found, as expected, that more sensors are on average always better, but that the return of higher accuracy with more electrodes diminishes quickly. By using 8 selected electrodes, they were able to achieve above 90% accuracy on a vocabulary of 10 words (as well as a class for 'silence'), while 15 electrodes were needed for 95% accuracy. Meltzner et al. [50] tested 12 different pairs of electrodes and found that no more accuracy was gained above 8. It seems therefore that the 8 electrodes of the Emotiv Epoc sensor should be enough for good results, but doubling the number would probably lead to better performance. Almost all other studies use either only electrodes on one half of the face or a non-symmetrical setup of electrodes, so an optimal EMG-headset for silent speech would likely follow that pattern.

5.4 THE EFFECT OF SIGNAL ARTIFACTS

Most biological electrical sensors are prone to artifacts. The used Emotiv EEG-headset possibly even more so as it is originally made for EEG brain measurements where all of the sensors are fixed relative to each other. As the same reference point is used, moving an electrode relative to the others can affect both the signal for that electrode as well as for the other electrodes, as seen in Figure 30. Early on in this project, two simple experiments were performed to uncover whether the sensor was capable of measuring muscle activity, not only relative movement of the electrodes. The first was by moving the tongue

inside the mouth while the face, and thus the sensor electrodes, were stationary. The second was by clenching the jaw so that facial muscles were active without any movement. Plots from both these tests are presented in Figures 31 and 32.

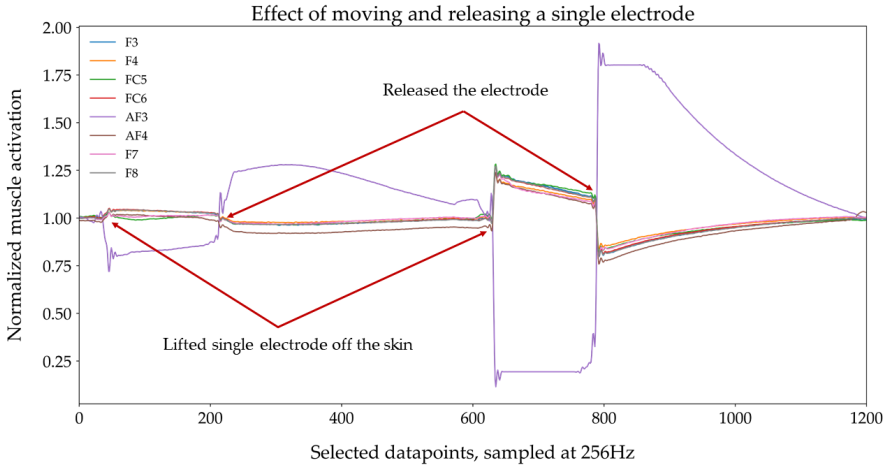


Figure 30.: A graph showing the effect on the Emotiv sensor signal when lifting and then releasing one of the electrodes.

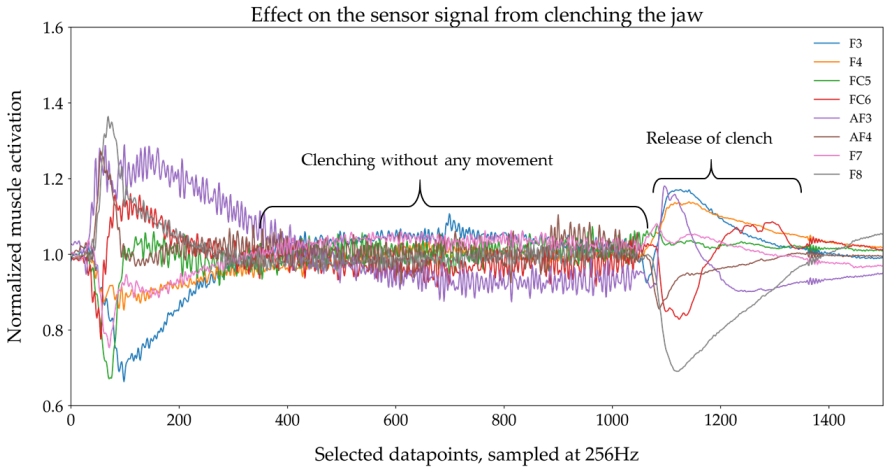


Figure 31.: The effect of clenching while wearing the Emotiv sensor. Note the high-frequent signal while clenching without any movement.

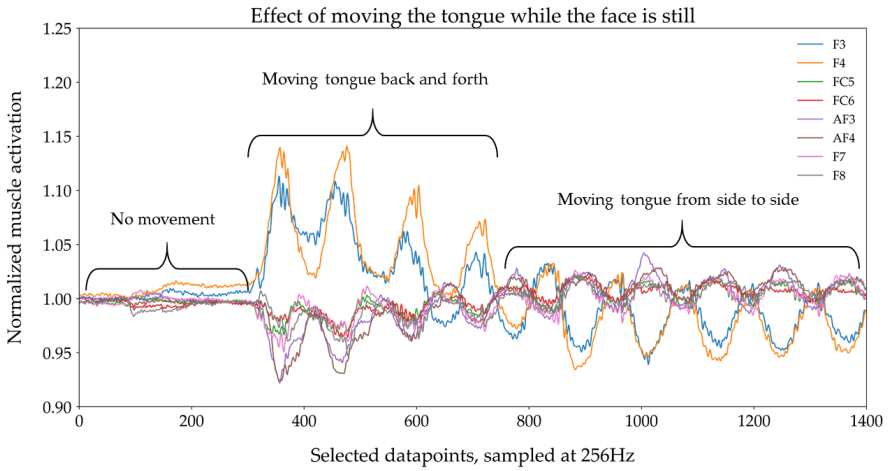


Figure 32.: The effect of moving the tongue while wearing the Emotiv sensor and keeping the face still. Note the more low-frequent signal of the movement, even though none of the sensor electrodes moved relative to each other. The two electrodes with the most visible activation (F3 and F4) are the ones placed on the throat.

There are additional artifacts present when using the sensor. The most apparent is the effect of facial movements during recordings that are not related to speech, such as swallowing and facial expressions. These artifacts are not exclusive to the Emotiv sensor and combine to a general challenge with EMG as the modality for silent speech. Nonetheless, the fact that the sensor picks up on all facial movements might in the future become a useful feature as it can supplement a silent speech interface (SSI) with information about the user's mood and focus.

6 CONCLUSION AND FUTURE DIRECTIONS

In this study, an Emotiv Epoc+ EEG headset has been used in a completely novel way as the interface for two different possible EMG silent speech interfaces (SSIs). Because this sensor is more readily available than medical-grade EMG sensors and results are comparable to other work using more expensive equipment, it has been proposed as a common platform for future EMG silent speech research. Both proposed SSIs were shown to work conceptually and can, with more time and effort, potentially be very useful in a wide range of situations. The highest single trial accuracy on Corpus 3 was 97.1% on a 10-word vocabulary, higher than most of the reported values of recent studies on the same topic, showing great potential towards EMG-to-text silent speech recognition. By synthesizing speech with intelligibility resulting in listeners correctly classifying 73.5% of the digits in the same 10-word vocabulary, functional EMG-to-speech is also shown to be possible with the Emotiv headset.

The objectives for this project were to:

(1) Analyze whether the chosen sensor can be used to recognize silently spoken single words from a small vocabulary.

(2) Enable a method for efficient collection of both EMG and audio datasets while using the Emotiv Epoc+.

(3) Discover types of classification and feature extraction methods that work well with the available data.

(4) Work on understanding the challenges connected to session independence and the potential for a direct EMG-to-speech solution.

and throughout this thesis, all four have been fulfilled. Much work remains to achieve higher recognition rates on larger vocabularies and with more speakers. Still, the results so far are well in line with previous research and create a solid foundation for future improvements. This thesis has thus contributed to a future where the immense powers of spoken language might soon be available for more people and in more situations through functional silent speech interfaces (SSIs).

A sound basis for two different SSIs has been created. For a more functional EMG-to-text SSI, a phoneme-based solution is probably needed, enabling a much larger vocabulary that can be used in real-time. Spelling out each character in a sentence works but is unsuitable for most applications. Better deep learning architectures, more training data, and smarter processing of the data on both sides of the model in a system pipeline would presumably increase performance further.

Because of limitations in time, the possibilities of an EMG-to-speech interface based on Corpus 5 were not explored as deeply as hoped for. Designing more suitable sequence-to-sequence models for transforming EMG data either to mel-spectra or directly to speech waveforms will most likely result in models that perform much better and can be used on an unlimited vocabulary. Looking more into smaller windows in time and better synchronization between EMG and audio data for the EMG-to-speech solution will also likely be necessary.

APPENDIX

A.1 LSTM AND GRU NETWORK STRUCTURES

Architecture 5: Layers of the LTSM-RNN that was used for classification.

```

from tensorflow.keras.layers import Input, LSTM, Dense, GlobalMaxPool1D
from tensorflow.keras.models import Model
from tensorflow.keras.optimizers import Adam

# Remove most of the data (leave every 8th timestep)
X = X[:,0:250:8,:]

x_train, x_test, y_train, y_test = train_test_split(X, Y, stratify=Y,
    test_size=0.2)

# Build the LSTM-RNN model
K = 10   # number of outputs (categories for classification)
M = 5    # number of hidden nodes

i = Input(shape=x_train[0].shape)
x = LSTM(M, return_sequences=True)(i)
x = GlobalMaxPool1D()(x)
x = Dense(K, activation='softmax')(x)

model = Model(i, x)
```

Architecture 6: Layers of the GRU-RNN that was used for classification.

```
from tensorflow.keras.layers import Input, GRU, Dense
from tensorflow.keras.models import Model
from tensorflow.keras.optimizers import Adam

# Remove most of the data (leave every 8th timestep)
X = X[:,0:250:8,:]

x_train, x_test, y_train, y_test = train_test_split(X, Y, stratify=Y,
    test_size=0.2)

# Build the GRU-RNN model
K = 10   # number of outputs (categories for classification)
M = 3    # number of hidden nodes

i = Input(shape=x_train[0].shape)
x = GRU(M)(i)
x = Dense(K, activation='softmax')(x)

model = Model(i, x)
```

A.2.1 *Corpus 1*

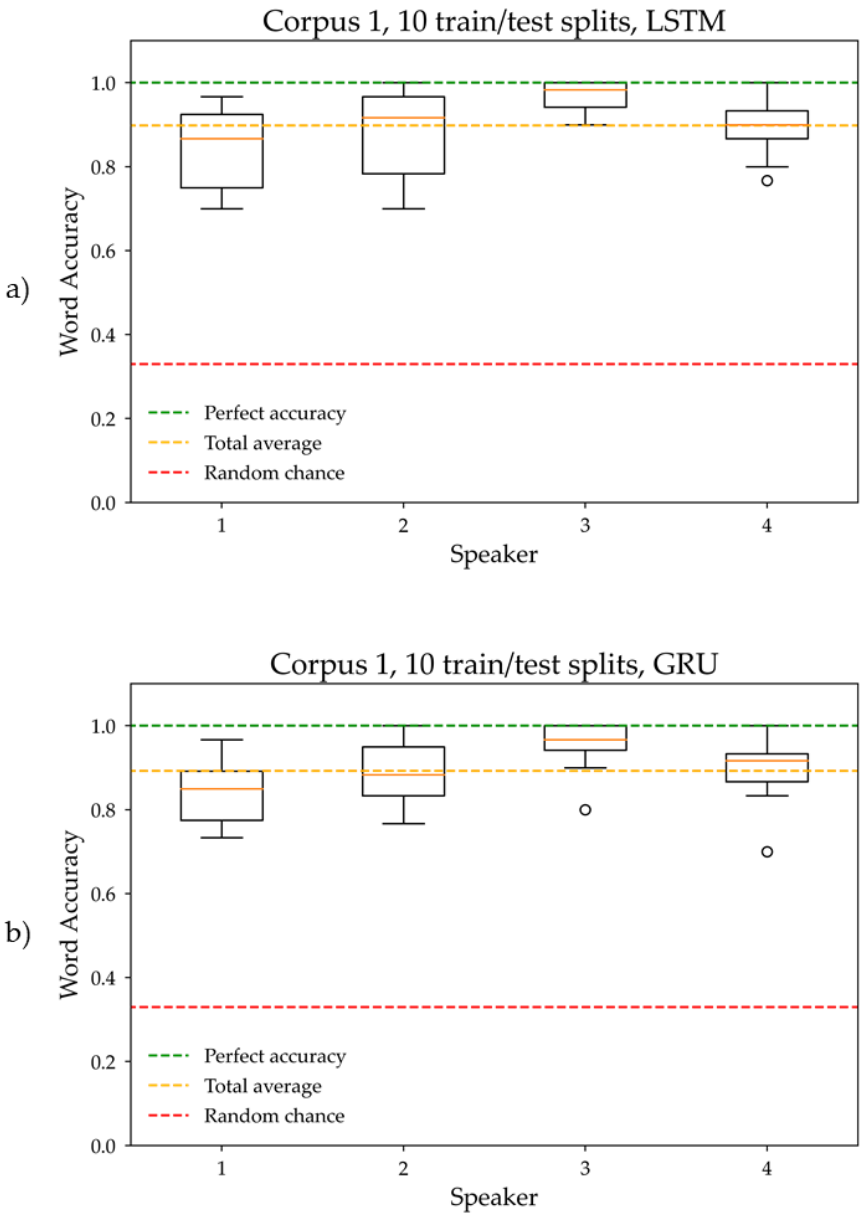


Figure 33.: Boxplots of the classification results for Corpus 1 using the final **a)** LSTM and **b)** GRU architectures 10 different train/test splits were performed with the median for each speaker as the solid orange line in each box. The box extends to lower and upper quartile, while the whiskers extend from the box to show the range of the data. Circles are regarded as outliers, but still calculated into the total average.

A.2.2 *Corpus 2*

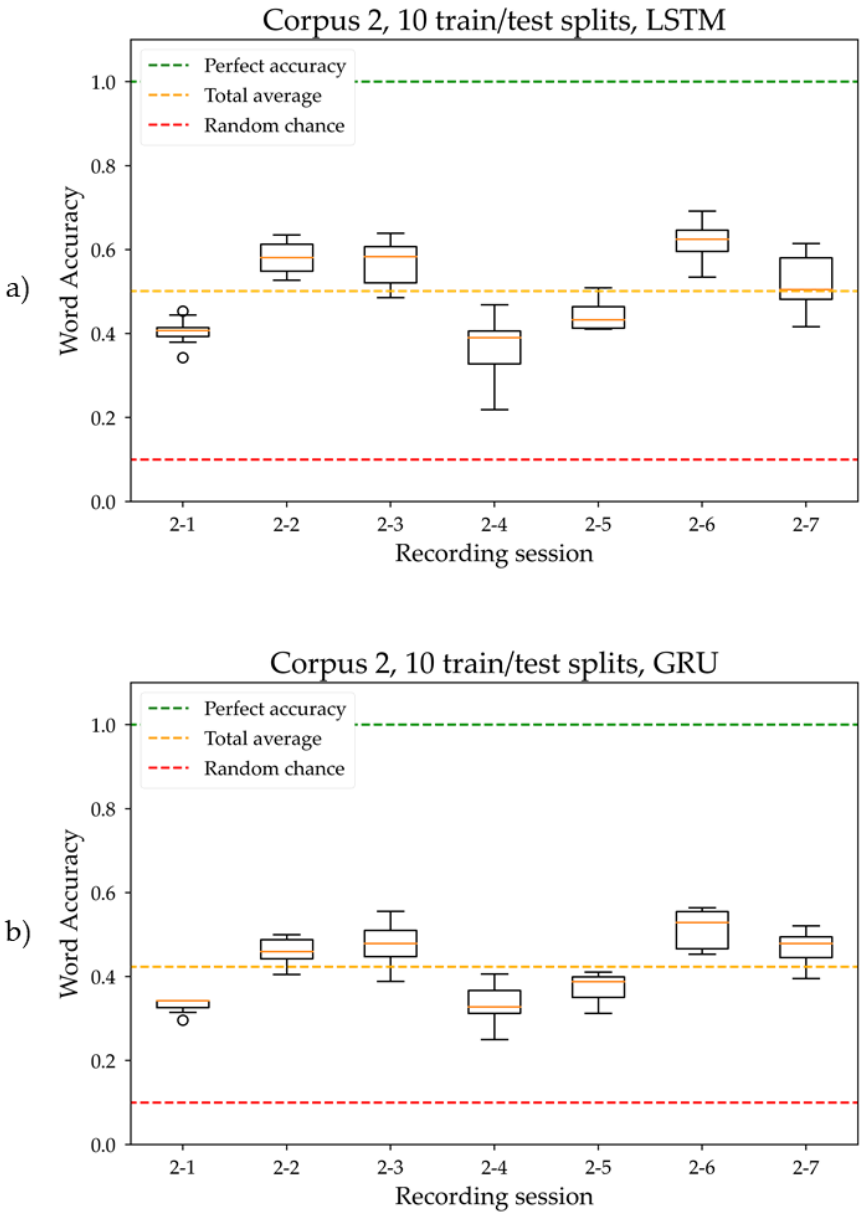


Figure 34.: Boxplots of the classification results for Corpus 2 using the **a)** LSTM and **b)** GRU architectures. Total average accuracies of 50.1% for the LSTM and 42.3% for the GRU are marked with the orange stippled lines.

A.2.3 *Corpus 3*

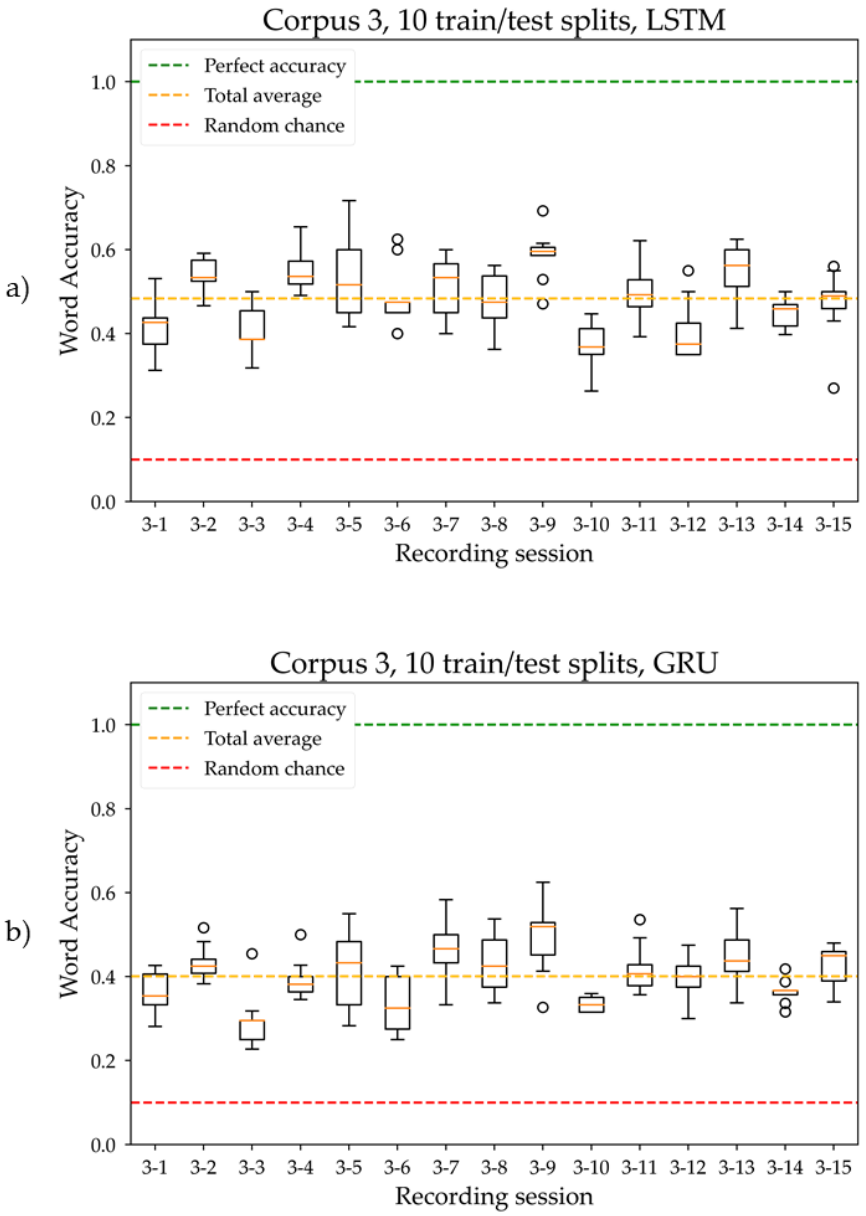


Figure 35.: Boxplots of the classification results for Corpus 3 using the **a)** LSTM and **b)** GRU architectures. Total average accuracies of 48.4% for the LSTM and 40.1% for the GRU are marked with the orange stippled lines.

A.2.4 *Corpus 4*

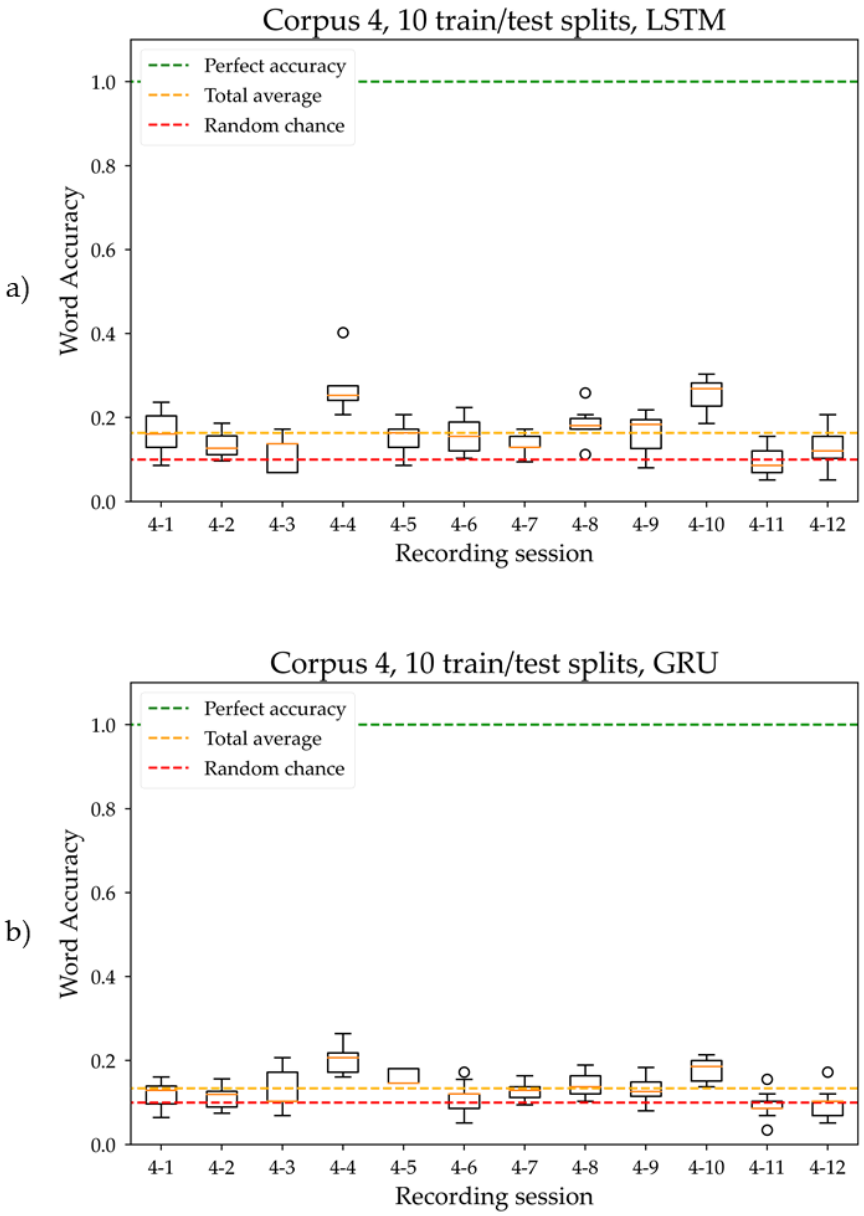


Figure 36.: Boxplots of the classification results for Corpus 4 using the **a)** LSTM and **b)** GRU architectures. Total average accuracies of 16.3% for the LSTM and 13.4% for the GRU are marked with the orange stippled lines.

A.3 EXTENDED FEATURE-METHOD TABLE

Table 18.: Mean recognition rate [in % (± standard deviation)] comparing the different feature extraction techniques and classification methods. Train and test data was randomly selected from the whole dataset of Corpus 3 and 10 train/test splits were performed or the CNNs and HMM, while only 2 train/test splits were used for the RNNs.

Features	CNN	sRNN	LSTM	GRU	HMM	CNN2
Raw data	15.3 (4.2)	10.0 (0)	10.0 (0)	10.0 (0)	52.9 (0.6)	74.4 (2.6)
Normalized	40.7 (5.6)	43.0 (10.2)	50.8 (2.8)	40.1 (5.1)	48.9 (1.4)	78.8 (1.5)
Minmax	67.6 (1.7)	52.1 (3.5)	55.5 (0.2)	40.9 (2.0)	56.9 (1.4)	81.3 (1.7)
Minmax + deltas	70.0 (2.2)	52.2 (2.3)	53.9 (1.1)	43.9 (1.2)	60.6 (0.9)	**85.4 (1.2)**
MFCCs	27.3 (2.3)	35.9 (1.4)	20.5 (2.6)	14.0 (4.0)	55.3 (1.1)	68.7 (1.7)
MFCCs + deltas	48.3 (2.8)	19.1 (9.1)	28.3 (1.2)	29.5 (1.3)	54.7 (1.3)	65.9 (1.8)
Activation	34.0 (5.5)	24.4 (9.0)	38.2 (3.5)	24.4 (9.0)	33.3 (0.8)	52.5 (1.2)

BIBLIOGRAPHY

[1] Achmad F. Abka and Hilman F. Pardede. "Speech recognition features: Comparison studies on robustness against environmental distortions". In: *2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. 2015, pp. 114–119. DOI: 10.1109/IC3INA.2015.7377757.

[2] Abraham Akinin, Akshay Paul, Jun Wang, Alessio Buccino, and Gert Cauwenberghs. "Biopotential Measurements and Electrodes". In: *Neural Engineering*. Ed. by Bin He. Cham: Springer International Publishing, 2020, pp. 65–96. ISBN: 978-3-030-43395-6. DOI: 10.1007/978-3-030-43395-6_2.

[3] A. D. Chan, C., K. Englehart, B. Hudgins, and D. F. Lovely. "Myo-electric signals to augment speech recognition". English. In: *Medical and Biological Engineering and Computing* 39.4 (2001), pp. 500–4. URL: https://search.proquest.com/docview/661512197?accountid=12870.

[4] Rubana H. Chowdhury, Mamun B. I. Reaz, Mohd Alauddin Bin Mohd Ali, Ashrif A. A. Bakar, K. Chellappan, and T. G. Chang. "Surface electromyography signal processing and classification techniques". In: *Sensors (Basel, Switzerland)* 13.9 (2013), pp. 12431–12466. ISSN: 1424-8220. DOI: 10.3390/s130912431.

[5] Ernest Cline. *Ready Player Two*. New York NY: Ballantine Books, 2020. ISBN: 9781524761332.

[6] *Communication systems - Technology*. Available at https://invisio.com/communication-systems/technology/ (04.12.2020). Invisio.

[7] The SciPy Community. *scipy.signal.find_peaks Documentation*. Available at https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html.

[8] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[9] B Denby, T Schultz, K Honda, T Hueber, J M Gilbert, and J S Brumberg. "Silent speech interfaces". In: *Speech Communication* 52 (2009), pp. 270–287. DOI: 10.1016/j.specom.2009.08.002.

[10] Bruce Denby, Shicheng Chen, Yifeng Zheng, Kele Xu, Yin Yang, Clemence Leboullenger, and Pierre Roussel. "Recent results in silent speech interfaces". In: *The Journal of the Acoustical Society of America* 141 (5), pp. 3646–3646. ISSN: 0001-4966. DOI: 10.1121/1.4987881.

[11] Lorenz Diener and Tanja Schultz. "Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion". In: *Proc. Interspeech 2018*. 2018, pp. 3162–3166. DOI: 10.21437/Interspeech.2018-2080.

[12] Lorenz Diener, Tejas Umesh, and Tanja Schultz. "Improving Fundamental Frequency Generation in EMG-to-Speech Conversion Using a Quantization Approach". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019, pp. 682–689. DOI: 10.1109/ASRU46091.2019.9003804.

[13] Emotiv. *Emotiv Epoc+ 14 Channel*. Available at https://emotiv.gitbook.io/epoc-user-manual/ and https://www.emotiv.com/setup/epoc/.

[14] *F8 2017: AI, Building 8 and More Technology Updates From Day Two*. Available at https://about.fb.com/news/2017/04/f8-2017-day-2/ (03.12.2020). Facebook, Apr. 19, 2017.

[15] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman. "Development of a (silent) speech recognition system for patients following laryngectomy". In: *Medical Engineering and Physics* 30 (4 May 2008), pp. 419–425. ISSN: 13504533. DOI: 10.1016/j.medengphy.2007.05.003.

[16] João Freitas, António Teixeira, Miguel Sales Dias, and Samuel Silva. *An Introduction to Silent Speech Interfaces*. Springer International Publishing. ISBN: 978-3-319-40173-7. DOI: 10.1007/978-3-319-40174-4.

[17] João Freitas, António Teixeira, Samuel Silva, Catarina Oliveira, and Miguel Sales Dias. "Detecting nasal vowels in speech interfaces based on surface electromyography". In: *PLoS ONE* 10 (6 June 2015). ISSN: 19326203. DOI: 10.1371/journal.pone.0127040.

[18] Masaaki Fukumoto. "SilentVoice: Unnoticeable voice input by ingressive speech". In: *UIST 2018 - Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (), pp. 237–246. DOI: 10.1145/3242587.3242603.

[19] David Gaddy and Dan Klein. *Digital Voicing of Silent Speech*. 2020. arXiv: 2010.02960 [eess.AS].

[20] Linda I. Garrity. "Electromyography: A review of the current status of subvocal speech research". In: *Memory Cognition* 5 (6), pp. 615–622. ISSN: 0090502X. DOI: 10.3758/BF03197407.

[21] William Gibson. *Neuromancer*. New York : Ace Science Fiction Books, 1984. URL: https://search.library.wisc.edu/catalog/999588118402121.

[22] Jose A. Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M. Martin Donas, Jose L. Perez-Cordoba, and Angel M. Gomez. "Silent Speech Interfaces for Speech Restoration: A Review". In: *IEEE Access* 8 (2020), pp. 177995–178021. DOI: 10.1109/access.2020.3026579. eprint: 2009.02110.

[23] Google. *Google Colaboratory*. Available at https://cloud.google.com/speech-to-text.

[24] Caroline Henton. "Challenges and Rewards in Using Parametric or Concatenative Speech Synthesis". In: *International Journal of Speech Technology* 5 (May 2002), pp. 117–131. DOI: 10.1023/A:1015416013198.

[25] G. Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

[26] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing". In: *Speech Communication* 55 (1 Jan. 2013), pp. 22–32. ISSN: 01676393. DOI: 10.1016/j.specom.2012.02.001.

[27] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. 1st. USA: Prentice Hall PTR, 2001. ISBN: 0130226165.

[28] Xuedong Huang, James Baker, and Raj Reddy. *A historical perspective of speech recognition*. 2014. DOI: 10.1145/2500887.

[29] Lilah Inzelberg, David Rand, Stanislav Steinberg, Moshe David-Pur, and Yael Hanein. "A Wearable High-Resolution Facial Electromyography for Long Term Recordings in Freely Behaving Humans". In: *Scientific Reports* 8.1 (2018), p. 2058. ISSN: 2045-2322. DOI: 10.1038/s41598-018-20567-y.

[30] Sadaf Iqbal, P.P. Muhammed Shanir, Yusuf Uzzaman Khan, and Omar Farooq. "Time Domain Analysis of EEG to Classify Imagined Speech". In: *Proceedings of the Second International Conference on Computer and Communication Technologies*. Ed. by Suresh Chandra Satapathy, K. Srujan Raju, Jyotsna Kumar Mandal, and Vikrant Bhateja. New Delhi: Springer India, 2016, pp. 793–800.

[31] Matthias Janke. "EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals". PhD thesis. Karlsruher Institut für Technologie (KIT), 2016. 153 pp. DOI: 10.5445/IR/1000054490.

[32] Matthias Janke and Lorenz Diener. "EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25 (12 Dec. 2017), pp. 2375–2385. ISSN: 23299290. DOI: 10.1109/TASLP.2017.2738568.

[33] Chuck Jorgensen and Kim Binsted. "Web browser control using EMG based sub vocal speech recognition". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. 2005, p. 294. DOI: 10.1109/hicss.2005.683.

[34] Chuck Jorgensen, Diana D. Lee, and Shane Agabon. "Sub Auditory Speech Recognition Based on EMG Signals". In: *Proceedings of the International Joint Conference on Neural Networks* 4 (2003), pp. 3128–3133. DOI: 10.1109/ijcnn.2003.1224072.

[35] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. *Towards Continuous Speech Recognition Using Surface Electromyography*. Tech. rep. 2006. URL: https://www.isca-speech.org/archive/interspeech_2006/i06_1592.html.

[36] Bouchard KE, Mesgarani N, Johnson K, and Chang EF. "Functional organization of human sensorimotor cortex for speech articulation". In: *Nature* 495.7441 (2013). ISSN: 1476-4687. DOI: 10.1038/NATURE11911.

[37] Arnav Kapur. "Human-Machine Cognitive Coalescence through an Internal Duplex Interface". MIT, 2018. URL: https://dspace.mit.edu/handle/1721.1/120883.

[38] Arnav Kapur, Shreyas Kapur, and Pattie Maes. "AlterEgo: A personalized wearable silent speech interface". In: *International Conference on Intelligent User Interfaces, Proceedings IUI* (Mar. 2018), pp. 43–53. DOI: 10.1145/3172944.3172977.

[39] Arnav Kapur et al. *Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia*. Apr. 2020, pp. 25–38. URL: https://www.media.mit.edu/publications/non-invasive-silent-speech-recognition-in-multiple-sclerosis-with-dysphonia/.

[40] Chanwoo Kim and Richard M. Stern. "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.7 (2016), pp. 1315–1329. DOI: 10.1109/TASLP.2016.2545928.

[41] Thomas Kluyver et al. “Jupyter Notebooks - a publishing format for reproducible computational workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by Fernando Loizides and Birgit Scmidt. Netherlands: IOS Press, 2016, pp. 87–90. URL: https://eprints.soton.ac.uk/403913/.

[42] Gautam Krishna, Co Tran, Mason Carnahan, and Ahmed Tewfik. “Continuous Silent Speech Recognition using EEG”. In: *arXiv* (Feb. 2020). URL: http://arxiv.org/abs/2002.03851.

[43] Stanley Kubrick and Arthur C. Clarke. *2001 - A Space Odessey*. 1968.

[44] Pradeep Kumar, Rajkumar Saini, Partha Pratim, Roy Pawan, Kumar Sahu, and Debi Prosad Dogra. “Envisioned speech recognition using EEG sensors”. In: (2018), pp. 185–199. DOI: 10.1007/s00779-017-1083-4.

[45] Yuet-Ming Lam, Philip Heng-Wai Leong, and Man-Wai Mak. “Frame-Based SEMG-to-Speech Conversion”. In: *2006 49th IEEE International Midwest Symposium on Circuits and Systems*. Vol. 1. 2006, pp. 240–244. DOI: 10.1109/MWSCAS.2006.382042.

[46] Pierre Lison and Jörg Tiedemann. “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. URL: https://www.aclweb.org/anthology/L16-1147.

[47] Siyuan Ma, Dantong Jin, Ming Zhang, Bixuan Zhang, You Wang, Guang Li, and Meng Yang. “Silent Speech Recognition Based on Surface Electromyography”. In: *2019 Chinese Automation Congress (CAC)*. 2019, pp. 4497–4501. DOI: 10.1109/CAC48633.2019.8996289.

[48] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. “Session independent non-audible speech recognition using surface electromyography”. In: *Proceedings of ASRU 2005: 2005 IEEE Automatic Speech Recognition and Understanding Workshop* 2005 (2005), pp. 307–312. DOI: 10.1109/ASRU.2005.1566521.

[49] Joseph G. Makin, David A. Moses, and Edward F. Chang. “Machine translation of cortical activity to text with an encoder–decoder framework”. In: *Nature Neuroscience* 23.4 (2020), pp. 575–582. ISSN: 15461726. DOI: 10.1038/s41593-020-0608-8.

[50] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. “Development of sEMG sensors and algorithms for silent speech recognition”.

In: *Journal of Neural Engineering* 15 (4), p. 046031. ISSN: 17412552. DOI: 10.1088/1741-2552/aac965.

[51] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. "Silent speech recognition as an alternative communication device for persons with laryngectomy". In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 25 (12 Dec. 2017), pp. 2386–2398. ISSN: 23299290. DOI: 10.1109/TASLP.2017.2740000.

[52] Denise Mitchell. *Surface Electromyography: Fundamentals, Computational Techniques and Clinical Applications.* Physical Medicine and Rehabilitation. Nova Science Publishers, Inc, 2016. ISBN: 9781536102024. URL: https://www.researchgate.net/publication/323799337_Surface_electromyography_Fundamentals_computational_techniques_and_clinical_applications.

[53] M. S. Morse, Y. N. Gopalan, and M. Wright. "Speech Recognition Using Myoelectric Signals With Neural Networks". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society Volume 13: 1991*. 1991, pp. 1877–1878. DOI: 10.1109/IEMBS.1991.684800.

[54] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. "Non-audible murmur recognition". In: *EUROSPEECH2003: 8th European Conference on Speech Communication and Technology, September 1-4, 2003, Geneva, Switzerland* (2003), pp. 2601–2604. ISSN: 1018-4074.

[55] 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*. 2016. URL: http://arxiv.org/abs/1609.03499.

[56] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. *WaveGlow: A Flow-based Generative Network for Speech Synthesis*. 2018. arXiv: 1811.00002 [cs.SD].

[57] Lazy Programmer. *Tensorflow 2.0: Deep Learning and Artificial Intelligence*. Available at https://www.udemy.com/course/deep-learning-tensorflow-2/.

[58] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. arXiv: 1712.05884 [cs.CL].

[59] Brendan Shillingford et al. "Large-Scale Visual Speech Recognition". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2019-September (2018), pp. 4135–4139. arXiv: 1807.05162.

[60] Mok Win Soon, Muhammad Ikmal Hanafi Anuar, Mohamad Hafizat Zainal Abidin, Ahmad Syukri Azaman, and Norliza Mohd Noor. "Speech recognition using facial sEMG". In: *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 2017, pp. 1–5. DOI: 10.1109/ICSIPA.2017.8120569.

[61] N. Sugie and K. Tsunoda. "A Speech Prosthesis Employing a Speech Synthesizer-Vowel Discrimination from Perioral Muscle Activities and Vowel Production". In: *IEEE Transactions on Biomedical Engineering* BME-32.7 (1985), pp. 485–490. DOI: 10.1109/TBME.1985.325564.

[62] Arthur R. Toth, Michael Wand, and Tanja Schultz. "Synthesizing speech from electromyography using voice transformation techniques". In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. ISCA, 2009, pp. 652–655. URL: https://dblp.org/rec/conf/interspeech/TothWS09.bib.

[63] Lev Vygotsky. *Thought and Language*. 1962. ISBN: 9780262220033.

[64] Eric J Wadkins. *A Continuous Silent Speech Recognition System for AlterEgo, a Silent Speech Interface*. 2019. URL: https://dspace.mit.edu/handle/1721.1/123121.

[65] Michael Wand. *Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling*. 2014. ISBN: 978-3-7315-0211-1. DOI: 10.5445/KSP/1000040667. URL: https://www.ksp.kit.edu/9783731502111.

[66] Xiaochen Wang, Mingxing Zhu, Han Cui, Zijian Yang, Xin Wang, Haoshi Zhang, Cheng Wang, Hanjie Deng, Shixiong Chen, and Guanglin Li. "The Effects of Channel Number on Classification Performance for sEMG-based Speech Recognition". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2020, pp. 3102–3105. DOI: 10.1109/EMBC44109.2020.9176260.

[67] Ron J. Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P. Kingma. *Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis*. 2021. arXiv: 2011.03568 [cs.CL].

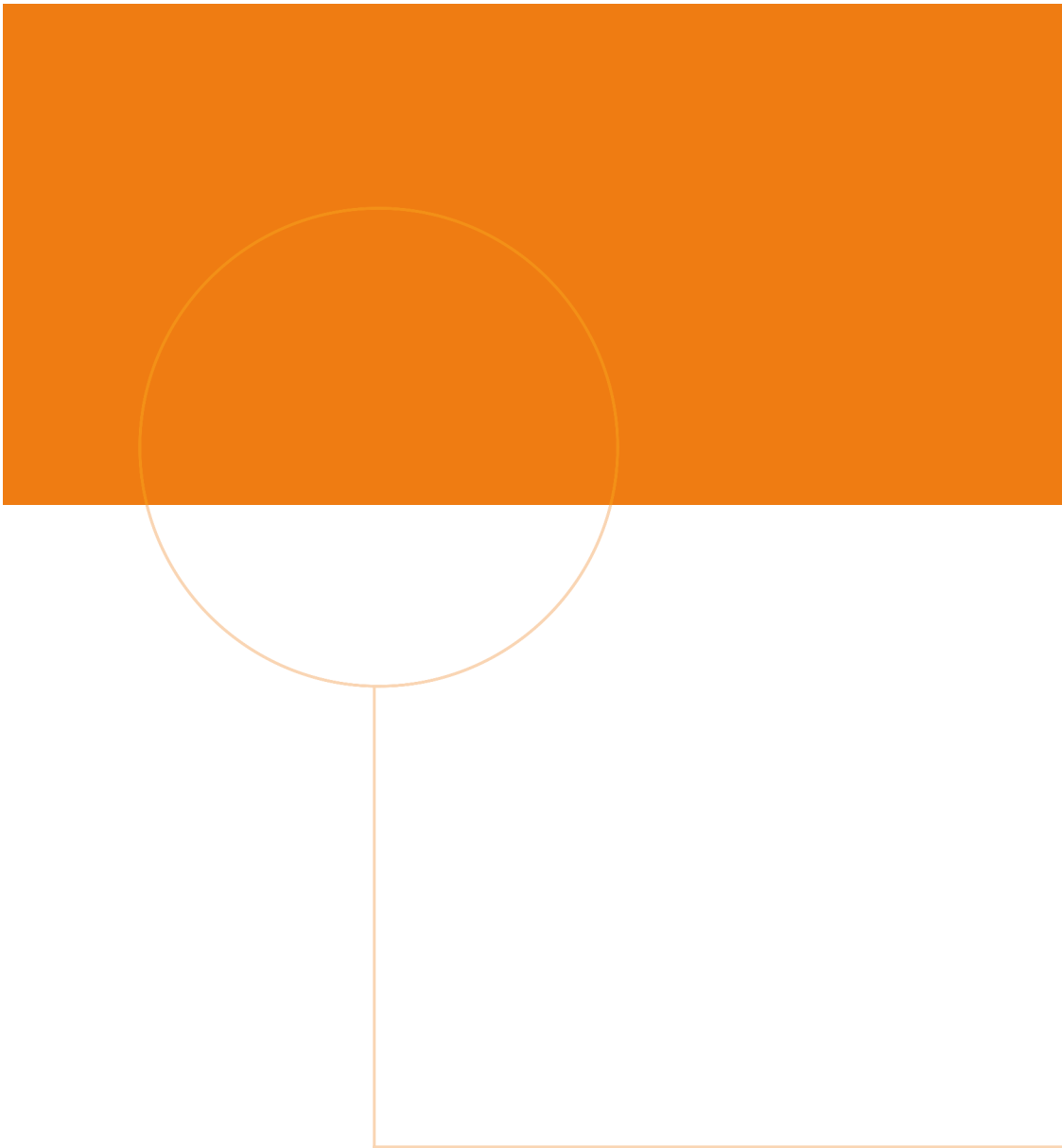
[68] Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. "High-performance brain-to-text communication via imagined handwriting". In: *bioRxiv* (2020). DOI: 10.1101/2020.07.01.183384.

[69] Dong Yu and Li Deng. *Automatic Speech Recognition*. Springer London. ISBN: 978-1-4471-5778-6. DOI: 10.1007/978-1-4471-5779-3.

[70] Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux. "End-to-End Speech Recognition from the Raw Waveform". In: *Proc. Interspeech 2018*. 2018, pp. 781–785. DOI: 10.21437/Interspeech.2018-2414.

[71] Ming Zhang, You Wang, Zhang Wei, Meng Yang, Zhiyuan Luo, and Guang Li. "Inductive conformal prediction for silent speech recognition". In: *Journal of Neural Engineering* (Mar. 2020). ISSN: 1741-2560. DOI: 10.1088/1741-2552/ab7ba0.

[72] Mingxing Zhu, Xiaochen Wang, Xin Wang, Cheng Wang, Zijian Yang, Oluwarotimi Williams Samuel, Shixiong Chen, and Guanglin Li. "The Effects of Electrode Locations on Silent Speech Recognition using High-Density sEMG". In: *2020 IEEE International Workshop on Metrology for Industry 4.0 IoT*. 2020, pp. 345–348. DOI: 10.1109/MetroInd4.0IoT48571.2020.9138289.



NTNU

Kunnskap for en bedre verden