Universität Bremen

Klassifikation paralinguistischer Merkmale aus EMG-Daten: Sprecheridentität und Sprachmodus

Bachelorarbeit am Cognitive Systems Lab
Prof. Dr.-Ing. Tanja Schultz
Fachbereich 3
Universität Bremen

von

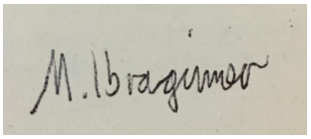
Magomed Ibragimov

Betreuer:
Dr. Dennis Küster
Lorenz Diener

Gutachter:
Prof. Dr.-Ing. Tanja Schultz
Dr. Tim Laue

Tag der Anmeldung: 3. April 2020
Tag der Abgabe: 25. September 2020

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich hoffe ein Bild meiner Unterschrift reicht hier aus. Ich kann aus Quarantäne Gründen nämlich nicht selber unterschreiben.



Bremen, den 25. September 2020

Zusammenfassung

Das Ziel dieser Bachelorarbeit ist es durch bioelektrische Signale der Gesichtsmuskulatur, den Sprachmodus sowie die Identität des Sprechers vorherzusagen. Dafür werden verschiedene Klassifikatoren in der Scikit-Learn Umgebung und der UKA-Korpus der CSL Bremen genutzt.

Durch Oberflächen-EMG wird die elektrische Muskelaktivität gemessen und digitalisiert. Diese Daten werden genutzt, um verschiedene Klassifikatoren zu trainieren und zu testen. Von den getesteten Klassifikatoren hat sich der LDA Klassifikator als der beste herausgestellt.

Es wurden Cross-Session sowie Cross-User Experimente für den Sprachmodus durchgeführt, bei denen sich Cross-Session als besser erwiesen hat. Am Ende wurde durch Frequenzbänder kontrolliert, ob der Fokus auf einen einzelnen Muskel im Gesicht zu besseren Ergebnissen führen kann. Das konnte in dieser Arbeit nicht bestätigt werden.

Insgesamt konnte eine Genauigkeit für die Sprecher-Erkennung mit über 90 Prozent und eine Genauigkeit von höchstens 59 Prozent für die Sprachmodus-Erkennung erreicht werden. Dies bestätigt, dass die Sprecher-Erkennung sehr gut klassifiziert werden kann. Der Sprachmodus hat eine niedrigere Genauigkeit erreicht, zeigt aber trotzdem, dass es möglich ist diesen vorherzusagen. Das deutet, dass die Erkennung von Unterschieden in der Sprachproduktion erkennbar ist. Die Genauigkeit ist nämlich deutlich über der Chance Wahrscheinlichkeit von 33 Prozent.

Inhaltsverzeichnis

1 Vorstellung **1**
1.1 Umfang der Arbeit . . . 2
1.2 Verwandte Arbeit . . . 2
1.3 Struktur . . . 3

2 Hintergrund **5**
2.1 Scikit-learn . . . 5
2.2 Muskel . . . 5
2.2.1 Motorische Einheit . . . 6
2.2.2 Muskelkontraktion . . . 7
2.2.2.1 Potenzialverlauf . . . 7
2.3 Sprachproduktion . . . 8
2.3.1 Relevante Muskeln . . . 8
2.3.2 Mechanismus . . . 10
2.4 EMG . . . 11
2.4.1 Digitalisierung . . . 11
2.5 Datenkorpus . . . 13
2.5.1 Informationen zu den Daten . . . 13
2.5.2 Sprachmodus . . . 14
2.6 Features . . . 15
2.6.1 Time Domain Features . . . 15
2.6.1.1 MAV . . . 15
2.6.1.2 MAV2 . . . 15
2.6.1.3 RMS . . . 15
2.6.1.4 IAV . . . 15
2.6.1.5 WL . . . 16
2.6.1.6 IEMG . . . 16
2.6.1.7 AAC . . . 16
2.6.1.8 ZC . . . 16
2.6.1.9 SSC . . . 16
2.6.1.10 WAMP . . . 17
2.6.2 Frequency Domain Features . . . 17
2.6.2.1 Mean Frequency . . . 17
2.6.2.2 Median Frequency . . . 17
2.7 Klassifikatoren . . . 17
2.7.1 LDA . . . 18
2.7.2 KNN . . . 18
2.7.3 DecisionTree . . . 19
2.7.4 RandomForest . . . 19

2.7.5 SVC 20

3 Versuchsaufbau **21**
3.1 Aufteilung der Daten 21
3.1.1 Nutzererkennung 21
3.1.2 Sprachmodus 22

4 Evaluation **23**
4.1 Sprecheridentität 23
4.1.1 Ungefiltert 23
4.1.1.1 Sessions 24
4.1.2 Filtered 29
4.1.2.1 Sessions 30
4.2 Sprachmodus 33
4.2.1 Cross-Session 33
4.2.1.1 Ungefiltert 34
4.2.1.2 Sessions-Ungefiltert 34
4.2.1.3 Filtered 38
4.2.1.4 Sessions-Filtered 39
4.2.1.5 Frequenzbänder 40
4.2.2 Cross-User 41
4.2.2.1 Ungefiltert 41
4.2.2.2 Filtered 44
4.2.2.3 Frequenzbänder 46

5 Diskussion **47**
5.1 Erkennung der Identität 47
5.1.1 Klassifikator 47
5.1.2 Große Sessions 47
5.1.3 Kleine Sessions 48
5.2 Sprachmodus 49
5.2.1 Cross-User 49
5.2.2 Cross-Session 49
5.2.3 Frequenzband 50
5.3 Datenqualität 50

6 Fazit **51**
6.1 Zusammenfassung 51
6.2 Zukünftige Arbeiten 51
6.3 Abschließende Worte 52

Literaturverzeichnis **53**

1. Vorstellung

Die gesprochene Sprache ist das wichtigste zwischenmenschliche Kommunikationsmedium. Wir kommunizieren aber nicht nur Fakten mit der Sprache, sondern auch paralinguistische Informationen. Jedoch kann es Situationen geben, in denen man Informationen über die Sprache weitergeben muss, dies jedoch aus verschiedenen Gründen nicht möglich ist. In Deutschland alleine leben nach dem Statistischen Bundesamt [Sta] über 300.000 Menschen mit Sprach- oder Sprechstörungen. Für viele dieser Personen ist das Einstellen einer privaten Hilfe nicht möglich. Für diese Menschen soll es möglich sein, auf Sprache basierende Schnittstellen ohne Probleme zu nutzen. Dies soll dadurch ermöglicht werden, dass biologisches Signale betrachtet werden, welche in den Muskeln während dem Reden produziert werden. Genauer geht es um die elektrischen Biosignale der Gesichtsmuskeln beim Produzieren der Sprache und ob diese genutzt werden können, um den Sprachmodus sowie die Identität zu bestimmen.

Eine Anwendungsweise dieser Signale sind heutzutage die sogenannten SSIs (Silent Speech Interfaces), die zum Beispiel im Militär [DCHM12] für strategische Zwecke genutzt werden. Eine weitere Möglichkeit stummen Menschen zu helfen ist die elektrische Stimulation der Mundmuskeln. In [SAD+19] zeigen Schultz et al., wie Funktionelle elektrische Stimulation (FES) genutzt werden können, um die Sprachmuskeln effektiv zu steuern. Des Weiteren wird durch Brain Computer Interfaces versucht die Sprache zu rekonstruieren. Also das Gedachte wird direkt in Sprache umgewandelt. Dazu wurden in dieser Doktorarbeit von Christian Emanuel Herf zwei verschiedene Technologien (fNIRS und ECoG)vorgestellt. [Her16]

Ich hoffe, dass diese Arbeit die Qualität und Genauigkeit solcher SSIs, durch das Vorhersagen der Identität und des Sprachmodus verbessern kann. Der Sprachmodus kann dabei vor der Generierung der eigentlichen Sprache erkannt werden[Sch19]. In dieser Arbeit wird geprüft, ob das Erkennen der Identität sowie des Sprachmodus durch Oberflächen-EMG Daten möglich ist. Dazu werden die Daten des UKA Korpus des CSL der Uni Bremen verwendet ([WJS14]).

1.1 Umfang der Arbeit

Es werden verschiedene Klassifikatoren in der Scikit-learn Umgebung verwendet, um die Identität sowie den Sprachmodus des Sprechers vorherzusagen. Zudem wird versucht durch das Entfernen von einzelnen Sessions und durch Filtern das Resultat so hoch wie möglich zu bekommen.

1.2 Verwandte Arbeit

Die Verwendung von EMG zur Spracherkennung geht zurück auf die 1980er Jahre. Die ersten Wettbewerbsfähigen Erkenner stammen jedoch jedoch von Chan et al. aus 2001, welche eine durchschnittliche Wortgenauigkeit von 93 Prozent für einem 10-Wort-Vokabular aus englischen Ziffern geschafft haben. Jorgensen et al. haben es 2003 ebenfalls geschafft gute Leistungen bei schweigendem Sprechen der Worte zu erreichen. Dies deutet darauf hin, dass diese Technologie für Silent Speech-Schnittstellen verwendet werden [WS11].

In den INTERSPEECH Konferenzen geht es um die Erkennung verschiedener paralinguistischer Merkmale. ([ISC]) In [JWS11] geht es zum Beispiel, um die Möglichkeit durch phonetische Entscheidungsbäume zwischen dem stummen und normalen Sprachmodus zu unterscheiden. In dieser Arbeit wird ebenfalls der UKA-Korpus der CSL Bremen verwendet. Hier wird gezeigt,dass es möglich ist einen Erkenner mit stummen sowie hörbaren Daten gleichzeitig zu trainieren. Das macht es einfacher während des Erkennungsprozesses den zu erkennenden Sprachmodus zu wechseln. Es werden hier zudem Genauigkeiten für einzelne Phoneme analysiert.

In [PCMDG+20] werden die verschiedenen Möglichkeiten, die Stimme einer Person durch Silent Speech Interfaces wiederherzustellen, untersucht.In dem Der Artikel gibt einen Überblick über die jüngsten Versuche, Sprache aus nicht-akustischen Biosignalen, die während der Sprachproduktion erzeugt werden, zu entschlüsseln.

Die Autoren präsentieren eine umfassende Liste von Sensortechnologien, die derzeit zur Erfassung von Biosignalen, einschließlich invasiver und nicht invasiver Techniken in Erwägung gezogen werden. Aus diesen Biosignalen kann Sprache dekodiert werden. Das passiert entweder durch automatische Spracherkennung oder durch direkte Sprachsynthese. Ein potenzieller Vorteil des letzteren Ansatzes ist, dass sie möglicherweise die Entschlüsselung von Sprache in Echtzeit ermöglicht. Dies könnte dazu beitragen in der Zukunft die Stimme einer Person wiederherzustellen.

Bei weiteren Arbeiten zum Thema SSIs gibt es zum Beispiel [JD17] sowie andere Arbeiten des CSL Bremen. Das CSL Bremen hat eine große Sammlung an Arbeiten in denen verschiedene Technologien in dieser Richtung ausprobiert wurden. In [WHH+13] zum Beispiel wird ein ICA Algorithmus genutzt um Artefakte in einem stummen Spracherkenner zu verringern. In [JWS10] zeigten Janke et al., dass EMG-Signale von hörbarer Sprache im Allgemeinen eine höhere spektrale Leistung als EMG-Signale von stummer Sprache, aufweisen. Mit diesen Informationen konnte eine Abbildung vom stillen zum hörbaren/geflüsterten EMG erreicht werden, was zu einer 12,3 prozentigen Verbesserung führte.

In [SWH+17] haben Schultz et al. die verschiedenen Biosignale bei der Erzeugung von Sprache beschrieben und zudem das Potenzial dieser Signale für das Bauen eines Biosignal basierten Spracherkenners analysiert. Das wäre ein Spracherkenner welcher alle möglichen Biosignale benutzt, um einen Spracherkenner so genau wie möglich zu gestalten. In [DAB+ar] wurden der Sprachmodus und die Sprecheridentität mit einem anderen Korpus des CSL vorhergesagt. Dieser Korpus wurde mit einem Array Elektrodensetup anstatt Einzelelektroden aufgenommen. In dieser Arbeit wird AUDEEP genutzt [SAS20].

1.3 Struktur

Der Rest des Dokuments ist folgendermaßen aufgebaut:

Kapitel 2 stellt vor worauf diese Arbeit basiert und stellt die nötigen biologischen sowie technischen Elemente dar.

In Kapitel 3 wird der Versuchsaufbau, mit der die späteren Versuche durchgeführt werden, vorgestellt.

In Kapitel 4 sind die Ergebnisse der in Kapitel 3 genannten Versuche zu sehen.

In Kapitel 5 werden die Ergebnisse noch einmal zusammengefasst und näher analysiert.

Am Ende in Kapitel 6 wird beschrieben, was die Funde meiner Meinung nach bedeuten. Zudem sind in diesem Kapitel Vorschläge für weitere Verbesserungen zu finden.

2. Hintergrund

2.1 Scikit-learn

Scikit-learn ([PVG+11]) ist die Software-Bibliothek welche für das Implementieren der verschiedenen Klassifikationsalgorithmen genutzt wurde. Es handelt sich hierbei um eine freie Software-Bibliothek zum maschinellen Lernen für die Programmiersprache Python.

2.2 Muskel

Muskeln können in drei Typen unterteilt werden: quergestreifte, Herz- und glatte Muskeln. Die beiden letzteren sind für diese Arbeit nicht von Interesse, da nur Gesichtsmuskeln, bei denen es sich um quergestreifte Muskeln handelt, beobachtet werden. Quergestreifte Muskulatur bildet die Skelettmuskulatur (2.10). Als Skelettmuskeln bezeichnet man diejenigen Muskeln, die für die Körperbewegungen zuständig sind. [tra]

Der Aufbau eines Muskels ist eine Spindel ähnliche Form, die im oberen und unteren Bereich dünner wird. Der Muskel besteht aus Muskelfaserbündeln, welche wiederum aus vielen Muskelfasern bestehen. Die Muskelfasern bestehen aus Myofibrillen. Die Myofibrillen bestehen abermals aus Filamenten, welche sich in zwei verschiedene Typen aufteilen lassen. Diese Filamente sind die Myosinfilamente und die Aktinfilamente. Der Aufbau ist in der Abbildung 2.1 zu sehen.

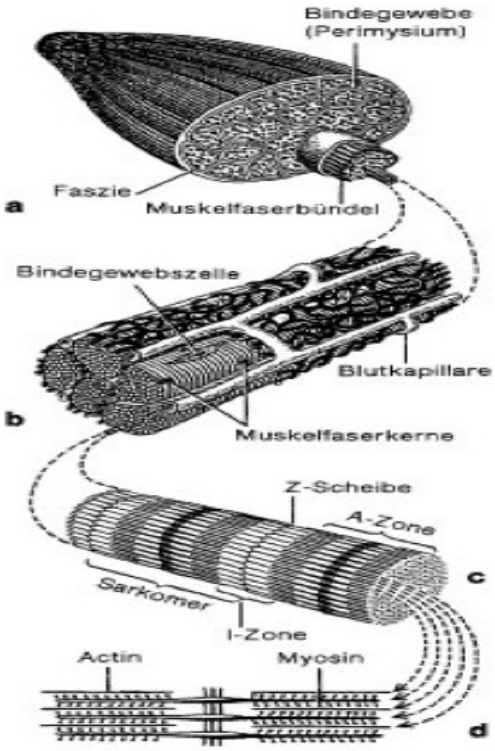


Abbildung 2.1: Aufbau eines Muskels [Spe]

2.2.1 Motorische Einheit

Eine Motorische Einheit [TM04] besteht aus einer Nervenzelle, einer Nervenfaser und einer Muskelfasern. Die Motorische Einheit ist für die Kontraktion der Muskeln durch Nervenimpulse zuständig. Dabei steuert jede motorische Nervenzelle jeweils mehrere Muskelfasern 2.2. Bei diesen Nervenimpulsen handelt es sich um die im Weiteren erklärten Aktionspotenziale.

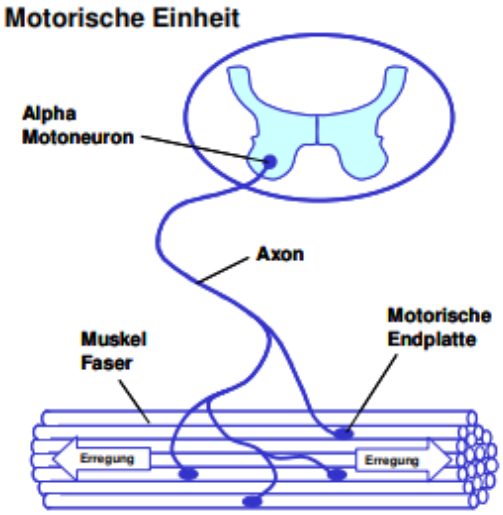


Abbildung 2.2: Aufbau einer Motorischen Einheit [Kon06]

2.2.2 Muskelkontraktion

Muskelkontraktionen sind notwendig für die menschliche Bewegung. Eine Muskelkontraktion [TM04] wird durch das Bilden einer ionischen Differenz zwischen dem Innen- und Außenraum einer Muskelzelle ausgelöst. Der Konzentrationsunterschied zwischen dem Inneren und Äußeren der Muskelzelle verursacht ein Ruhepotenzial. Diese Potenzialdifferenz wird durch eine Ionenpumpe aufrechterhalten. Über das Axon kommt ein Reiz von der Motorischen Einheit an die motorische Endplatte (Verbindung zwischen den Synapsen) und der Muskelfaser. Wenn es zu einem Reiz kommt, kommt es zu einem Na+ Einstrom und einem K+ Ausstrom. Dies bewirkt eine Depolarisation, die sofort durch rückwärts Austausch von Ionen durch die Ionenpumpen umgekehrt wird (Repolarisation). In 2.3 ist die Muskelkontraktion aus einer chemischen Perspektive zu sehen.

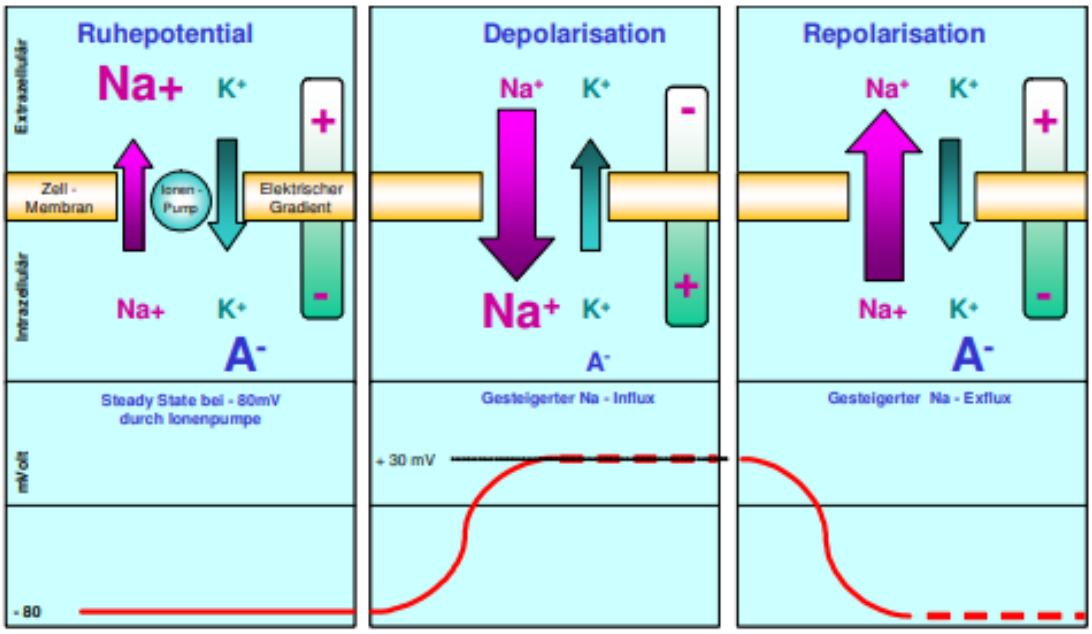


Abbildung 2.3: Das Aktionspotenzial aus der Perspektive der Ionen [Kon06]

2.2.2.1 Potenzialverlauf

Aktionspotenziale können in vier Phasen eingeteilt werden (siehe 2.3):

1. In der Anfangsphase treibt ein Reiz die negative Spannung in Richtung null.

2. Beim Überschreiten einer bestimmten Na+ Konzentration, beschleunigt sich die Depolarisation sehr stark. Das führt dazu, dass die Membranpotenzial sehr schnell von positiv zu negativ wechselt.

3. Nach dem Erreichen des Maximums bewegt sich das Aktionspotenzial zum Anfangswert, also dem Ruhepotenzial, zurück.

4. Das Ruhepotenzial wird unterschritten(Hyperpolarisation) und schließlich wieder erreicht. [Kon06]

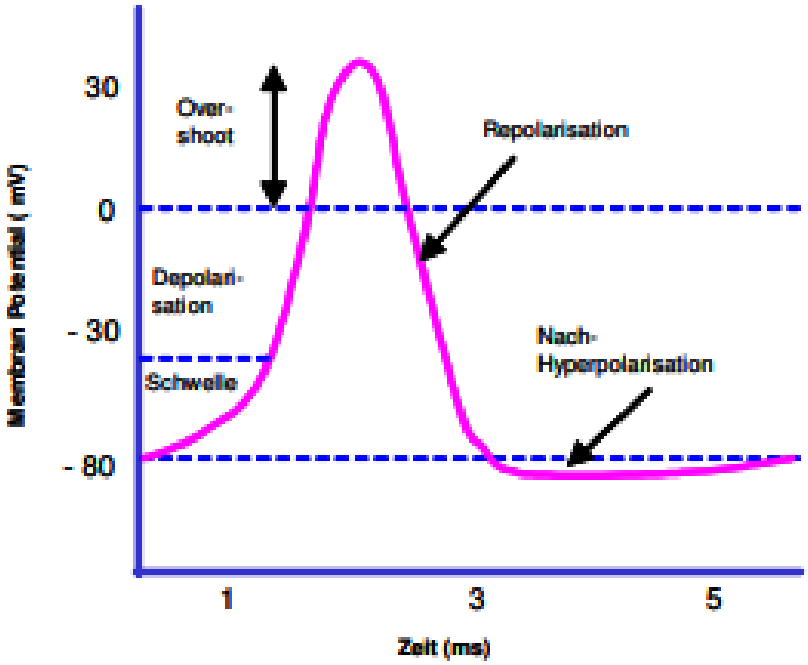


Abbildung 2.4: Verlauf des Aktionspotenzials in den verschiedenen Phasen [Kon06]

2.3 Sprachproduktion

2.3.1 Relevante Muskeln

In diesem Teil werden die relevanten Muskeln für die Sprachbildung aufgelistet. Die Informationen in diesem Teil kommen aus der Diplomarbeit von Matthias Janke [Jan16]. Die Positionen der Muskeln im Gesicht sind in der Abbildung 2.5 zu sehen.

Zunge

Die Zunge ist eine Verbindung aus intrinsischen und extrinsischen Muskeln. Sie nimmt den größten Teil der Mundhöhle ein und hat damit großen Einfluss auf die Vokalartikulation. Die Vokale sind durch die Position der Zunge charakterisiert. Es werden keine Daten der Zunge für diese Arbeit genutzt.

Orbicularis oris

Orbicularis oris ist ein runder Muskel. Dieser zieht sich zusammen, um die Lippen zu bewegen.

Levator labii superioris

Levator labii superioris zieht die Oberlippe nach oben.

Zygomaticus major und minor

Zygomaticus major und minor ziehen die Mundwinkel nach oben und nach hinten zurück, beim Lächeln zum Beispiel.

Depressor anguli oris

Der Depressor anguli oris zieht die Mundwinkel nach unten.

Depressor labii inferioris

Depressor labii inferioris senkt die Unterlippe ab.

Risorius

Risorius zieht die Lippenecken zurück.

Mentalis

Mentalis zieht die Unterlippe nach vorne und ihre Kontraktion führt zu einer Vorwölbung der Lippenwinkel und zum Bewegen der Unterlippe.

Buccinator

Buccinator komprimiert die Kontrolle und zieht den Mundwinkel zurück.

Platysma

Platysma ein großer Muskel, der unter dem Kiefer und im Nacken bis zur oberen Brust reicht. Dieser Muskel drückt auf den Unterkiefer und hilft dabei, die Lippen zu senken.

Masseter

Der Masseter ist ein kräftiger Muskel, der den Kiefer schließt. Der Masseter Muskel ist in der Abbildung leider nicht zu sehen. Es handelt sich dabei um den Kaumuskel.

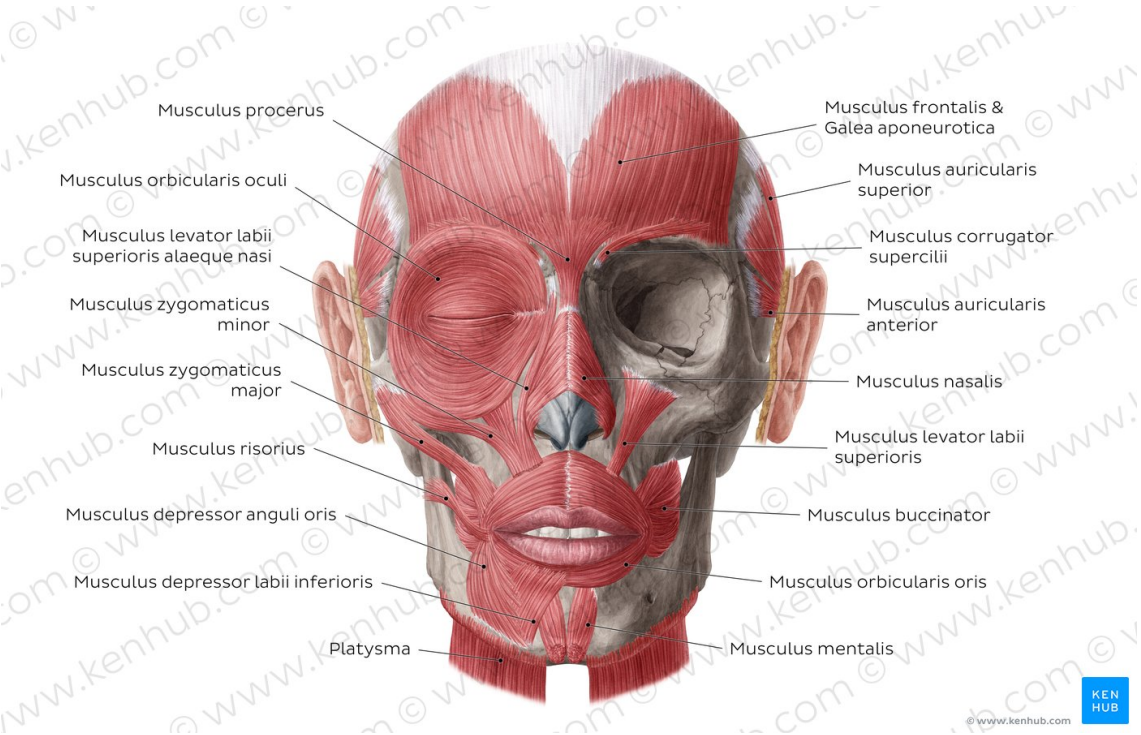


Abbildung 2.5: Die Muskeln im Gesicht eines Menschen. [ken]

2.3.2 Mechanismus

Abbildung 2.6 zeigt den Querschnitt der Anatomie des Sprachsystems innerhalb des menschlichen Oberkörpers. Die groben Komponenten des Systems sind die Lunge, die Luftröhre, der Kehlkopf, der Rachenraum, der Mund und die Nase. Die Rachen- und Mundhöhle können zu einer Einheit, die als Vokaltrakt bezeichnet wird, zusammengefasst werden und die Nasenhöhle wird oft als Nasaltrakt bezeichnet. Weitere für die Sprachproduktion entscheidende Organe, sind die Stimmbänder, die Lippen, die Zunge, die Zähne und das Velum. Der Unterkiefer ist sowohl für grobe als auch feine Bewegungen verantwortlich, welche die Größe und Form des Vokaltrakts sowie die Positionen der anderen Artikulatoren beeinflussen. [DFGM15]

Der Nasaltrakt, Rachenraum und der Mund sind wichtig für die Sprachproduktion. In der Arbeit von Matthias Janke [Jan16] wird die Sprachentwicklung nach der Reihenfolge des Luftstroms erklärt. Dabei wird von Anfang des Luftstroms in der Lunge bis zum Ende des Luftstroms im Nasaltrakt unterschieden. Dabei ist die Lunge die Quelle für die Luft. Die Luft strömt in den Kehlkopf zu den Stimmbändern, von da aus in den Rachen. Danach weicht ein Teil der Luft aus dem Mund hinaus, woraus sich die gesprochene Sprache ergibt. Der Rest der Luft verlässt den Körper durch die Nase. Die Position des Kehlkopfs und Rachen ist in 2.7 zu sehen.

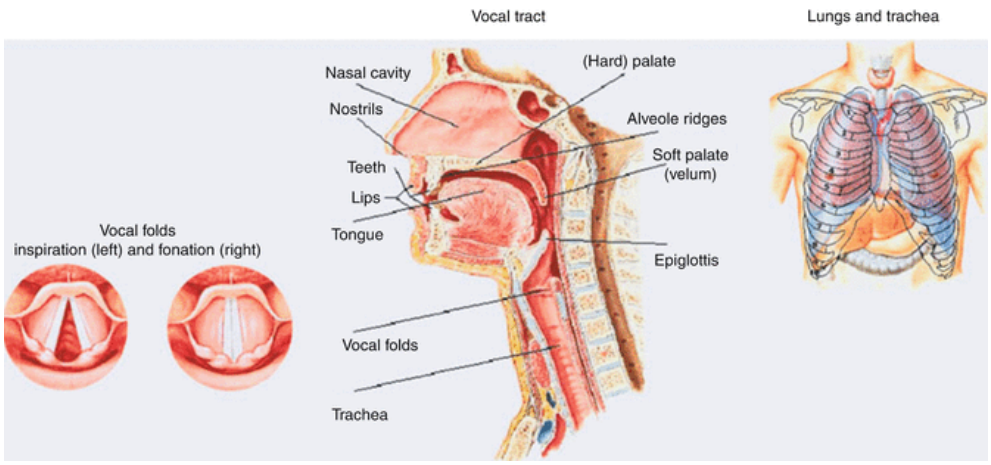


Abbildung 2.6: Die Einzelteile der Menschlichen Sprachproduktion [DFGM15]

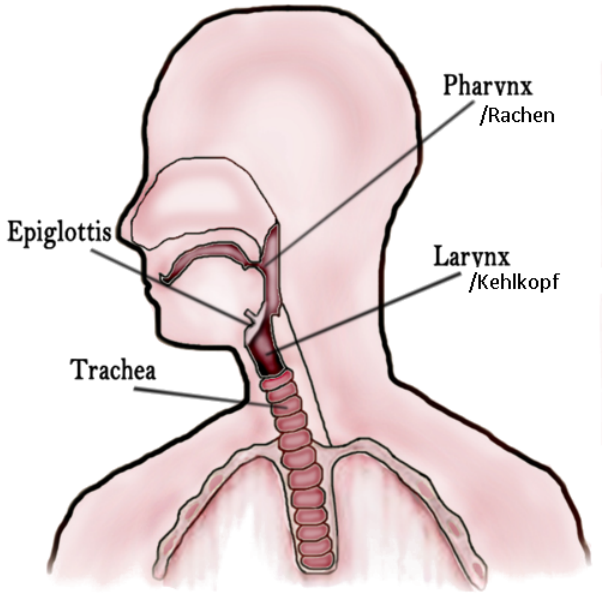


Abbildung 2.7: Positionen des Pharynx und Larynx. [Wikd]

2.4 EMG

In der Elektromyografie (EMG) geht es um das Messen der Muskelfunktion durch die Aufzeichnung des elektrischen Signals, welches vom relevanten Muskel stammt. Diese elektrischen Signale beziehen sich auf die zuvor beschriebenen Aktionspotenziale.

Es wird zwischen zwei Arten von EMG unterschieden, der intramuskulären EMG und der Oberflächen-EMG. In der intramuskulären EMG werden Nadelelektroden durch die Haut in das Muskelgewebe eingeführt. Dies ist nützlich, wenn man Aktionspotenziale einzelner Muskelfasern messen will. Zum Beispiel um durch das EMG-Signalmuster Muskelkrankheiten zu erkennen. Oberflächen-EMG ist hingegen die Aufzeichnung von Signalen mit Hilfe von auf der Haut angebrachten Oberflächenelektroden, es ist also im Vergleich zur intramuskulären EMG nicht invasiv. Oberflächenelektroden haben zudem den Nachteil, dass nur oberflächliche Muskeln abgeleitet werden können. [Kon06].

2.4.1 Digitalisierung

In 2.8 ist ein rohes EMG-Signal zu sehen. Die Grundlinie ist das vordefinierte Grundlinienrauschen. Die drei Kontraktionssalven stammen aus drei Kontraktionen des Bizeps. Das EMG-Signal hat eine stochastische Natur, ist also nicht in ihrer exakten Form reproduzierbar. Dieses Rohsignal wird dann vom Analogen ins Digitale umgewandelt. In 2.9 kann man sehen was für einen Einfluss das Wählen der Abtastrate auf die Digitalisierung des Rohsignals hat. Zudem sieht man das Ergebnis der Digitalisierung mit verschiedenen Abtastfrequenzen. [Kon06]

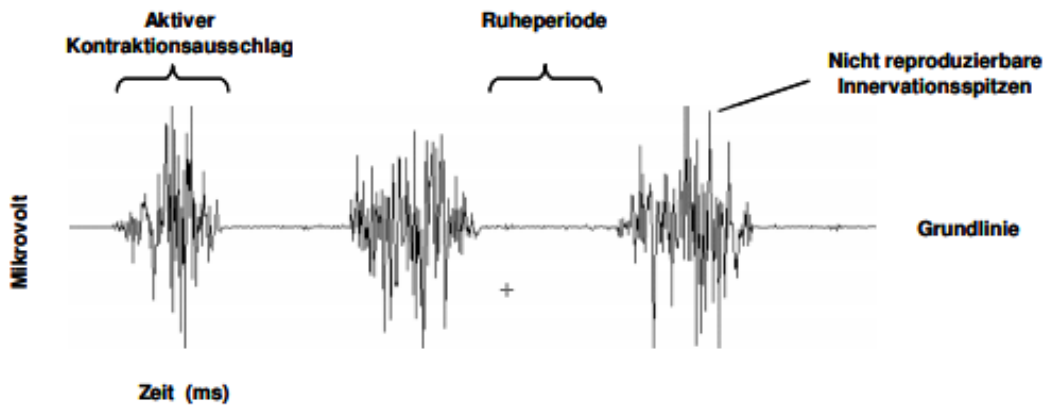


Abbildung 2.8: Die Roh-EMG-Signalaufzeichnung dreier Kontraktionssalven. [Kon01]

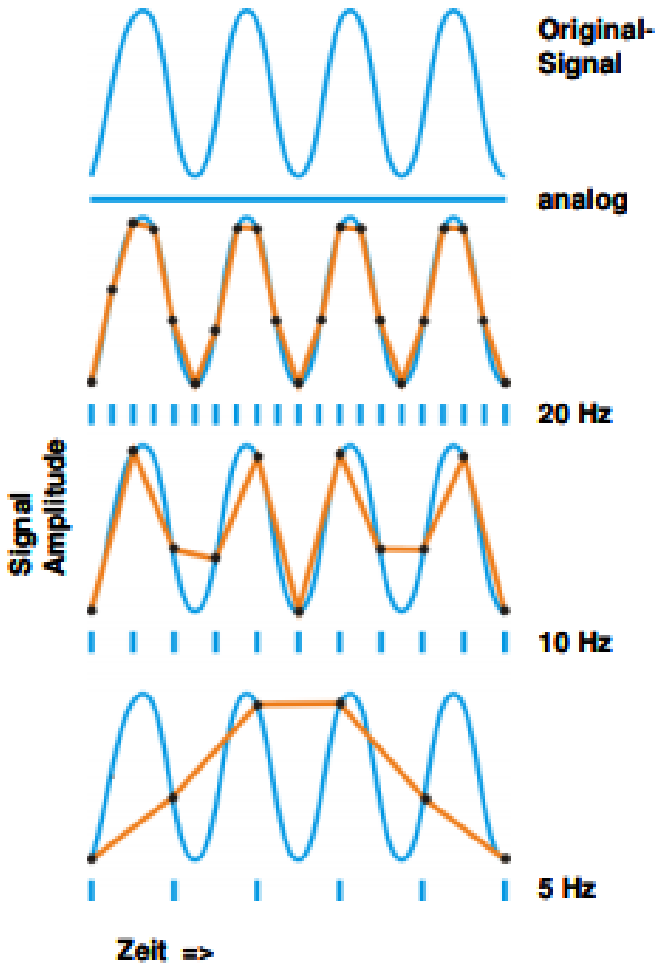


Abbildung 2.9: Der Einfluss der A/D-Abtastfrequenz auf das digitalisierte Signal. [Kon01]

2.5 Datenkorpus

Der Datensatz, den ich für diese Arbeit verwendet habe, ist der [WJS14] UKA-Korpus des CSL in Bremen. Hierbei handelt es sich um eine Sammlung von Oberflächen EMG Aufnahmen von verschiedenen Sprechern in verschiedenen Sprachmodi. Kanal 5 wurde in dieser Arbeit nicht entfernt.

Die Elektrodenpositionen für die Aufnahmen, sowie die relevanten Muskeln sind in der Figur 2.10 zu sehen. Die EMG-Daten wurden mit einer Sechskanal-Elektrodenanordnung aufgezeichnet. 2.10 zeigt die Positionierung der Elektroden, die das EMG-Signal von sechs artikulatorischen Muskeln aufzeichnen. Eine zusätzliche Erdelektrode wurde am Handgelenk der Probanden platziert. [WJS14].

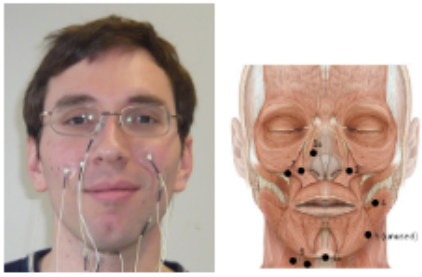


Abbildung 2.10: Elektrodenpositionierung für den EMG-UKA Korpus [WJS14]

2.5.1 Informationen zu den Daten

Die Daten wurden in große Sessions, Multi-Mode-Sessions und kleine Sessions aufgeteilt 2.2. Die Multi-Mode Sessions sind für die Bestimmung des Sprachmodus relevant, denn diese liefern 50 Aufnahmen für jeden Sprachmodus. Die großen Sessions beinhalten mehr als 500 Aufnahmen für den hörbaren Modus. Für die Aufteilung der Sessions nach Sprachmodi sowie die Anzahl der Sessions, durchschnittliche Länge und gesamte Länge der Sessions siehe 2.1. Die großen Sessions wurden in der Sprachmodus Erkennung nicht verwendet, da sie nur aus hörbaren Aufnahmen bestehen [WJS14].

Tabelle 2.1: Datenmenge im EMG-UKA-Korpus. [WJS14]

	Anzahl der Sessions		
Sprecher	Gesamt	Groß	Multi-Mode
1	3	0	3
2	33	1	15
3	1	0	1
4	2	0	2
5	1	0	1
6	1	0	1
7	2	0	2
8	20	1	7
Gesamt	63	2	32

Tabelle 2.2: Zusammenfassung der Sitzungen im gesamten EMG-UKA-Korpus.[WJS14]

Teilmenge	#Sprecher	#Sessions	Durchschnitt	Gesamt
Hörbar(Klein)	8	61	3:08	3:11:34
Flüsternd(Klein)	8	32	3:22	1:42:42
Stumm(Klein)	8	32	3:19	1:46:20
Hörbar(Groß)	2	2	27:02	54:04
Gesamt Menge an Daten ist 7:32 Stunden				

2.5.2 Sprachmodus

Der Sprachmodus wird in dieser Arbeit in drei verschiedene Modi aufgeteilt. Die Sprecher wurden angewiesen, alle Sprachmodi so zu produzieren, wie sie sich am natürlichsten anfühlen [WJS14]. Die englischen sowie die deutschen Namen für die verschiedenen Sprachmodi, wurden in dieser Arbeit verwendet.

Hörbare Sprache(Audible Speech)

Bei der hörbaren Sprache handelt es sich um die normale Stimme die von den Sprechern benutzt wird.

Geflüsterte Sprache (Whispered Speech)

Geflüsterte Sprache wird erzeugt, wenn die Stimmbänder nicht vibrieren, sondern zusammengezogen werden. Dies führt dazu, dass die normale Grundfrequenz des Sprechers durch einen zischenden Ton ersetzt wird. Im EMG-UKA-Korpus werden hörbare und geflüsterte Sprache als getrennte Sprachmodi betrachtet, obwohl geflüsterte Sprache auch hörbar ist.

Stumme Sprache (Silent Speech)

Stumme Sprache bedeutet, dass der Sprecher zwar normale artikulierende Bewegungen ausführt, aber keine Geräusche aus seinem Mund kommen lässt.

2.6 Features

Die Feature-Wahl habe ich auf die Analysen der Features von [Phi12] basiert. Zudem habe ich noch die zwei Spektral Features hinzugenommen. Die von mir in dieser Arbeit gewählten Features sowie ihre mathematische Schreibweise sind im folgendem Teil angegeben. Insgesamt wird in dieser Arbeit ein Featureset von 10 Time Domain Features und 2 Frequency Domain Features genutzt. Es werden für jedes Feature ein Feature Array für jeden Kanal erstellt und anschließend in ein Datenset kombiniert. Bei der Fensterung habe ich mich auf ganze Utterances bezogen. Also jedes Feature wird für jeden Kanal für jede Utterance ein Mal ausgerechnet. Mit einer strengeren Fensterung konnten keine besseren Ergebnisse erzielt werden.

2.6.1 Time Domain Features

Features im Zeitbereich(Time Domain Features) sind schnell und einfach implementiert, da diese keine Transformation benötigen. Es werden die rohen Daten genutzt, um die untenstehenden Berechnungen anzuwenden. Das steht im Gegensatz zu den Frequency Domain Features, welche eine Transformation der Daten benötigen.

2.6.1.1 MAV

Der mittlere absolute Wert (Mean Absolute Value; MAV) ist ein Mittelwert des absoluten Wertes der EMG-Signalamplitude in einem Segment des Signals, der definiert werden kann als

$MAV = \frac{1}{N}\sum_{i=1}^{N}|x_i|$.

wobei x_i das EMG-Signal in dem Teil i repräsentiert und N die Länge des EMG-Signals ist. Das ist für die anderen Features genauso.

2.6.1.2 MAV2

Die MAV2-Funktion ist der absolute Mittelwert aller MAV Werte des gesamten EMG-Signals.

2.6.1.3 RMS

Das mittlere Quadrat der Wurzel (Root Mean Square; RMS) wird als Gauß'scher Zufallsprozess modelliert. Die mathematische Definition des RMS-Merkmals ist $RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}$.

2.6.1.4 IAV

Bei dem IAV in dieser Arbeit handelt es sich um das Integral der EMG Werte. Das Merkmal ist in dieser Arbeit anders als das IAV Merkmal in [Phi12], wo es als ein anderer Name für MAV beschrieben wird.

$IAV = \sum_{i=1}^{N}|x_i|$

2.6.1.5 WL

Die Wellenformlänge (WL) ist ein Maß für die Komplexität des EMG Signals. Es ist definiert als Länge der EMG-Wellenform über die Zeit. Es kann berechnet werden mit der Formel:

$WL = \sum_{i=1}^{N-1} |x_{i+1} - x_i|$.

2.6.1.6 IEMG

Der Integrated EMG (IEMG) ist definiert als eine Summierung von Absolutwerten der EMG-Signalamplitude, die so berechnet wird:

$IEMG = \sum_{i=1}^{N} |x_i|$.

2.6.1.7 AAC

Die durchschnittliche Amplitudenänderung (Average Amplitude Change; AAC) ist sehr ähnlich zum WL Feature, außer dass die Wellenlänge gemittelt wird. Das sieht man auch ganz einfach an der Formel für dieses Feature:

$AAC = \frac{1}{N} \sum_{i=1}^{N-1} |x_{i+1} - x_i|$.

2.6.1.8 ZC

Das Zero Crossing ist die Anzahl der Amplitudenwerte des EMG-Signals, welche den Null-Amplitudenpegel überschreiten. Um Niederspannungsschwankungen und Hintergrundgeräusche zu vermeiden, wird eine Schwellenwertbedingung implementiert und die Berechnung ist definiert als,

$\sum_{i=1}^{N-1} [f(x_i * x_{i+1}) \bigcap |x_i - x_{i+1}| \geq grenze]$;

Wobei $f(x) = \begin{cases} 1, & \text{if } x \geq grenze \\ 0, & \text{otherwise} \end{cases}$

Die Grenze ist da, damit sehr niedrige Spannungsänderungen nicht aufgefangen werden. Die Grenzwerte sollten im 50 µV bis 100 mV Bereich liegen.

2.6.1.9 SSC

Der Vorzeichenwechsel (SSC) beschreibt die Häufigkeit, mit der die Steigung des EMG-Signals das Vorzeichen wechselt, also die Anzahl der Änderungen zwischen positiver und negativer Steigung. Dies kann mathematisch ausgedrückt werden als

$SSC = \sum_{i=2}^{N-1} [f[(x_i - x_{i-1}) * (x_i - x_{i+1})]]$;

Wobei $f(x) = \begin{cases} 1, & \text{if } x \geq grenze \\ 0, & \text{otherwise} \end{cases}$

2.6.1.10 WAMP

Die Willison-Amplitude(WAMP) ist ein Maß für die Frequenzinformation des EMG-Signals so wie in dem ZC Feature näher erklärt wurden. Sie beschreibt die Anzahl an Unterschieden , zwischen der Signalamplitude zweier benachbarter Segmente , welche einen vordefinierten Schwellenwert überschreiten. Darüber hinaus steht dieses Feature im Zusammenhang mit der Zündung von Aktionspotenzialen in den motorischen Einheiten (MUAP) und der Muskelkontraktionskraft. Die Definition lautet wie folgt

$WAMP = \sum_{i=1}^{N-1} [f(|x_n - x_{n+1}|)]$;

Wobei $f(x) = \begin{cases} 1, & \text{if } x \geq grenze \\ 0, & \text{otherwise} \end{cases}$

2.6.2 Frequency Domain Features

Bei den Frequenz- oder Spektralbereichsmerkmalen (Frequency Domain Features) wird das Frequenzspektrum des EMG-Signals analysiert. Dafür muss zuerst eine Fourier-Transformation der Daten stattgefunden haben.

Diese zwei Features habe ich in einer Arbeit zur Erkennung von Handsymbolen auf Basis von EMG-Signalen gefunden und habe sie deshalb in diese Arbeit eingebunden. Für diese Features werden die Daten zuerst durch eine Fourier-Transformation verändert. [Lee19]

2.6.2.1 Mean Frequency

MNF ist die Durchschnittsfrequenz, welche als Summe aus dem Produkt des EMG-Leistungsspektrums und der Frequenz geteilt durch die Gesamtsumme der Spektrumsintensität berechnet wird. Das Leistungsspektrum wird durch eine Fourier Transformation erreicht. Die Berechnung dieses Features ist:

$MNF = \sum_{j=1}^{M} f_j P_j / \sum_{j=1}^{M} P_j$

Wobei f_j die Frequenz des Spektrums für den Frequenzbereich j ist, Pj das EMG-Leistungsspektrum des Frequenzbereiches j ist und M die Länge des Frequenzbereichs j ist. Das ist bei der Median Frequency genau so.

2.6.2.2 Median Frequency

Median Frequency (MDF) ist eine Frequenz, bei der das Spektrum in zwei Bereiche mit gleicher Amplitude aufgeteilt wird. Die Median Frequency kann mit dieser Formel beschrieben werden:

$\sum_{j=1}^{MDF} P_j = \sum_{j=MDF}^{M} P_j = \frac{1}{2} \sum_{j-1}^{M} P_j$

2.7 Klassifikatoren

In dieser Arbeit wurden mehrere verschiedene Machine Learning Klassifikatoren genutzt und die Resultate miteinander verglichen [SSBD14]. Alle Klassifikatoren wurden mit den Standardeinstellungen der Bibliothek genutzt.

2.7.1 LDA

Die Lineare Diskriminanzanalyse (LDA) funktioniert, indem sie die Dimensionalität des Datensatzes reduziert und alle Datenpunkte auf eine Linie projiziert. Dann kombiniert sie diese Punkte in Klassen auf der Grundlage ihrer Entfernung von einem ausgewählten Punkt oder Schwerpunkt.

Die lineare Diskriminanzanalyse ist ein linearer Klassifizierungsalgorithmus und wird am besten verwendet, wenn die Daten eine lineare Beziehung aufweisen.

In 2.11 ist ein sehr grobes Beispiel für einen LDA-Algorithmus zu sehen. Das Datenset wird hierbei durch die Linie, welche mit den Sortierkriterien erstellt wird, in zwei Klassen geteilt, wobei alles über der Linie zu einer Klasse und alles unter der Linie zu einer anderen Klasse gehört.

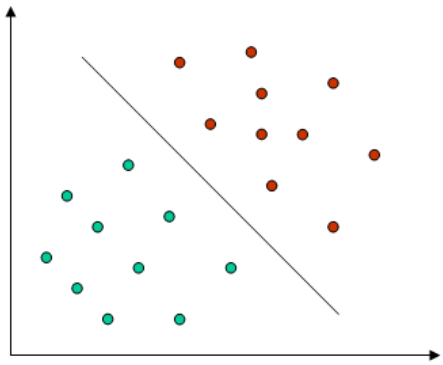


Abbildung 2.11: Sehr grober Vorgang bei einem LDA. Die verschiedenfarbigen Punkte repräsentieren verschiedene Klassen. [Wika]

2.7.2 KNN

K-Nearest Neighbors arbeitet, indem es den Abstand von einem Testbeispiel zu den bekannten Werten eines Trainingsbeispiels berechnet. Die Gruppe von Datenpunkten, die den kleinsten Abstand zwischen den Trainingspunkten und dem Testpunkt ergeben, ist die ausgewählte Klasse. In der Abbildung 2.12 sieht man Datenpunkte aus drei verschiedenen Klassen. In 2.13 ist der Gebrauch eines KNN Algorithmus zu sehen. Man kann sehen, dass die meisten Datenpunkte, die Nachbarn waren, auch zur selben Klasse gehören.

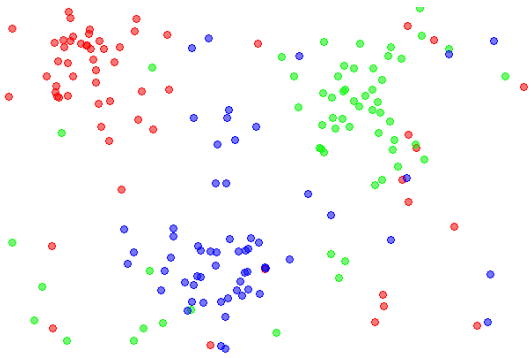


Abbildung 2.12: Datenpunkte von 3 Klassen. Die Klassen hier sind Rot, Grün und Blau.[Wikb]

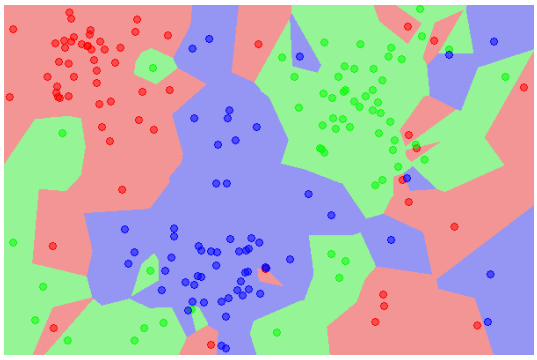


Abbildung 2.13: KNN angepasst an die Daten von 2.12[Wikc]

2.7.3 DecisionTree

Ein DecisionTree-Klassifikator funktioniert, indem er einen Datensatz auf der Grundlage verschiedener Kriterien in immer kleinere Teilmengen zerlegt. Zur Unterteilung des Datensatzes werden verschiedene Sortierkriterien verwendet, wobei die Anzahl der Beispiele mit jeder Unterteilung kleiner wird.

Sobald das Netzwerk die Daten auf ein Beispiel herunter gebrochen hat, wird das Beispiel in eine Klasse eingeordnet. Wenn mehrere DecisionTree Klassifikatoren miteinander verknüpft werden, werden sie RandomForest-Klassifikator genannt.

2.7.4 RandomForest

Randomforest ist ein überwachter maschineller Lernalgorithmus, der auf Ensemble-Lernen basiert. Ensemble-Lernen ist eine Art des Lernens, bei dem man verschiedene Arten von Algorithmen oder denselben Algorithmus mehrfach miteinander verbindet, um ein besseres Vorhersagemodell zu bauen. Der RandomForest-Algorithmus kombiniert mehrere Entscheidungsbäume.

2.7.5 SVC

SVCs(Support Vector Machines) arbeiten, indem sie eine Linie zwischen den verschiedenen Clustern von Datenpunkten ziehen, um diese in Klassen zu gruppieren. Die Klassen werden dann je nach den Abständen der Punkte zu den Linien eingeteilt.

Die Abbildungen 2.14 und 2.15 zeigen einen Datensatz vor und nach dem Trainieren eines oben erklärten SVC-Klassifikators. Es wird der Abstand der Punkte zur gezeichneten Linie gemessen. Dadurch wird ermittelt, dass Blau zu einer Klasse und Gelb zu einer anderen Klasse gehört.

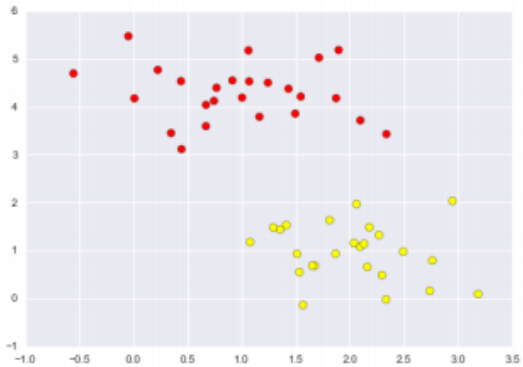


Abbildung 2.14: Daten für die Durchführung des SVC Algorithmus. Die gelben und blauen Punkte stehen für Datenpunkte verschiedener Klassen. Die Einheiten der Achsen sind unwichtig für die Darstellung. [Van16]

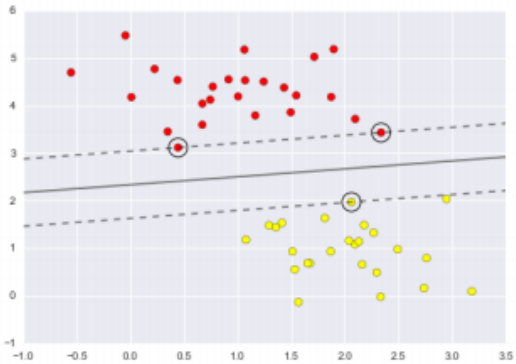


Abbildung 2.15: Ein SVM-Klassifikator, der an die Daten angepasst ist, mit Margins (gestrichelte Linien) und den Stützvektoren (Kreise) dargestellt. Die Einheiten der Achsen sind unwichtig für die Darstellung. [Van16]

3. Versuchsaufbau

3.1 Aufteilung der Daten

Die Daten der Sprecher mit weniger als 2 Sessions werden in dieser Arbeit nicht verwendet. Demnach werden die Daten der Sprecher 1,2,4,7 und 8 verwendet. In den Versuchen zu dem Sprachmodus werden alle Sessions die nur audible Dateien enthalten entfernt.

3.1.1 Nutzererkennung

Für die Nutzererkennung werden alle Sessions außer einer genutzt, um den Klassifikator zu trainieren. Danach wird der trainierte Klassifikator auf der letzten Session getestet. Aus den Ergebnissen wird die durchschnittliche Genauigkeit der trainierten Klassifikatoren, sowie die Standardabweichung zwischen den Genauigkeiten der Klassifikatoren angeguckt. Zudem werden die Versuche mit den zwei größten Sessions und keiner anderen Session wiederholt. Zum Beispiel habe ich die Daten aller Sessions außer der ersten Session von Nutzer 1 verwendet und habe damit den Klassifikator trainiert. Dann habe ich den Klassifikator mit den Daten von der ersten Session von Nutzer 1 getestet.

Dies wird mit mehreren Klassifikatoren durchgeführt und die Resultate werden miteinander verglichen. Die verwendeten Klassifikatoren sind in Kapitel 2.6 zu sehen. Für den SVC Klassifikator werden die Kernel Linear,Poly,Gaussian und Sigmoid verwendet. In Kapitel 4 ist die durchschnittliche Genauigkeit der Erkennung für jeden Klassifikator zu sehen. Dazu sind in Kapitel 4 noch Resultate der einzelnen Sessions zu sehen. Es wird zudem versucht durch das entfernen von Daten ein besseres Ergebnis zu erreichen. Am Ende wird eine Confusion Matrix Analyse für einzelne Sessions sowie für die Daten an sich durchgeführt. Diese Versuche werden je einmal mit den ungefilterten Daten und einmal mit den gefilterten Daten durchgeführt. Es wurde ein Bandpassfilter mit Grenzfrequenzen von 10 Hz und 200 Hz verwendet um die Daten zu Filtern.

3.1.2 Sprachmodus

Für die Erkennung des Sprachmodus sind die Daten in zwei verschiedene Weisen aufgeteilt. Es wird ein Klassifikatoren auf den Daten von allen außer einem Nutzer trainiert und danach auf dem letzten Nutzer getestet(Cross-User). Bei der anderen Variante werden für jeden Nutzer alle Sessions außer einer genutzt, um den Klassifikator zu trainieren und danach wird der Klassifikator auf der letzten Session getestet(Cross-Session). Bei beiden Varianten wird am Ende die durchschnittliche Genauigkeit der trainierten Klassifikatoren sowie die Standardabweichung zwischen den Genauigkeiten der Klassifikatoren angeguckt. Diese Versuche werden je einmal mit den ungefilterten Daten und einmal mit gefilterten Daten durchgeführt. Für die gefilterten Daten wurde ein Bandpassfilter mit Grenzfrequenzen von 10 Hz und 200 Hz verwendet.

Dies wird mit mehreren Klassifikatoren durchgeführt und die Resultate miteinander verglichen. Für die Cross-Session Versuche werden die Genauigkeiten der einzelnen Sessions ebenfalls angegeben. Zudem habe ich alle Sessions herausgefiltert die nur den Sprachmodus *hörbar* beinhalten. Am Ende wird das Signal in verschiedene Frequenzbänder geteilt und die Experimente nochmal durchgeführt um zu sehen, ob es einen bestimmten Muskel gibt der mit einer besseren Genauigkeit vorhersagen kann als die anderen Muskeln. Hierfür wird ein LDA verwendet.

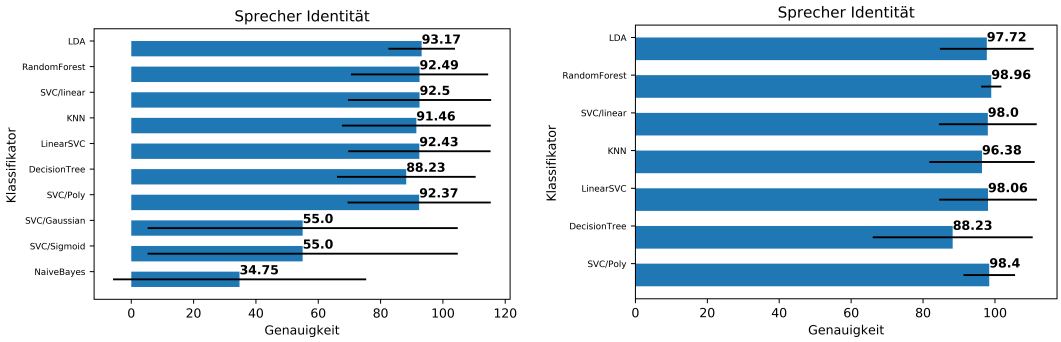
4. Evaluation

4.1 Sprecheridentität

Hier wurde die Genauigkeit der Sprecheridentität mit verschiedenen Klassifikatoren analysiert. Zudem wurde eine Confusion-Matrix Analyse durchgeführt. Bei den Confusion-Matrix Analysen sowie bei der Betrachtung der einzelnen Sessions wurde ein LDA Klassifikator genutzt, weil dieser die höchste Genauigkeit gezeigt hat. Dies wurde einmal mit gefilterten und einmal mit ungefilterten Daten durchgeführt.

4.1.1 Ungefiltert

Bei komplett ungefilterten Daten kommt man bei der Erkennung der Sprecheridentität zu sehr guten Ergebnissen für die Genauigkeit sowie die Standardabweichung. Die Ergebnisse sind in 4.1(a) zu sehen. Die Ergebnisse für die LDA sind hier am Besten. Die Genauigkeit ist hier am höchsten und die Standardabweichung am geringsten. Die Ergebnisse für die SVC Klassifikatoren mit dem Kernel Gaussian und Sigmoid sind genau bei der Chance Genauigkeit von 55 Prozent. Bei der Aufteilung der Ergebnisse kann man sehen das alle Sessions Entweder 100 oder 0 sind, wodurch sich die Genauigkeit von 55 Prozent und eine sehr hohe Standartabweichung ergibt. Die Genauigkeit von NaiveBayes ist hier mit 34.75 am geringsten. Man sieht in 4.2, dass die Erkennungsrate für die Sprecher 1,2 und 8 , sehr hoch sind. Sprecher 4 hat eine 50 prozentige Erkennungsrate und Sprecher 7 eine nahe 0 prozentige Erkennungsrate. Bei diesen Tests viel auf, dass die Genauigkeit für die Sprecher mit weniger Daten vergleichbar viel niedriger ist als die Genauigkeit für die anderen Sprecher. Deshalb wurde ein zweiter Versuch durchgeführt, in Dem nur die Daten der Sprecher 2 und 8 genutzt wurden. Dabei waren die Resultate deutlich besser als in dem vorherigen Versuch (siehe 4.1(b)).



(a) Sprecher 1,2,4,7 und 8

(b) Sprecher 2 und 8

Abbildung 4.1: Durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedenen Klassifikatoren bei der Erkennung der Sprecheridentität.

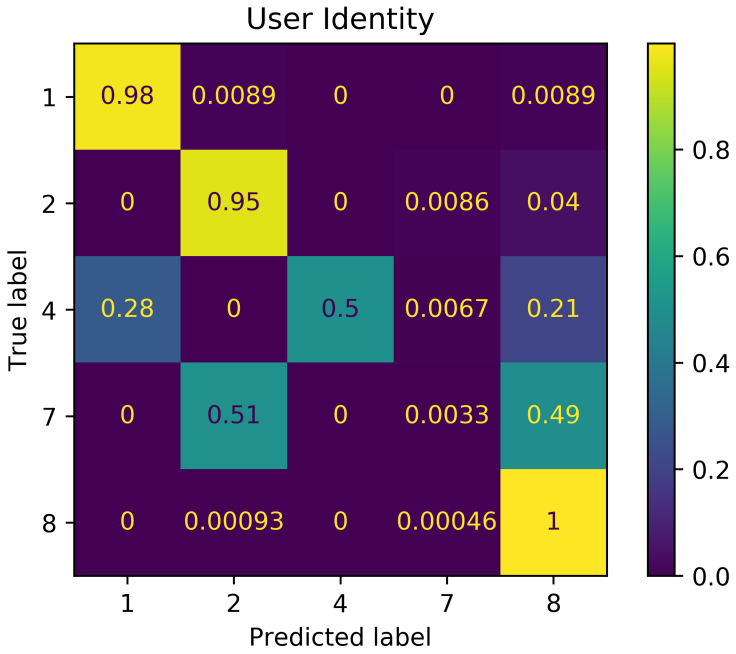


Abbildung 4.2: Confusion-Matrix der Erkennung der Sprecheridentität mit ungefilterten Daten.

4.1.1.1 Sessions

Bei den Genauigkeiten der einzelnen Sessions sieht man, dass es eine ziemlich gleichmäßige Genauigkeit zwischen allen Sessions gibt. Die meisten Sessions von Sprecher 2 und 8 haben eine Genauigkeit von 100 Prozent deshalb wurde in der Tabelle 4.1 die Sessions mit 100 prozentiger Genauigkeit entfernt. Es existieren jedoch einzelne Sessions die deutlich geringere Genauigkeiten aufweisen. In Sprecher 2 sieht man zum Beispiel in den Sessions 29 und 26 eine viel geringere Genauigkeit als bei allen anderen Sessions. In 4.3 sieht man, dass in über 73 Prozent der Fälle Sprecher 2 mit Sprecher 8 verwechselt wird. In Session 26 sieht man, dass mit 32 Prozent

ebenfalls eine hohe Verwechslung-rate mit Sprecher 8 vorhanden ist zudem ist in dieser Session noch eine 17 Prozent Verwechslung-rate mit Sprecher 7. Die restlichen Sessions von Sprecher 2 haben entweder eine Genauigkeit von 100 Prozent oder haben sehr wenige Verwechslungen mit dem anderen Sprechern.

Tabelle 4.1: Genauigkeiten aller Sessions bei der Sprecher Erkennung.

Session	Resultat in Prozent
1001	96
1002	99,33
1003	99,33
2004	96,67
2008	99,33
2026	51
2027	93
2029	26,67
2031	96
4001	100
4002	0
7001	0,7
7002	0
8005	98
8014	98

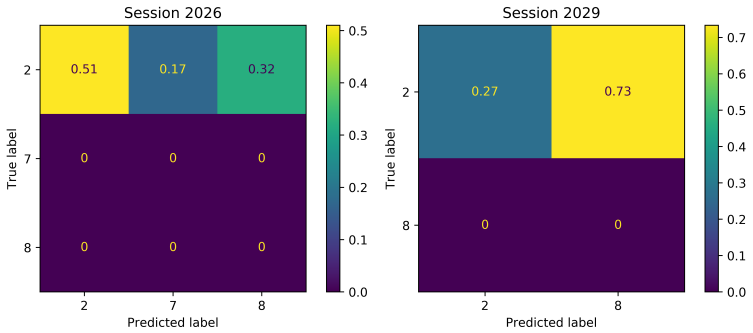


Abbildung 4.3: Confusion-Matrix der Session 29 und 26 von Sprecher 2

Bei Sprecher 8 mit einer vergleichbaren Anzahl an Sessions sind solche Ausnahmen nicht zu sehen. Bei Sprecher 4 in Session 2 und in beiden Sessions von Sprecher 7 sieht man eine weit niedrigere Genauigkeit als in den Sessions der anderen Sprecher. Für Sprecher 4 ist die Genauigkeit bei der zweiten Session 0 Prozent und bei der ersten Session sehr hoch (4.1). In 4.4 sieht man wie sich die Genauigkeit von 0 in dieser Session ergibt. Sprecher 4 wird hier zu 57 Prozent mit Sprecher 1 und zu 42 Prozent mit Sprecher 8 verwechselt.

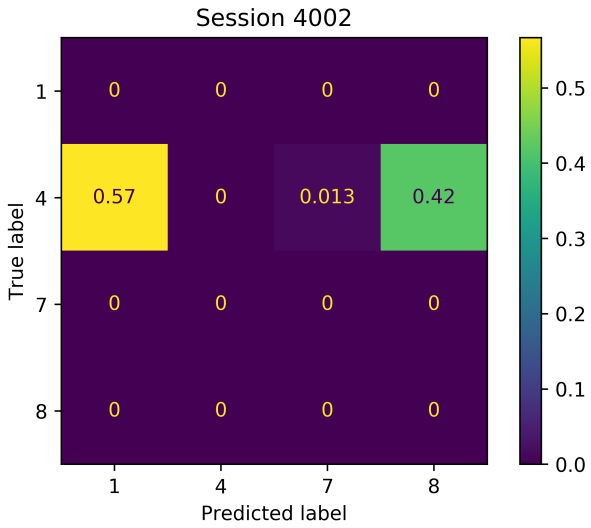


Abbildung 4.4: Confusion-Matrix der Session 2 von Sprecher 4

Ich vermute, dass es nicht nur auf die Anzahl der Sessions, sondern auch auf die Qualität der Sessions ankommt, denn bei Sprecher 4 kann man einen großen Unterschied zwischen der ersten und der zweiten Session sehen. Zudem hat Sprecher 7 eine niedrigere Genauigkeit in beiden Sessions, also scheint die Anzahl an Sessions alleine auch eine Rolle zu spielen. In 4.5 sieht man, warum die Genauigkeit von Sprecher 7 so niedrig ist. In Sprecher 7 werden nur die Sprecher mit den meisten Sprechern, also 2 und 8, vorhergesagt.

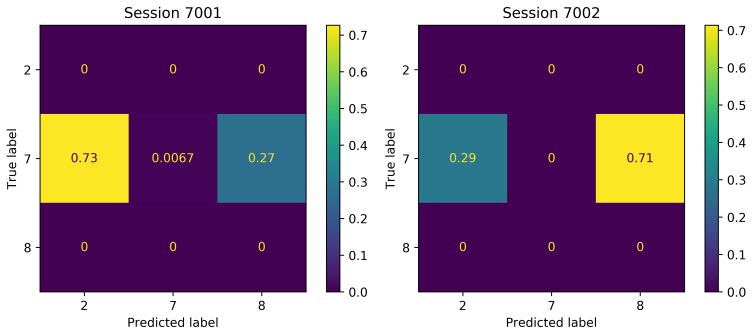


Abbildung 4.5: Confusion-Matrix der beiden Sessions von Sprecher 7

Die Sprecher mit den meisten Sessions haben nämlich eine sehr hohe Genauigkeit (siehe Sprecher 2 und 8). Jedoch ist in Sprecher 1 die Genauigkeit mit über 95 Prozent für jede Session, trotz weniger Daten sehr hoch wodurch ich mehr auf die Qualität der Sessions als Hauptmerkmal tendiere. Bei diesen Tests viel auf, dass die Genauigkeit für die Sprecher mit weniger Daten vergleichbar viel niedriger sein kann als die Genauigkeit für die anderen Sprecher. Deshalb habe wurde ein zweiter Versuch durchgeführt in dem nur die Daten der Sprecher 2 und 8 beobachtet wurde. Dabei waren die Resultate deutlich besser als in dem vorherigen versuch (siehe 4.1(b)). Um den Effekt der Sessionqualität auf die Genauigkeit zu bestimmen wurden im späteren Verlauf die Sessions mit einer niedrigen Genauigkeit entfernt und es wurde erneut getestet.

Session 29 entfernt

Man sieht in 4.6 das die Erkennungsrate für Sprecher 2 um 3 Prozent von 95 auf 98 Prozent steigt. Für Sprecher 7 steigt die Verwechslung-rate mit Sprecher 2 um 5 Prozent von 51 zu 56 Prozent. Ich gehe davon aus das es steigt, weil Sprecher 2 jetzt einen gleichmäßigeren Datensatz besitzt. Die anderen Sprecher bleiben gleich.

Die Ergebnisse sowie der Vergleich zu der Version mit Session 29 ist in der Tabelle 4.2 zu sehen. Es werden die vorherigen Ergebnisse mit den neuen verglichen und in der letzten Zeile gezeigt wie es sich geändert hat.

In den vorherigen Tests ist aufgefallen, dass Session 29 beim Sprecher 2 eine viel niedrigere Genauigkeit aufweist als die anderen Sessions desselben Sprechers. Aus diesem Grund wurde hier ein weiterer Test durchgeführt in dem Session 29 aus Sprecher 2 entfernt wurde. Bei den Ergebnissen dieses Versuches ist eine Steigerung der Genauigkeit, in Session 26 von Sprecher 2 von 51 auf 58 Prozent zu beobachten. Genauer kann man in 4.7 sehen das die Verbesserung der Genauigkeit auf die verringerte Anzahl der Verwechslungen von Sprecher 2 und Sprecher 8 zurückzuführen ist. Gleichzeitig ist aber auch ein Fall der Genauigkeit in den Sessions 27 und 31 zu sehen. Die niedrigere Genauigkeit liegt an der höheren Verwechslung-rate von Sprecher 2 mit Sprecher 7 (4.8(a) und 4.8(b)) in Sprecher 1 kann man eine leichte Entwicklung der Genauigkeiten ins Negative in 4.4 kann man sehen das Sprecher 2 am häufigsten mit Sprecher 1 verwechselt wird. Den ersten beiden Sessions beobachten (4.2). Bei Sprecher 4 ist wiederum ein leichter Fall der Genauigkeit für die erste Session von 100 Prozent auf 99.33 Prozent zu sehen(4.2) die zweite Session bleibt jedoch bei 0, ich gehe davon aus das bei dieser Session ein Fehler bei der Aufnahme passiert ist. Die Genauigkeit bei Sprecher 7 ändert sich nicht.

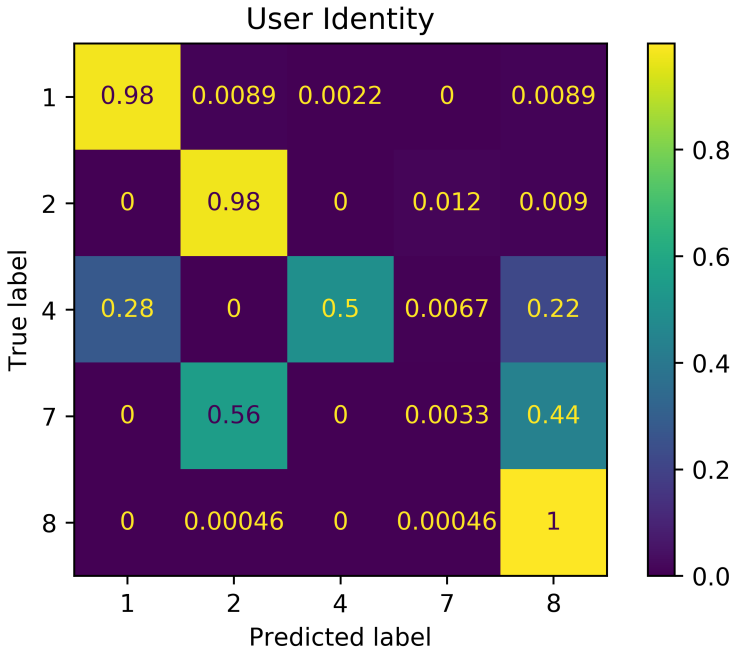


Abbildung 4.6: Confusion-Matrix der Erkennung Sprecheridentität mit ungefilterten Daten.

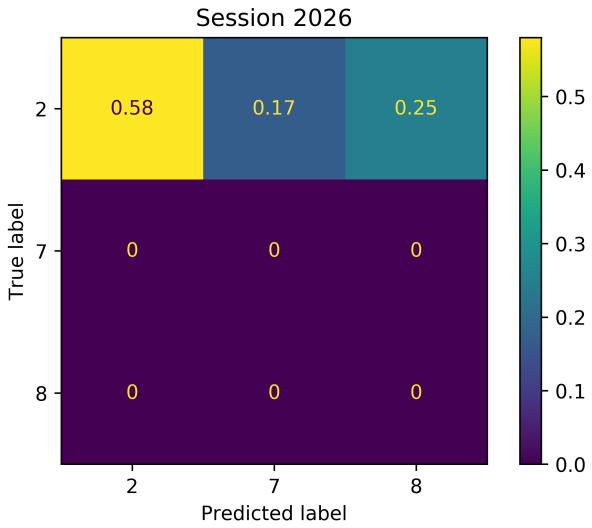
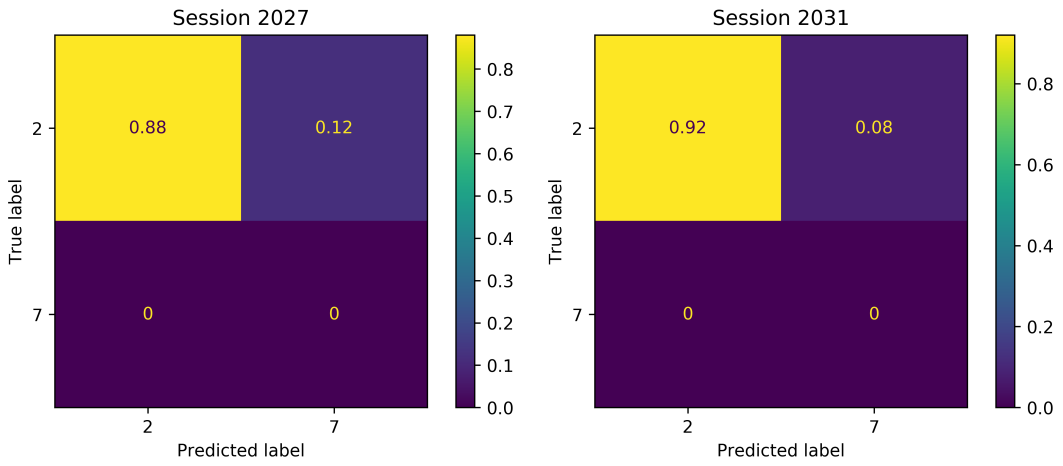


Abbildung 4.7: Erkennungsraten verschiedener Sprecher bei der Klassifizierung der Sprecheridentität in Session 26 von Sprecher 2.

Tabelle 4.2: Genauigkeiten aller Sessions bei der Sprecher Erkennung. Es werden die Resultate der gefilterten Daten mit den Resultaten der ungefilterten Verglichen.

Session	Resultat	Resultat	Veränderung
1001	96	96.67	-0.67
1002	99,33	98.67	-0.67
1003	99,33	98.67	-0.67
2004	96,67	96.67	0
2008	99,33	99.33	0
2026	51	58	+7
2027	93	88	-5
2031	96	92	-4
4001	100	99.33	-0.67
4002	0	0	0
7001	0,7	0,7	0
7002	0	0	0
8005	98	100	+2
8014	98	98	0

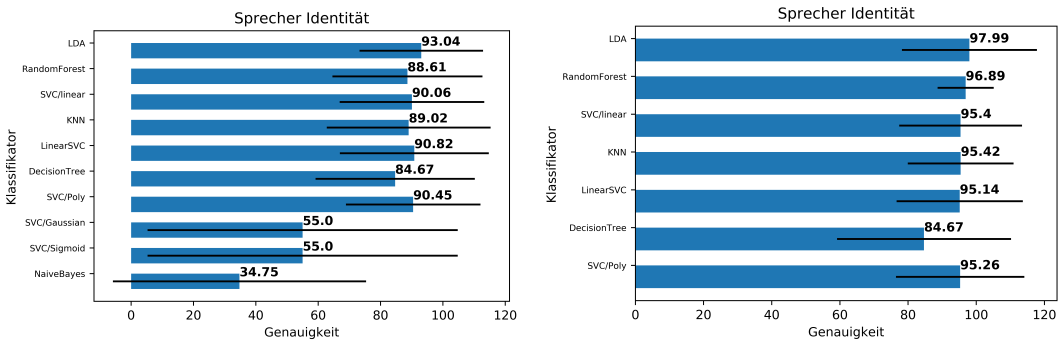


(a) Session 27 (b) Session 31

Abbildung 4.8: Die Erkennungsraten verschiedener Sprecher bei der Klassifizierung der Sprecheridentität in Session 27 und 31 von Sprecher 2.

4.1.2 Filtered

Hier habe ich einen Bandpassfilter(10-200Hz) auf die Daten angewendet. In der Abbildung 4.9(a) kann man sehen, dass die Resultate leicht niedriger sind als bei der Variante ohne Filter. Zudem habe ich hier auf die Klassifikatoren mit einer sehr niedrigen Genauigkeit verzichtet. Man sieht in 4.10, das die Erkennungsraten sehr ähnlich zu den Erkennungsraten in der ungefilterten Version sind 4.2. Lediglich Sprecher 1 hat mit 81 Prozent eine weit niedrigere Erkennungsrate als in der ungefilterten Version. Sprecher 2 hat eine Erkennungsrate von 96 Prozent. Sprecher 4 hat eine 43 prozentige Erkennungsrate und eine 45 prozentige Verwechslungsrate mit 8. Sprecher 7 hat eine 36 prozentige Erkennungsrate und eine 36 prozentige Verwechslungsrate mit 2 und 28 prozentige Verwechslungsrate mit 8.



(a) Sprecher 1,2,4,7 und 8 (b) Sprecher 2 und 8

Abbildung 4.9: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedener Klassifikatoren bei der Erkennung der Sprecheridentität.

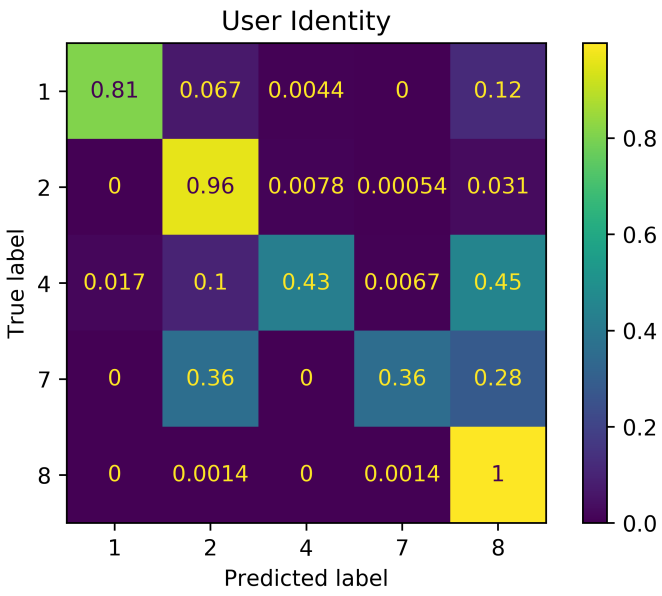


Abbildung 4.10: Die Confusion-Matrix der Erkennung Sprecheridentität mit gefilterten Daten.

Bei diesen Tests viel auf, dass die Genauigkeit für die Sprecher mit weniger Daten vergleichbar viel niedriger ist als die Genauigkeit für die anderen Sprecher. Deshalb wurde ein zweiter Versuch gemacht, indem nur die Daten von Sprecher 2 und 8 verwendet wurden. Dabei waren die Resultate deutlich besser als in dem vorherigen versuch (siehe 4.9). Die Resultate hier sind leicht niedrigerer als bei der ungefilterten Variante.

4.1.2.1 Sessions

In der Tabelle 4.3 sieht man die Genauigkeit für jede Session. Die Sessions mit 100 Prozent Genauigkeit wurden entfernt. Die meisten der Sessions für Sprecher 8 und 2 hatten hier eine Genauigkeit von 100 Prozent. Es gibt jedoch einzelne Sessions die deutlich geringere Genauigkeiten aufweisen. In Sprecher 2 sieht man zum Beispiel in den Sessions 29 und 26 eine viel geringere Genauigkeit als bei allen anderen Sessions. Ich gehe, wie bei der ungefilterten Version, davon aus das in diesen Sessions Fehler bei der Aufnahme passiert sind. Bei Sprecher 8 mit einer vergleichbaren Anzahl an Sessions sind solche Ausnahmen nämlich in Sprecher 8 auch nicht zu sehen.

Tabelle 4.3: Genauigkeiten aller Sessions bei der Sprecher Erkennung mit gefilterten Daten.

Session	Resultat in Prozent
1001	66
1002	81,33
1003	95,33
2002	98
2015	98
2026	68
2029	26,67
4001	85.33
4002	0
7001	36
7002	34.67
8006	96

In 4.11 sieht man das in über 73 Prozent der Fälle Sprecher 2 mit Sprecher 8 verwechselt wird. In Session 26 sieht man mit 32 Prozent ebenfalls eine hohe Verwechslungsrate mit Sprecher 8 und zudem noch eine 17 Prozent Verwechslungs-rate mit Sprecher 7. Die restlichen Sessions von Sprecher 2 haben entweder eine Genauigkeit von 100 Prozent oder haben nur sehr wenige Verwechslungen mit dem anderen Sprechern.

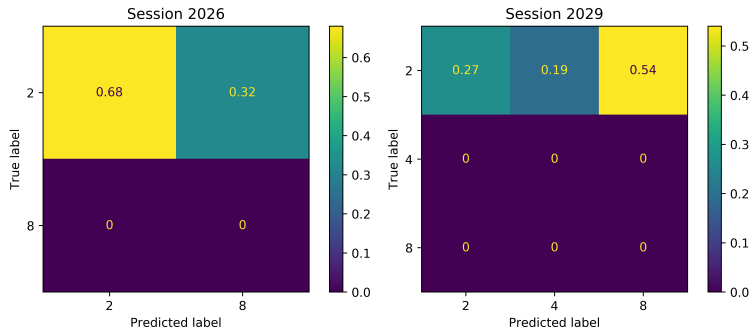


Abbildung 4.11: Confusion-Matrix der Session 29 und 26 von Sprecher 2

Für Sprecher 4 in Session 2 und für Sprecher 7 in beiden Sessions sieht man eine weit niedrigere Genauigkeit als in den Sessions der anderen Sprecher, dasselbe lässt sich für Sprecher 1 weniger stark betrachten. Hier hat die erste Session eine Genauigkeit von 66 Prozent und die zwei anderen Sessions jeweils eine Genauigkeit von 81.33 und 95.33 Prozent. Die Genauigkeit ist hier niedriger als bei der ungefilterten Version. Bei Sprecher 4 ist die Genauigkeit bei der zweiten Session 0 und bei der ersten Session sehr hoch. In 4.12 sieht man die Zusammensetzung dieser Genauigkeit. Anders als bei der ungefilterten Variante haben wir hier nämlich in der zweiten Session kaum Sprecher 1 vorhersagen, dafür sind jetzt 90 Prozent der vorhersagen Sprecher 8.

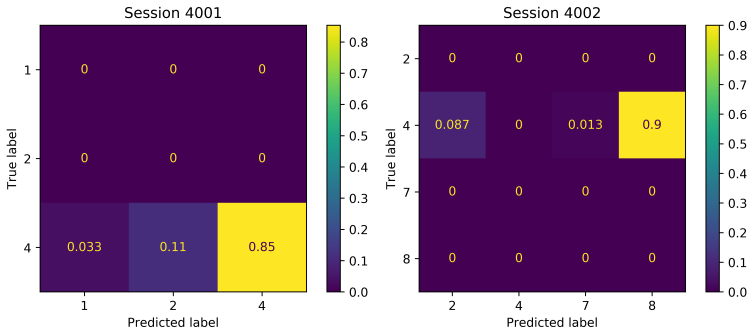


Abbildung 4.12: Die Confusion-Matrix der beiden Session für Sprecher 4

Meine Vermutung ist das es nicht nur auf die Anzahl der Sessions, sondern auch auf die Qualität der Sessions ankommt die Gründe hierfür sind dieselben wie für die gefilterte Variante. Bei diesen Tests ist aufgefallen, dass die Genauigkeit für die Sprecher mit weniger Daten vergleichbar viel niedriger ist als die Genauigkeit für die anderen Sprecher.

Deshalb wurde ein zweiter Versuch durchgeführt in dem nur die Daten der Sprecher 2 und 8 genutzt wurden. Dabei waren die Resultate deutlich besser als in dem vorherigen Versuch (siehe 4.1(b)) haben sich aber nicht zu sehr verändert, denn die Fehler in der Genauigkeit passieren größtenteils, wenn Sprecher 2 und Sprecher 8 verwechselt werden. Wie man in der Confusion Matrix von Session 29 sowie Session 26 von Sprecher 2 sehen kann. Die kleinen Unterschiede im 95–99 Prozent Bereich belaufen sich meisten auch auf das Vertauschen von Sprecher 2 und Sprecher 8.

Session 29 entfernt

Bei den Ergebnissen dieses Versuchs ist eine Steigerung der Genauigkeit, in Session 26 von Sprecher 2 von 68 auf 87 Prozent, zu beobachten(4.4). In 4.13 sieht man, dass das Steigern der Genauigkeit daher kommt das Sprecher 8 seltener geraten wird. Nähere Analysen wurden hier weggelassen, weil sie sehr ähnlich zu der ungefilterten Variante sind.

Tabelle 4.4: Genauigkeiten aller Sessions bei der Erkennung der Sprecheridentität. Es werden die Resultate der gefilterten Daten mit den Resultaten der ungefilterten verglichen.

Session	Resultat(filter)	Resultat ohne	Veränderung
1001	66	61.33	-4.67
1002	81,33	82	-0.67
1003	95,33	91.33	-4
2001	99.33	98	-1.33
2002	98	100	+2
2026	68	87	+19
2027	93	100	+7
2031	96	100	+4
4001	85.33	97.33	+12.67
4002	0	0	0
7001	36	36	0
7002	34.67	34,67	0
8006	96	96	0

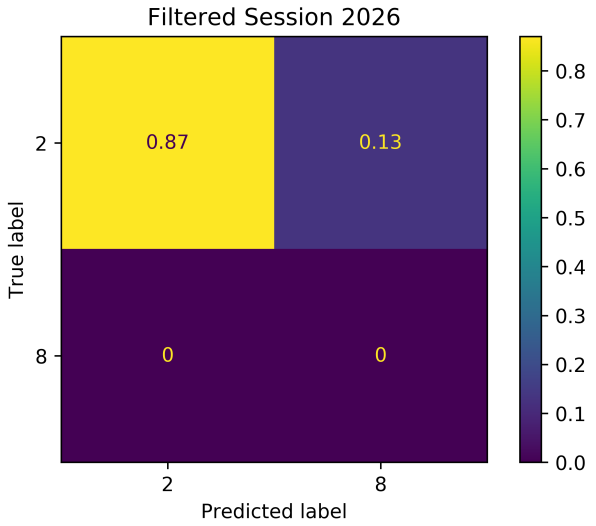


Abbildung 4.13: Die Confusion-Matrix der 26. Session von Sprecher 2

4.2 Sprachmodus

4.2.1 Cross-Session

In(4.14 und 4.17) wurden die Daten für die Experimente so aufgeteilt das alle Sessions, welche alle drei Sprachmodi enthalten genutzt wurden. Es wurde mit allen Sessions außer einer Trainiert und auf der ausgelassenen Session wurde getestet. Die Genauigkeiten in den unteren Graphen ist dabei der durchschnittliche Wert zwischen den Resultaten von allen Sessions.

4.2.1.1 Ungefiltert

Hier(4.14) ist eine sehr gleichmäßig verteile Genauigkeit über alle Klassifikatoren zu betrachten wobei der Lda, der Randomforest sowie der LinerSVC die höchsten Werte vorweisen. Der LDA-Klassifikator mit 59.6 Prozent bietet die höchste Genauigkeit. Der KNN-Klassifikator und der DecisionTree Klassifikator haben die niedrigsten Genauigkeiten wobei DecisionTree mit 46.88 Prozent die niedrigste Genauigkeit hat.

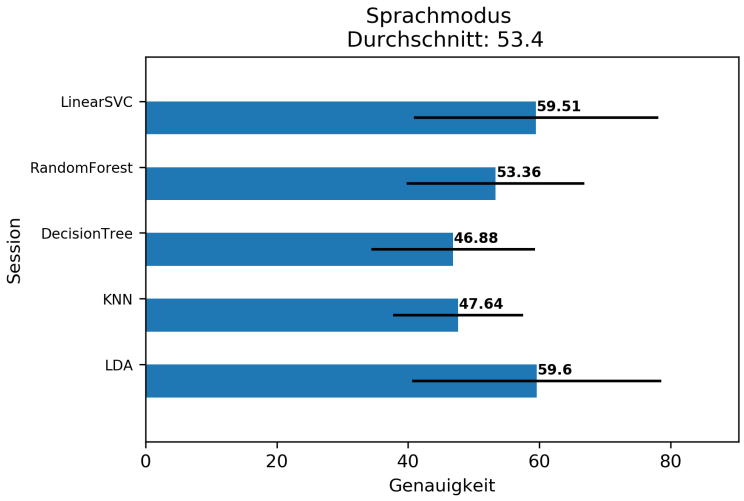


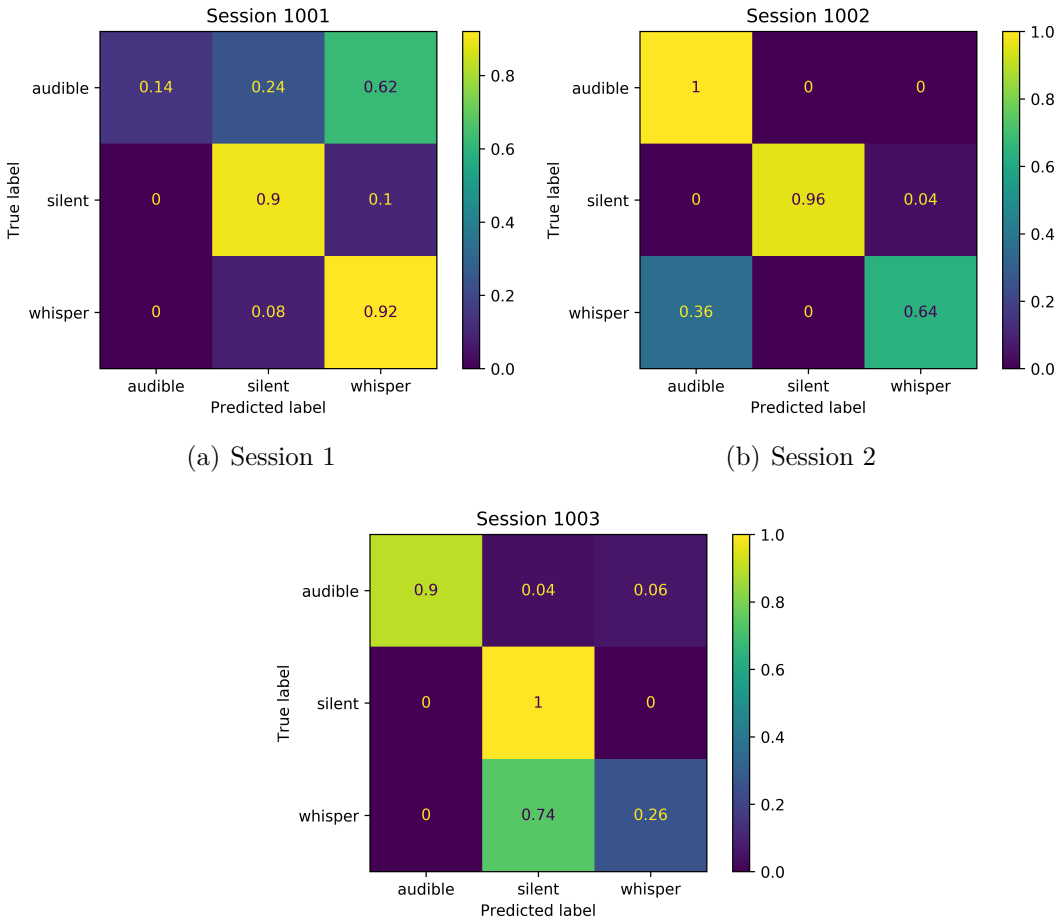
Abbildung 4.14: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedener Klassifikatoren bei der Erkennung des Sprachmodus. Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8

4.2.1.2 Sessions-Ungefiltert

Hier werden die Genauigkeiten der verschiedenen Sessions von jedem Sprecher und der Durchschnittswert aller Sessions präsentiert. Die Resultate sind in der Tabelle 4.5 zu sehen. Die höchste Genauigkeit ist in Session 2 von Sprecher 1 mit 86.67 Prozent, die niedrigste Genauigkeit ist Session 1 von Sprecher 4 mit 12 Prozent. Das passt zu den Resultaten der Sprechererkennung, in der diese Session ebenfalls die niedrigste Genauigkeit hatte. Die höchste Durchschnittsgenauigkeit hat ebenfalls Sprecher 1 mit 74.67 Prozent. In 4.15 sieht man das die Klassifizierungen für silent in jeder Session über 90 Prozent genau sind.In Session 1 ist zudem die Genauigkeit für whisper mit 92 Prozent sehr hoch und in Session 2 ist mit 90 Prozent die Genauigkeit für audible sehr hoch. Für Sprecher 7 (4.16(b)) wurde in Session 1 alles als audible klassifiziert und in Session 2 alles als whisper klassifiziert. In Sprecher 4 (4.16(a)) scheinen die Klassifizierungen einfach zufällig verteilt zu sein.

Tabelle 4.5: Genauigkeiten aller Sessions bei der Sprachmodus Erkennung mit ungefilterten Daten.

Session	Resultat in Prozent
1001	65.33
1002	86.66
1003	72
2001	66.67
2003	75.33
2004	43.33
2005	85.33
2006	79.33
2007	78,66
2008	79,33
2009	60,66
2010	74
2012	78
2013	22,67
2028	52
2029	46,67
2031	63.22
2032	64
4001	12
4002	26,67
7001	34,67
7002	33,33
8002	58
8003	64
8010	42
8016	75,33
8017	42
8018	64
8019	58

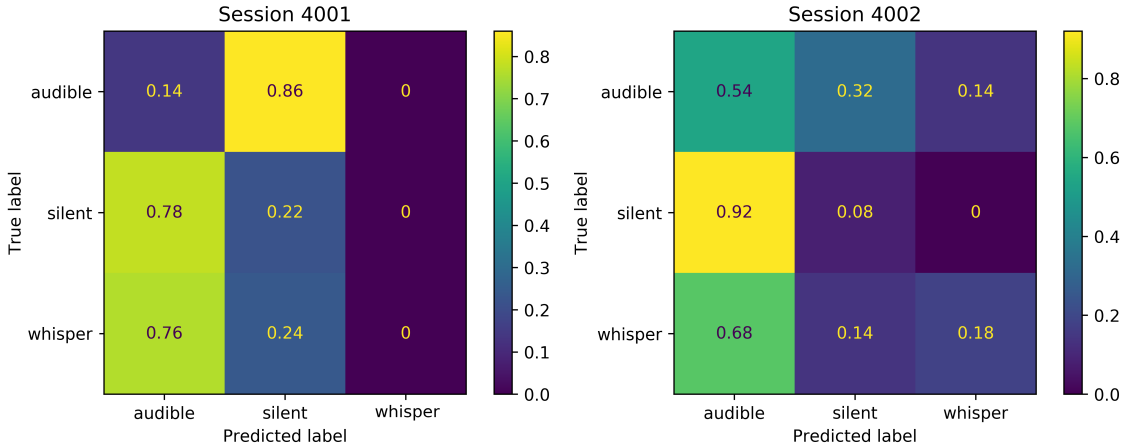


(a) Session 1

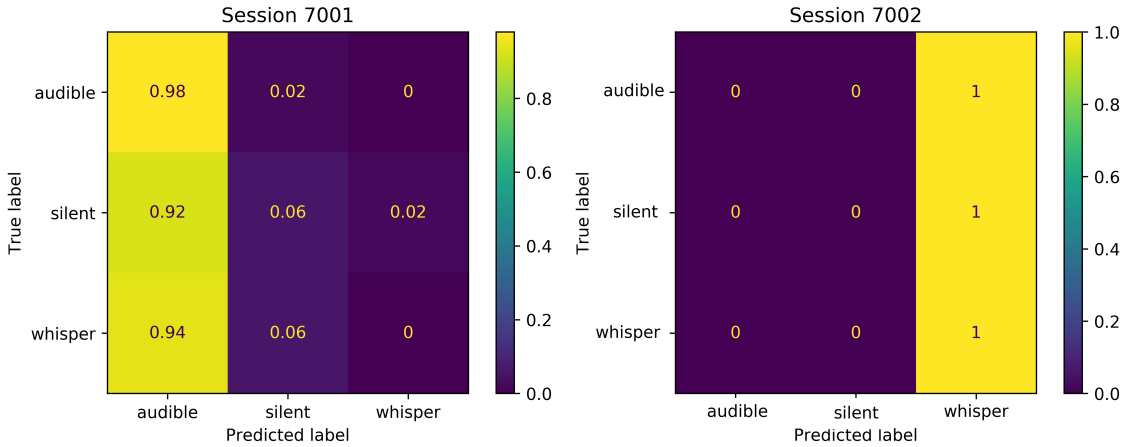
(b) Session 2

(c) Session 3

Abbildung 4.15: Erkennungsraten von Sprecher 1 für die Vorhersage des Sprachmodus mit ungefilterten Daten.



(a) Sprecher 4



(b) Sprecher 7

Abbildung 4.16: Die Erkennungsraten der Sprecher 4 und 7 für die Vorhersage des Sprachmodus mit ungefilterten Daten.

4.2.1.3 Filtered

Hier(4.17) wurde einen Bandpassfilter angewendet. Dieser filtert alles unter 10Hz und alles über 200Hz. Diese Filter Grenzen wurden mit Absprache meiner Tutoren gewählt und sollen unwichtige Artefakte und Hintergrundrauschen rausfiltern, um damit die Genauigkeit zu erhöhen. Es ist eine ähnliche Verteilung der Genauigkeit wie in (4.14) zu betrachten wobei der Lda, der Randomforest sowie der LinerSVC die höchsten Werte vorweisen. Hier hat wieder der LDA-Klassifikator mit 55.04 Prozent die höchste Genauigkeit. Der KNN-Klassifikator hat die niedrigste Genauigkeit mit 41.12 Prozent. Es lässt sich eine allgemein niedrigere Genauigkeit als bei der ungefilterten Variante in 4.14 beobachten.

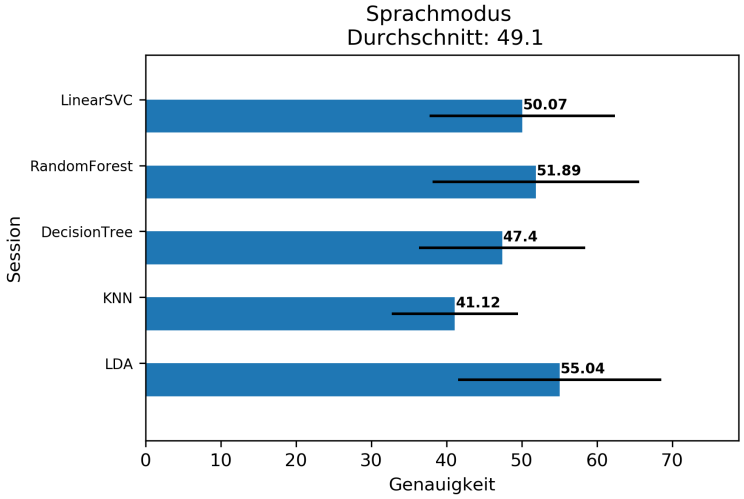


Abbildung 4.17: Die durchschnittliche Genauigkeit, sowie die Standardabweichung verschiedener Klassifikatoren bei der Erkennung des Sprachmodus. Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8. Zudem wurde hier ein Bandpassfilter von 10 bis 200Hz verwendet.

4.2.1.4 Sessions-Filtered

Hier wurde sich die Genauigkeiten der verschiedenen Sessions von jedem Sprecher und der Durchschnittswert aller Sessions angeguckt (4.6). Die höchsten erreichte Genauigkeit war in Session 3 von Sprecher 1 mit 84.67 Prozent die niedrigste Genauigkeit ist Session 1 von Sprecher 4 mit 18.67 Prozent. Das passt zu den Resultaten der Sprechererkennung in der diese Session ebenfalls die niedrigste Genauigkeit hatte. Die höchste Durchschnittsgenauigkeit hat Sprecher 8 mit 74.43 Prozent.

Tabelle 4.6: Genauigkeiten aller Sessions bei der Sprachmodus Erkennung mit gefilterten Daten.

Session	Resultat in Prozent
1001	84.67
1002	75.33
1003	63.33
2001	66
2003	66
2004	42.67
2005	68.67
2006	69.33
2007	66
2008	61.33
2009	56
2010	62.67
2012	62
2013	34.67
2028	43.33
2029	48
2031	34.67
2032	42,67
4001	41,33
4002	18,67
7001	36
7002	46,67
8002	62,67
8003	54,67
8010	58,67
8016	64
8017	51,33
8018	61,33
8019	66,61

4.2.1.5 Frequenzbänder

In [vB01] wird die Verbindung zwischen verschiedenen Frequenzbändern in der EMG-Aufnahme und die Verbindung zu verschiedenen Muskelgruppen dargestellt. In diesem Teil wurde kontrolliert, ob es ein erkennbares Frequenzband gibt welches bessere Resultate vorzeigt. Damit wäre nämlich bewiesen, dass ein bestimmter Muskel für die Vorhersage des Sprachmodi größere Wichtigkeit ist als die anderen. Das Signal wurde in Frequenzbänder der Größe 10 aufgeteilt und ihre Genauigkeit in den verschiedenen Frequenzbereichen betrachtet. Hier wurde ein LDA verwendet. Die Resultate sind in 4.18 zu sehen. Es ist eine gleichmäßige Verteilung von Genauigkeiten im 46–51 Prozent Bereich zu beobachten mit den niedrigsten Werten an den beiden Enden mit 34.98 Prozent im 290-300Hz Bereich und 41.22 Prozent im 0–10 Hz Bereich. Der höchste Wert ist im 10-20Hz Bereich mit 52.49 Prozent zu beobachten. Im 260-270Hz Bereich lässt sich jedoch ebenfalls eine Genauigkeit von 52.11 Prozent beobachten. Aufgrund der nur sehr kleine Unterschiede zwischen den verschiedenen Frequenzbändern lässt sich allerdings kein Frequenzbereich als der beste Frequenzbereich eingrenzen. Die besten Frequenzbereiche sind die Bereiche von 10-40Hz 70-130Hz und 260-290Hz eingrenzen.

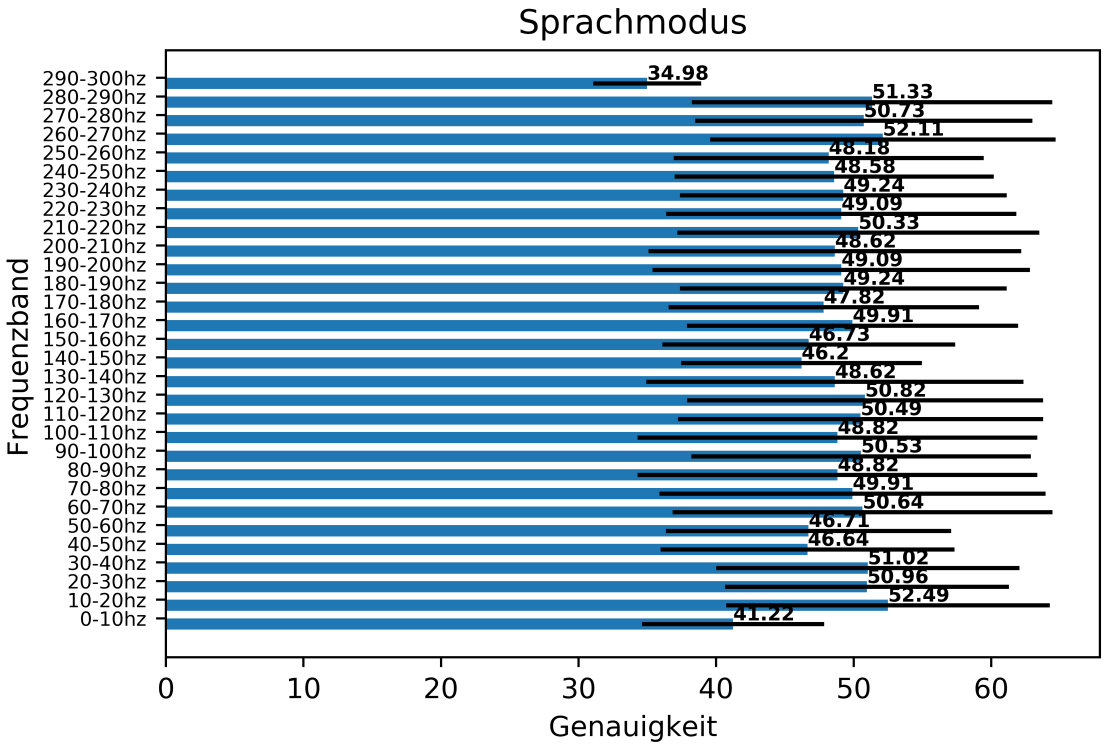


Abbildung 4.18: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedener Frequenzbänder bei der Erkennung des Sprachmodus. Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8. Es handelt sich um Frequenzbänder 0-300Hz in 10Hz Schritten.

4.2.2 Cross-User

Hier wurden die Daten für die Experimente so aufgeteilt das alle Sprecher, welche mehr als eine Session enthalten, genutzt wurden. Es wurde mit allen dieser Sprecher außer einem trainiert und auf dem ausgelassenen wurde getestet. Die Genauigkeiten in den unteren Graphen(4.19 und 4.21) sind dabei die durchschnittlichen Werte zwischen den Resultaten von allen Sprechern.

4.2.2.1 Ungefiltert

Hier(4.19) wurden die Daten genutzt, ohne sie davor zu filtern. Es ist eine ähnliche Verteilung der Genauigkeit wie in (4.14) zu betrachten wobei der LDA, der Randomforest sowie der LinerSVC die höchsten Werte vorweisen. Der LDA-Klassfikator bietet hier wieder, mit 50.04 Prozent, die höchste Genauigkeit. Der KNN-Klassifikator hat die niedrigste Genauigkeit mit 40.24 Prozent. Bei den Confusion-Matrix Analysen (4.20) wurden ein LDA verwendet.

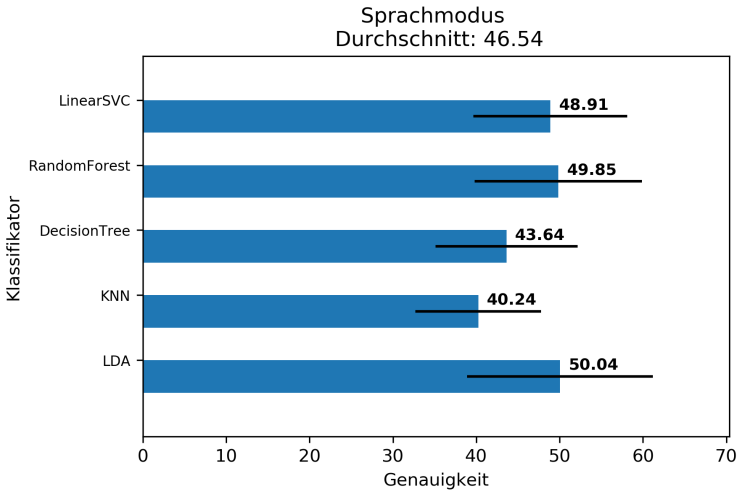


Abbildung 4.19: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedener Klassifikatoren bei der Erkennung des Sprachmodus.Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8

In 4.20(a) sieht man das bei Sprecher 1, bei audible und whisper zu 99 Prozent whisper vorhergesagt.

In 4.20(b) sieht man das bei Sprecher 2, größtenteils audible vorhergesagt wird. Es wird zu 83–86 Prozent audible vorhergesagt und zu 13–14 Prozent whisper.

Sprecher 3 (4.20(c)) ist der erste Sprecher, bei der eine Verteilung der Vorhersagen zu sehen ist, in der nicht eine der Möglichkeiten immer wieder ausgewählt wurde. Stadtessen sieht man eine 92 prozentige Genauigkeit für silent und eine 84 prozentige Genauigkeit für whisper. Lediglich audible hat man 10 Prozent eine niedrige Genauigkeit.

Bei Sprecher 4 (4.20(d)) ist eine Verteilung der Vorhersagen zu sehen, die etwas besser als die von den ersten beiden Sprechern ist. Man kann eine 81 prozentige Genauigkeit für whisper. Für die restlichen Zustände lässt sich eine sehr zufällige Verteilung erkennen. Audible wird entweder als audible oder als whisper erkannt und silent wird größtenteils, nämlich zu 49 Prozent als whisper erkannt.

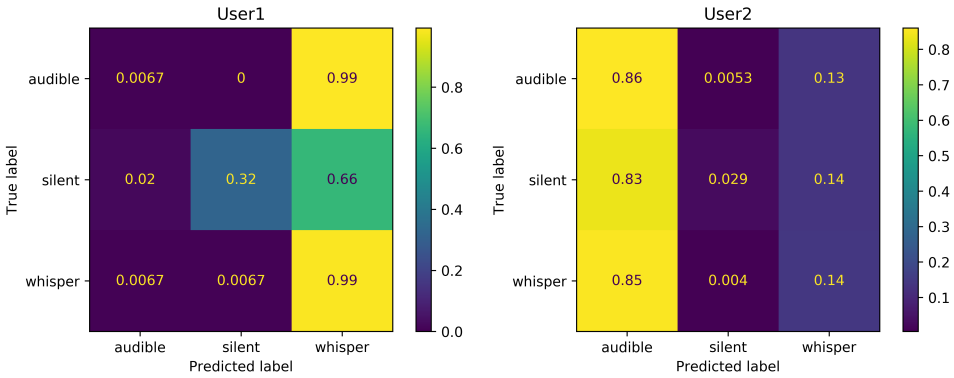
Bei Sprecher 5 (4.20(e)) ist eine richtige Vorhersage von audible zu 82 Prozent und silent zu 94 Prozent zu sehen. Lediglich der whisper Sprachmodus hat bei diesem Sprecher mit 32 Prozent keine gute Genauigkeit.

Für Sprecher 6 (4.20(f)) ist eine richtige Vorhersage von silent zu 98 Prozent zu sehen. Die zwei anderen Sprachmodi werden nicht gut erkannt. Whisper wird zu 74 Prozent als silent eingestuft und audible zu 58 Prozent als whisper.

Bei Sprecher 7 (4.20(g)) ist eine richtige Vorhersage von silent in 99 Prozent der Fälle zu beobachten. Die zwei anderen Sprachmodi werden nicht gut erkannt. Whisper wird zu 63 Prozent mit silent verwechselt und zu 37 Prozent als audible. Audible zu 55 Prozent als silent und 45 Prozent als audible. Der whisper Sprachmodus wird bei diesem Sprecher überhaupt nicht vorhergesagt.

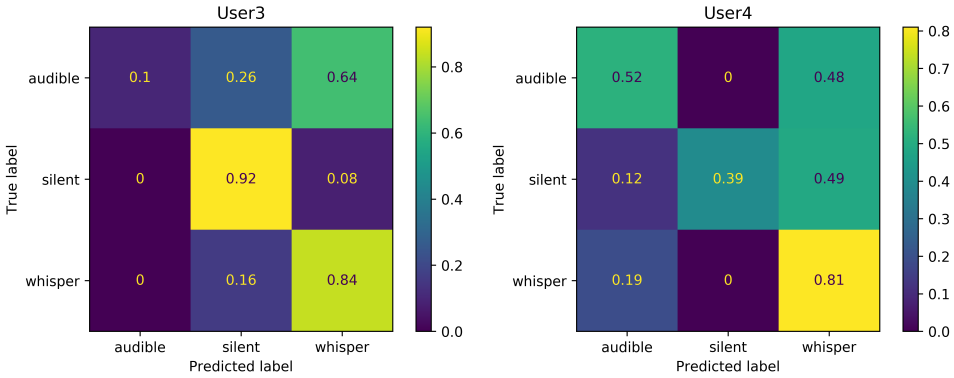
Für Sprecher 8 (4.20(h)) wird zum Großteil nur silent vorhergesagt egal um welchen Sprachmodus es sich handelt.

Es fällt auf, dass sich bei dem Sprecher 3 sowie bei Sprecher 5–8 eine über 90 Prozent Genauigkeit für den silent Sprachmodus erkennen lässt, wobei Sprecher 8 in allen Bereichen silent klassifiziert hat. Zudem haben Sprecher 1,3 und 4 eine über 90 Prozent Genauigkeit für den whisper Sprachmodus.



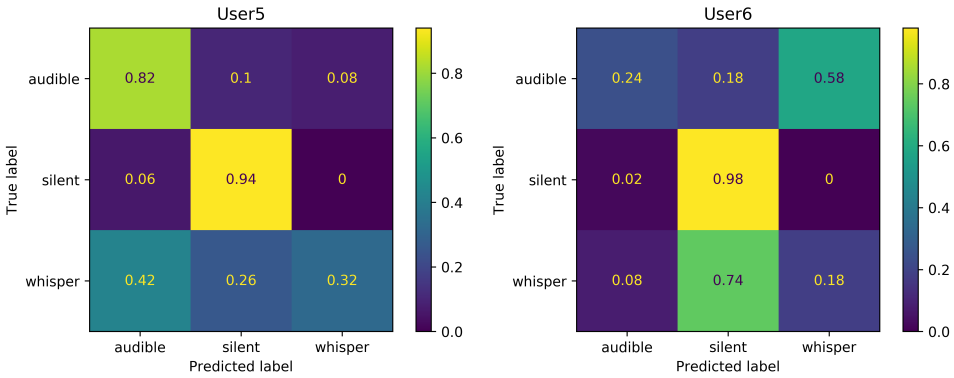
(a) Sprecher 1

(b) Sprecher 2



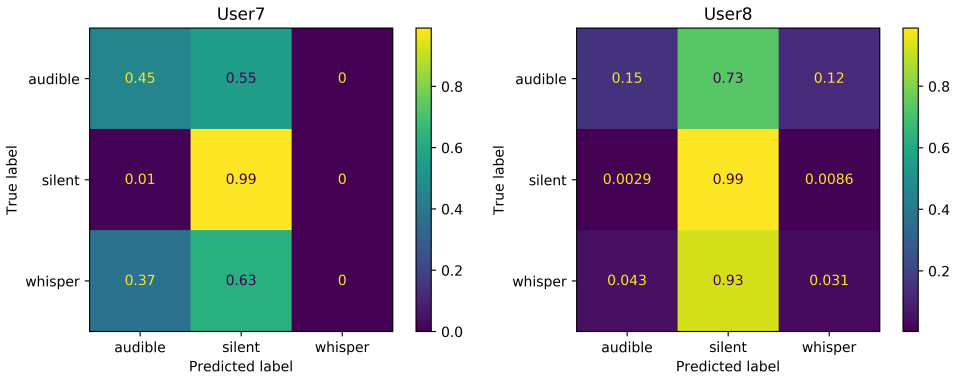
(c) Sprecher 3

(d) Sprecher 4



(e) Sprecher 5

(f) Sprecher 6



(g) Sprecher 7

(h) Sprecher 8

Abbildung 4.20: Die Erkennungsraten der Sprecher für die Vorhersage des Sprachmodus mit ungefilterten Daten.

4.2.2.2 Filtered

Hier(4.21) wurde der selbe Bandpassfilter auf die Daten angewendet wie in dem Sprachmodus. Es ist eine ähnliche Verteilung der Genauigkeit wie in (4.14) zu betrachten wobei der Lda, der Randomforest sowie der LinerSVC die höchsten Werte vorweisen und der LDA-Klassfikator mit 49.13 Prozent die höchste Genauigkeit bietet. Der KNN-Klassifikator hat die niedrigste Genauigkeit mit 38.78 Prozent. Es lässt sich eine allgemein niedrigere Genauigkeit als bei der ungefilterten Variante in 4.19 beobachten.

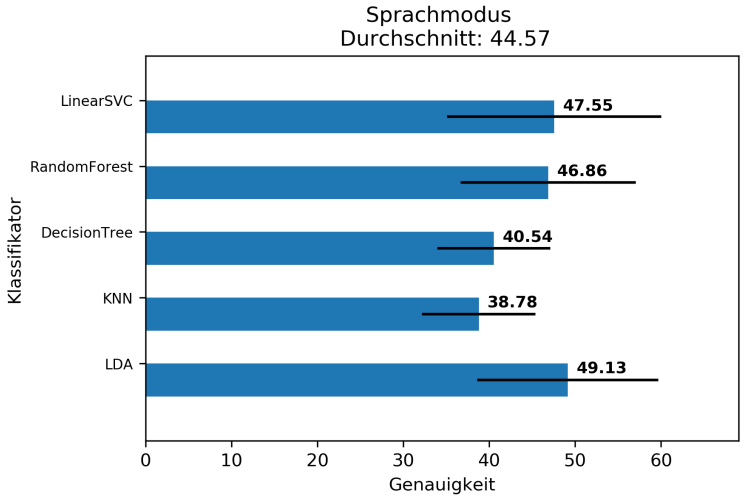
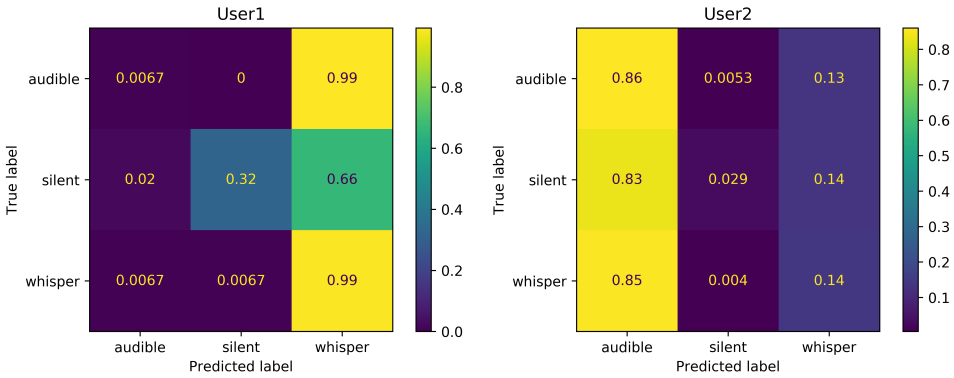


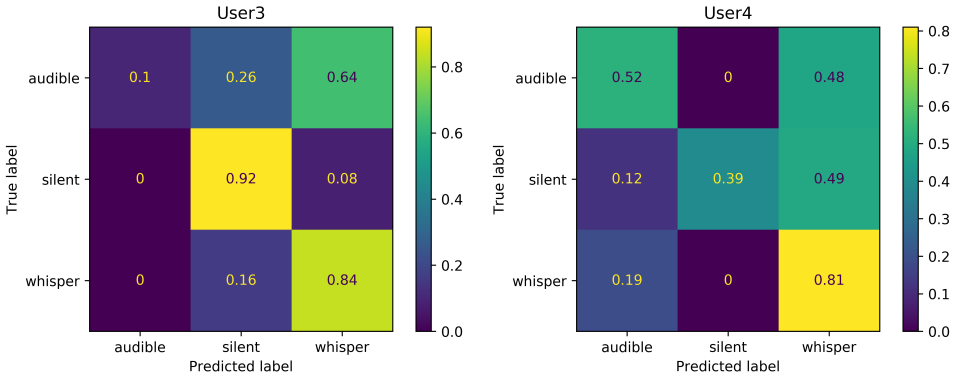
Abbildung 4.21: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedener Klassifikatoren bei der Erkennung des Sprachmodus.Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8 zudem wurde ein 10-200hz filter verwendet.

Die Confusion-Matrix Analyse für die gefilterten Daten (4.22) liefert identische Resultate zu den ungefilterten Resultaten.



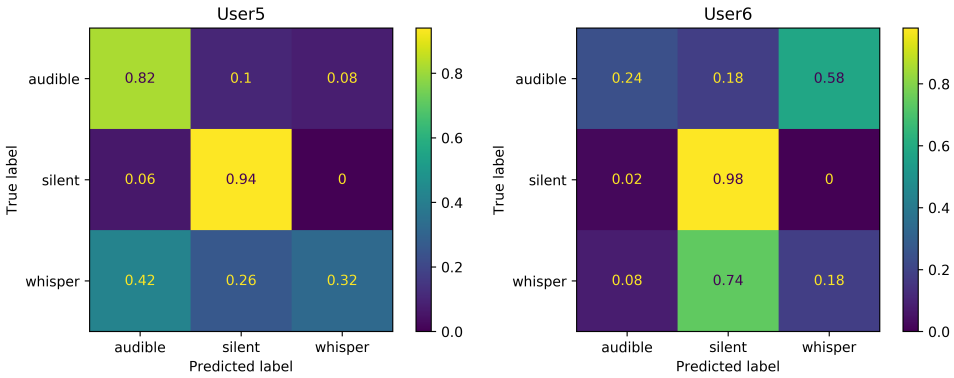
(a) Sprecher 1

(b) Sprecher 2



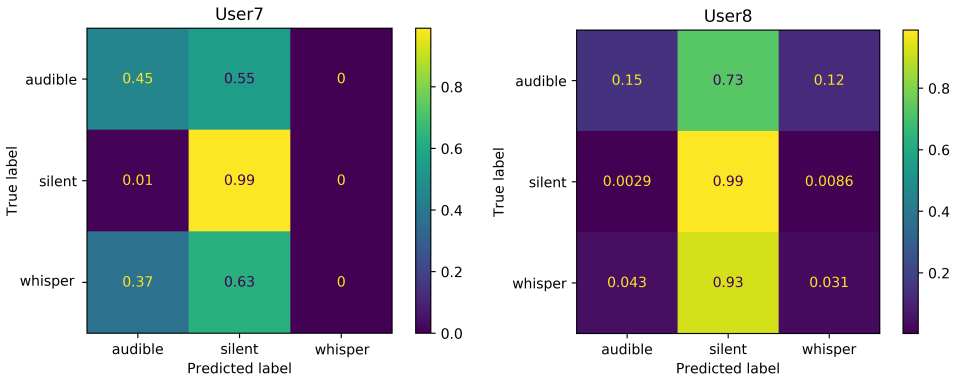
(c) Sprecher 3

(d) Sprecher 4



(e) Sprecher 5

(f) Sprecher 6



(g) Sprecher 7

(h) Sprecher 8

Abbildung 4.22: Die Erkennungsraten der Sprecher für die Vorhersage des Sprachmodus mit gefilterten Daten.

4.2.2.3 Frequenzbänder

Hier wird derselbe Versuch wie in der Cross-User Version durchgeführt. Die Resultate sind in 4.23 zu sehen. Es ist eine ziemlich gleichmäßige Verteilung der Genauigkeiten im 44–50 Prozent Bereich zu beobachten mit den niedrigsten Werten an den beiden Enden mit 33.1 Prozent im 290-300Hz Bereich und 36.02 Prozent im 0–10 Hz Bereich. Der höchste Wert ist im 120-130Hz Bereich mit 50.41 Prozent zu beobachten. Es lässt sich eine höhere durchschnittliche Genauigkeit in den Frequenzbereichen um 190-290Hz beobachten. Aufgrund der nur sehr kleine Unterschiede zwischen den verschiedenen Frequenzbändern lässt sich allerdings kein Frequenzbereich eingrenzen, in dem die Resultate signifikant höher sind.

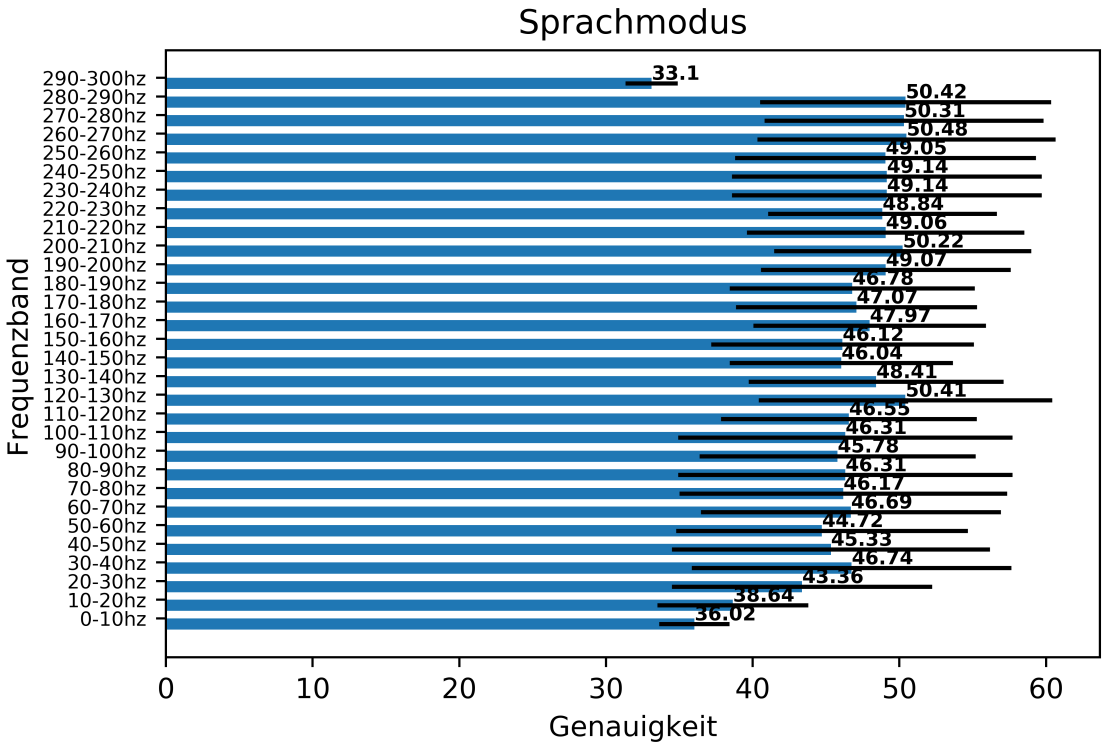


Abbildung 4.23: Die durchschnittliche Genauigkeit, sowie die Standardabweichung der verschiedenen Frequenzbänder.Hier werden die Daten aller Sprecher verwendet die mindestens 2 Sessions besitzen, also die Sprecher 1,2,4,7 und 8. Es handelt sich um Frequenzbänder 0-300hz in 10hz Schritten. Hier wird die Cross-User Aufteilung verwendet.

5. Diskussion

5.1 Erkennung der Identität

5.1.1 Klassifikator

Während dieser Arbeit wurden verschiedene Klassifikatoren eingesetzt. Diese Klassifikatoren lassen sich generell in 2 Gruppen einteilen. Die erste Gruppe sind die Klassifikatoren die zu brauchbaren Resultaten im Bezug auf die Genauigkeit der Vorhersagen geführt haben. Mit brauchbar sind hier Resultate von über 80 Prozent gemeint. Zu dieser Gruppe gehören LDA,RandmForest,KNN,DecisionTree sowie die SVC Klassifikatoren mit dem Kernel Poly und linear. Von diesen Klassifikatoren hat sich der LDA Klassifikator als der beste in den meisten Versuchen herausgestellt. Der RandomForest Klassifikator war für den Versuch am besten in dem nur 2 Nutzer miteinander verglichen wurden. Für alle anderen Versuche hat der LDA-Klassifikator jedoch ein besseres Resultat in der Durchschnittsgenauigkeit und der Standartabweichung geliefert. In die Zweite Gruppe gehören jene Klassifikatoren, welche zu Resultaten geführt haben, die am Zufallswert oder darunter liegen. Zu dieser zweiten Gruppe gehören die SVC Klassifikatoren mit den Kerneln Gaussian und Sigmoid sowie der NaiveBayes Klassifikator.

5.1.2 Große Sessions

Die Resultate für die einzelnen Sessions sind in (4.1) für die ungefilterten Resultate und in (4.3) für die Resultate mit den gefilterten Daten zu sehen.

Sprecher 2 Mit den vorliegenden Daten konnte die Sprecheridentität mit einer Genauigkeit von 100 Prozent für die meisten Sessions erkannt werden. Für die Sprecher mit vielen Sessions wurde eine so ziemlich hundertprozentige Genauigkeit erreicht. Die eine Ausnahme war Session 29, in der die Sprecheridentität nur mit einer Genauigkeit von 27 Prozent erreicht werden konnte. Da diese Genauigkeit bei der gefilterten sowie bei der ungefilterten Methode gleich niedrig war, gehe ich davon aus, dass es bei der Aufnahme dieser Session zu einem Fehler gekommen ist. Dieser

Fehler hat dazu gesorgt, dass die Identität von Sprecher 2 sehr oft mit Sprecher 8 verwechselt wurde. Diese Verwechslungsrate steigt sogar noch weiter an, wenn man nur die Sessions der Sprecher 2 und 8 nutzt. Session 29 und 26 von Sprecher 2 haben ebenfalls eine niedrigere Genauigkeit. Diese Genauigkeit konnte jedoch bei den Versuchen mit den gefilterten Daten stark erhöht werden, weshalb ich darauf schließe, dass die Ungenauigkeit in Session 26 an der Qualität von Session 29 hängt.

Sprecher 8 Bei Sprecher 8 konnten ebenfalls hohe Genauigkeiten für alle Sessions beobachtet werden. Der Unterschied zu Sprecher 2 ist jedoch, dass es keine Session mit einer sehr niedrigen Genauigkeit gibt. Dadurch die Durchschnittsgenauigkeit herunterzieht.

5.1.3 Kleine Sessions

Sprecher 1 In Sprecher 1 konnte ein Abstieg der Genauigkeit vor und nach dem Filtern observiert werden. Ich gehe davon aus, dass die Aufnahmebedingungen dieser 3 Sessions zu einer erhöhten Einzigartigkeit der Daten, im Vergleich zu den anderen, geführt haben. Für Sprecher 1 konnten auch bei den gefilterten Daten einer Durchschnittsgenauigkeit von über 80 Prozent erzielt werden.

Sprecher 4 Für Session 1 kann man ebenfalls einen Abstieg der Genauigkeit vor und nach dem Filtern beobachten. Ich gehe hier wie in Sprecher 1 davon aus das die Aufnahmebedingungen dieser 3 Sessions zu einer erhöhten Einzigartigkeit der Daten geführt hat. Eine weitere erwähnenswerte Beobachtung ist die Steigung der Genauigkeit von Session 1 Sprecher 4 um 12 Prozent nach dem Entfernen von Session 29 von Sprecher 2. Der Grund dafür ist, dass Sprecher 2 weniger oft geraten wird (**??**). Ich gehe davon aus das der Grund für die Entwicklung ist, dass durch das Entfernen von Session 29 die Daten von Sprecher 2 uniformer werden und es dadurch zu weniger Möglichkeiten kommt in denen Sprecher 2 mit Sprecher 4 vertauscht werden kann. Session 2 in Sprecher 4 hat in jedem Versuch eine Genauigkeit von 0 Prozent. Session 2 hat zwar eine Genauigkeit von 0 Prozent, wenn man Session 1 zum Trainieren benutzt jedoch führt das Training mit Session 2 dazu das Session 1 eine sehr hohe Genauigkeit aufweist.

Sprecher 7 In Sprecher 7 ist die Genauigkeit nach dem Filtern weit höher als vor dem Filtern. Im Durchschnit steigt die Genauigkeit von 0 auf über 35 Prozent. Da die Wahrscheinlichkeit bei 0 liegt, gehe ich davon aus, dass beim Aufnehmen der EMG-Signale Artefakte im 200-300Hz oder 0-10Hz Bereich mitaufgenommen werden. Welche dazu geführt haben, dass die Daten ohne gefiltert zu werden, unbrauchbar sind. Man müsste hier eventuell einen schmaleren Filtern einsetzen.

5.2 Sprachmodus

5.2.1 Cross-User

Bei den Cross-User Versuchen konnten leider keine guten Resultate für die Funktionsfähigkeit der Sprachmodus Erkennung festgestellt werden. Es wurde zwar eine Durchschnittsgenauigkeit von 50 Prozent mit einem LDA erreicht. Beim näheren Betrachten der Daten stellt sich nämlich heraus, das es quasi zufallsbedingt ist. Es gibt einige einzelne vielversprechende Resultate wie zum Beispiel Sprecher 3 (4.20(a) / 4.20(c)) mit den silent und whisper Werten. Sowie Sprecher 5 (4.20(b) / 4.20(d)) mit den audible und silent Werten. Sonst wird eine der Sprachmodi am meisten Klassifiziert. Dies führt dazu, dass die Genauigkeit nahe 33 Prozent liegt, dass scheint vor allem die Sprecher mit mehr als 2 Sessions zu betreffen. Es fällt auf das sich bei dem Sprecher 3 sowie bei Sprechern 5–8 eine über 90 Prozent Genauigkeit für den silent Sprachmodus erkennen lässt, wobei Sprecher 8 in allen Bereichen silent klassifiziert hat. Zudem haben Sprecher 1,3 und 4 eine über 90 Prozent Genauigkeit für den whisper Sprachmodus. Es scheint als die Sprecher 1,3 und 4 beim Flüstern fundamental andere Daten erzeugt haben ,als die anderen Sprecher. Andersherum scheint es so als waren Sprecher 3 und 5–8 beim Aufnehmen der Daten *anders* stumm waren als der Rest der Sprecher.

Ich vermute die schlechten Ergebnisse liegen an den Unterschieden in den Aufnahmebedingungen zwischen den verschiedenen Sessions und an der niedrigen Qualität der EMG Aufnahmen. Diese 2 Punkte führen dazu, dass beim Nutzen der Daten aller Sprecher keine qualitativ hochwertige Bestimmung eines Sprachmodus möglich ist. Die genannten Beispielen, bei denen es einigermaßen gut Funktioniert hat, sind wahrscheinlich pures Glück. Um diesen Versuch besser durchführen zu können, hätte man von der Aufnahme an mehr auf die Uniformität der Aufnahmebedingungen achten sollen. Ich glaube nicht, dass man hieraus einen Schluss für die Machbarkeit einer Cross-User Erkennung ziehen kann.

5.2.2 Cross-Session

Bei den Cross-Session Versuchen sind teilweise viel bessere Resultate zu sehen als bei der Cross-User Variante. Mit 86 Prozent Genauigkeit ist Session 2 von Sprecher 1 ein sehr gutes Ergebnis. Diese Ergebnisse sinken um einiges nach dem Filtern. Sprecher 1 hat zudem die höchste Durchschnittsgenauigkeit, obwohl dieser nur aus 3 Sessions besteht. Ich gehe daher davon aus, dass die Menge der Daten nicht der wichtigste Punkt für die Erkennung des Sprachmodus ist. Eine kleine Anzahl hochwertiger Daten scheint besser geeignet zu sein als eine große Anzahl mittelwertiger Daten. Das Session-abhängige Training liefert für einige Sessions sehr gute Resultate, für manche andere, wie Session 13 von Sprecher 2 ist es trotzdem sehr gering. Der Grund für diese große Diskrepanz liegt hier wahrscheinlich wieder an den verschiedenen Aufnahmebedingungen zwischen verschiedenen Sessions. Ich nehme an, dass bei einer gleichmäßig hohen Qualität und bei normalisierten Aufnahmebedingungen die Resultate für jede Sessions in den Cross-Session Experimenten über 80 Prozent erreichen könnten.

5.2.3 Frequenzband

Durch die Aufteilung des Signals in verschiedene Frequenzbänder konnte leider kein Frequenzbereich erkannt werden, welcher eine höhere Genauigkeit für die Erkennung des Sprachmodus bietet.

Siehe4.18 für die Cross-Session Resultate und 4.23 für die Cross-User Resultate.

5.3 Datenqualität

Durch das Entfernen von Session 29 konnte in der Sprechererkennung insgesamt eine bessere Genauigkeit erreicht werden wobei manche Sessions eine schlechtere Genauigkeit gezeigt haben 4.4. Die Probleme, welche durch die Qualität der Dateien entstehen, könnten vielleicht gelöst werden indem beim Aufnähmen dieser Sessions darauf geachtet wird dass eine gleichmäßige Qualität zwischen den Sessions herrscht und die einzelnen Aufnahmen so hochwertig wie möglich sind. Hier müsste man zum Beispiel darauf achten das kein Kanal Kaputt ist.

Aufnahmetechnologie

Eine Möglichkeit zur Verbesserung wäre es, einen neuen Korpus aufzunehmen mit einem Array Aufnahmesetup anstelle von Einzelelektroden. Dies sorgt nach ([DRVSar]) nämlich für bessere Resultate.

6. Fazit

6.1 Zusammenfassung

Die Ergebnisse in Kapitel 5 zeigen, dass die Ziele für diese Arbeit zum Teil erfüllt wurden. Die Erkennung der Sprecheridentität funktioniert sehr gut und stellt damit für Menschen, welche ihre Stimme verloren haben, eine Möglichkeit dar ihre Identität per sEMG-Signal zu bestätigen.

Die Ergebnisse für den Sprachmodus waren weniger gut. Sie zeigen jedoch, dass es möglich ist den Sprachmodus zu klassifizieren. Es deutet zudem an, dass die Erkennung von Unterschieden in der Sprachproduktion erkennbar ist. Die Erkennung des Sprachmodus müsste mit besseren Daten noch einmal getestet werden. Um zu kontrollieren, ob die niedrigere Genauigkeit an der Datenqualität liegt.

6.2 Zukünftige Arbeiten

In der Zukunft müsste man diese Versuche für den Sprachmodus nochmal mit besseren Daten durchführen. Zudem könnten neue Aufnahmen gemacht werden, in denen bestimmte Muskelgruppen fokussiert werden. Um zu gucken, ob es einen Unterschied in der Genauigkeit bei der Wahl verschiedener Muskelgruppen gibt.

Eine Möglichkeit zur Verbesserung, wäre es einen neuen Korpus aufzunehmen mit einem Array Aufnahmesetup anstatt Einzelelektroden. In ([DRVSar]) wird erklärt, dass das Wechseln auf ein Array Aufnahmesetup die Resultate verbessert. Beim Aufnähmen dieser Sessions sollte darauf geachtet werden, dass eine gleichmäßige Qualität zwischen den Sessions herrscht und die einzelnen Aufnahmen so hochwertig wie möglich sind.

Eine weitere Idee für die Aufnahmen wäre es, die verschiedenen Sprachmodi nicht in derselben Session aufzunehmen, sondern es Blockweise zu tun. Indem man einzelne

Blöcke zum Trainieren und andere zum Testen aufnimmt. Meine Hoffnung ist, dass durch das Aufnehmen in Blöcken eine bessere Datenqualität für die Trainingsdaten entsteht, weil die Sprecher nicht von Sprachmodus zu Sprachmodus wechseln müssen.

Nach [DAB+ar] sind gelernte Featuresets besser als selbstgemacht Featuresets. Es ist also eventuell eine höhere Erfolgsrate möglich, wenn man ein zusätzliches gelerntes Featureset für die Daten übernimmt. Das gelernte Featureset wurde hier mit dem AUDEEP toolkit gebaut.

Man könnte andere Klassifikatoren ausprobieren wie zum Beispiel AUDEEP in [DAB+ar]. Es sind auch andere Algorithmen möglich, wie die zuvor genannten phonetischen Entscheidungsbäume[JWS11].

In [JWS10] zeigten Janke et al., dass EMG-Signale von hörbarer Sprache im Allgemeinen eine höhere spektrale Leistung als EMG-Signale von stummer Sprache, aufweisen. Mit diesen Informationen konnte eine Abbildung vom stillen zum hörbaren/geflüsterten EMG erreicht werden, was zu 12,3 prozentiger Verbesserung führte. Demnach würde ich die selben Versuche nochmal mit der Abbildungen der Sprachmodi machen.

6.3 Abschließende Worte

Die Leistung von Schnittstellen für stumme Sprache, die auf der sEMG-Signal Klassifizierung basieren, wird mit der Zeit immer besser. Eine hohe Genauigkeit bei der Vorhersage des Sprachmodus würde zu einer höheren Qualität von Silent EMG to Speech Anwendungen führen. Die zusätzliche paralinguistische Information erlaubt es, solchen Anwendungen die Kadenz sowie die Lautstärke der Stimme eines Patienten besser nachmachen zu können. Ich hoffe, dass die Ergebnisse in dieser Arbeit dabei helfen, dem Ziel näher zu kommen, stummen Menschen den Gebrauch von Sprachschnittstellen zu vereinfachen.

Literaturverzeichnis

[DAB+ar] Lorenz Diener, Shahin Amiriparian, Catarina Botelho, Kevin Scheck, Dennis Küster, Björn W. Trancoso, Isabel Schuller, and Tanja Schultz. Towards silent paralinguistics: Deriving speaking mode and speaker id from electromyographic signals. In *INTERSPEECH 2020 - 21st Annual Conference of the International Speech Communication Association*, 2020 (to appear).

[DCHM12] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner. Signal processing advances for the mute semg-based silent speech recognition system. In *MILCOM 2012 - 2012 IEEE Military Communications Conference*, pages 1–6, Oct 2012.

[DFGM15] Laura Docio-Fernandez and Carmen García Mateo. *Speech Production*, pages 1493–1498. Springer US, Boston, MA, 2015.

[DRVSar] Lorenz Diener, Mehrdad Roustay Vishkasougheh, and Tanja Schultz. CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion. In *INTERSPEECH 2020 - 21st Annual Conference of the International Speech Communication Association*, 2020 (to appear).

[Her16] Christian Herff. *Speech Processes for Brain-Computer Interfaces*. PhD thesis, University of Bremen, 2016.

[ISC] Interspeech website https://www.isca-speech.org/iscaweb/index.php/online-archive.

[Jan16] Matthias Janke. *EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals*. PhD thesis, Karlsruher Institut für Technologie, 2016.

[JD17] M. Janke and L. Diener. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2375–2385, 2017.

[JWS10] Matthias Janke, Michael Wand, and Tanja Schultz. Spectral energy mapping for emg-based recognition of silent speech. In *First International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications*, 2010. Side event of Biosignals 2010 conference.

[JWS11] Matthias Janke, Michael Wand, and Tanja Schultz. Investigations on speaking mode discrepancies in emg-based speech recognition. In

12th Annual Conference of the International Speech Communication Association, 2011. Interspeech 2011.

[ken] Bild über die muskeln im menschlichen gesicht aufgerufen am 17.9.2020(aufgerufen am https://www.kenhub.com/de/library/anatomie/die-mimische-muskulatur).

[Kon01] Peter Konrad. Eine praxisorientierte einführung in die kinesiologische elektromyographie. In *EMG-FIBEL*, 2001. EMG Fibel.

[Kon06] Peter Konrad. A practical introduction to kinesiological electromyography. In *The ABC of EMG*, 2006. EMG book.

[Lee19] Aiden Lee. Machine learning - emg project for prosthesis patients. https://gitlab.com/AidenLee94/machine-learning-emg-project-patients, 2019.

[PCMDG+20] José Pérez-Córdoba, Juan Martín-Doñas, Angel Gomez, Alejandro Gomez-Alanis, and Jose Gonzalez Lopez. Silent speech interfaces for speech restoration: A review. 09 2020.

[Phi12] Angkoon Phinyomark. Feature reduction and selection for emg signal classification. *Expert Systems with Applications*, 39(8):7420–7431, 2012.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[SAD+19] Tanja Schultz, Miguel Angrick, Lorenz Diener, Dennis Küster, Moritz Meier, Dean Krusienski, Christian Herff, and Jonathan Brumberg. Towards restoration of articulatory movements: Functional electrical stimulation of orofacial muscles. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3111–3114, July 2019.

[SAS20] N Cummins S Amiriparian, M Freitag and B Schuller. https://github.com/auDeep/auDeep, 2020.

[Sch19] Tanja Schultz. Biosignal processing for human-machine interaction. in: Keynote presented at interspeech 2019, september 2019 graz, austria. 09 2019.

[Spe] Bild über den aufbau von einem skelettmuskel aufgerufen am 17.9.2020(https://www.spektrum.de/lexika/images/bio).

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms*, 2014. ML Book.

[Sta] $statistikderschwerbehindertenmenschen2019(aufgerufenam16.9.2020https:$ $//www.destatis.de/de/themen/gesellschaft - umwelt/gesundheit/behinderte-menschen/_inhalt.html\#$sprg233848).

[SWH+17] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271, 2017.

[TM04] R. Merletti T. Moritani, D. Stegeman. Basic physiology and biophysics of emg signal generation. In *ELECTROMYOGRAPHY Physiology, Engineering,and NoninvasiveApplications*, 2004. EMG Book.

[tra] Aufbau von muskeln aufgerufen am 22.9.2020(https://www.trainingsworld.com/training/muskelaufbau/physiologie-aufbau-funktion-skelettmuskulatur-2379200).

[Van16] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media, Inc., 1st edition, 2016.

[vB01] A. van Boxtel. Optimal signal bandwidth for the recording of surface emg activity of facial, jaw, oral, and neck muscles. psychophysiology. *Pubmeds*, 38:22–34, 2001.

[WHH+13] Michael Wand, Adam Himmelsbach, Till Heistermann, Matthias Janke, and Tanja Schultz. Artifact removal algorithm for an emg-based silent speech interface. In *International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan*, 2013. EMBC 2013.

[Wika] Bild der lda aufgerufen am 12.9.2020(https://upload.wikimedia.org/wikipedia/de/a/a0/diskriminanzfunktion.png).

[Wikb] Bild einer datensatzes nach einer anpassung von einem knn aufgerufen am 12.9.2020(https://commons.wikimedia.org/wiki/file:data3classes.png).

[Wikc] Datensatz für knn aufgerufen am 12.9.2020(https://commons.wikimedia.org/wiki/file:map1nn.png).

[Wikd] Positionen des pharynx und larynx aufgerufen am 17.9.2020$(https: //commons.wikimedia.org/wiki/file = throat_diagram.png.$

[WJS14] Michael Wand, Matthias Janke, and Tanja Schultz. The emg-uka corpus for electromyographic speech processing. In *The 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014. Interspeech 2014.

[WS11] Michael Wand and Tanja Schultz. Session-independent emg-based speech recognition. pages 295–300, 01 2011.