Neuroinformatische Analyse von Sprachsignalen zur Evaluierung einer Stottertherapie

Markus Neviadomski

DIPLOMARBEIT

zur Erlangung des Grades eines Diplom-Informatikers an der

Fachhochschule Schmalkalden

Fachbereich Informatik

Referent: Prof. Dr. rer. nat. Martin Golz

Koreferent: Dipl.-Inform. (FH) David Sommer

28. Dezember 2006

Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die aus anderen Quellen entnommenen Stellen als solche gekennzeichnet habe. Diese Diplomarbeit hat in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegen.

Schmalkalden, am 29. Dezember 2006

Markus Neviadomski

Inhaltsverzeichnis

Kurzfassung

Das Institut der Kasseler Stottertherapie bietet in erster Linie eine Intensivtherapie für stotternde Menschen nach dem Fluency-Shaping-Ansatz. Daneben wird jedoch auch an Grundlagenforschung (siehe MRT-Forschungen von Prof. Neumann am Uni-Klinikum Frankfurt) und Evaluation des Therapieansatzes gearbeitet. Hier sieht das Institut in Zukunft einen stark wachsenden Bedarf, um die Evaluation der vorhandenen Therapieformen und langfristig auch die Behandlung der Patienten zu verbessern.

Neuroinformatische Methoden sind relativ junge Methoden zur Untersuchung von hochkomplexen Problemstellungen. Gerade im medizinischen Umfeld stellt die Neuro-Informatik aber entscheidende Werkzeuge und Hilfsmittel bereit, alte Problemstellungen aus einem neuen Blickwinkel zu betrachten. Diese Arbeit beschäftigt sich im Wesentlichen mit der Frage, ob diese Werkzeuge auch bei der Analyse von Sprechdaten stotternder Menschen helfen können. Obwohl Spracherkennungssysteme in der heutigen Zeit bereits komplexe Sachverhalte gut bewältigen, gibt es praktisch kaum Erfahrung mit „problembehafteter“ Sprache.

An dieser Stelle setzt diese Arbeit an und entwickelt einen Ansatz, mit dessen Hilfe sich globale Merkmale aus Sprechdaten extrahieren lassen. Dabei stellt die Sprechtechnik, die im Institut vermittelt wird, das entscheidende Hilfsmittel zur Unterscheidung zwischen gestotterter und nicht gestotterter Sprache dar. Die Merkmale selbst werden mit Hilfe von 2 Mustererkennungsverfahren daraufhin untersucht, inwieweit sie zur Unterscheidbarkeit von Aufnahmen mit Stotteranteilen und Aufnahmen ohne solche Anteile beitragen.

Kapitel 1

Einleitung

1.1 Die Kasseler Stottertherapie

Die Kasseler Stottertherapie (KST) ist eine seit 10 Jahren angebotene Sprechtherapie für Erwachsene und mittlerweile auch für ältere Kinder, in der stotternde Patienten eine neue, verbesserte Sprechweise erlernen. Die Therapie arbeitet nach dem Prinzip des Fluency-Shaping [27], eine in den USA bereits seit langem erfolgreich angewandte Therapiemethode. Dabei wird eine neue, weiche Sprechweise eingeübt, die zu Beginn der Therapie stark verlangsamt angewendet wird. Durch ständiges Training erreicht der Klient damit eine wesentlich flüssigere und natürliche Sprechweise, ohne das Stottern komplett zu heilen. Eine Heilung vom Stottern ist bisher generell nicht möglich. Durch diese Therapieform und beständiges Üben, auch nach der Therapie, ensteht eine langfristige, stabile Verbesserung der Sprechflüssigkeit [16, S. 7].

Die KST wird seit 1999 auf wissenschaftlich belegbare Erfolge hin untersucht. Die Ergebnisse sind in der Studie der Universität Kassel(UniK), Fachbereich Psychologie, dokumentiert und wissenschaftlich anerkannt. Hierfür wurden von allen Klienten des Jahres 1999 Sprechdaten (siehe auch 2.3) erhoben und dabei die flüssig und die unflüssig gesprochenene Silben ausgezählt [6]. Damit kann die KST als bisher einzige Stottertherapie im deutschsprachigen Raum ein wissenschaftlich belegbares Langzeitergebnis vorweisen.

Die in dieser Studie durchgeführte Evaluation hat den Nachteil, das ein hoher personeller Aufwand betrieben werden muss, um zu gesicherten Ergebnissen zu gelangen. Mit den Erkenntnissen der Arbeit der Therapeuten und der Studie der UniK ausgestattet, wurde dieses Diplomthema angeregt.

Die Arbeit ist dabei Bestandteil eines Projektes, das die qualitative und quantitative Erweiterung der Qualitätssicherungsmaßnahmen in der Logopädie im Bereich Stottern zum Ziel hat. Als langfristiges Ziel soll ein Paket aus organisatorischen Maßnahmen und Werkzeugen entstehen, das auf ein-

fache Art und Weise die objektive Bemessung des Therapiefortschrittes für eine große Anzahl an Klienten ermöglicht. Aus den Überlegungen, daß die Sprechtechnik, die durch eine Fluency-Shaping-Therapie vermittelt wird, gewisse Auffälligkeiten zeigt und angeregt durch gute Erfahrungen mit neuroinformatischen Methoden an der FH Schmalkalden, entstand die Idee zu dieser Arbeit. Ziel ist die Entwicklung verschiedener Verfahren und deren Prüfung, ob diese Bestandteil einer möglichen Automatisierung und Vereinfachung der bisherigen händischen Auswertung werden können. Dabei sollen die Ideen und Ansätze bestehender Evaluierungen nach Möglichkeit Verwendung finden.

In Bereichen mit einem hohen Anteil an sozialwissenschaftlichen Kompetenzen ist der Einsatz von technischen Methoden und neuen Verfahren mit erheblichen Schwierigkeiten verbunden. Deshalb ist in diesen Bereich bisher wenig Fortschritt zu erkennen und man kann sich lediglich auf Erkenntnisse aus Bereichen stützen, die an ähnlichen, im Detail jedoch grundlegend anderen Problemstellungen arbeiten. Das wären zum Beispiel der Bereich Spracherkennung, der eine lange Tradition hat und in dem langjährige Erfahrungswerte existieren. Daneben gibt es vereinzelte Forschungsprojekte, die sich mit anderen, technischen Aspekten der Stimme beschäftigen, so zum Beispiel in der Robotik und in der Medizin zur Verbesserung von Sprachsynthesesytemen.

1.2 Ziele

In dieser Arbeit sollen Standardverfahren aus der Signalverarbeitung und der Neuroinformatik auf ihre Tauglichkeit für die Anwendung gestestet werden. Dafür soll ein Verfahren zur Extraktion von herausstechenden Merkmalen der gestotterten Sprache entwickelt und die daraus gewonnenen Merkmale mit den Standardverfahren der Mustererkennung getestet werden. In der Vorüberlegung ergaben sich dabei ganz konkrete Fragestellungen, die es zu analysieren galt:

- Gibt es Verfahren, mit denen aussagekräftige Merkmale, die auf Stottern schließen lassen, zuverlässig extrahiert werden können?
- Lässt eine Aufnahme eines gesprochenen Satzes Rückschlüsse auf den Therapieerfolg zu?
- Welche Merkmale der eingesetzten Sprechtechnik lassen sich als Indiz für flüssiges Sprechen heranziehen?

Eine Maßgabe der Kasseler Stottertherapie war, die durchzuführende Arbeit möglichst weit an das Arbeitsumfeld in der Therapie anzulehnen, etwaige positive Resultate können somit anschließend praxisnah verwertet werden.

Dazu wird im nächsten Kapitel auf grundlegende Vorraussetzungen und die Bedingungen eingegangen, in deren Rahmen die Arbeit entstanden ist. Damit möchte ich solides Vorwissen erzeugen, das für das weitere Verständnis unbedingt nötig ist. Im daran anschließenden Kapitel werden die angewandten Methoden der Merkmalserkennung und deren Resultat beschrieben. Daran schließt sich mit Kapitel 4 eine Beschreibung der verwendeten Klassifikationsverfahren an, während die Diskussion der Ergebnisse sich im nächsten Abschnitt findet. Mit einem kritischen Blick zurück und gleichzeitiger Vorausschau soll die Schlussbemerkung das Interesse auf zukünftige Möglichkeiten im Gebiet der medizinischen Sprachanalyse lenken.

Kapitel 2

Ausgangssitutation in der Stottertherapie

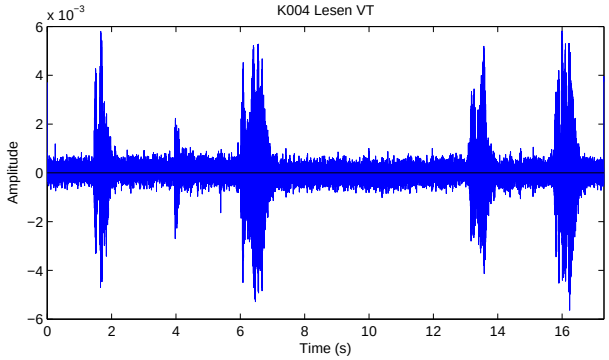
Dieses Kapitel beschreibt die Ausgangssituation, die zu Beginn der Arbeit in der Kasseler Stottertherapie (KST) vorzufinden war. Ausgehend von den medizinisch ausreichend beschriebenen Störungsbildern werden die momentan eingesetzten Evaluierungsmethoden kurze Erwähnung finden, um zur eigentlichen Problemstellung hinzuführen. Im Vorfeld der eigentlichen Signalanalyse muss klar definiert werden, welche Merkmale für das Projektziel relevant sein könnten und im Zeitrahmen zu bearbeiten sind. Darauf gehe ich am Schluss dieses Kapitels ein. Dabei soll der Bogen zwischen der seit mehreren Jahren durchgeführten Evaluation der KST durch Prof. H. Euler von der Universität Kassel(UniK) [6] und dem in dieser Arbeit beschriebenen Verfahren geschlagen werden.

2.1 Auswahl der Probanden

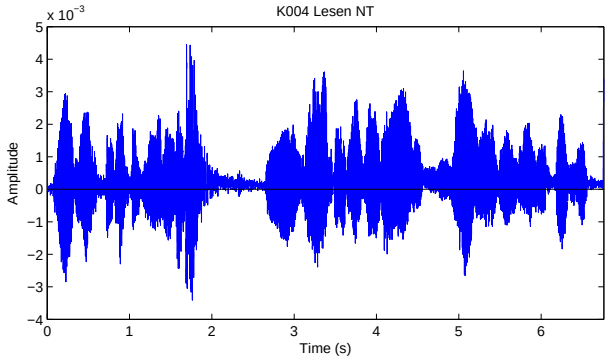
Von allen Klienten, die ihre Intensivtherapie im Jahr 2005 absolviert haben, wurden 20 Klienten für diese Arbeit ausgewählt, von denen vollständige[1] Therapiedaten vorliegen. Diese Auswahl wurde im ersten Schritt zufällig getroffen und dann dahingehend überprüft, ob sowohl der reale Altersschnitt [17], der an der Therapie teilnehmenden Personen als auch die tatsächlich vorhandene Breite an Störungsbildern[2] durch die Auswahl repräsentiert wurde. Die zufällige Auswahl wurde solange wiederholt, bis diese Kriterien erfüllt waren. Die Altersspanne reicht dabei von einem 15-jährigen Jugendlichen bis zu einem 67-jährigen Erwachsenen (s. Anhang 2.1), mit deutlichem Schwerpunkt bei den 20-30-jährigen. Das Verhältnis von männlichen zu weiblichen Probanden liegt bei $17 : 3 \approx 6 : 1$, was nahe bei dem

[1] Vollständig in dem Sinne, das die komplette Therapiehistorie zu Beginn der Studie bekannt war.

[2] Die Kriterien sind in Abschnitt2.4 beschrieben.



(a)



(b)

[Qual | i | täts | sich | er |ung | | der | Kas | se | ler | | Un | i | ver | si | tät]

(c)

Abbildung 2.1: Gegenüberstellung von 2 Aufnahmen mit demselben Text des gleichen Klienten vor(a) und nach(b) der Therapie und dessen Transkription(c).

Verhältnis von 5 zu 1 in der Gesamtmenge der vom Stottern Betroffenen liegt [27, S. 11] [2, S. 117] [22, S. 36]. Dieser Alters- und Geschlechterdurchschnitt findet sich genauso in der Klientel der KST wieder [16].

Aus den Evaluationsergebnissen der KST [16, Abb. 3] ergibt sich, dass der größte prozentuale Unterschied in der Sprechflüssigkeit zwischen den Zeitpunkten vor und unmittelbar nach der Therapie auftritt. In Abbildung 2.1 sind beispielhaft die Vor- und Nachtherapieaufnahmen desselben Klienten gezeigt. Es wird jeweils der gleiche Teil des Satzes abgebildet, wodurch die Unterschiede in der Artikulation besonders gut sichtbar werden. Die Transkription des Satzes ist darunter abgebildet.

Von den 20 Klienten werden somit 40 Videoaufnahmen analysiert. Dazu wird aus den Videodaten die Tonspur in Stereo und mit 44,1kHz gesampelt zur weiteren Verarbeitung extrahiert.

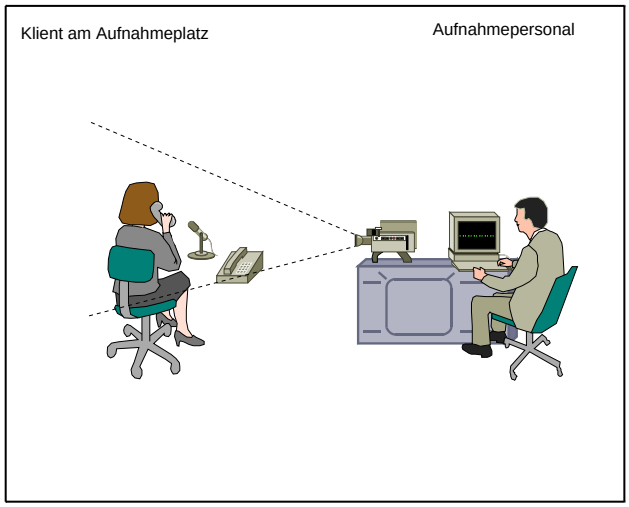


Abbildung 2.2: Versuchsaufbau für Audio/Video-Aufnahmen in der Kasseler Stottertherapie

2.2 Versuchsaufbau

Das in dieser Arbeit verwendete Audiomaterial wurde nicht explizit dafür angefertigt, sondern aus dem Therapiedatenbestand der KST entnommen. Das kommt dem praxisnahen Charakter der Arbeit zugute, birgt aber auch einige Probleme bezüglich der Qualität der Aufnahmen. Diese werden an einem modifizierten PC-Arbeitsplatz erstellt, der sich in einem abschließbaren Büro befindet. Somit werden äußere Störeinflüsse weitgehenst vermieden. Der Aufnahmeplatz ist mit einem Telefon und Mikrofon für den Klienten ausgestattet. Über eine Videokamera werden Bild- und Ton-Signal digital an den PC übertragen und dort in einem vorgegebenen Format (MPEG Layer 2, 44,1kHz, Stereo) abgelegt. Als Aufnahmegeräte kommen digitale Sony-SVHS-Kameras zum Einsatz, der PC ist mit Pinnacle Studio v.10[3] ausgestattet. Die Tonspur wurde händisch aus den Videoaufnahmen extrahiert und ohne Resampling im Microsoft-Wave-Format gespeichert.

Der Proband sitzt dem aufnehmenden Therapeuten direkt gegenüber und wird von diesem durch die Aufnahmeprozedur geleitet. Durch die ständige Beobachtung des Probanden durch den Therapeuten wird eine alltagsähnliche Situation erzeugt und ein gewisser Erfolgsdruck auf den Klienten ausgeübt. Nachweislich ist die Stottersymptomatik wesentlich geringer, wenn der Proband sich unbeobachtet fühlt [17]. Der Therapeut ist angehalten, während der Aufnahme nicht zu reden, so dass nur der Proband aufgezeichnet wird. Die Zusammenstellung der verwendeten Hard- und Software-Komponenten hat sich im Laufe der Therapie als relativ robust und für

[3] http://www.pinnaclesys.com

Therapiezwecke als qualitativ ausreichend erwiesen. Für Aufnahmen zur Signalverarbeitung ist die Tonqualität in der Regel ausreichend, aber nicht als gut zu bezeichnen. Hinzu kommt, daß durch die Aufnahmesituation bedingt, nicht alle Nebengeräusche ausgeschlossen werden können. An dieser Stelle wäre ein abgeschottetes Aufnahmestudio, in dem sich der Proband befindet, deutlich günstiger. Nur so können immer identische Aufnahmebedingungen, die als Grundvoraussetzung einer aussagekräftigen Klassifkation zum Beispiel bei Umapathy genannt sind [37, S. 423], realisiert werden. Ein derartiges Szenario ist wiederum aus therapeutischer Sicht nicht realisierbar, da die Stresssituation für den Probanden wesentlich geringer und somit die Aufnahme hinsichtlich des Störungsbildes nicht objektiv genug wäre [17].

2.3 Vorhandenes Datenmaterial

Ein wichtiges Kriterium, welches von Anfang an von der KST gestellt wurde, ist die Verwendung des bestehenden Datenmaterials und die direkte Integration möglicher Projekterfolge in die laufende Therapie. Dadurch mussten auch die in Abschnitt 2.2 genannten Nachteile bei der Aufnahme in Kauf genommen werden.

Im Rahmen der Therapiedurchführung und zur Evaluation werden standardmäßig folgende Daten erhoben:

– Sprechdaten vor der Therapie zum diagnostischen Vorgespräch

– Sprechdaten nach Abschluss der Therapie

– Sprechdaten 1 Jahr, 3 Jahre und 5 Jahre nach der Therapie

Dabei umfasst jede Datenerhebung zwei standardisierte Situationen und zwei alltagsrelevante Sprechsituationen, um situationsabhängige Schwankungen in der Sprechunflüssigkeit zu minimieren [6, S. 73]:

– diagnostisches Therapeutengespräch

– Telefonat mit einer unbekannten Person

– Lesen eines Standardtextes[4]

– Passanteninterview mit 11 Standard-Fragen und einer freien Frage

Für die vorliegende Arbeit wurde der Lesetext als Datenquelle gewählt, um auf einer einheitlichen Datenbasis aufbauen zu können. Die phonetische Varianz ist selbst bei gesunden Sprechern aus unterschiedlichen Regionen

[4] siehe Anhang A.1

Tabelle 2.1: Stottersymptomatik nach Natke [27, S. 25], ergänzt um die in der KST und anderer Literatur [9] verwendeten Begrifflichkeiten

	n. Natke		n. KST
Kernverhalten	Repetation	„ke-ke-kann“	Klonische Blocks
	Prolongationen	„ffffast“	Prolongationen
	Blocks	„- - - - kann“	Tonische Blocks

noch sehr groß. Bei den vielfältigen Ausprägungen von möglichen Sprechstörungen stellt dies bereits ein nicht zu vernächlässigendes Problem dar [40]. Aus organisatorischen Gründen ist es nur möglich gewesen, Daten zu verwenden, die im Jahre 2005 erhoben worden sind.

2.4 Störungsbilder

Generell unterscheidet man in der Sprachtherapie drei grundlegend verschiedene Störungsbilder, die im Sprechverlauf i.d.R. sehr gemischt auftreten (Tabelle 2.1). In meiner Arbeit verwende ich ausschließlich die Begrifflichkeiten, die auch in der KST Anwendung finden. Zu diesen kommt eine sehr differenzierte Sekundär-Symptomatik. Abgrenzen muss man an dieser Stelle andere Störungen wie Sprachentwicklungstörungen, Aussprachestörungen (z.B. Stammeln) oder Redestörung (z.B. Poltern) [9, S. 6-8]. Diese Beeinträchtigungen weisen gelegentlich ein für Laien kaum vom Stottern unterscheidbares Störungsbild auf. Das ist in meiner Arbeit aber ein zu vernachlässigender Faktor, da durch die Auswahl der Probanden in der KST diese Störungsbilder weitgehend ausgesiebt werden [17]. Hinzu kommt eine sehr differenzierte Sekundär-Symptomatik, die für die Analyse der Sprachsignale keinerlei Bedeutung hat [31, S. 994]. Diese Sekundärsymptomatik kann aber entscheidend zur Klassifikation eines Stotterereignisses beitragen, ist aber noch wesentlich komplexer als die Audiosignale [17].

2.5 Lösungsmöglichkeit zur Kontrolle des Therapieverlaufs

Die bisher anerkannten Verfahren zur Evaluierung des Therapieerfolges von Stottertherapien basieren im Kern alle auf der Errechnung von Kenngrößen aus händisch ermittelten Stotterereignissen anhand aller Probanden. Dabei werden Stottereignisse bzw. gestotterte und nichtgestotterte Silben händisch transkripiert, oder unterstützt durch Software[5] händisch ermittelt. Die so

[5] http://www.fluencymeter.de

gewonnenen Daten repräsentieren zwar ein sehr genaues Ergebnis, zeigen aber immer nur eine Momentaufnahme der Symptomatik. Eine Verlaufserstellung ist nur unter hohem personellen und organisatorischem Aufwand möglich. Dieses Vorgehen liegt auch der Evaluationsstudie der Universität Kassel zugrunde [6].

2.6 Ausgewählte Merkmale

Bei der Auswahl an relevanten Merkmalen orientierte sich meine Suche an den Ergebnissen der Studie der Universität Kassel und Expertenaussagen[6] [17], auf deren Know-How innerhalb der KST zurückgegriffen werden konnte.

Man unterscheidet generell zwischen Merkmalen im Zeitbereich und im Frequenzbereich. Für Anwendungen aus dem Bereich der Sprachanalyse und Spracherkennung wird im Allgemeinen mit Frequenzbereichsanalysen gearbeitet. In der Spracherkennung zielen die Analysen daraufhin ab, unabhängig vom Sprecher bestimmte Muster zu erkennen, die für Phoneme oder Lautgruppen eindeutig sind [30, S. 312]. Ein prominentes Beispiel dafür findet sich in der Bestimmung der Formanten. Die in der Arbeit vorliegende Problemstellung ist zwar zwingend sprecherunabhängig zu halten, soll aber aus Gründen des Aufwandes auch unabhängig von Worten oder bestimmten Phonemen sein. Damit scheidet die Formantenanalyse aus [24].

2.6.1 Statistische Merkmale im Zeitverhalten

Ein sofort augenfälliges Merkmal des Stotterns ist die erhebliche Abweichung von der normalen Sprechgeschwindigkeit. Im Allgemeinen treten zwei Extreme auf: Ausdauernde Blockaden, gefolgt von sehr schnell gesprochenen Abschnitten, oder sehr langsame Sprechgeschwindigkeit durch dauerhafte Blockaden und Wiederholungen. Über normales Sprechtempo herrscht in der Literatur auch keine einheitliche Vorstellung. Nach Fiedler liegt diese

[6] Dr. A.W.v. Gudenberg und das Therapeutenteam der KST

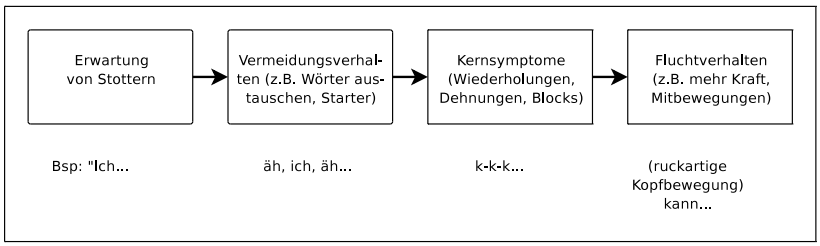


Abbildung 2.3: Schematischer Ablauf eines Stottereignisses (nach [28] und [31])

zwischen 150 und 200 Silben pro Minute (SpM) [8, S. 52]. In neuerer Literatur werden hingegen 170 bis 210 SpM angenommen [3, S. 111].

2.6.2 Pausenverhalten

Das Pausenverhalten eines Sprachsignals ist als Merkmal im Zeitbereich einzuordnen und beschreibt die Länge, Häufigkeit und Verteilung der Sprechpausen. Besonders gut ist das bereits in Abb.2.1 zu sehen, wo sowohl die Gesamtdauer der Aufnahmen als auch das Verhältnis von Sprache zu relativer Stille massiv differieren. Wie in Abschnitt 2.4 beschrieben, sind stille Abschnitte einer Aufnahme nicht nur als Atem- und Gedankenpausen zu verstehen, sondern können auch ein Stotterereignis darstellen[7]. Dies ist ohne Kenntnisse, die auf einer höheren Ebene als die reine Zeitserie des Sprachsignals angesiedelt sind, nicht zu differenzieren, weshalb ich diese Ansätze hier nicht weiter vertiefen kann.

Aus dem scheinbar simplen, sofort augenfälligen Merkmal ist durch wenige Überlegungen ein hoch komplexer Vorgang von Erregung und Hemmung des Sprechers entstanden. Aus diesem Grund möchte ich mich in dieser Arbeit auf die quantitativ messbaren Merkmale beschränken, die durch die Voice-Active-Detection berechnet werden (Abschnitt 3.2). Auch die Forschungsergebnisse der KST belegen, daß sich durch Pausenreduktion und Änderung der Artikulationsgeschwindigkeit[8] die relative Sprechgeschwindigkeit steigert [16, S. 8].

2.6.3 Anschwingverhalten

Ein entscheidendes Element der Sprechtherapie in der KST ist die als weicher Einsatz bezeichnete Dehnung in Verbindung mit einem sanften Stimmeinsatz jeder Silbe bzw. eines jeden Phonems am Wortanfang[9]. Diese Form der Lautbildung wird neben der Silbendehnung als zentrales Element der Fluency-Shaping-Therapie der KST von den Klienten trainiert. Dies erfolgt zu großen Teilen mit der Software Flunatic[10], die dem Übenden direktes Feedback der Aussprache in Form einer sog. Stimmkurve liefert (Anhang A.2). Die Abbildungen zeigen sehr schön das Anschwingverhalten, und markieren es gegebenenfalls rot, wenn zu hart gesprochen wird.

Aus den Gegenüberstellungen in Abbildung 2.4 geht hervor, wie sich das Anschwingverhalten mit Einsatz der Sprechtechnik in den Aufnahmen von einem steilen, abrupten Anstieg vor der Therapie in einen sanften, weichen Anstieg nach der Therapie ändert.

[7] Das Störungsbild des tonischen Stottern (sog. Hänger) kann im Extremfall bis zu mehreren Minuten andauern.

[8] Im extremsten Fall beträgt die Silbendehnung bis zu 2 Sekunden pro Silbe

[9] Das Phonem entspricht der kleinsten lautlichen Einheit [30, S. 11]

[10] http://www.flunatic.de

Auch hier liegt aus Sicht der Signalverarbeitung ein sehr augenfälliges Merkmal vor, welches bei Erstellung des Merkmalsraumes schwerpunktmäßig Beachtung finden wird.

Ein weiteres, für die Sprechtechnik der Therapie charakteristisches Merkmal ist die sog. Silbenbindung. Dabei werden während des Sprechens die Stimmbänder konstant in Schwingung gehalten und typischerweise erst am Wortende wieder abgesetzt. Dadurch wird das gesamte Wort mit einer deutlichen Grundschwingung hinterlegt, die gerade bei verlangsamter Sprechweise sehr auffällig wirkt. Bei Normalsprache ist diese Grundschwingung nicht konstant vorhanden. Dieses Merkmal hat allerdings eine wensentlich höhere sprachliche Variabilität als das Anschwingverhalten, so daß eine Konzentration auf den weichen Einsatz mehr Erfolg verspricht.

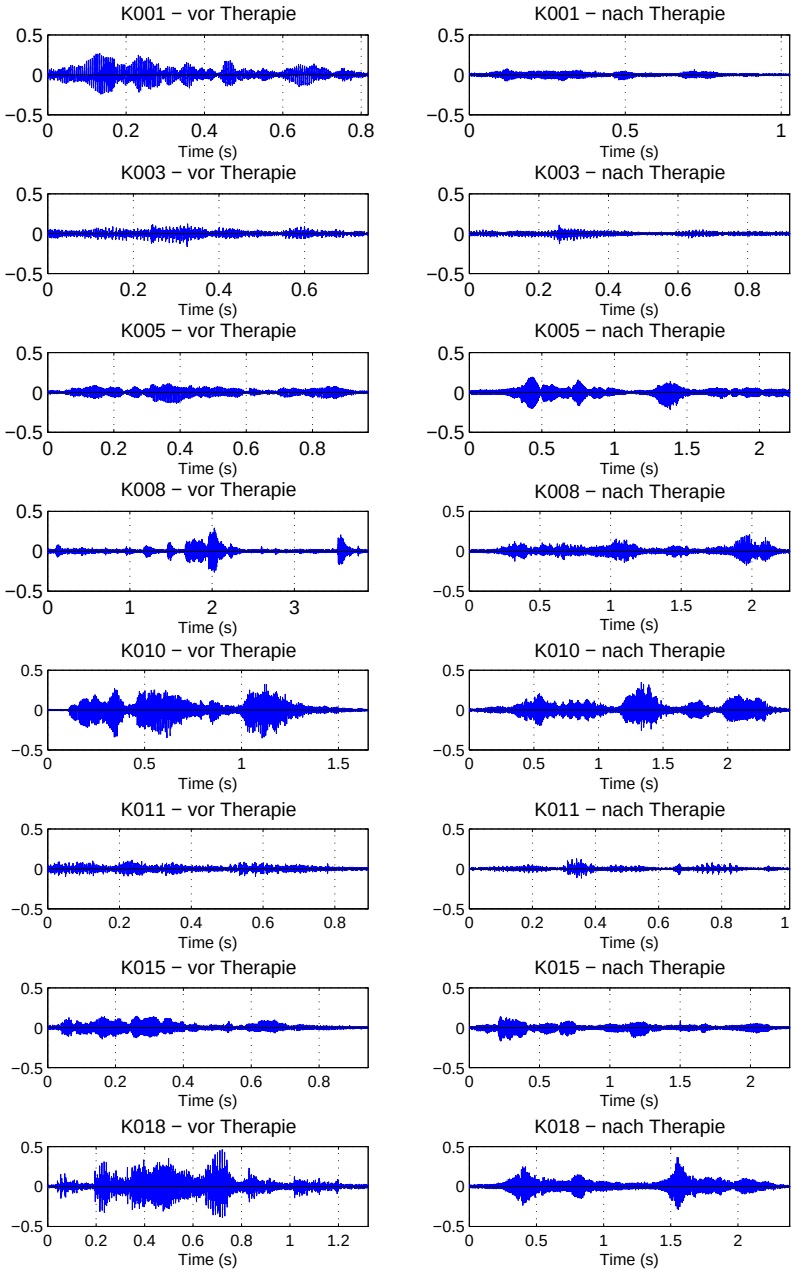


Abbildung 2.4: Gegenüberstellung von Aufnahmen mit dem gesprochenen Wort „Universität“ von ausgewählten Klienten vor(links) und nach(rechts) der Therapie.

Kapitel 3

Merkmalsvektoren durch Merkmalsextraktion

Die Merkmalsextraktion erzeugt in der Regel aus komplexen und hochdimensionalen Daten einen Merkmalsraum, auf den die Mustererkennung angewandt werden kann. Die Merkmale müssen letztendlich möglichst frei von störenden Einflüssen und Verfälschungen sein. Dazu ist in jedem Fall eine Vorbehandlung des Signals erforderlich, um eine einheitliche Ausgangsbasis für die weitere Signalverarbeitung zu schaffen. Zur weiteren Verarbeitung sollte ein möglichst störungsfreies Signal vorliegen, deshalb entledigt man sich äußerer Störeinflüsse (Brummen, etc). Die eigentliche Merkmalsextraktion in dieser Arbeit stützt sich auf 3 Säulen:

– Merkmale im Zeitbereich

– Wavelet-Transformierte Teilbereiche („Region of interest“) der Aufnahme

– Cepstral-Transformation über die „Region of interest“.

Um diese Merkmale zu erhalten, war es im Rahmen dieser Arbeit nötig, eine Pausenerkennung und eine modifizierte Silbendetektion zu entwickeln. Mit der Pausenerkennung werden die Informationen im Zeitbereich erfasst, diese Verfahren werden in der Literatur auch als „Voice-Activity Detection“ beschrieben. Mit einer modifizierten Silbendetektion soll die Region of Interest genau markiert werden, die für das Anschwingverhalten relevant ist. Dazu wurde ein neuer Algorithmus entwickelt, da die gängigen Silbendetektionen nur die Maxima der Silben in der Funktion des Signals zur Zeit finden. Alle in diesem Kapitel beschriebenen Verfahren wurden mit Hilfe der Software MATLAB im Rahmen dieser Arbeit entwickelt, die verwendeten Skripte befinden sich auf der beiliegenden CD-ROM. Teilweise bilden andere Verfahren und Ideen die Grundlage für die Algorithmen, dies ist an den betreffenden Stellen entsprechend vermerkt.

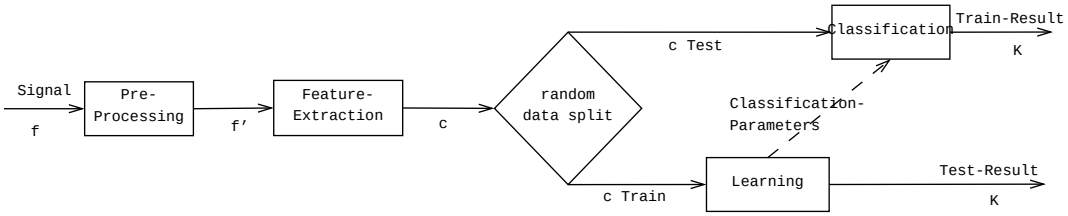


Abbildung 3.1: vereinfachte Architektur eines Klassifikationssystems mit Signal f, vorverarbeitetem Signal f', Merkmalsvektoren c und Klassifikationsergebnis K [29] [14].

3.1 Vorverarbeitung

3.1.1 Segmentierung

Die im Folgenden beschriebenen Verfahren Voice-Activity-Detection und die Halbsilbenerkennung beruhen auf der Signalenergie, die wiederum von der Lautstärke des gesprochenen Signals abhängig ist. Da die Lautstärke langsam im Zeitverlauf variiert, nimmt man eine grobe Segmentierung in Abschnitte von mehreren Sekunden vor. Bei Normalsprechern wird eine Segmentlänge von 2 Sekunden angesetzt [10, S. 5]. In diesem besonderen Fall, der bei dieser Arbeit vorliegt, wird diese auf 10 Sekunden ausgedehnt. Das wurde notwendig, da das Störungsbild der tonischen Blocks sich als abrupte, lange Pause im Signal darstellt. Solche Pausen wären bei einer Segmentlänge von 2 Sekunden nur schwer bis gar nicht zu erfassen. In dieser Stufe der Vorverarbeitung bleibt das Signal selbst unberührt, es erfolgt keine Gewichtung.

3.1.2 Mono-Signal

Liegen die Aufnahmen als Stereo-Signale vor, erfolgt eine Transformation in ein Mono-Signal (Einspur-Aufnahme) durch Mittelwertbildung beider Kanäle. Die Stereo-Aufzeichnung erfolgt komplett durch ein in der Kamera integriertes Mikrofon und nicht durch 2 gesonderte Mikrofone. Dadurch ergeben sich keine relevanten Unterschiede zwischen beiden Kanälen, so daß durch die Transformation das Signal nicht merklich verfälscht wird.

3.1.3 Downsampling und Noise-Filterung

Die Samplerate der Ausgangsignale ist durch die Vorgaben bei der Erstellung der Aufnahmen auf 44,1kHz festgesetzt. Um auch zukünftig Sprachsignale, die mit anderer Aufnahmetechnik realisiert werden, direkt mit der durchgeführten Auswertung vergleichen zu können, wird ein Downsampling auf $f_{sample} = 16000$Hz durchgeführt. Diese Samplerate wird vielfach zur Sprachanalyse verwendet [7] [29], um alle im Sprachbereich relevanten Frequenzen

bis etwa 4 KHz zu erfassen, gleichzeitig aber das Datenvolumen möglichst klein zu halten. Beim Downsampling ist das nach seinen Entdeckern Nyquist und Shannon benannte Abtasttheorem

$$f_{sample} > 2 \cdot f_{max} \tag{3.1}$$

zu beachten. Aus diesem Grund liegt die Abtastrate in der Sprachsignalanalyse bei mindestens 8kHz und in der Regel werden 16kHz nicht überschritten [30, S. 105].

Bedingt durch qualitative Schwankungen bei der Aufnahmetechnik ist auf den Aufnahmen ein mehr oder weniger starkes Rauschen zu finden. Um die Erkennungsleistung der Merkmalsextraktion zu verbessern, wird eine Filterung eingesetzt, die auf der Wavelet-Paket-Transformation(WPT)[1] basiert und mit einem amplitudenabhängigen, gleitenden Schwellwert arbeitet [26]. Diese Filterung bietet den Vorteil, unterschiedliche Störungen durch Umgebungslärm sowohl in einem großen als auch schmalen Frequenzband zu filtern, ohne das eigentliche Signal stark zu verfälschen [1, S. 2].

3.1.4 Fensterung

Ein Sprachsignal ist ein stochastisches Signal und als solches nur sehr schwer zu charakterisieren. Viele gängige Verfahren zur Charakterisierung eines Signals setzen Stationarität voraus, darunter alle auf der Fourier-Transformation aufbauenden Verfahren. Von einem stationären Signal kann man ausgehen, wenn der Erwartungswert und die Kovarianz eines Signals konstant ist. Es reicht also aus, genau eine Periode zu charakterisieren, um das komplette Signal zu beschreiben. Durch Fensterung des Signals werden diese kurzen Abschnitte erzeugt. Bei langsam instationären Signalen, bei denen der Erwartungswert sich nur langsam verändert und keine großen Sprünge ausweist, erzeugt die Fensterung eine quasi-Stationarität. Je kürzer die Segmentlänge der Fensterung ist, umso geringer ist die Wahrscheinlichkeit, daß das Signal die Stationaritäts-Bedingung verletzt.

Eine Verringerung der Fensterlänge hat jedoch zur Folge, das ein Signal-Segment immer weniger Informationen enthält. So gehen unter Umständen wichtige Informationen verloren, die die Komplexität eines Signals beschreiben. Die Folge ist eine erhöhte Unsicherheit durch Streuung zwischen mehreren Segmenten. So führt bei der Spektralanalyse eine kurze Segmentlänge zu gröberer Diskretisierung im Frequenzbereich und einer besseren Auflösung im Zeitbereich. Umgekehrt verhält es sich bei großen Fensterlängen.

Für die Mustererkennung erweist sich eine kurze Fensterlänge jedoch wieder als Vorteil. Je kürzer das Fenster gewählt wird, umso mehr Daten liegen für die Lernverfahren vor. Daraus ergibt sich eine bessere Generalisierungsfähigkeit der Algorithmen. Eine Überlappung der Segmente führt zu

[1] Wavelet-Transformation und WPT siehe Abschn. 3.5

einer weiteren Vergrößerung der Datenmenge, bringt aber einen entscheidenden Nachteil mit sich. Es entstehen dabei Segmente, die untereinander ähnlich sind, wobei der Grad der Ähnlichkeit vom Umfang der Überlappung abhängig ist. Bei einer nahezu vollständigen Überlappung nahe 100% geht auch die Ähnlichkeit gegen 100%. Bei zufälliger Partitionierung in Test- und Traningsmenge durch die Mustererkennungsverfahren finden sich ähnliche Segmente in beiden Teilmengen. Das führt zu einer Verzerrung in der Schätzung der Fehlerrate, die dadurch zu gering geschätzt wird. Klassifikatoren mit hoher Adaptivität können den Fehler dabei auf bis zu 0% schätzen [38, S. 21] [14].

In der Literatur gibt es unterschiedliche Angaben für die optimale Segmentlänge, diese reichen von 10ms für Spracherkennung bis hin zu 80ms für semantische Untersuchungen [29, S. 266]. Um für den Anwendungsfall dieser Arbeit eine optimale Segmentgröße zu finden, werden die Mustererkennungsverfahren auch über eine logarithmierte Skala getestet. Der geringste Wert für die Fenstergröße T_F wird dabei mit 5ms angenommen, der größte Wert mit $500ms$. Dabei wurde erwartet, daß die optimale Segmentgröße im unteren Bereich der Werte liegt. Als Referenzfenstergröße zur Entwicklung der weiteren Methoden wurde jeweils mit einer Fenstergröße von $T_F = 30ms$ und einer Überlappung von $T_{upd} = 15ms$ gearbeitet.

3.1.5 Gewichtung der Segmente

Die Wichtung ist nötig, um Diskontinuitäten zu vermeiden. Diese entstehen ansonsten durch die zugrundeliegende Periodisierung in Folge der diskreten Fouriertransformation. Das hat jedoch eine Verzerrung der Leistungsdichtefunktion im Spektralbereich zur Folge. Ein ungewichtetes Periodogramm[2] ist frei von Verzerrung, aber mit einer sehr hohen Varianz behaftet. Diese kann im Extremfall bis zu 100% betragen. Durch Modifikation (Wichtung mit einer Fensterfunktion) verringert man die Varianz, nimmt jedoch eine gewissen Unschärfe (Bias) in Kauf. Dieser Nachteil kann über eine Bias-Varianz-Steuerung verringert werden [20, S. 29].

3.2 Voice Activity Detection

Die Sprachaktivitätsdetektion (Voice-Activity-Detection, VAD) wird in dieser Arbeit dazu verwendet, die im Abschnitt 2.6 aufgeführten Merkmale im Zeitbereich zu erkennen, um so eine Abschätzung über das Pausenverhalten des Sprechers durchzuführen. Die im Folgenden beschriebenen Überlegungen stützen sich teilweise auf Kapitel drei der Diplomarbeit von L. Fuss [10], in dem drei verschiedene Methoden der VAD gegenübergestellt werden. Die VAD kann auf zwei verschiedene Ansätze aufgebaut werden,

[2] Gibt die spektrale Leistungsdichtefunktion an.

hier wird ausschließlich die Teilung in Sprache und Sprechpause verwendet[3]. In der Sprechpause besteht das Signal lediglich aus Rauschen.

Voice-Activity-Detection kommt in vielen verschiedenen Bereichen zum Einsatz, besonders in der Datenübertragung wie Mobilfunk. Auch in der Spracherkennung ist sie elementarer Bestandteil. Durch die verbreitete Anwendung gibt es eine Vielzahl von Ansätzen zur Realisierung. In diesem, sich recht einfach darstellenden Fall, daß nur eine Unterscheidung zwischen Sprachsignal und Rauschen zu implementieren war, und daß durch die Aufnahmebedingungen eine recht rauscharme Aufnahme vorliegt, wurde für diese Arbeit die Energieschwellmethode für die VAD ausgewählt. Nach Fuss, L. [10, S. 31] sind bei den genannten Bedingungen hinreichend gute Ergebnisse zu erwarten.

Das Signal wird, wie im vorherigen Kapitel beschrieben, gefenstert und anschliessend jedes Segment mit einer Hanning-Funktion gewichtet. Die Hanning-Funktion stellt einen vernünftigen Mittelwert als Gewichtung dar, so daß sie in dieser Arbeit zum Einsatz kam. Von jedem Signalfenster i wird die Energie nach Gleichung 3.2 und ein relativer Schwellwert K_{EN} berechnet.

$$E_i = \frac{1}{n} \sum_{k=1}^{n} (x_i)^2 \tag{3.2}$$

Der Schwellwert ergibt sich aus einem variablen Parameter p_{EN}, der zu optimieren ist, und der maximalen Energieamplitude des Signals.

$$K_{EN} = p_{EN} \cdot \max_{i \in \{1,,,,,F\}} (E_i) \tag{3.3}$$

Überschreitet die Energie eines Fensters die Schwelle K_{EN}, wird dieser Teil des Sprachsamples als Sprache markiert, der Rest des Signals als Rauschen. In Bild 3.2 ist ein Ausschnitt des Spachsignals und dessen VAD mit relativem Schwellwert ($p_{EN} = 0{,}1$) gezeigt. Man sieht, daß der Sprachanteil sehr gut markiert wird. Für das vorliegende Audiomaterial hat sich der genannte Wert $p_{EN} = 0{,}1$ als praktikabel erwiesen. Das wurde durch mehrfaches Variieren p_{EN} und einer repräsentativen Auswahl von Aufnahmen anhand grafischer Auswertung ermittelt.

Die errechnete Energieamplitude wird anschließend durch Faltung mit einer Gaussfunktion tiefpassgefiltert, um ihre hochfrequenten Bestandteile zu eliminieren. Die verwendete Filterfunktion 3.4 mit Standardabweichung σ und Erwartungswert μ zeigt einen idealen Gaussfilter. Gute Resultate wurden bei einer Cut-Off-Frequenz von $F_{CutOff} = 500Hz$ erzielt.

$$f_G(k) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{(t-\mu)^2}{2\sigma^2}} \tag{3.4}$$

[3]Ein anderer Ansatz unterscheidet noch nach stimmhafter und stimmloser Sprache, dies kann hier vernachlässigt werden.

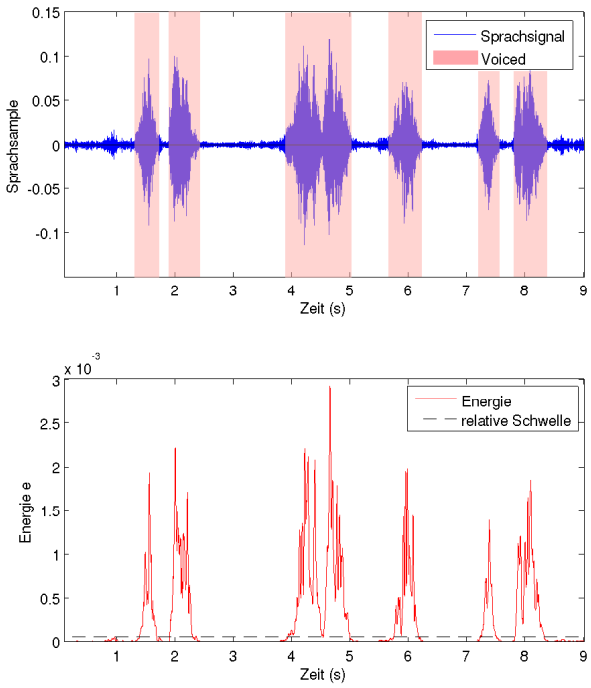


Abbildung 3.2: Ergebnis der Voice-Activity-Detection. Das obere Bild zeigt die Signalamplitude, farbige Flächen stellen die „voiced-parts“ dar. Im unteren Bild die geglättete Energieamplitude mit der Schwelle $p_{EN} = 0{,}01$

3.3 Fluency-Daten

Aus dem Ergebnis der VAD werden dann für jedes 10-Sekunden-Segment der Aufnahme die therapierelevanten Fluency-Daten (Sprachflüssigkeitsdaten) bestimmt. Dabei handelt es sich um Größen der deskriptiven Statistik, die mit den 1. und 2. Momenten der Verteilungsdichtefunktion korrespondieren:

– Verhältnis gesprochener Sprache zu Rauschen in Prozent

– Mittelwert der Pausenlänge

– Median-Wert der Pausenlänge

– Standardabweichung der Pausenlängen

– extrapolierte Sprechgeschwindigkeit in Silben pro Sekunde[4]

[4]Die Sprechgeschwindigkeit kann erst anhand der Ergebnisse der Silbendetektion be-

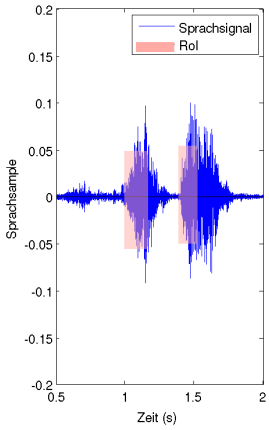


Abbildung 3.3: Region of Interest(RoI) einer Silbe

Dabei ist der Median der Pausenzeiten als sehr robust gegen Ausreißer zu verstehen, womit ein hinreichend genauer Wert ermittelt werden kann [14]. Diese und ähnliche Werte werden im Allgemeinen zur Evaluierung von Sprechdaten herangezogen und sind hinreichend untersucht [17]. Eine sehr detaillierte Auflistung findet sich im Handbuch zur Software FluencyMeter von Christian W. Glück [13, S. 30ff].

3.4 Silbendetektion

Das wichtigste Merkmal der in der KST gelehrten Sprechtechnik ist der weiche Stimmeinsatz, als das Anschwingverhalten einer jeden Silbe. Diese Sprechweise weicht gerade in den ersten Tagen nach der Therapie massiv sowohl von einer normalen Sprechweise und noch mehr von der Sprechweise vor der Therapie ab. Über die Silbendetektion soll dieses Merkmal zugänglich gemacht werden. Es ist davon auszugehen, daß jede Silbe, egal ob am Anfang eines Wortes oder am Ende weicher begonnen wird. Der Ausklang einer Silbe ist weniger von Bedeutung, da dieser sich automatisch dem Anschwingen angleicht. Eine weich begonnene Silbe wird also auch weich beendet [17]. Dadurch kann diese Arbeit sich auch auf die Erkennung der Silbe bis Scheitelpunkt[5] der Amplitude beschränken. Im folgenden wird also Halbsilbendetektion synonym benutzt, der erkannte Bereich soll demzufolge als Halbsilbe bezeichnet werden. Dabei ist aus therapeutischer Sicht vor allem die Silbe am Wortanfang relevant, deshalb ist es weniger wichtig, jede Silbe korrekt zu erkennen [17]. Vielmehr wird Wert darauf gelegt, den relevanten

rechnet werden, diese ist im folgenden Abschnitt3.4 beschrieben.

[5]gemeint ist das lokale Maximum der geglätteten Energieeinhüllenden

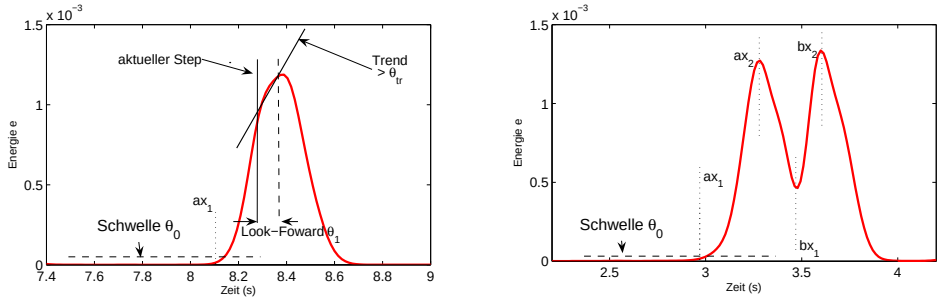


Abbildung 3.4: Parametrisierung der Halbsilbendetektion im linken Bild mit Energieschwelle θ_0, Look-Forward θ_1 und Trend $> \theta_{tr}$. Im rechten Bild eine Silbe mit zwei Maxima und den detektierten Marken x_1 für Silbenanfang und x_2 für Silbenmaximum.

Teil der erkannten Silben genau einzugrenzen (Abb. 3.3).

3.4.1 Ablauf der Silbendetektion

Die in dieser Arbeit entwickelte Halbsilbendetektion basiert wie auch die Voice-Activity-Detection auf der Energie des Signals. Im Prinzip werden zwei Marken benötigt, die erste kennzeichnet den Beginn der Silbe und kennzeichnet damit quasi das Ende eines von der Voice-Activity-Detection erkannten stillen Bereiches. Die zweite Marke kennzeichnet das Maximum, den Peak der Silbe. Dazu sind zwei Parameter nötig:

– der Schwellwert θ_0 für den Startwert bzw. die erste Marke

– Trend über eine bestimmte Signallänge (look-foward) $\theta_t r$

– der Mindest-Wert für den Trend $\theta_t r - min$

Aus der Signalenergie lässt sich näherungsweise die Einhüllende des Ausgangssignals darstellen, sofern die Energieamplitude mit einem Tiefpass geglättet wird. Das Energiesignal wird mit der in Abschnitt 3.2 verwendeten Energiefunktion berechnet und anschliessend tiefpass-gefiltert. Die Cut-Off-Frequenz des Tiefpass-Gaussfilters bleibt ebenso bei 500Hz, dieser Wert hat sich als praktikabler Mittelwert erwiesen. Eine zu starke Glättung würde kurze Peaks verschwinden lassen und so einen Informationsverlust zur Folge haben. Als weitere Information wird durch lineare Regression der Trend eines kurzen Abschnittes der Energieamplitude vorausschauend berechnet („look-forward“, s. Abb 3.4). Einen kurzen Ausschnitt mit dem Verlauf der Energiefunktion zeigt Abbildung 3.5 mit dem mittigen Graph.

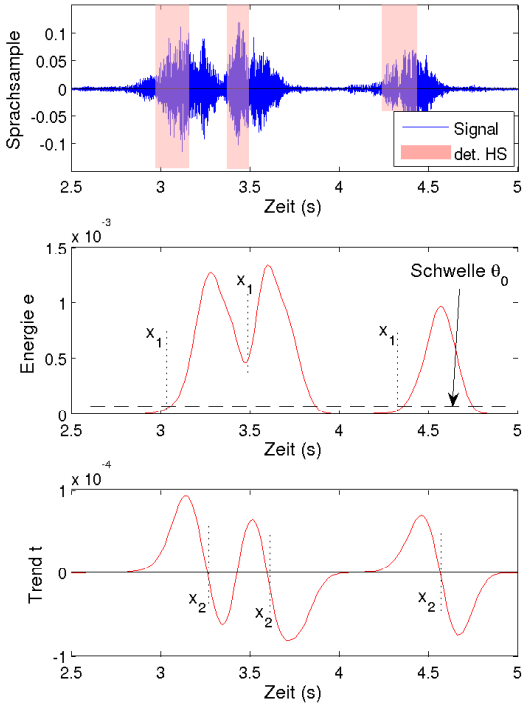


Abbildung 3.5: Silbendetektion eines Signalabschnitts (oben) mit den zwei Hilfsfunktionen Energie (Mitte) und Trend (unten).

3.4.2 Probleme der Silbendetektion

Ein Problem für eine gute Silbenerkennung stellen Silben mit 2 lokalen Maxima dar. Hier wird durch den verwendeten Algorithmus nur das erste Maximum korrekt erkannt, die Erkennung des zweiten Maximum ist von einigen Faktoren abhängig. Die auftretenden Möglichkeiten und ihre Einschätzung zeigt Tabelle 3.1 Eine Möglichkeit, um den Fehler der Detektion von mehreren Maxima als unabhängige Silbe auszugleichen, wird stark durch die von den Klienten umgesetzte Sprechtechnik behindert. Die Einbeziehung der Sprechgeschwindigkeit könnte hier Abhilfe schaffen, da 2 Maxima, die zu nah zusammenliegen, zu einer Silbe zusammengefasst werden könnten. Die Sprechgeschwindigkeit kann aber zur Laufzeit erst anhand der detektierten Silben geschätzt werden. Damit unterliegt sie dem gleichen Fehler, der korrigiert werden soll und kann somit nicht als Maß herangezogen werden. Bei einer stark verlangsamten Sprechweise liegen die Maxima zu weit auseinander, als daß eine derartige Methode Erfolg hätte. Eine robuste Lösung wäre hier nur möglich, wenn die Sprechgeschwindigkeit durch ein anderes System

Tabelle 3.1: Auftretende Variationen des Signals einer Silbe und dessen Einschätzung für die Silbendetektion.

Konstellation	**Einschätzung**
1 Maximum	wird korrekt erkannt
2 Maxima gleicher Höhe mit Abstand $< \theta_1$	1. Maxima korrekt erkannt, 2. Maxima wird übergangen
2 Maxima gleicher Höhe mit Abstand $> \theta_1$	2 Maxima detektiert, Zuordnung zu einer Silbe nicht möglich. Zur Berechnung des Anschwingverhalten kein Problem, erzeugt aber Ungenauigkeiten in der Sprechgeschwindigkeits-Schätzung
2 Maxima unterschiedlicher Höhe mit Abstand $> \theta_1$	2 Maxima detektiert, mit gleicher Problematik wie oben.

vorab korrekt erkannt wird und zur Laufzeit des Algorithmus Verwendung findet.

3.4.3 Validierung der Silbendetektion

Rauschen kann durchaus für zusätzlich detektierte Silben sorgen, sofern es einen Pegel stark über dem unteren Schwellwert θ_0 hat. Mit Hilfe der VAD und über die Silbenlänge (Gesamtsilbe nicht unter 80ms lang) werden durch den entwickelten Algorithmus falsch erkannte Halbsilben verworfen. Eine wesentlich verbesserte Methode beschreibt Fuss [10], die aber für die Problemstellung dieser Arbeit nicht erforderlich scheint. Dort wird mit insgesamt 3 Schwellen gearbeitet, um Silben mit mehr als einem lokalen Maximum sicher zu erkennen. Bei der oben beschriebenen Methode treten an dieser Stelle relativ viele Fehler auf, die in Tabelle 3.1 benannt werden. Die Gesamtfehlerrate für die Detektion wurde durch Auszählen ermittelt und ist für 4 prominente Beispiele hier aufgeführt (siehe Tab. 3.2). Auffällig bei der Bestimmung der Fehlerrate ist die starke Divergenz, das heisst das Auseinanderfallen der Fehlerrate, zwischen verschiedenen Aufnahmen. Hier zeigt sich auch mathematisch, daß sich die individuellen Ausprägungen jedes Sprechers bei einer Sprechstörung zu verstärken scheinen. Die Fehlerrate ε bei einer Anzahl Silben S errechnet sich wie folgt,

$$\varepsilon = \frac{|S - S_d + S_f|}{S} \tag{3.5}$$

Tabelle 3.2: Fehlerquote der Halbsilbenerkennung für die ersten 10 Sekunden der Aufnahmen von 2 Klienten

	K001 VT	K001 NT	K002 VT	K002 NT
tatsächl. Silben S	42	31	17	16
erkannte Halbsilben S_d	35	22	15	15
davon doppelt erk. S_{dop}	0	1	1	8
n. erkannte Halbsilben	7	9	2	1
Fehl-Erkennung S_m	0	1	1	0
Fehlerrate ε	16,7%	33,9%	20,6%	31,25%

wenn S_d die Anzahl detektierter Halbsilben und S_f die Anzahl der fehlerhaft erkannten Halbsilben ist. Die fehlerhaften Halbsilben wiederum ergeben sich aus der Summe der fälschlicherweise markierten Silben und der halbierten Anzahl der doppelt erkannten Silben S_{dop}.

$$S_f = S_m + \frac{1 + S_{dop}}{2} + 0,5 \tag{3.6}$$

Als doppelt erkannte Halbsilben werden solche gezählt, bei denen das erste und das zweite Maximum jeweils als extra Silbe markiert werden.

Die höhere Fehlerrate bei Aufnahmen, die nach der Therapie angefertigt werden, lässt sich vermutlich auf die langsamere Sprechgeschwindigkeit zurückführen. An dieser Stelle sind sicher noch Verbesserungen möglich, zum Beispiel die direkte Berücksichtigung der relativen Sprechgeschwindigkeit. Insgesamt sind aber auch andere Algorithmen sehr fehleranfällig, in der Literatur werden Fehlerraten von etwa 12% für gelesene und etwa 20% Fehler für Silbenerkennung bei spontan gesprochener Sprache genannt [10]. Bei diesen Beispielen wird immer eine gesunde, störungsfreie Sprache zugrunde gelegt, im Vergleich zu dem hier vorliegenden Datenmaterial also Idealzustand.

3.5 Wavelet-Transformation

Aus den detektierten Halbsilben müssen Informationen gewonnen werden, die als Eingabevektoren für die Klassifikationsverfahren geeignet sind. Dazu werden die in diesem und im nächsten Abschnitt beschriebenen Transformationen eingesetzt. Die gebräuchlichste Transformation, um aus einem zeitdiskreten Signal spezifische Informationen zu gewinnen, ist die diskrete Fourier-Transformation. Die Grundannahme dabei ist, daß das Signal als abgetastetes, periodisches Signal vorliegt[6]. Die diskrete Fourier-Transformation kann

[6]Der bekannteste Fall eines periodischen Signals wird durch die Sinus- oder Kosinus-Schwingung repräsentiert.

jedoch nur mit Signalen umgehen, die ihre Struktur langfristig ändern, also in einem kurzen Zeitfenster quasi-stationär sind [33, S. 33f]. Die für diese Arbeit vorliegenden Daten hingegen sind hochgradig instationär, so daß die Fourier-Analyse nicht geeignet erscheint.

3.5.1 Grundlagen der Wavelet-Transformation

Abhilfe verspricht an dieser Stelle die sogenannte Wavelet-Transformation, die lokale Eigenschaften eines Signals (kurze Störungen, abrupte Änderungen im Frequenzbereich, etc.) durch Basisfunktionen mit zeitlich lokaler Ausdehnung gut erfasst. Eine ausführliche Darstellung, auf der auch dieser Abschnitt basiert, findet sich im Handbuch der Matlab-Wavelet-Toolbox [26]. Die mathematischen Hintergründe sind der guten Arbeit von S. Schlagner und U. Strehlau entnommen [33].

Eine Fourier-Transformation stellt im Gegensatz dazu ein Signal als Summe von wenigen, einfachen und harmonischen Schwingungen dar. An Stelle dieser periodischen Sinus- und Kosinus-Funktionen der Fourier-Transformation treten aperiodische Basis-Wavelets. Das Signal wird durch Translation und Skalierung dieser irregulären und meist asymmetrischen Basis- oder „Mother"-Wavelets[7] in seine Bestandteile zerlegt. Diese Wavelets sind als irreguläre, oftmals viele Freuenzen enthaltende Funktionen definiert. Es liegt auf der Hand, daß Instationaritäten und abrupte Änderungen des Ausgangssignals hiermit besser zu erfassen sind. Ein aperiodisches Signal wäre nur durch die Summe einer unendlich großen Anzahl von harmonischen Schwingungen zu beschreiben. Das prädestiniert die Wavelet-Transformation zur Erkennung von lokalen Eigenheiten eines Signals. Mathematisch wird ein Wavelet $\Psi(t) \in L^2(\mathbb{R})$ als rasch abklingende, oszillierende Funktion beschrieben, welche in einem kleinen Zeitbereich ungleich null ist und einen endlichen Energiebetrag aufweist, wobei $\Psi(0)$ die Fourier-Transformierte zur Frequenz $\omega = 0$ ist:

$$\hat{\Psi}(0) = \int_{-\infty}^{\infty} \Psi(t)dt = 0 \tag{3.7}$$

3.5.2 Typische Basis-Wavelets

Durch die typischen Eigenschaften eines Wavelets existiert eine quasi unendliche Anzahl verschiedener Wavelets und Wavelet-Familien. Als Wavelet-Familie bezeichnet man ein Wavelet, das über verschiedene Momente verfügt. Dies wird auch als Ordnung des Wavelets bezeichnet [33, S. 36]. Die einfachste Form stellt das Haar-Wavelet dar, das im Grunde einer zeitlich begrenzten Rechteckfunktion entspricht. Die Abb. 3.6 zeigt noch weitere, wichtige Basis-Wavelets als Funktion $\Psi(t)$.

[7] Ingrid Daubechies bezeichnet Wavelets als „kleine Wellchen" [33, S. 33].

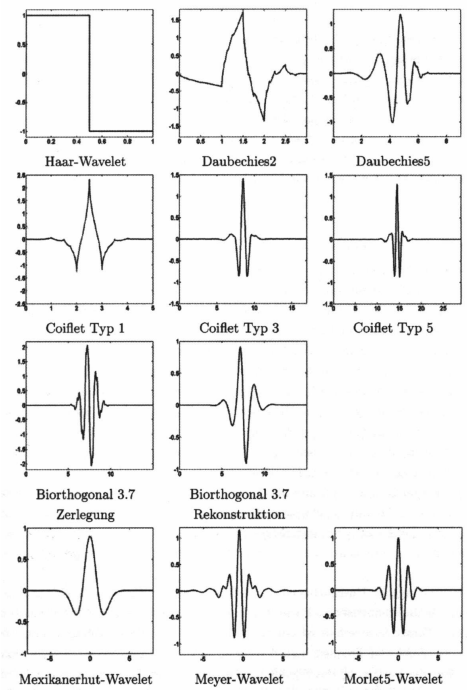


Abbildung 3.6: Darstellung einiger Wavelets als Funktion $\Psi(t)$, entnommen aus [33]

3.5.3 Diskrete Wavelet-Transformation

Die beliebige Zerlegung eines Signals S in jede mögliche Kombination aus Wavelet-Koeffizienten würde eine nicht zu bewältigende Datenmenge erzeugen. Die diskrete Wavelet-Transformation (DWT) und ihre Implementation durch Mallat löst das Problem durch eine Filterbank mit Hoch- und Tiefpass-Filter (Abb. 3.7). Der Filtervorgang erzeugt aus dem Signal S ein hochfrequentes, aber wenig skaliertes Signal, das als Wavelet-Detail-Koeffizient cD bezeichnet wird. Der Tiefpass ergibt den um cD verringerten Approximations-Koeffizienten cA als niederfrequentes, aber hoch skaliertes Signal. Das Downsampling verwirft jeden zweiten Datenpunkt des Approximations- und des Detail-Signals und korrigiert so die Anzahl der Samples auf die Größe des Original-Signals.

3.5.4 Wavelet Decomposition Tree

Wird der Filterprozess der DWT beliebig auf die jeweilige Approximation angewendet, erhält man den sogenannten Wavelet-Dokompositions-Baum. Dadurch wird das Signal bei n Iterationen in n niedrig-frequente Detail-Komponenten zerlegt, die mit jeder Approximations-Stufe in der Auflösung

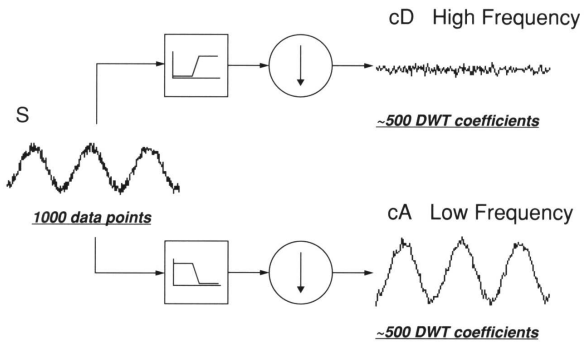


Abbildung 3.7: Darstellung eines Approximationschrittes der Diskreten Wavelet-Transformation eines Signals S in seine Approximation cA und Detail cD. [26]

sinkt. So werden Grobstrukturen im Signal anschaulich von Feinstrukturen getrennt und können einzeln analysiert werden. Eine noch bessere und flexiblere Analyse erlaubt die Zerlegung zum sogenannten Wavelet-Paket-Baum (WPT). Hier wird nicht nur die Approximation des Signals als Ausgangssignal für die nächste Zerlegungsstufe verwendet, sondern auch das Detail nochmals in beide Bestandteile gefiltert. Dies beansprucht wesentlich mehr Rechenleistung, erzeugt aber eine höhere Genauigkeit bei der Rekonstruktion des Signals durch Umkehrung der Zerlegung. Der WPT kommt daher hauptsächlich bei Filteralgorithmen und Kompressionsverfahren zur

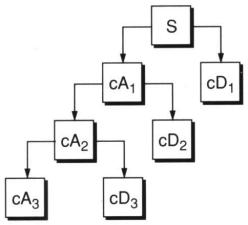


Abbildung 3.8: Schema einer 3-Stufigen einfachen Wavelet-Transformation [26]

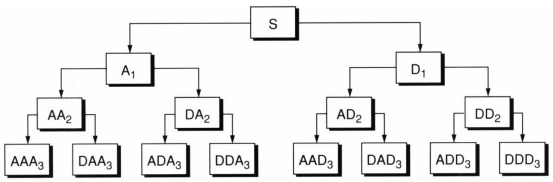


Abbildung 3.9: Schema einer Wavelet-Paket-Zerlegung in 3 Stufen [26]

Anwendung [8].

3.5.5 Angewandte Wavelet-Transformation

Für diese Arbeit wurde die Wavelet-Transformation auf die bei der Halbsilbendetektion markierten Bereiche angewandt. Dabei kam eine einfache Wavelet-Zerlegung mit 8 Approximationsstufen zum Einsatz. Eine Rekonstruktion des Signals ist hier nicht nötig, so dass kein offensichtlicher Grund bestand, die rechenintensivere Wavelet-Paket-Transformation einzusetzen. Mit 8 Approximations-Stufen erhält man nach vollständiger Zerlegung eines Signals mit der ursprünglichen Länge von $1sec = 16000Samples$ ein Approximations-Signal mit 63 Datenpunkten. Durch diese geringe Länge kann davon ausgegangen werden, daß alle relevanten Informationen des Ausgangssignals in den Detail-Koeffizienten enthalten sind und eine weitere Zerlegung keinen Informationsgewinn mehr bringt.

Als Basis-Wavelet wurde das Wavelet 3. Ordnung der Daubechies-Familie ausgewählt (db3-Wavelet). Wavelets der db-Familie werden als recht universelle Wavelets für die Signalverarbeitung beschrieben. Ab der zweiten Ordnung sind sie beliebig glatt und mit beliebig vielen verschiedenen Momenten konstruierbar [33, S. 38]. Genauere Untersuchungen zu den verwendeten Wavelets kommen auch zu dem Schluss, auf keinen Fall

> „unendlich Energie in die Wahl des Wavelets zu verschwenden. Man sollte sich nicht von der Tatsache blenden lassen, daß jedes Signal sein optimales Wavelet besitzt. Dies trifft zwar zu, doch geht damit auch der ganze Vorteil eines allgemeinen Verfahrens, wie es die Wavelets darstellen, verloren. Wer wirklich darauf aus ist, ein Signal im Hinblick auf eine ganz bestimmte Anwendung zu analysieren, sollte lieber gleich einen anderen Zugang wählen.“(Quelle [21])

3.6 Cepstral-Transformation

Eine andere Form der spektralen Informationsgewinnung stellt die Cepstralanalyse dar. Die folgende Beschreibung der Cepstral-Transformation orientiert sich weitgehenst an den Gedanken von Paulus, D. [29] und Paulus, E. [30]. Als Implementation wurde auf die Auditory Toolbox [34] von Malcolm Slaney zurückgegriffen. Dabei handelt es sich um eine frei verfügbare Matlab-Toolbox für Audio-Signalverarbeitung. Sie befindet sich auch auf der beiliegenden CD.

[8] Der in Abschn. 3.1.3 genutzte Filter basiert auf der WPT

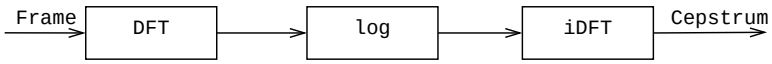


Abbildung 3.10: Schematische Darstellung einer einfachen Cepstral-Transformation

3.6.1 Entstehung des Cepstrum

Der Name selbst verrät schon ein wenig über die Enstehung des Cepstrum. Dazu wird das Leistungsdichtespektrum $S(f)$ eines Signales $s(f)$ logarithmiert und anschliessend die inverse Fourier-Tranformierte gebildet. Die Komponenten $c_\nu^{(\tau)}$ des komplexen Cepstrum $c^{(\tau)} = (c_0^{(\tau)}, c_1^{(\tau)}, ..., c_{M-1}^{(\tau)})^T$ erhält man also wie folgt:

$$c_\nu^{(\tau)} = \frac{1}{M} \sum_{\mu=0}^{M-1} (logF_\mu^{(\tau)}) exp\left(\frac{2\pi i \nu \mu}{M}\right) \tag{3.8}$$

Das in der Praxis gebräuchlichere reelle Cepstrum erhält man, in dem man statt der komplexen Funktion $logS(f)$ die reellwertige, absolute Funktion $log|S(f)|$ verwendet:

$${}^r c_\nu^{(\tau)} = \frac{1}{M} \sum_{\mu=0}^{M-1} (log|F_\mu^{(\tau)}|) exp\left(\frac{2\pi i \nu \mu}{M}\right) \tag{3.9}$$

Der Wertebereich des Cepstrum ist aber in beiden Fällen immer reell-wertig, darum erscheint der Name des komplexen Cepstrum etwas unglücklich gewählt. Man erhält also einen zum Zeitbereich parallelen Wertebereich, die Quefrenz. Daraus resultiert die sprachliche Entstehung des Wortes Cepstrum, nämlich durch Vertauschung der Anfangsbuchstaben aus Spektrum. Der Vorteil dieses Algorithmus liegt in der Anwendung des Logarithmus, was zur Folge hat, daß aus multiplikativen Anteilen im Spektrum additive Anteile im Cepstrum werden [7]. Dadurch können im Cepstrum die verschiedenen Anteile eines Signals gut erkannt und behandelt werden. Im Prinzip ergibt jede Frequenz eine Spitze, wobei gilt: Je höher eine Frequenz, umso weiter rechts ist die Spitze im Cepstrum angesiedelt. Niedrig-frequente Schwingungen manifestieren sich im linken Bereich. Das erlaubt zum Beispiel eine einfache cesptrale Filterung, die sogenannte Lifterung, in dem die entsprechenden Quefrenzen subtrahiert werden und eine Rücktransformation des Signales erfolgt. Damit bietet sich das Cepstrum besonders für Applikationen an, in denen Signale mit typischen, immer wiederkehrenden Schwingungen verarbeitet werden [7].

Das Kurzzeitcepstrum $c(\tau, t)$ erhält man, indem statt des Frequenzspektrums $S(f)$ ein Kurzzeitspektrum $S(f, t)$ mittels diskreter Fourier-Transformation erzeugt wird. Ebenso wie das Leistungsdichtespektrum

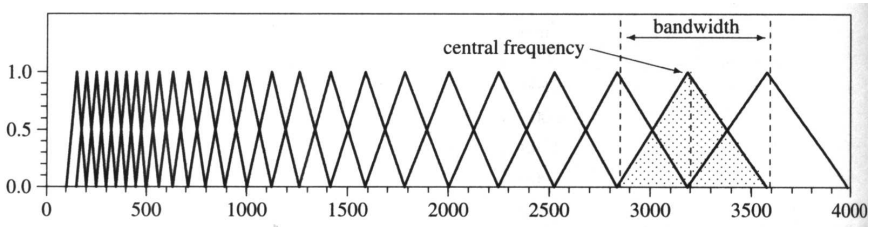


Abbildung 3.11: Mel-Filterbank aus Dreiecksfiltern mit 25 Bändern [29, S. 276]

ist das Kurzzeitcepstrum symmetrisch, so daß es genügt, die Werte c_0, c_1,....,$c_{M/2}$ zu behalten. Diese Werte werden in der Literatur als Cepstral-Koeffizienten bezeichnet [30, S. 315].

Damit bietet sich das Cepstrum besonders für Applikationen an, in denen Signale mit typischen, im besten Falle bekannten Schwingungsanteilen verarbeitet werden. Als prominentes Beispiel dient die Spracherkennung über charakteristische Formantenfrequenzen oder das Filtern von störenden Anteilen bestimmter Systeme wie Telefonleitungen.

3.6.2 Mel-Skala

Die Leistungsfähigkeit der menschlichen Spracherkennung und Verarbeitung durch Gehörtrakt und Gehirn ist immer wieder eine Herausforderung, wenn es darum geht, die Funktionalität dieser Organe wenigstens teilweise nachzubilden. Da liegt es nahe, bereits bei der Vorverarbeitung sich zumindest grob an der Funktionsweise des menschlichen Gehörs zu orientieren. Dieser Gehörorientierte Ansatz berücksichtigt das unterschiedliche Auflösungsvermögen des Ohres für unterschiedliche Frequenzbänder. Empirische Untersuchungen haben zur sogenannten Mel-Skala geführt. Bis etwa 700Hz bleibt das Frequenzauflösungsvermögen konstant, mit steigender Frequenz nimmt es logarithmisch ab. Für eine Filterung bedeutet dies, daß die lineare Frequenz-Achse f wie folgt in die Mel-Skala f_{mel} transformiert wird [29, S. 275]:

$$f_{mel} = 2595log\left(1 + \frac{f}{700Hz}\right) \tag{3.10}$$

Realisiert wird die Mel-Skalierung durch eine Gruppe von überlappenden Dreiecksfiltern (Abb. 3.11), deren Kanäle umso breiter werden, je höher die Mitten-Frequenz wächst.

3.6.3 Mel-Frequency-Cepstral-Coefficients

Eine Kombination von Mel-Filterbank und Cepstral-Transformation führt regelmäßig zu einer verbesserten Analyseleistung in der Sprachsignalverar-

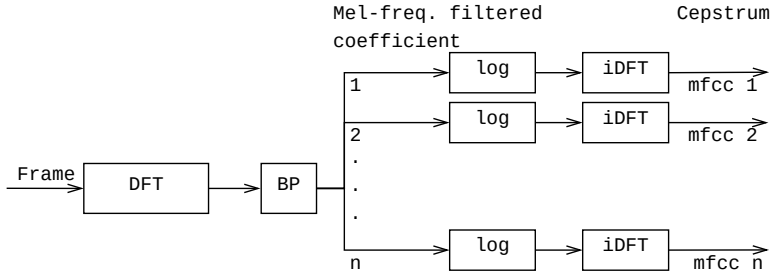


Abbildung 3.12: Schematische Darstellung einer Cepstral-Transformation mit Mel-gefilterten Koeffizienten

beitung. Dadurch erlangt man eine hohe Robustheit gegenüber Störungen und anderen unerwünschten Einflüssen [30, S. 316].

Dazu gewichtet man das Spektrum mit der Mel-Filterbank und erhält das Mel-Spektrum (MFC) und berechnet von jedem logarithmierten Filter-Koeffizienten die inverse Fouriertransformierte (Abb. 3.12). Auf diese Art und Weise erhält man bei K Filtern entsprechend K Mel-gefilterte Cepstral-Komponenten (MFCC). Je nach Feinheit der Mel-Filterbank stellen diese Komponenten einen sehr schmalbandigen Frequenz-Bereich des originalen Signals dar, dessen Informationsgehalt in Abhängigkeit vom analysierten Signal stark schwanken kann. Entscheidend ist jedoch, daß charakteristische Frequenzen besser erkannt werden können, als im ungefilterten Cepstrum. Typische Schwingungen finden sich nämlich immer in der gleichen MFCC-Komponente.

3.6.4 Angewandte Cepstral-Transformation

Durch die Eigenheiten der Cepstral-Transformation lassen sich gerade harmonische Bestandteile wie die Sprachgrundfrequenz oder Formantenfreqeunzen im Cepstrum eines Signal erkennen. Eine weitere periodische Komponente eines gesunden Sprachsignals ist die Glottisschwingung (Abb. 3.13), die Gall in einer Periodizität des Sprachsignals wiederfindet [11, S. 18f]. Die Glottis, also die Stimmlippen[9], ist das Organ, welches die Grundschwingungen der Sprache erzeugt. Allerdings gibt es auf diesem Gebiet noch keine Untersuchungen, inwieweit solche Periodizitäten bei gestotterter Sprache vorhanden sind oder sogar komplett fehlen. Das hochgradig individuelle Störungsbild lässt jedoch vermuten, daß diese Frage keinesfalls global zu beantworten ist. Umso größer ist das Interesse dieser Arbeit, wenigstens ansatzweise eine Aussage über die Verwertbarkeit der Cepstralanalyse auf diesem Gebiet treffen zu können.

[9] Umgangssprachlich, aber medizinisch nicht korrekt, wird die Glottis als Stimmbänder bezeichnet.

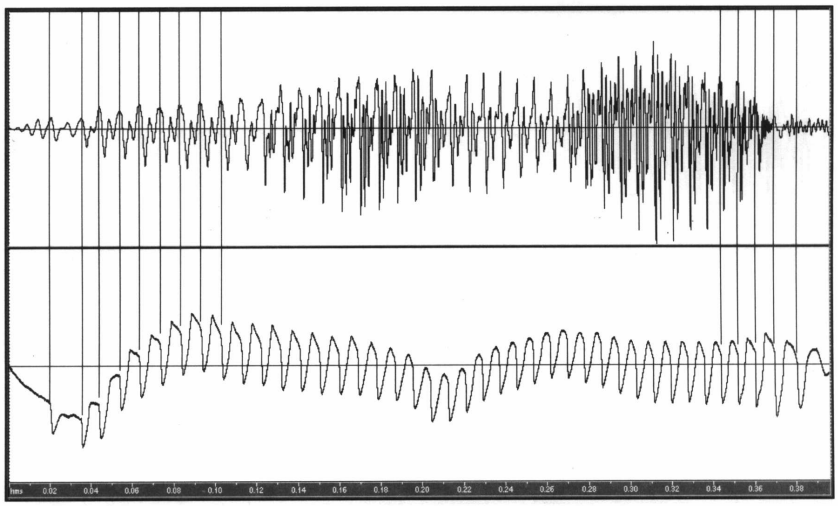


Abbildung 3.13: Sprachsignal eines gedehnten Lautes (oben) mit Elektro-Glottogramm (unten) [11, S. 19]

Zur Anwendung kommt in dieser Arbeit die Funktion `[ceps,...] = mfcc(input,...)` der Auditory-Toolbox [34, S. 29]. Diese verwendet einen 13-stufigen Mel-Filter. Um möglichst alle Merkmale in der Mustererkennung zu berücksichtigen, werden alle 13 Cepstralkoeffizienten ohne Bewertung oder Gewichtung in die Merkmalsvektoren mit eingebracht. Eine Zwischenverarbeitung birgt das Risiko, relevante Merkmale unbeabsichtigt zu übergehen.

3.7 Merkmale aus Wavelet- und Cepstral-Transformation

Das Ziel der Merkmalsextraktion, einen dem Ausgangssignal gegenüber verringerten Merkmalsraum zu erzeugen, wurde durch die bisher angewandten Methoden nicht erreicht. Die Aufspaltung in mehrere Komponentensignale durch Wavelet- und Cepstral-Transformation haben das Gegenteil, also einen erweiterten Merkmalsraum erzeugt. Danach werden für alle Dekompositions-Level des Wavelet-paket-tree und für alle 13 Cepstral-Koeffizienten folgende Werte berechnet:

– Energie relativ zur Anzahl der diskreten Werte (Samplelänge)

– Entropie

– log. Energie

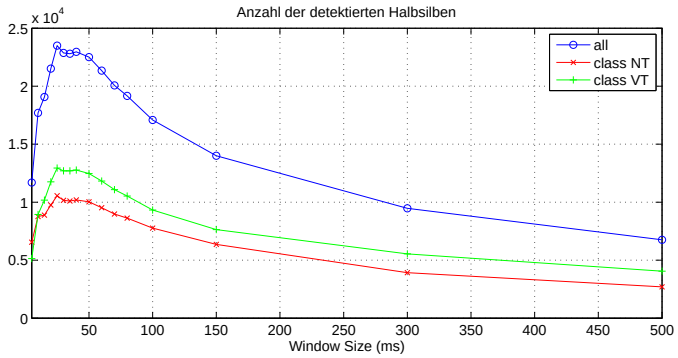


Abbildung 3.14: Extrahierte Merkmalsdatensätze in Abhängigkeit zur Fenstergröße

Von entscheidender Bedeutung ist, daß die Berechnung für jeden Merkmalsvektor identisch erfolgt und keine Anpassung des Berechnungsverfahrens an die Daten durchgeführt wird. Eine solche Veränderung der Ausgangssituation würde in der Mustererkennung für nicht zu korrigierende Verfälschungen sorgen und somit das komplette Ergebnis unbrauchbar machen.

3.8 Klassifikation

Die in dieser Arbeit verwendeten Methoden zur Klassifikation sind überwacht lernende Verfahren und benötigen demzufolge eine Klasseninformation für jeden Merkmalsvektor. Diese Information liegt jeder aufgenommenen Datei bei, muss also nur in die Merkmalsvektoren übertragen werden. Eine fehlerbehaftete Einschätzung und Klassifizierung durch Experten ist somit ausgeschlossen. Um das Ziel[10] der Diplomarbeit zu erreichen, wird mit den zwei Klassen, die vermutlich die größte Distanz im Merkmalsraum zueinander haben, gearbeitet. Es wird jeweils „Vor der Therapie“ und „Nach der Therapie“ unterschieden und als Klasse festgelegt (siehe Ziele der Arbeit: Abschn. 1.2). Daraus ergibt sich die in Abb. 3.14 dargestellte Verteilung der Merkmalsvektoren auf die Klassen. Von der Klasse „Vor der Therapie“ liegen generell etwas weniger Merkmalsvektoren vor, da die Aufnahmen dieser Klasse tendenziell immer etwas kürzer sind. Die Sprechgeschwindigkeit ist dort im Mittel geringfügig höher als direkt nach der Therapie.

[10] Lassen sich Indizien und Merkmale finden, mit deren Hilfe eine Klassifikation nach flüssiger und gestotterter Sprache möglich ist? (Abschnitt 1.2)

3.9 Normalisierung

Der Wertebereich der Merkmalsvektoren differiert, bedingt durch die unterschiedlichen Verfahren, sehr stark. Die Klassifikatoren können jedoch oftmals mit solch einem Wertebereich nur sehr schlecht umgehen. Deshalb erfolgt eine Normalisierung aller Merkmale in einen Bereich zwischen 0 und 1. Dafür stehen eine große Auswahl an Skalierungsmethoden zur Verfügung, wie zum Beispiel lineare, logarithmische oder z-Skalierung. Für diesen Merkmalsraum wurde eine lineare Skalierung gewählt, da insgesamt sehr wenig Wissen über das Verhalten der extrahierten Merkmale in den Randgebieten des Wertebereiches vorhanden ist. Die lineare Skalierung verändert nicht den relativen Bezug der einzelnen Werte zueinander, wie dies in andern Fällen durchaus erwünscht sein kann (vgl. logarithmierte Skalierung). Die Normalisierung wurde außerdem innerhalb des 95%-Quantils durchgeführt, um eine gleichmäßigere Verteilung der Werte zu erreichen.

Kapitel 4

Klassifikatoren

Um auf die Fragen, die sich zu Beginn des Projektes stellten[1], eine aussagekräftige Antwort zu finden, werden im Rahmen dieser Arbeit 3 verschiedene Klassifikatoren auf die gewonnenen Daten angewendet:

– Lineare Diskriminanzanalyse

– Lernende Vektorquantisierung

– Support-Vector-Maschine

Für alle drei Verfahren existieren an der Fachhochschule Schmalkalden fertig implementierte Simulatoren, auf die zurückgegriffen werden konnte. Diese Software zählt zwar nicht zu den Standardwerkzeugen, die in der Industrie verbreitet sind, es liegen jedoch hinreichend große Erfahrungen damit vor, so daß keine Validierung der Werkzeuge nötig ist [14].

4.1 Vorbedingungen

Die extrahierten Merkmalsvektoren werden durch die Klassifikation einer definierten Menge an Klassen zugeordnet. Die dazu erforderlichen Klassifikatoren und ihre theoretische Grundlage möchte ich in diesem Kapitel beschreiben. Die Auswertung der Klassifkation wird im Kapitel 5 ausführlich erläutert. Der Klassenraum, der in dieser Arbeit Anwendung findet, besteht aus zwei Klassen: „Vor Therapie“ und „Nach Therapie“. Die verwendete Menge an Merkmalsvektoren muss

1.) repräsentativ sein.

2.) Die Merkmalsvektoren einer Klasse müssen sich in einem hinreichend kompakten Gebiet befinden.

[1] siehe Einleitung, Abschnitt 1.2

Je eher vor allem das zweite Postulat erfüllt ist, umso mehr lassen sich die Ergebnisse des Trainingsprozesses generalisieren. Zur Abschätzung der Generalisierungsfähigkeit ist die Klassifikationsrate ein geeignetes Maß [12, S. 93].

Sind alle Vorbedingungen erfüllt, können die Klassifikationsmodelle erstellt werden. Die Erstellungsvorschriften sind für neuronale Netze (z.Bsp LVQ-Kl.) genauso gültig wie für Verfahren, die regelbasiertes, maschinelles Lernen implementieren[2]. Dazu werden aus der Gesamtmenge M zwei zufällige, disjunkte Teilmengen M_{tr} und M_{te} gebildet. Die Menge M_{tr} dient zum Trainieren des Klassifikators, d.h. zur Erzeugung der optimalen Trennfunktion, während mit der Menge M_{te} die Trennfunktionen auf ihre Generaliserungsfähigkeit hin geprüft werden. Aus Validierungsgründen muss dieser Vorgang mehrfach wiederholt werden. Gängige Kreuzvalidierungsverfahren[3] sind unter anderem:

– Leave-m-out m aus n, m < n-1 [4]

– Leave-one-out 1 aus n

Bei m aus n wird die Testmenge zufällig mit m Elementen aus der Gesamtmenge erstellt, somit verbleiben $n-m$ Elemente in der Trainingsmenge M_tr. Gebräuchliche Teilungsverhältnisse sind dabei $M_{tr} = 50, M_{te} = 50$ oder das in dieser Arbeit verwendete[5] Separationsverhältnis 80% zu 20%. Bei der verwendeten Teilung muss die etwas geringere Generalisierungsfähigkeit des Klassifikators durch eine mehrfache Zahl an Wiederholungen kompensiert werden [38, S. 93]. Im zweiten Fall 1 aus n wird genau ein Element der Menge M_{te} zugeordnet, das stellt damit einen Spezialfall der ersten Variante dar. Der Klassifikator wird mit den restlichen Elementen, die sich in M_{tr} befinden, trainiert. Dann wird die Trennfunktion mit dem einen Element der Testmenge überprüft und die Klassifikation bestimmt. Dieser Vorgang wird für alle n Elemente wiederholt [12, S. 93f] [38, S. 39]. Nach allen Trainings- und Testdurchläufen können die Klassifikations- und die Reklassifikationsrate bestimmt werden. Dabei stellt die Klassifikationsrate die wichtigste Größe dar, sie drückt die Generalisierungsfähigkeit des Klassifikators aus. Sie kann damit als die Erkennungsrate des trainierten Klassifikators angesehen werden, denn sie beschreibt letztendlich, wie gut unbekannte Merkmalsvektoren klassifizert werden. Sie ergibt sich aus der Summe der richtigen Klassifikationen k_i der Klasse i zur Anzahl der Gesamtklassifikationen m, wobei q_i die a-priori

[2] z.Bsp. SVM-Klassifikatoren

[3] Aus Platzgründen beschränke ich mich auf die Beschreibung der zwei Verfahren, die in dieser Arbeit zur Anwendung kommen. Weiterführende Literatur ist unter [32] und [18] zu finden.

[4] Wird in der Literatur auch als Boostrap bezeichnet.

[5] Der verwendete oLVQ1-Klassifikator wird mit $M_{tr} = 80\%, M_{te} = 20\%$ und der SVM-Klassifikator nach dem Leave-one-out Verfahren validiert.

– Wahrscheinlichkeit der Klasse i ist:

$$K = \sum_{i=1}^{c} q_i \frac{k_i}{m_i} \tag{4.1}$$

Die Reklassifikationsrate errechnet sich analog dazu aus den Klassifikationen der Trainingsmenge M_tr. Sie beschreibt die Anpassungsfähigkeit des Klassifikators an die Trainingsmenge und lässt damit Rückschlüsse darauf zu, wie kompakt die einzelnen Klassen im Merkmalsraum liegen und wie gut diese separabel sind.

4.2 Lineare Diskriminanzanalyse

Die lineare Diskriminanzanalyse (LDA) zählt zu den überwachten Lernverfahren und wurde von Fisher bereits 1936 entwickelt. Damit ist es möglich, eine 2-Klassen Problematik linear zu trennen. Das Prinzip der LDA basiert auf der Transformation des Merkmalsraumes durch eine lineare Abbildung. Dabei soll die Kovarianz der Vektoren innerhalb einer Klasse minimiert und die Kovarianz zwischen den Klassen maximiert werden [19, S. 67f]. Das soll an dieser Stelle nicht weiter vertieft werden, eine ausführliche Beschreibung findet sich bei Hentschel [19]. Die Grundannahme des LDA-Klassifikators wird allerdings im Kapitel zu der Support Vector Maschine wieder aufgegriffen werden.

4.3 Learning Vector Quantization (LVQ)

Die lernende Vektor-Quantisierung(LVQ) ist ebenfalls ein überwacht lernender Klassifikator. Das heißt, die Klasseninformation für jeden Merkmalsvektor muss bekannt sein, da sie in der Trainingsphase mit berücksichtigt wird. Er besteht aus einem einschichtigen neuronalen Netz, dem oftmals eine Schicht Eingabeneuronen vorgeschaltet wird (siehe Abb: 4.1). Die aktiven Neuronen werden häufig als Kohonen-Neuronen bezeichnet. Eine ungeordnete Menge an sogenannten Codebook-Vektoren soll den Eingaberaum $X_1, X_2, \cdots, X_n$ möglichst optimal abdecken [23] [39], andernfalls entsteht eine hohe Anzahl toter Neuronen. Zuviele tote Neuronen bewirken ggf. nicht genügend verwendbare Codebook-Vektoren, so daß die Klassifikationsrate sinkt.

4.3.1 oLVQ1-Klassifikator

Das oLVQ1-Verfahren (optimized learning vector quantization) basiert auf dem LVQ1-Klassifikator, bei dem der Eingabevektor X_n mit jedem Gewichtsvektor (Codebook-Vektor) W_j eines jeden Kohonen-Neurons j verglichen wird. Das Gewinnneuron W_c ist das Neuron mit dem größten Ähnlichkeitsmaß, das meist durch die Vektordifferenz $X - W_j$ ausgedrückt wird.

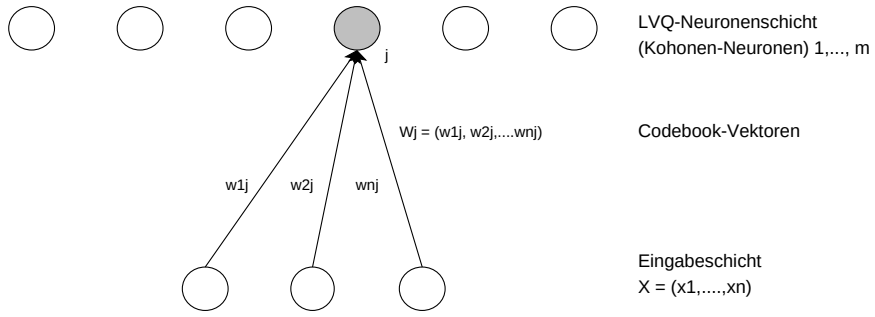


Abbildung 4.1: Netzstruktur der LVQ

Im Prinzip ist der LVQ-Algorithmus ein „Nearest-Neighbourg"-Klassifikator, der zusätzlich ein Lernverfahren anwendet, um die Gewichtsvektoren W_j zu modifizieren [39, S. 172].

Im Vergleich zum LVQ1 ist oLVQ1 ein auf schnelle Konvergenz optimierter Klassifikator. Hierzu wird jedem Gewichsvektor $W_c(t)$ eine eigene Lernrate $\alpha_c(t)$ zugewiesen, während beim LVQ1 eine Lernrate für alle Neuronen verwendet wird [39, S. 172]. Dadurch erhält man das in (4.2) beschriebene Lernverfahren.

$$W_c(t+1) = \begin{cases} W_c(t) + \alpha_c(t)[X/t) - W_c(t)] & \text{falls Klasse}(W_c) \equiv \text{Klasse}(X) \\ W_c(t) - \alpha_c(t)[X/t) - W_c(t)] & \text{falls Klasse}(W_c) \neq \text{Klasse}(X) \end{cases}$$

$$W_j(t+1) = W_j(t) \quad \text{falls} \quad j \neq c \tag{4.2}$$

Dabei muss noch sichergestellt werden, daß die Lernfaktoren $\alpha_c(t)$ sich in den Grenzen $0 < \alpha_c(t) <= 1$ bewegen, da sonst die Anpassung des Codebook-Vektors im Lernvorgang nicht mehr korrekt gegeben ist. Nach [39, S. 176] ergibt sich die optimale Änderung der Lernfaktoren (Lernschrittweite $\alpha_c(t)$) für den Lernschritt $t$wie folgt:

$$\alpha_c(t) = \frac{\alpha_c(t-1)}{1+s(t)\alpha_c(t-1)}$$

$$\text{wobei gilt} \quad s(t) = \begin{cases} +1 & \text{falls Klassifikation korrekt} \\ -1 & \text{falls Klassifikation nicht korrekt} \end{cases} \tag{4.3}$$

Die Lernschrittweite wird im Verlaufe des Trainings immer weiter monoton verringert, um so das optimale Gewicht einzustellen. Als Startwert für $\alpha_c(t)$ schlägt Zell den Wert 0,3 vor, da größere Werte einen schnelleren Lernfortschritt am Anfang bewirken [39, S. 177].

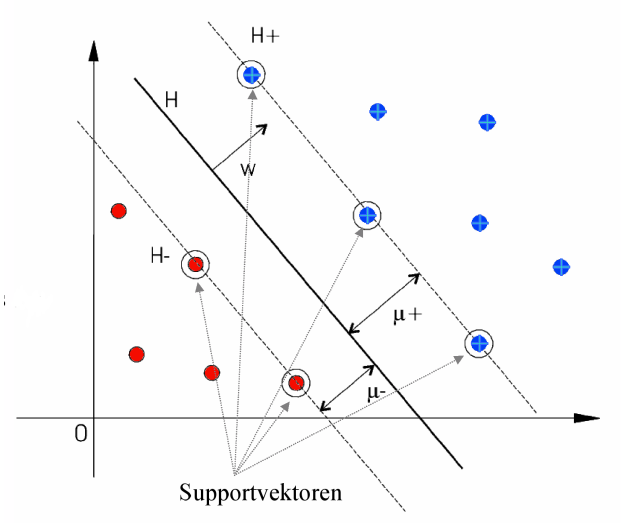


Abbildung 4.2: Schematische Darstellung der Berechnung der Trennfunktion der linearen SVM nach Sommer, D. [35]

4.4 Support-Vector-Maschinen (SVM)

Auch die Support-Vektor-Maschine gehören zu den überwachten Lernverfahren, die im einfachsten Fall mit dem Maximum-Margin-Klassifikator arbeitet. Damit sind nur linear separable Problemstellungen lösbar, man spricht auch von einer Linearen SVM. In der Trainingsphase wird die beschränkte Anzahl an Eingabevektoren gesucht, die entscheidend zur Klassifikation beitragen. Diese werden als sogenannten Support-Vektoren (Abb. 4.1) bezeichnet. Die optimale Trennfunktion (vgl. Gleichung 4.4) wird durch eine Hyperebene repräsentiert, die zu allen Supportvektoren den maximalen Abstand hat (Maximum-Margin-Klassifikator) [4, S. 9-21] [5].

$$f(x) = \sum_{i=1}^{n} w_i x_+ b \tag{4.4}$$

Um auch nicht-lineare Problemstellungen lösen zu können, wurde die SVM weiter entwickelt und eine Unschärfekonstante eingeführt, die die maximale Zahl falsch klassifizierter Merkmalsvektoren beeinflusst. Diese SVM wird auch als Soft-Margin-SVM bezeichnet. Nach Christianini und Shawe-Taylor besteht das Ziel darin, bei minimierter Anzahl an Fehlklassifikationen den Margin zu maximieren [4, S. 103]. Eine weitere Verbesserung stellt die Anwendung einer Kern-Funktion[6] dar, durch die Merkmalsvektoren durch eine nicht-lineare, feste Funktion in einen linear separablen Merkmalsraum über-

[6] zum Beispiel Polynom-, Gauss- oder Sigmoid-Kerne

führt werden. Dort kann mit einer linearen SVM eine eindeutige Trennfunktion errechnet werden, die zu einer erfolgreichen Klassifikation führt. Ausführliche Beschreibungen und der mathematische Hintergrund findet sich bei Christianini und Shawe-Taylor [4]. Durch die Berechnung einer Kernfunktion liefert eine lineare SVM immer ein eindeutiges Ergebnis und bietet durch die freie Wahl aller Parameter und der Kernfunktion die Möglichkeit, ein Problem optimal zu lösen.

Kapitel 5

Ergebnisse der Mustererkennung

In diesem Kapitel sollen die durchgeführten Klassifikationstests und deren Ergebnisse vorgestellt werden. Dazu wurden die Merkmalsvektoren in unterschiedlichen Konstellationen mit den in Kapitel 4 beschriebenen Klassifikationsverfahren kombiniert und die Klassifikationsraten ermittelt. Die hochspezifischen Sprechdaten, die in dieser Arbeit untersucht wurden, sind bisher nicht mit automatischen Klassifikationsmethoden analysiert worden. Auch nach ausgiebiger Recherche in der Literatur und bei Fachleuten ließ sich kein Hinweis finden. Deshalb liegt es nahe, mit erprobten Standardverfahren der Mustererkennung, die auf anderen Forschungsgebieten (vgl. Mikroschlaferkennung [25]) sehr erfolgreich angewendet werden, zu arbeiten. Damit sollen die Fragen, die sich zu Beginn der Arbeit gestellt haben, möglichst umfassend beantwortet werden. Allerdings lässt sich bereits jetzt sagen, daß das nur ein kleiner Teil aller Analysemöglichkeiten ist, der auf die vorliegenden Daten angewendet werden kann. Schon bei der Entwicklung der Algorithmen für die Merkmalsextraktion sind viele Ideen entstanden, die dringend einer weiteren Bearbeitung bedürfen.

5.1 Allgemeines

Um überhaupt eine Aussage darüber treffen zu können, ob die in der Merkmalsextraktion gewonnen Daten irgendeine Aussagekraft besitzen und eine Klassifikation erlauben, wurden die beschriebenen Klassifikatoren mit willkürlich gewählten Parametern mit der Datenmenge trainiert. Für einen ersten Test ist zusätzlich zum oLVQ1 und SVM-Klassifikator noch mit der linearen Diskriminanzanalyse(LDA) nach Fisher getestet worden. Nachdem Klarheit darüber herrscht, ob überhaupt eine Klassifikation möglich ist, müssen die optimalen Parameter der Klassifikatoren ermittelt werden.

Im nächsten Schritt soll dann der Einfluss der verschiedenen Merkmale

auf das Klassifikationsergebnis untersucht werden. Im Auschlussverfahren werden nur Merkmale bestimmter Extraktionsverfahren in die Merkmalsvektoren aufgenommen und von diesen anschliessend die Klassifikationsrate ermittelt. Die Versuchsreihen wurden komplett mit dem oLVQ1-Klassifikator absolviert, da dieser sehr schnell konvergiert und relativ wenig CPU-Zeit benötigt. Die Testreihen wurden jeweils über die Fenstergröße als Parameter der Merkmalsextraktion durchgeführt, um das Verhalten der einzelnen Merkmale zu erfassen. Dazu wurde die Fenstergröße T_F von 5ms bis 500ms variiert.

Für eine ausreichend genaue Untersuchung ist es in jedem Fall erforderlich, weitere Parameter der Merkmalsextraktion zu variieren und den optimalen Wert des Parameters zu finden. Ausserdem lassen sich auf Grundlage der durchgeführten Berechnungen mit anderen Verfahren weitere Merkmale ermitteln. Das hätte jedoch den Rahmen dieser Arbeit gesprengt, so daß die Analysen als erste Versuche angesehen werden müssen, die eine eher grobe Richtung vorgeben oder ausschliessen.

Ausserdem kommt erschwerend hinzu, daß die Klassifikatoren, insbesondere der SVM-Klassifikator äußerst rechenintensiv sind. Dadurch wird viel CPU-Zeit benötigt, um ein Ergebnis zu erhalten. Der Vorteil der hier verwendeten Implementation liegt in seiner Modularität und Parallelisierbarkeit auf einem Rechencluster. Die durchgeführten Berechnungen konnten fast ausschlieslich auf dem Client-Server basierten Cluster-System der FH Schmalkalden erledigt werden.

Ist eine Klassifikation prinzipiell möglich, ist es von sehr großem Interesse, eine Aussage über die Generalisierungsfähigkeit der Analyse zu treffen. Dazu übergibt man zum Testen des Klassifikators die Daten eines Probanden, der dem Klassifikator vorher nicht bekannt war. Zu dieser Validierung, wie der Klassifikator mit unbekannten Probanden umgeht, wird der SVM-Klassifikator verwendet werden. Dieser erzeugt in der Regel genauere Resultate als der oLVQ1 zu leisten vermag, benötigt aber wesentlich mehr Rechenleistung [36]. Hier ist also eine höhere Klassifikationsrate zu erwarten. Damit lässt sich dann auch eine Aussage darüber treffen, welche Methode des maschinellen Lernens sich besser zur Klassifikation der vorliegenden Problemstellung eignet.

5.2 Lineare Diskiminanzanalyse

Ein Testlauf mit der linearen Diskriminanzanalyse(LDA) nach Fisher brachte kein brauchbares Ergebnis (siehe Tabelle 5.2). Der LDA-Klassifikator erkennt nur Problemstellungen gut, die linear trennbar sind. Für diese Problemstellung zeigt sich der Klassifikator überfordert, also liegt hier keine triviale Problemstellungen vor [36]. Eine Klassifikationsrate bei nahezu genau 50% zeigt dies deutlich (Tab. 5.2). Bei dem hinreichend komplexen Merk-

Tabelle 5.1: Fenstergr. 30ms getestet mit Fisher-LDA Klassifikator

	Klassifikation	Reklassifikation
Median	49.9656%	50.0086 %
Std-Error	0.48868%	0.12217%

malsraum, wie ihn diese Daten auszuweisen scheinen, erstaunt ein solches Ergebnis nicht.

5.3 oLVQ1: Anzahl der Neuronen und Test der Fenstergröße

Eine Testreihe mit der Segmentlänge von 25ms wurde durchgeführt, um die optimale Anzahl an Neuronen, die für den oLVQ1-Klassifikator nötig sind, zu bestimmen. Dabei wurden alle Merkmalsvektoren herangezogen und von zwei bis 500 Neuronen mit einer Schrittweite von 2 die Klassifikationsrate bestimmt. Aufgrund der hohen Anzahl an Merkmalsvektoren ist bei einer geringen Neuronenzahl eine hohe Steigerung der Lernrate zu erkennen, die sich im Verlauf schnell abflacht. Das lässt auf einen sehr kompakten Merkmalsraum mit einer gewissen Überdeckung schliessen. Da der oLVQ1-Klassifikator nicht eineindeutig ist und die Eingabeneuronen bei jedem Versuch zufällig angeordnet werden, führt man jeden Versuch mehrfach durch. So erhält man die mittlere Klassifikationsrate und deren Standardabweichung, die ein weiteres Indiz für einen recht homogenen Merkmalsraum ist. Da sie sich konstant im Bereich um +-2% bewegt, kann der Lernvorgang als sehr stabil angesehen werden.

Auch ein Over-Learning ist im getesteten Bereich nicht zu verzeichnen, dies würde sich durch ein Abfallen der Klassifikationsrate bei höherer Neuronenzahl bemerkbar machen. Auch das starke Ansteigen der Reklassifikationsrate bei wenigen Neuronen deutet daraufhin, daß die zwei Klassen bis auf einen kleinen Bereich der Überdeckung sehr gut trennbar sind.

Ab etwa 150 Neuronen pendelt sich die Lernrate bei etwa 75% ein und bleibt im folgenden sehr stabil mit einem leicht positiven Trend. Aus diesem Grund wurden alle weiteren Berechnungen mit 400 Neuronen durchgeführt, die mittlere Klassifikationsrate liegt in dem Fall bei 75,8% (Abb. 5.1).

5.4 Optimierung der Merkmalsextraktion

Um die Leistungsfähigkeit der Merkmalsextraktion zu überprüfen, wird der oLVQ1-Klassifikator über die variierende Fenstergröße getestet. Da bei Ana-

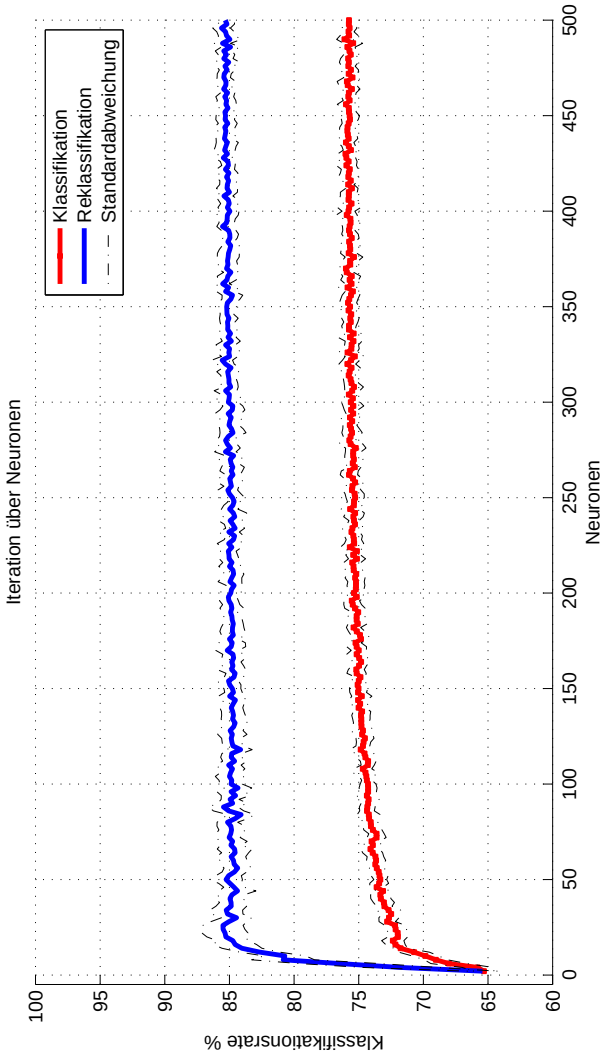


Abbildung 5.1: oLVQ1-Screening zur Bestimmung der optimalen Neuronenzahl: Fenstergröße: 25ms, alle Merkmale getestet, 40 Wiederhohlungen für jeden Test von 2-500 Neuronen, Schrittweite 2

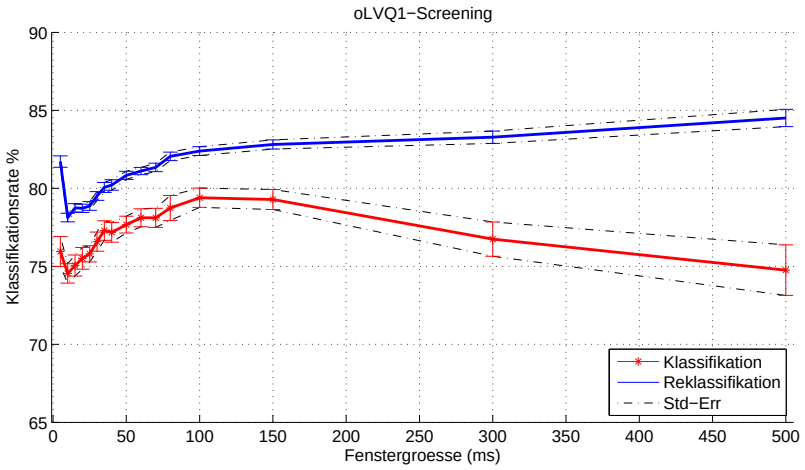


Abbildung 5.2: oLVQ1 – Klassifikations- und Reklassifikationsrate für alle Merkmale ermittelt. Für jede Fenstergröße von 5 ms bis 500 ms mit 400 Neuronen über 40 Durchläufe ermittelt.

lysen im Sprachbereich generell mit kleinen Fenstergrößen gearbeitet wird, sollte auch in diesem Fall die Klassifikationsrate bei einer kleinen Fenstergröße am höchsten sein. In der Literatur werden für Sprachsignalverarbeitung Fenstergrößen von etwa 10ms bis 80ms genannt. Um trotzdem alle Eventualitäten auszuschliessen, wurden alle Daten über die schon angesprochene logarithmierte Skala von 5ms bis 500ms verwendet.

Da es eine Vielzahl von möglichen Merkmalskombinationen und Parametern gibt, die die Gesamtklassifikation jeweils unterschiedlich beeinflussen, wäre es nötig, jede mögliche Kombination zu untersuchen. Insgesamt liegen 181 Merkmale je Vektor vor, um diese alle zu prüfen, wären immense Rechenkapazitäten nötig. Sämtliche Berechnungen konnten auf dem Cluster der Fachhochschule durchgeführt werden, wobei nicht immer alle Rechner zur Verfügung standen. Deshalb beschränkt sich diese Arbeit auf eine erste Auswahl, um die wichtigsten Merkmale zu prüfen.

Die Ergebnisse des oLVQ1-Klassifikators mit allen Merkmalen in Abb. 5.2 zeigen sehr deutlich, daß die Erwartung hoher Klassifikationsraten bei kleinen Fenstergrößen nicht zutreffend ist. Ein Minumim stellt sich bei einer Fenstergröße von 10ms ein, die Klassifikationsrate liegt bei 74,5%. Ab da ist ein relativ konstanter Anstieg auf eine Klassifikationsrate von 79,4% bei Segmentlängen von 100ms zu verzeichnen. Damit bewegt sich die optimale Fenstergröße oberhalb des in der Literatur für Sprachverarbeitung empfohlenen Wertes von bis zu 80ms [29, S. 266]. Die Ursache dafür liegt mit hoher Wahrscheinlichkeit in den hochgradig instationären Ausgangssignalen, die eine deutlich höhere Varianz aufweisen, als Sprachaufnahmen eines sogenannten Normal-Sprechers. Ab 100ms fällt die Klassifikationsrate

wieder ab, was zu erwarten war. Die Merkmalsermittlung mit einem derart breiten Signalfenster führt dazu, daß feine Ausprägungen des Signals nicht mehr aufgelöst werden. Relevante Merkmale werden somit nicht mehr erkannt.

Das Zufallsverhalten des oLVQ1-Klassifikators wird durch eine mehrfache Wiederholung des Lernens und anschliessendem Test ermittelt. Die Standardabweichung aller Tests lässt Rückschlüsse auf diese Größe zu. In diesem Test beträgt die Standardabweichung der Klassifikationsrate nie mehr als 1,6%, was auf eine sehr stabile Klassifikation schliessen lässt.

Mit einer Klassifikationsrate von nahezu 80% wird eindeutig gezeigt, daß die in dieser Arbeit durch Merkmalsextraktion generierten Merkmale die Klasseninformation einer Aufnahme mit Einschränkungen beschreiben. Damit kann die anfangs gestellte Frage nach Informationen über den Therapieerfolg eindeutig beantwortet werden[1]. Mit weiterführenden Optimierungen jedes einzelnen Prozessschrittes der im Laufe der Arbeit entwickelten Mustererkennungskette ist es sicher möglich, sich einer Klassifikationsrate von über 90% anzunähern.

Um die Frage zu beantworten, welche Merkmale zu einer guten Klassifikation beitragen, werden die Merkmalsvektoren nach den 3 Methoden der Merkmalsextraktion aufgespalten. Das ergibt drei Merkmalsräume, in denen die Vektoren jeweils nur alle Zeitbereichsmerkmale, alle Merkmale aus der Wavelet-Transformation oder alle Merkmale aus der Cepstral-Transformation beinhalten.

Eine Untersuchung des Trends einer Halbsilbe mittels Regressionsverfahren brachte kein aussagekräftiges Ergebnis. Hinzu kommt, daß diese Information mit Hilfe der Wavelet-Dekomposition in den hohen Approximations-Stufen wesentlich genauer beschrieben wird. Deshalb wurde dieser Ansatz, der im Anfangsstadium der Arbeit genauer untersucht wurde, als unzureichend verworfen.

5.4.1 Merkmale im Zeitbereich

Generell kann man sagen, daß in der Sprachsignalverarbeitung die Zeitbereichsmerkmale eher eine Randbedeutung haben. Die wichtigen und die Entscheidung maßgeblich beeinflussenden Merkmale sind ausschlieslich im Frequenzbereich angesiedelt [29, S. 265]. Allerdings werden bei Sprachverarbeitung bevorzugt Merkmale benötigt, die möglichst Sprecher-unabhängig sind. Zur guten Klassifikation von den vorliegenden Therapieaufnahmen wird aber eine gewisse Sprecherabhängigkeit benötigt. Das stellt eine sehr diffizile Forderung dar: Einerseits müssen auch hier Merkmale sprecherunabhängig sein, um eine gute Generalisierung der Algorithmen zu erreichen und auf der anderen Seite soll genau die Sprechweise klassifiziert werden. Damit einher

[1] siehe Abschnitt 1.2: Lässt eine Aufnahme eines gesprochenen Satzes Rückschlüsse auf den Therapieerfolg zu?

geht auch eine gewisse Abhängigkeitsgrad von individuellen Sprechmustern. Die Generalisierung der Klassifikatoren wird später noch untersucht werden, wobei diese Arbeit nur einen geringen Ausschnitt dieser speziellen Problemstellung darstellen kann.

Bei den hier vorliegenden Daten stellt das Pausenverhalten, also ein Merkmal im Zeitbereich, ein entscheidendes Kriterium für die Klassifikation dar. Die hohe Individualität wurde bereits in der Anfangsbetrachtung in Abschnitt 2.6.3 gezeigt, aus diesem Grunde ist nur eine moderate Klassifikationsrate zu erwarten.

Es ergibt sich mit 80,1% sogar eine geringfügig höhere Klassifikationsrate, als wenn alle Merkmale zur Klassifikation herangezogen werden (Abb. 5.3). Diese Klassifikationsrate wird bei einer Segmentgröße von 35ms erreicht, was der Forderung einer kurzen Segmentlänge zur Verarbeitung von Sprachdaten nachkommt. Allerdings ist das Maximum in dem Bereich nicht eindeutig. Zwischen 5ms und 80ms Segmentlänge schwankt die Klassifikationsrate beständig zwischen etwa 76% und dem Maximum. Das lässt auf gewisse Streuungen in der Merkmalsextraktion schliessen, die sich vor allem im Zeitbereich bemerkbar machen. Bei anderen Merkmalen tritt dieser Effekt nicht auf. Das zeigen die Grafiken zur Wavelet- und Cepstralanalyse im folgenden Abschnitt.

5.4.2 Wavelet-Komponenten

Die Wavelet-Transformation wurde aufgrund ihrer robusten Eigenschaften bei hochgradig instationären Signalen für diese Arbeit ausgewählt. Hierfür wurde ein Approximations-Level von 8 Dekompositionen gewählt, so

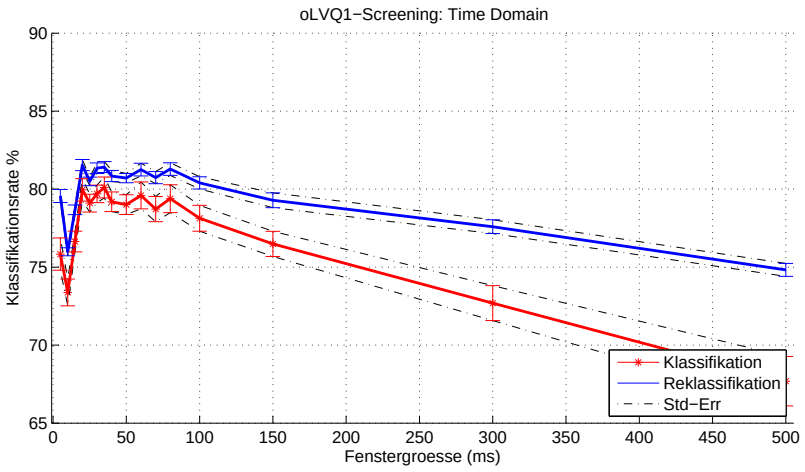


Abbildung 5.3: oLVQ1 – Klassifikationsrate über alle Zeitbereichsmerkmale, 400 Neuronen, von 5 ms bis 500 ms mit je 40 Iterationen

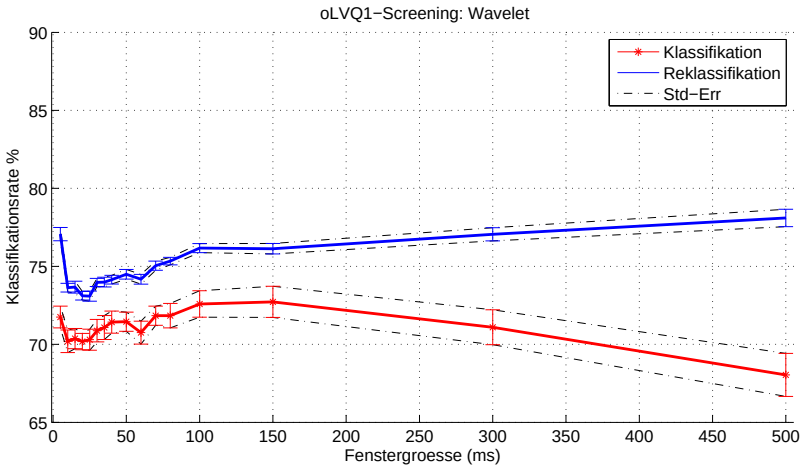


Abbildung 5.4: oLVQ1 –Klassifikationsrate über alle Merkmale der Wavelet-Transformation, 400 Neuronen, von 5 ms bis 500 ms mit je 40 Iterationen

daß dem Klassifikator die Energie, Entropie und logarithmierte Energie von 8 Approximationen und 8 Detail-Signalkomponenten als Eingabe zur Verfügung stehen. Durch ihre Eigenschaften, gerade feine Details im Signal gut hervorzuheben, ist zu erwarten, daß Merkmale aus der Wavelet-Transformation einen entscheidenden Beitrag zur Klassifikation leisten.

Umso mehr erstaunt es, daß die Klassifikationsrate nie über 72,6% steigt. Diese ist erst bei einer Segmentlänge von 100ms erreicht (Abb. 5.4), wobei generell die Klassifikationsrate mit diesen Merkmalen am stabilsten zur variierten Fenstergröße bleibt. Das deutet auf robuste Merkmale hin, deren Potential noch nicht voll ausgeschöpft wurde. Eine mögliche Erklärung für die hohe Relevanz großer Segmentlängen könnte in der stark gedehnten Sprechweise der „Nach Therapie“-Aufnahmen zu finden sein. Eine geringe Segmentlänge erreicht zwar eine hohe Auflösung im Zeitbereich, geht aber mit einer sinkenden Auflösung im Frequenzbereich einher. Durch gedehnte Sprache sinkt also der Informationsgehalt bei gleichbleibender Segmentlänge. Das kann wiederum zu einer Verschlechterung der spektralen Merkmalsextraktion führen. Dieser Sachverhalt bedarf in jedem Fall einer weiteren Untersuchung, um die Klassifikationsrate weiter zu steigern.

5.4.3 Cepstral-Komponenten

Die Cepstral-Transformation stellt in gewisser Hinsicht den Gegenpart zur Wavelet-Transformation dar, da hier die Nachteile der diskreten Fourier-Transformation unter Umständen voll zum Tragen kommen können. Durch die hochgradig instationären Signale kann eine reine Fourier-basierte Ana-

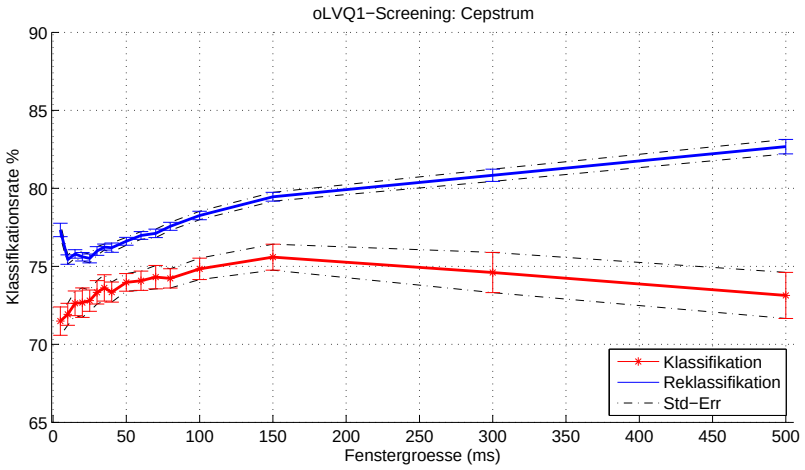


Abbildung 5.5: oLVQ1 –Klassifikationsrate über alle Merkmale der Cepstral-Transformation aus allen Koeffizienten, 400 Neuronen, von 5 ms bis 500 ms mit je 40 Iterationen

lyse schnell an ihre Grenzen stoßen (siehe Abschnitt 3.5) und dementsprechend wenig zur Klassifikation beitragen. Weiterhin sind keine Untersuchungen bekannt, inwieweit sich Stottereereignisse durch das Vorhandensein beziehungsweise das Ausbleiben von harmonischen Grundschwingungen (siehe Abschnitt 3.6.4) im Sprachsignal manifestieren.

Das Verfahren der Cepstral-Tranformation basiert auf der diskreten Fourier-Transformation und ist somit nicht robust gegen instationäre Signale. Daher ist fraglich, inwieweit dieses Verfahren überhaupt zu einer robusten Klassifikation beitragen kann.

Die Ergebnisse sprechen jedoch eine deutliche Sprache. Für die durchgeführten Klassifikationstests tragen die Merkmale, die mit Hilfe der Cepstral-Transformation gewonnen werden, entscheidend mit zum Ergebnis bei. Die Klassifikationsrate liegt mit maximal 75,6% um 4% über dem Ergebnis, das allein mit Hilfe der Wavelet-Transformation erzielt wurde. Die Standardabweichung zeigt sehr stabile Merkmale. Mit 1,5% Abweichung liegt sie in etwa gleich auf mit der Standardabweichung der Wavelet-Merkmale von 1,4% liegt. Das Interessante ist, daß die Merkmale der Cepstral-Transformation ihre maximale Klassifikationsrate erst bei einer Segmentlänge von 150ms erreichen. Auch hier kann nur die Besonderheit der gedehnten Sprechweise als Erklärung dienen, denn in der Literatur wird die Cepstral-Transformation auch mit teilweise sehr kleinen Segmentlängen von unter 10ms erfolgreich angewendet [7].

5.5 Einfluss einzelner Mel-Cepstral-Komponenten

Angeregt durch den erstaunlich hohen Einfluss der Mel-gefilterten Cepstral-Komponenten auf das Klassifikationsergebnis soll dieser Punkt hier näher beleuchtet werden. Dazu wird der oLVQ1-Klassifikator jeweils mit Merkmalsvektoren trainiert, deren einzigster Bestandteil jeweils eine Cepstral-Komponente ist. Durch die 13-stufige Mel-Filterbank entstehen also 13 einzelne Klassifikationstests, die den Einfluss einer jeden Cepstral-Komponente zeigen.

Im Sprachbereich sind die Merkmale, die für einzelne Phoneme charakteristisch sind, eher im unteren Frequenzbereich angesiedelt. Filtert man aus einer Sprachaufnahme also die hohen Frequenzen heraus, wird man den Sprecher trotzdem noch verstehen, wenn auch etwas verzerrt. Filtert man hingegen alle tieferen Frequenzen heraus, sind die artikulierten Laute und Worte kaum noch zu verstehen. Überträgt man diese Betrachtung auf die einzelnen Mel-gefilterten Cepstral-Komponenten, sollten also die niedrigen Komponenten am ehesten zur Klassifikation beitragen.

Ob diese Tatsache auch bei einer allgemeinen Untersuchung, die den komplexen Vorgang der Sprach- und Worterkennung ausser Acht lässt, übertragbar ist, soll in diesem Abschnitt untersucht werden.

Betrachtet man die 13 Ergebnisse der oLVQ1-Klassifikation für jede Komponente, so ist schnell ersichtlich, daß einige wenige Komponenten sehr viel zur Klassifikation beitragen und Andere nur wenig aussagekräftige Informationen beinhalten. Der Verlauf der Klassifikationsrate über die Fenstergröße ändert sich nur in sehr geringem Maße und weist die selbe Charakteristik auf, wie die mit Hilfe aller Cepstral-Komponenten ermittelte Klassifikationsrate (Abb. 5.5). Die Ergebnisse der drei Cepstral-Komponenten, die das Verhalten am besten charakterisieren, zeigen die Abbildungen 5.6, 5.7 und 5.8[2]. Wie zu erwarten, liegt der Hauptgehalt der Informationen in der ersten Cepstralkomponente, deren Bandpassfilter der Mel-Filterbank von 0Hz bis 266Hz reicht [34, S. 29]. Allein die Informationen aus dieser Komponente reichen für eine Klassifikationsrate von nahezu 70%. Die Komponente, die am wenigsten zur Klassifikation beiträgt, kommt auf lediglich etwa 62% (Abb. 5.7). Interessant ist auch noch, daß in höheren Komponenten (12 und 13) wieder mehr Informationen zu finden sind (Abb. 5.8).

Abschliessend kann man feststellen, daß sich die Informationen der einzelnen Komponenten sehr ähnlich sind, jede trägt ihren Teil zur Klassifikation bei. Keine Komponente erzeugt Merkmale im Merkmalsvektor, die die Klassifikation negativ beeinflussen. In der Summe erreicht man die beste Klassifikationsrate also mit Hilfe aller Merkmale. Zieht man jetzt noch die Ergebnisse von Zeitbereichs- und Waveletanalyse hinzu, lässt sich auch die

[2]Die Abbildungen der oLVQ1-Tests aller Cepstral-Komponenten finden sich auf der beiliegenden DVD.

dritte Fragestellung dieser Arbeit beantworten[3]. Letztendlich steuern alle Merkmale ihren Teil zur Klassifikation bei, es gibt keine vollständig redundanten Informationen und keine Merkmale, die die Klassifikation stark negativ beeinträchtigen und weggelassen werden müssten.

[3]Welche Merkmale der eingesetzten Sprechtechnik lassen sich als Indiz für flüssiges Sprechen heranziehen?

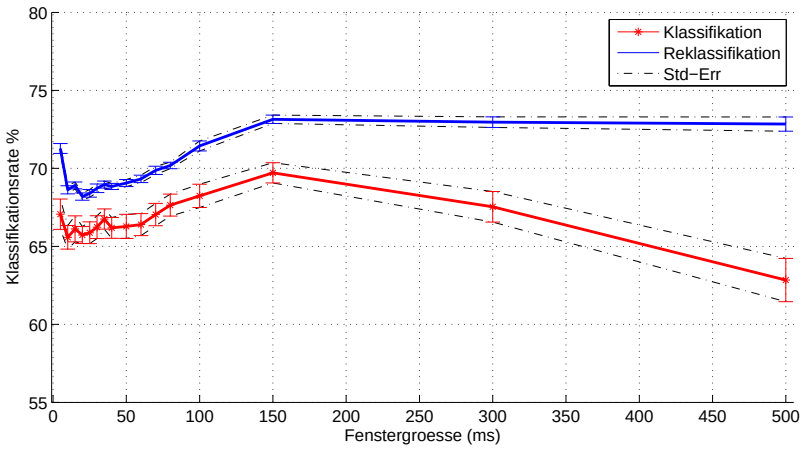


Abbildung 5.6: oLVQ1–Klassifikationsrate der Merkmale des 1. Cepstralkoeffizienten über die Segmentlänge von 5ms bis 500ms mit je 40 Wiederholungen

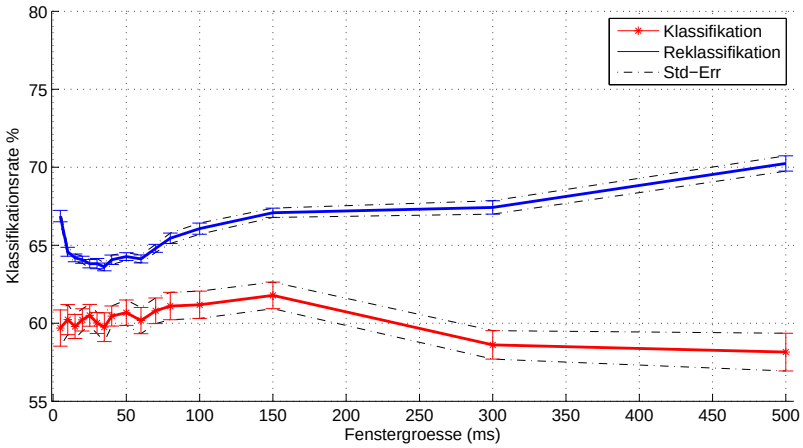


Abbildung 5.7: oLVQ1–Klassifikationsrate der Merkmale des 5. Cepstralkoeffizienten über die Segmentlänge von 5ms bis 500ms mit je 40 Wiederholungen

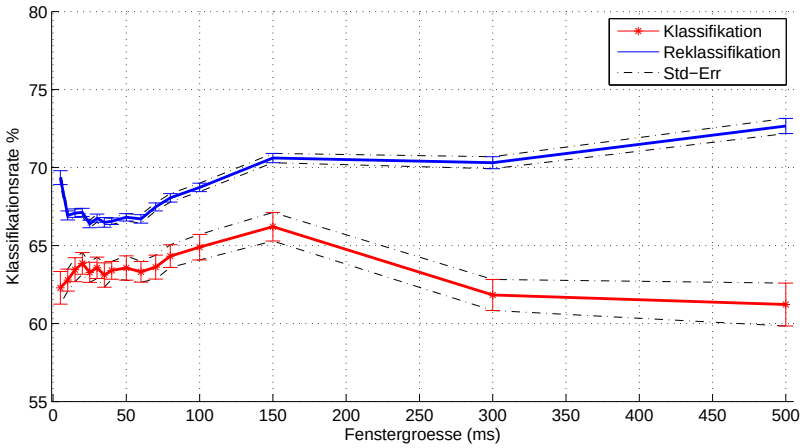


Abbildung 5.8: oLVQ1–Klassifikationsrate der Merkmale des 12. Cepstralkoeffizienten über die Segmentlänge von 5ms bis 500ms mit je 40 Wiederholungen

5.6 Optimale Parameter der Support-Vector-Maschine

Die SVM als eindeutiges Lernverfahren liefert bei gleichen Eingabevektoren immer exakt die gleiche Trennfunktion. Das Ergebnis ist im Gegensatz zum oLVQ1-Klassifikator nicht von einer Zufallsgröße abhängig. Die SVM liefert ein sehr genaues Ergebnis und ist in der Lage, auch Klassen mit starker Überlappung im Merkmalsraum sehr gut zu trennen [36]. Die Klassifikationsrate sollte hier also höher sein, als die mit dem oLVQ1-Klassifikator erreichte. Jedoch steigt der Rechenaufwand quadratisch an, je mehr Eingabevektoren verwendet werden und je größer der Merkmalsraum ist. Gleiche Merkmalsvektoren haben immer die gleiche Lage im Merkmalsraum, so daß die Trennfunktion nur von den Parametern der SVM abhängig ist. Das ist einmal der Paramter c der das Soft-Margin-Verhalten beeinflusst und der Parameter γ des verwendeten Gauss-Kernels. Zur Bestimmung der optimalen Parametrisierung variiert man jeden Parameter über einen größeren Bereich und bestimmt mittels Kreuzvalidierung die Klassifikationsrate. Hier wurden beide Parameter im Bereich von -2 bis 2 mit der Schrittweite 0.1 variiert und mittels leave-one-out die Gegenprobe durchgeführt. Im Ergebnis erreicht man jedoch mit der SVM auch nur eine Klassifikationsrate von

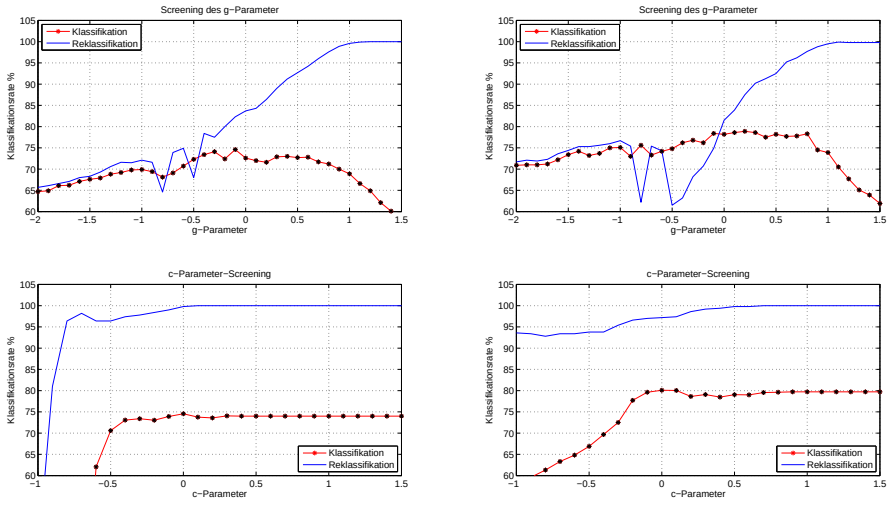


Abbildung 5.9: Screening über γ-Parameter (oben) und c-Parameter (unten) für die Segmentlängen 25ms (links) und 100ms (rechts). Bei einer Segmentierung mit 25ms hat die Merkmalsextraktion die meisten Merkmalsvektoren ausgegeben (ca. 23500), während bei 100ms der oLVQ1-Klassifikator das beste Ergebnis erzielt hat. Alle SVM-Screenings wurden mit einer Gauss-Kernel-Funktion durchgeführt.

80,5% über alle Merkmale. Diese erzielt man auch bei einer Segmentgröße von 100ms, während sich als Optimum beider Parameter $\gamma = 0,2$ und $c = 0$ ergibt (siehe Abb. 5.9). Das stellt im Vergleich zum oLVQ1-Klassifikator (79,4%) einen derart minimalen Anstieg dar, daß man gar nicht von einer Verbesserung reden kann.

Diese praktisch nicht vorhandene Steigerung zum oLVQ1-Klassifikator deutet auf einen sehr kompakten Merkmalsraum hin, in dem sich die Klassen nicht ohne weiteres trennen lassen. Auch eine Änderung der Skalierung (logarithmiert, exponentiell) brachte keine Verbesserung. Diese Eigenschaften des Merkmalsraumes lassen sich mit einem sehr inhomogenen Probandenfeld erklären.

Wie in den Anfangsbetrachtungen schon ausgeführt, spiegelt die Auswahl der Klienten für diese Studie möglichst repräsentativ die gesamte Bandbreite an Störungsbildern wieder, die in der KST therapiert werden. So gibt es beispielsweise Klienten, die trotz Stotter-Ereigniss in der Lage sind, mehr oder weniger flüssig zu sprechen[4]. Ein Klient mit sehr massiver Symptomatik hingegen erzielt in der Therapie einen starken persönlichen Fortschritt, erreicht aber nicht das Flüssigkeits-Niveau dieser Klienten.

Eine derart hohe Individualität der Probanden stellt eine große Herausforderung für die Klassifikatoren dar. Auch in anderen Bereichen, so zum Beispiel die Mikroschlafdetektion bei Fahrermüdigkeit, ist diese hochgradige Individualität festgestellt worden [15].

5.7 Validierung der Klienten

Eine wichtige Information ist, den Einfluss eines jeden Probanden auf das Klassifikationsergebnisses zu kennen. Das lässt ausserdem Rückschlüsse auf die Generalisierungsfähigkeit der trainierten Klassifikatoren für unbekannte Probanden zu. Um diese Informationen zu erhalten wurde der SVM-Klassifikator wegen seiner prinzipiell besseren Robustheit ausgewählt, um ein Leave-one-out über alle 20 Probanden durchzuführen.

Der Datensatz mit einer Segmentlänge von 100ms wurde derart geteilt, daß die Trainingsmenge alle Merkmalsvektoren von 19 Klienten beinhaltet, während alle Merkmalsvektoren des 20ten Klienten die Testmenge darstellen. Bei derart großen Datensätzen, insgesamt sind etwa 17600 Merkmalsvektoren in beiden Mengen enthalten, benötigt die SVM mehr Rechenleistung, als zur Verfügung stand. Die Entscheidung fiel auf eine reduzierte Menge, die im Verhältnis von 80 zu 20 für die Teilmengen T_{Tr} und T_{Te} gebildet wird. Per zufälliger Auswahl werden also aus der Trainingsmenge 1000 Vektoren und aus der Testmenge 250 Vektoren entnommen und mit der SVM klassifiziert. Die SVM wird dabei auf die ermittelten Optima der Parameter

[4] Zum Beispiel Klient K011 und K014, siehe Anhang A.3

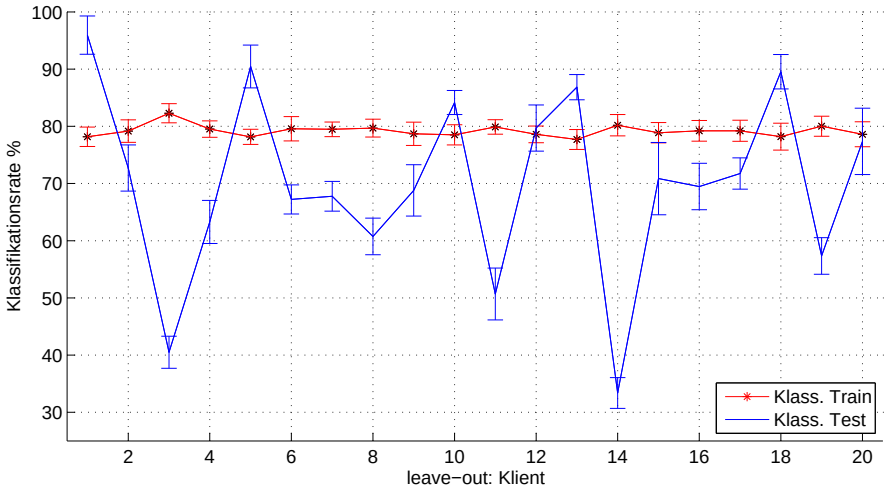


Abbildung 5.10: Leave-one-out über alle 20 Klienten und Bestimmung der Klassifikationsrate mit SVM-Klassifikator und 40 Durchläufe. Der in x-Richtung aufgetragene Klient stellt die Testmenge.

γ und c eingestellt. Um eine eventuelle Streuung in den Daten zu berücksichtigen, wird dieser Vorgang 40 mal wiederholt. Das Ergebnis zeigt, daß es einige wenige Klienten gibt, deren Daten entscheidend zur Generalisierung des Klassifikators beitragen (siehe Abb. 5.10). Das äussert sich in einem massiven Einbruch in der Klassifikationsrate der Testdaten, so wie es bei Klient K003, K011, K014 und K019 der Fall ist. Gleichzeitig ist ein leichtes Maximum in der Klassifikationsrate der Trainingsdaten zu erkennen. Das zeigt, daß an dieser Stelle eine etwas stärkere Adaption an die Testdatensätze erfolgt. Insgesamt zeigt die Abbildung 5.10 nochmals sehr schön die hohe Individualität der Probanden.

Vergleicht man die Abbildung mit der Tabelle in Anhang A.3, so lässt sich ein Zusammenhang zwischen Therapiefortschritt und Einfluss des Klienten auf die Generalisierungsfähigkeit des Klassifikators feststellen. Die Tabelle im Anhang trägt den Therapiefortschritt im Hinblick auf die Sprechflüssigkeit und die korrekte Anwendung der Sprechtechnik anhand einer Standard-Skala für jeden Klient auf. Die Klienten K003, K011, und K014 weisen nach der Therapie eine hohe Sprechflüssigkeit auf, die sich der eines Normalsprechers annähert und wenden gleichzeitig die Sprechtechnik nur in geringem Maße an. Hier kann also mit hoher Wahrscheinlichkeit von nahezu stotterfreier Sprechweise ausgegangen werden.

Die Klienten, bei denen eine hohe Klassifikationsrate der Testdaten zu verzeichnen ist, nämlich K001, K005, K010, K013 und K018, haben alle eine entgegengesetzte Gemeinsamkeit. Hier zeigt die Tabelle A.3 einen relativ

geringen Fortschritt bezüglich der Sprechflüssigkeit, dafür aber eine starke Anwendung der Sprechtechnik. Die Grundannahme der Kasseler Stottertherapie, das sich bei hinreichender Anwendung der Sprechtechnik eine flüssigere Sprechweise als ohne Anwendung der Sprechtechnik einstellt, spiegelt sich in dem Ergebnis hier so nicht wieder. Dieses Faktum bedürfte einer weiteren Untersuchung, zumal die Grundannahme der KST wissenschaftlich ausreichend evaluiert ist [6].

Kapitel 6

Fazit

Zu Beginn dieser Untersuchung war keineswegs abzusehen, ob das Projekt vom Umfang her überhaupt soweit realisierbar war, daß ein abschliessendes Urteil möglich ist. Hinzu kam, daß keinerlei Voruntersuchungen der vorhandenen Daten bekannt war. Der Aufbau einer kompletten Musterverarbeitungskette mit Datenaufbereitung, Merkmalsextraktion und maschinellen Lernverfahren gestaltete sich dementsprechend umfangreich und war mit hohem Einarbeitungsaufwand verbunden. Eine Eigenart derartiger Projekte ist auch, daß aussagekräftige Ergebnisse erst vorliegen, wenn alle Teilkomponenten relativ fehlerfrei arbeiten. Die eigentlichen Probleme lagen wie so oft in Detailfragen, die es zu lösen galt. So mussten möglicherweise relevante Merkmale aufwändig von Hand herausgearbeitet werden, um sich einen Eindruck von ihrer Tauglichkeit zu verschaffen. Nach diesen Vorgaben wurde die komplette Merkmalsextraktion entwickelt und ihre optimalen Parameter ermittelt.

Gerade in der Anfangsphase waren umfangreiche Analysen und Beratungen unter Teilnahme von Fachtherapeuten und Dr. Wolff von Gudenberg nötig, um das nötige Fachwissen über die Erkrankung des Stotterns, die Symptomatik und die therapeutischen Maßnahmen zu gewinnen. Ohne diese Hintergrundinformationen ist eine solche Arbeit mit dem Charakter einer Machbarkeitsstudie nicht durchführbar.

6.1 Diskussion der Ergebnisse

Trotz anfänglicher Skepsis bezüglich der Aussagekraft der Audio-Daten konnte ein Zusammenhang zwischen bekannten Therapiedaten der Klienten und den aus den Audiodaten extrahierten Merkmalen hergestellt werden. Dazu muss noch angemerkt werden, daß es sich bei Anschwing- und Pausenverhalten ausschlieslich um lokale Merkmale der Audiosignale handelt. Den Merkmalen fehlt zudem jeder Bezug zu einer sprachlichen Ebene, so daß die Beziehung der Merkmale zur Semantik der Sprache offen bleibt.

Trotzdem konnte die dafür beachtliche Klassifikationsleistung erzielt werden. Auch wenn diese Korrelation sich nicht vollständig in Grundannahmen der Kasseler Stottertherapie einfügt, so ist es doch nachweislich möglich, therapierelevante Rückschlüsse aus den Audiodaten zu treffen. Obwohl diese Arbeit mehr neue Fragen eröffnet als beantwortet werden konnten, lässt sich ein positives Fazit ziehen. Es wird die prinzipielle Machbarkeit solcher Analysen aufgezeigt und die Basis für weitere, vertiefende Arbeiten in diese Richtung gelegt. Insbesondere im Bereich der Wavelet-Analyse steckt vermutlich ein starkes Optimierungspotential, so daß als Zielstellung für zukünftige Untersuchungen eine Klassifikationsrate von mindestens 90% angestrebt werden kann.

6.2 Zukünftige Untersuchungen

Um die Klassifikationsrate dauerhaft zu verbessern, entstanden schon im Verlauf dieser Arbeit weitere Ideen. Zur weiteren Optimierung der Signalverarbeitung, sollte die Wavelet-Dekomposition überarbeitet und optimiert werden. Wavelets werden im Allgemeinen als sehr geeignet für die Verarbeitung problematischer Signale beschrieben. Umso erstaunlicher ist das relativ schlechte Ergebnis, das in dieser Arbeit mit Wavelets erzielt wurde. An dieser Stelle scheint also noch viel Optimierungspotential verborgen zu sein.

Eine weitere Optimierungsmöglichkeit dürfte in der Halbsilbenerkennung liegen, wenn die Signalbereiche links- und rechtsseitig der Region-of-Interest auf eigene Merkmale hin untersucht werden. Zu guter Letzt führt die normale Weiterentwicklung der EDV-Technik auch in Zukunft zu einem rasanten Zuwachs an Rechenleistung pro Prozessor. Je mehr davon zur Verfügung steht, umso genauere Berechnungen lassen sich sowohl im Bereich der Merkmalsextraktion als auch für die Klassifikatoren durchführen. Das lässt wiederum auf bessere Ergebnisse für die Klassifikation hoffen.

6.3 Ausblick zur Analyse von Therapiedaten

Aus therapeutischer Sicht steht die Analyse der Korrelation der Therapeuteneinschätzung mit der Generalisierungsfähigkeit der Klassifikatoren im Vordergrund der zukünftigen Arbeit. Inwieweit diese Informationen bereits jetzt zur Evaluation des Therapiefortschrittes nutzbar sind, möchte der Autor nicht abschliessend beurteilen. Dies obliegt den entsprechenden Fachleuten. Allerdings ist ein sehr eindeutiger Zusammenhang zwischen den Daten ersichtlich, so daß eine Überführung der Ergebnisse in eine praktische Anwendung sicherlich möglich ist, wenn gleichzeitig die Grundlagen in der Musterkennung und Merkmalsextraktion weiterentwickelt werden.

Anhang A

Kasseler Stottertherapie

A.1 Lesetext für Klienten der KST

Qualitätssicherung der Kasseler Universität

Zu einem neuerlichen Besuch der FDP-Landtagsfraktion an der Universität Gesamthochschule Kassel waren die Fraktionsvorsitzende und hochschulpolitische Sprecherin, Ruth Wagner, und ihr Stellvertreter, Dieter Posch. In Gesprächen mit der Hochschulleitung und Professoren des Fachbereichs E-Technik wurde die aktuelle Situation der Hochschule besprochen. Im Mittelpunkt stand aber die Besichtigung des Neubaus der E-Technik in der Wilhelmshöher Allee. Bei einem Architektenwettbewerb hatte ein Student des Fachbereichs Architektur der GhK den Preis bekommen, der auch tatsächlich umgesetzt worden ist und der eine großzügige, funktionale und kostengünstige Lösung zum Ergebnis hatte. In einer Verbindung von öffentlichen Bauverfahren und privatem Baumanagement sowie einer Begutachtung durch private Architekten wurde ein Verfahren an der Hochschule erstmals realisiert, das die FDP-Landtagsfraktion vor einigen Jahren zur Beschleunigung der Bauverfahren im hessischen Landtag eingebracht hat. So ist ein interessanter, funktionaler und gleichzeitig ästhetisch wirkender und dabei kostengünstiger Bau entstanden.

A.2 Flunatic!-Screenshots

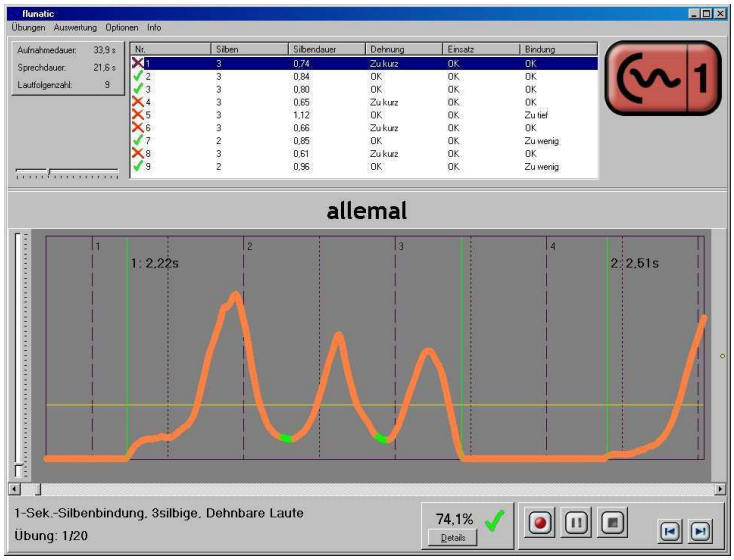


Abbildung A.1: Flunatic!-Bildschirm im Übungsmodus mit Protokoll und direktem Feedback.

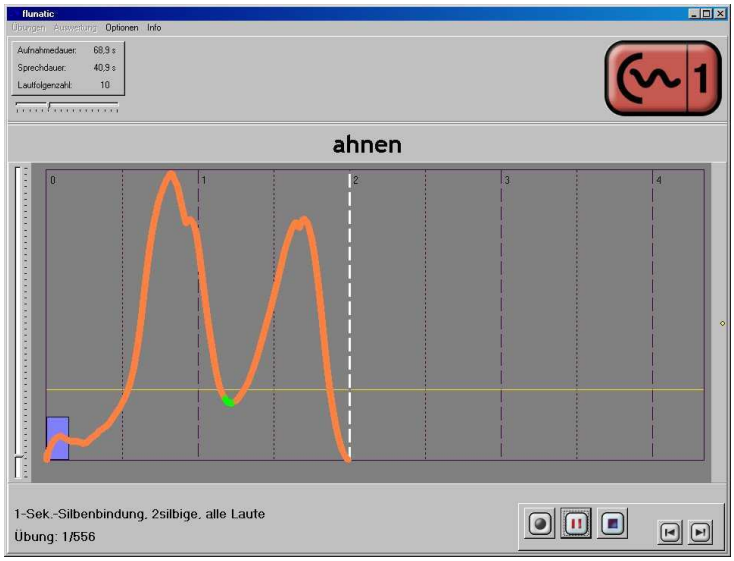


Abbildung A.2: Flunatic!-Bildschirm mit Kontrollbox für weichen Einsatz

A.3 Probandenübersicht

Die Tabelle stellt eine kurze Übersicht der anonymisierten Probandendaten, die für die Analysen verwendet wurden, zur Verfügung. Die Werte in den Spalten Sprechtechnik und Flüssigkeit beruhen auf einer subjektiven Einschätzung des Probanden durch Fach-Therapeuten. Dabei wurde eine Skala von 1 (sehr gut) bis 5 (sehr schlecht) mit der Schrittweite 0,5 zugrunde gelegt.

Tabelle A.1: Subjektive Einschätzung der Probanden durch Therapeuten

			Sprechtechnik		Flüssigkeit	
Klient	Alter(J)	m/w	v. Therapie	n. Therapie	v. Therapie	n. Therapie
K001	20	m	5	2,5	4	3
K002	27	m	5	3	5	4
K003	18	m	5	4	5	1,5
K004	29	m	5	2	5	2,5
K005	23	w	5	1,5	4	2
K006	24	w	5	2	4,5	3,5
K007	43	m	5	3,5	3,5	2
K008	20	m	5	1,5	5	1,5
K009	16	m	5	3	5	2
K010	22	m	5	2	5	1,5
K011	53	m	5	4	1,5	1,5
K012	37	m	5	3	5	3,5
K013	15	m	5	2,5	5	4
K014	31	m	5	3,5	1,5	1,5
K015	19	w	5	3	3,5	2
K016	16	m	5	2,5	3	2,5
K017	26	m	5	1,5	2	1,5
K018	20	m	5	1,5	4,5	2
K019	67	m	5	3	3	2,5
K020	17	m	5	2	5	1

Literaturverzeichnis

[1] Bahoura, M. und J. Rouat: *Wavelet Noise Reduction: Application to speech enhancement*. Techn. Ber. G7H 2B1, Universite du Quebec à Chicoutimi, Chicouti, Quebec, Canada, 2 11. Kopie auf CD-ROM (wavelet-noise-reduction-application.pdf).

[2] Bloodstein, O.: *A Handbook on Stuttering*. Singular Publishing Group, Inc., San Diego, London, 5 Aufl., 1997.

[3] Boberg, E. und D. Kully: *Comprehensive stuttering program*. College Hill Press, San Diego/Cal., 1995.

[4] Christianini, N. und J. Shawe-Taylor: *Support-Vector-Machines and other kernel-based leanring methods*. Cambridge University Press, Cambridge, 3 Aufl., 2002.

[5] Eschler, K.: *Support Vector Maschines*, 2006. Kopie auf CD-Rom (07_SVMs_Vortrag.pdf).

[6] Euler, H. und A. Wolff v. Gudenberg: *Die Kasseler Stottertherapie - Ergebniss einer computergestützten Biofeedbacktherapie für Erwachsene*. Sprache Stimme Gehör, 24:71–79, 2000. Kopie auf CD-ROM (KST-Forschungsergebnisse.pdf).

[7] Euler, S.: *Grundkurs Spracherkennung*. Vieweg, Wiesbaden, 1 Aufl., 2006.

[8] Fiedler, P.: *Sprechstunde oder Psychotherapie? Wege oder Umwege in der erfolgreichen Behandlung erwachsener Stotternder*. Jugend und Volk, Wien, 1988.

[9] Fiedler, P. und R. Standop: *Stottern - Ätiologie, Diagnose, Behandlung*. Psychologie Verlags Union, Weinheim, 4 Aufl., 1998.

[10] Fuss, L.: *Sprechgeschwindigkeit beschreibende Merkmale in der Emtionserkennung*. Diplomarbeit, Universität Karlsruhe, Institut für Nachrichtentechnik, Automation und Robotik, Juni 2006. Kopie auf CD-Rom (da_fuss.pdf).

[11] Gall, V. und R. Berg: *Feinstrukturen von Stimme und Sprache*. Edition Wötzel, Frankfurt am Main, 1 Aufl., 1998.

[12] Gerber, R.: *Ein Untersuchungssystem für Aufmerksamkeitsverluste auf Basis von Eyetrackingmessungen und Klassifikation mit prototypvektorbasierten Neuronalen Netze*. Diplomarbeit, Fachhochschule Schmalkalden, Fachbereich Informatik, 2003. Kopie auf CD-Rom (2003_Gerber_Rene.pdf).

[13] Glück, D. C. W.: *Handbuch zu FluencyMeter basic*. URL, http://www.fluencymeter.de, 2002.

[14] Golz, M.: *Persönl. Gespräch*, 2006. Fachhocfhschule Schmalkalden.

[15] Golz, M. und D. Sommer: *Detection of Strong Fatigue During Overnight Driving*. Biomedizinische Technik, 50(suppl. vol. 1):479 – 480, 2005.

[16] Gudenberg, A. Wolff v.: *Die Kasseler Stottertherapie - Evaluierung einer computergestützten Intensivtherapie*. Forum Logopädie, 20:6 –11, Mai 2006. Kopie auf CD-Rom (KST_FL3_2006.pdf).

[17] Gudenberg, A. Wolff v. und F. Jassens: *Persönl. Mitteilungen*, 2006. Institut der Kasseler Stottertherapie, Bad Emstal.

[18] Haykin, S.: *Neuronal Networks - A Comprehensive Foundation*. Personal Education International, USA, 1999.

[19] Hentschel, G.: *Merkmalsextraktion durch kanonische Kontextkorrelationsprojektion*. URL, http://phobos.imib.rwth-aachen.de/lehmann/seminare/bv_2005-04.pdf. Kopie auf CD-Rom (bv_2005-04.pdf.pdf).

[20] Holzbrecher, M.: *Signaldekomposition mittels Independet Component Analysis (ICA) zur Extraktion von Artefakten in EEG-Signales*. Diplomarbeit, Fachhochschule Schmalkalden, Dezember 2006. Kopie auf CD-Rom (diplomarbeit_MH.pdf).

[21] Hubbard, B. B.: *Wavelets: Die Mathematik der kleinen Wellen*. Birkhäuser Verlag, 1997. französische Originalausgabe: „Ondes et Ondelettes“, Paris, 1995.

[22] Kehoe, T. D.: *Stuttering: Science, Therapy & Practice*. Casa Future Technologies, Boulder, 3 Aufl., 1998.

[23] Kohonen, T.: *Self-Organizing Maps*. Springer, Berlin, 3 Aufl., 2001.

[24] Kremer, G.: *Untersuchung der Sprecher-individualität höherer Formanten*. URL, http://www.ims.uni-stuttgart.de/lehre/studentenarbeiten/fertig/studienarbeit_gerhard_kremer.pdf, 2004. Kopie auf CD-Rom (studienarbeit_gerhard_kremer.pdf).

[25] M., G.: *Tagungsband 2004: Tag der Forschung*, 2004.

[26] Misiti, Y. und G. Oppenheim: *Matlab-Wavelet Toolbox Users's Guide*. The MathWorks Inc., Natick, 1 Aufl., 1996.

[27] Natke, U.: *Stottern - Erkenntnisse, Theorien, Behandlungsmethoden*. Verlag Hans Huber, Bern, 1 Aufl., 2000.

[28] Natke, U.: *Erkenntnisse über das Stottern.*. Bd. 205, S. 6 – 14, Düsseldorf, 2001. Bundesarbeitsgemeinschaft Hilfe für Behinderte.

[29] Paulus, D. W. und J. Hornegger: *Applied Pattern Recognition - Algorithm and Implementation in C++*. Vieweg, Wiesbaden, 5 Aufl., 2006.

[30] Paulus, E.: *Sprachsignalverarbeitung – Analyse, Erkennung, Synthese*. Spektrum, Akadem. Verlag, Heidelberg, Berlin, 1 Aufl., 1998.

[31] Ptok, M., U. Natke und H. M. Oertle: *Stottern - Pathogenese und Therapie*. Deutsches Ärzteblatt, 103(18):C993 – C997, 2006.

[32] Russel, S. und P. Norvig: *Artificial Intelligence - A Modern Approach*. Personal Education International, USA, 2003.

[33] Schlagner, S. und U. Strehlau: *Fourer-Analyse versus Wavelet-Analyse*. Shaker Verlag, Aachen, 2004.

[34] Slaney, M.: *Auditory Toolbox: A Matlab Toolbox for Auditory Modelling Work*. Techn. Ber. #1998-10, Interval Research Corporation, 1998. Kopie auf CD-ROM (AuditoryToolboxTechReport.pdf).

[35] Sommer, D.: *Support-Vector-Maschine*. Kopie auf CD-Rom (SVM_ger.pdf).

[36] Sommer, D.: *Persönl. Mitteilungen*, 2006. Fachhochschule Schmalkalden.

[37] Umapathy, K., S. Krishnan und D. G. Parsa, V. andb Jamieson: *Discrimination of Pathological Voices Using a Time-Frequency Approach*. IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, 52(3):421 – 430, 2005.

[38] Wilke, J.: *Detektion von Müdigkeitszuständen in elektrophysiologischen Signalen aus Nachtfahrt-Simulationen unter Verwendung ausgewählter Methoden der Neuroinformatik und des maschinellen Lernens*. Diplomarbeit, Fachhochschule Schmalkalden, Fachbereich Informatik, März 2005. Kopie auf CD-Rom (diplomarbeit_jens_wilke.pdf).

[39] Zell, A.: *Simulation neuronaler Netze*. Addison-Wesley, Bonn, 1 Aufl., 1994.

[40] Zückner, H., U. Dr. Natke und W. Heil: *FAQ Stottern*. URL, http://www.stottermodifikation.de/faq.html, 2006. Kopie auf CD-Rom (FAQ-Stottern.pdf).