
Universität Karlsruhe (TH)
Forschungsuniversität · gegründet 1825

Institut für Theoretische Informatik

Diplomarbeit

Erkennung von lautlos und kontinuierlich gesprochener Sprache mittels Elektromyografie

cand. inform. Matthias Walliczek

Juni 2006

Betreuer:

Prof. Dr. rer. nat. Alexander Waibel

Dipl.-Inform. Florian Kraft

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfsmittel angefertigt habe. Alle verwendeten Quellen sind als solche kenntlich gemacht und im Literaturverzeichnis aufgeführt.

Karlsruhe, den 29. Juni 2006

Matthias Walliczek

Danksagung

Mein besonderer Dank gilt meinem Betreuer, Herrn Dipl.-Inform. Florian Kraft für die freundliche und sehr gute fachliche Betreuung und dafür, dass er mich bei Problemen stehts nach besten Kräften unterstützt hat.

Bedanken möchte ich mich auch bei Frau Dr. Tanja Schultz und Szu-Chen Jou, die mir mit Fachwissen gerne zur Seite standen.

Ebenfalls bedanken möchte ich mich bei Herrn Prof. Dr. Alexander Waibel.

Mein Dank gilt weiterhin Peter Osztotics und Irene Hassinger, die sich bereitwillig als Testpersonen zur Verfügung gestellt haben.

Der Mensarunde und meiner Familie danke ich für alles Nichtfachliche.

Und schließlich möchte ich mich bedanken bei all denen, die ich hier leider nicht mehr einzeln aufzählen kann, die mir mit Tipps und Gesprächen oft sehr weitergeholfen haben.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Nachdem die akustische Spracherkennung für viele Jahrzehnte lediglich in Forschungsinstituten, Testsystemen und für Spezialaufgaben eingesetzt wurde, zeichnet sich aktuell ihr Durchbruch für die Nutzung in Produktivsystemen ab, die von der gesamten Bevölkerung genutzt werden können. So bietet beispielsweise die Deutsche Bahn eine kostenlose Fahrplanauskunft per Telefon an, bei der Start, Ziel und Zeit der Reise per Spracherkennung eingegeben werden. Zahlreiche Mobiltelefone erlauben eine Sprachwahl, bei der zuvor bis zu 30 Namen aufgenommen werden und anschließend per Sprachkommando gewählt werden können.

Doch trotz allem bestehen für die akustische Spracherkennung zwei prinzipielle Einschränkungen: Zum einen ist bei zu starken Nebengeräuschen eine zuverlässige Erkennung nicht mehr möglich. Geringfügige Nebengeräusche können zwar noch ausgefiltert werden, ab einer bestimmten Signalstärke überlagern die Nebengeräusche das eigentliche Signal jedoch zu stark, um noch getrennt werden zu können. Dies ist beispielsweise in einem Düsenjet der Fall.

Zum anderen muss der Sprecher laut sprechen und dadurch Geräusche erzeugen, durch die andere Menschen gestört werden können. In einer Konferenz beispielsweise würden es andere Teilnehmer als störend empfinden, wenn man seine Notizen per Spracherkennung diktieren würde, hier bleiben also nur Stift und Notizblock als Speichermedium.

Gesucht ist also ein Erkennungssystem, das auch stimmlose Sprache erkennen kann. Ein erster Ansatz dafür wäre „Lippenlesen", also die Interpretation der Mund- und insbesondere Lippenbewegungen. Wie das McGurk-Experiment [MM76] zeigt, wird diese Technik unbewusst von fast allen Menschen eingesetzt, um die akustische Spracherkennung zu unterstützen und gerade bei vielen Nebengeräuschen zu verbessern. Auch bei der automatischen Spracherkennung wird Lippenlesen bereits erfolgreich eingesetzt [DMW94]. Allerdings geht es dabei immer nur um die Unterstützung der akustischen Spracherkennung; eine Spracherkennung ausschließlich auf Basis des Lippenlesens funktioniert weder beim Menschen noch beim Computer, da nur ca. 15% der deutschen Laute anhand des Lippenbildes unterscheidbar sind. Selbst geübte Lippenleser können bei bekannten Themen nur bis zu 30% eines Textes von den Lippen ablesen [Bun].

Offenbar stehen fürs Lippenlesen zu wenig Informationen über den Status des Vokaltraktes zur Verfügung, insbesondere die Zungenbewegung ist bei geschlossenem Mund

nicht erkennbar. Um auch diese Bewegungen erkennen zu können, braucht man ein Verfahren, mit dem man auch nicht sichtbare Muskelbewegungen erfassen kann: Die Elektromyografie.

1.2 Spracherkennung mittels Elektromyografie

Die Elektromyografie als Verfahren zur Erkennung von Muskelaktionspotentialen wurde als Erstes im medizinischen Bereich eingesetzt, um Muskelerkrankungen (Myopathien) zu erkennen und Prothesen zu steuern. So ist die Hauptanwendung der Elektromyografie „die Erkennung von Myopathien und Neuropathien, das heißt die Feststellung, ob eine Krankheit muskuläre und/oder nervliche Ursachen hat." [Wikc]. Dadurch können die Ursachen für Schwächungen, Lähmungen, unwillkürliche Zuckungen und abnormale Muskelenzymspiegel gefunden werden. Bei der Diagnose von neuromuskulären Krankheiten wie Amyotrophe Lateralsklerose, Neuropathien, Nervenverletzungen und Muskelschäden kann die Elektromyografie behilflich sein [Wikb].

Bei der Elektromyografie werden die elektrischen Impulse erfasst, die bei jeder Muskelkontraktion erzeugt werden. Da auch die Gesichtsmuskulatur entsprechende Signale freisetzt, werden bereits seit den 90er Jahren Systeme erforscht, die alleine auf Basis dieser Signale erkennen, welche Wörter der Sprecher ausgesprochen hat. Da dafür keine akustischen Signale mehr benötigt werden, ist es auch möglich, stimmlose Sprache zu erkennen, bei der der Sprecher die Wörter lautlos mit dem Vokaltrakt formt. Mögliche Einsatzszenarien reichen vom lautlosen Diktieren während einer Konferenz bis zum lautlosen Telefonieren, bei der die erkannten Muskelbewegungen erst beim Gesprächspartner von einer Computerstimme zu hörbaren Wörtern oder Sätzen umgewandelt werden. Ebenso wäre es möglich, die erkannte Sprache in eine beliebige andere Sprache zu übersetzen und erst dann per Lautsprecher auszugeben.

Allerdings ist es bis dahin noch ein langer Weg; momentane Spracherkenner auf Basis der Elektromyografie (im Folgenden EMG-Spracherkennungssysteme genannt) können bislang lediglich einzelne Wörter oder feststehende Phrasen erkennen. Das hat jedoch zur Konsequenz, dass nur die Wörter oder Phrasen erkannt werden können, die auch trainiert wurden. Sollen zusätzliche Wörter oder Sätze erkannt werden, so müssen diese zusätzlich trainiert werden. Da bei dem momentanen Versuchsaufbau ein mehrere Aufnahmesitzungen umfassendes Training problematisch ist, müssen so zu Beginn jeder Aufnahmesitzung sämtliche Wörter oder Sätze neu trainiert werden.

1.3 Aktueller Stand in der Spracherkennung

Zum Vergleich: Akustische Spracherkennungssysteme werden bereits im Produktiveinsatz für zunehmend mehr Aufgaben eingesetzt: Angefangen von spracherkennungsfähigen automatischen Telefonhotlines für die Gepäckauskunft von Fluggesellschaften über Bestellhotlines für Telefongesellschaften bis zur automatischen Fahrplanauskunft für die Buslinien. Das Vokabular geht dabei deutlich über einfache ja/nein-Antworten hinaus und erlaubt auch die Erkennung von beliebigen Familien- und Ortsnamen.

In der Forschung ist man noch weiter: Aktuelle Systeme sind in der Lage, beliebige kontinuierliche Sprache eines beliebigen Sprechers zu erkennen. So sind beispielsweise bei einem englischen Text mit einer Perplexität[1] von 86,0 bei einem untrainierten Sprecher in der Broadcast-News-Domain[2] Wortfehlerraten[3] von 20,4% erreichbar [PSF+05].

1.4 Ziele dieser Diplomarbeit

In dieser Arbeit sollte als erster Schritt versucht werden, durch eine bessere Vorverarbeitung auch bei einem größeren Vokabular die einzelnen Wörter zuverlässiger unterscheiden zu können.

An diese Voruntersuchung schloss sich als Schwerpunkt dieser Arbeit das Ziel an, durch den Übergang zu kleineren akustischen Einheiten mit begrenzter Trainingsdatenmenge die Erkennung eines größeren Vokabulars zu ermöglichen. Es sollten also Einheiten gefunden und trainiert werden, die Bestandteil von mehreren Wörtern sind, so dass letztendlich auch die Erkennung von untrainierten Wörtern möglich ist.

Als letzter Schritt sollte schließlich getestet werden, ob mit diesen Veränderungen auch kontinuierliche EMG-Sprache erkannt werden kann.

Nicht Thema dieser Arbeit waren Untersuchungen der Sitzungsunabhängigkeit: In dieser Diplomarbeit wurden nur jeweils Daten derselben Aufnahme zum Training und

[1] Die Perplexität ist ein Maß dafür, wie viele Wörter ein Erkenner im Schnitt gleichzeitig erkennen können muss. Eine niedrige Perplexität bedeutet, dass nur wenige Wörter zu einem Zeitpunkt zu unterscheiden sind, was bedeutet, dass die Erkennungsaufgabe leicht ist. Eine hohe Perplexität heißt, dass die Verwechslungsgefahr höher ist und damit auch die Fehlerwahrscheinlichkeit und die Schwierigkeit der Erkennungsaufgabe [Rog03].

[2] Broadcast-News-Domain bedeutet, dass es sich um Aufnahmen von Nachrichtensendungen handelt, bei denen ein Nachrichtentext von einem Sprecher vorgelesen wird.

[3] Die Wortfehlerrate als Maß für die Erkennungsleistung bei kontinuierlicher Sprache ist definiert als $\frac{\#\text{Auslassungen}+\#\text{Vertauschungen}+\#\text{Einfuegungen}}{\#\text{zuerkennendeWorte}}$.

Testen verwendet, weil sich die EMG-Signale durch minimal abweichende Elektrodenpositionen oder Veränderungen im Hautwiderstand zwischen verschiedenen Aufnahmesitzungen stark unterscheiden. Die Konsequenz daraus ist, dass insgesamt nur vergleichsweise wenig Daten zur Verfügung standen, da die Dauer einer Aufnahmesitzung durch die Haftfähigkeit der Elektroden auf wenige Stunden beschränkt war. Aus diesen Gründen wurde zunächst nur die Erkennung von insgesamt 32 Wörtern getestet - im Vergleich zu den bisher verwendeten 10 Ziffern ist das aber eine quantitative Verbesserung. Bei Tests mit kontinuierlicher Sprache wurde weiterhin mit einem größeren Vokabular gearbeitet.

Des Weiteren wurde in dieser Arbeit ausschließlich mit vorgelesener Sprache gearbeitet; spontane Sprache sollte nicht untersucht werden. Die Testpersonen wurden außerdem angewiesen, mit möglichst neutralem Gesichtsausdruck zu sprechen. Sie mussten weiterhin durch einen Button manuell Beginn und Ende der Sprachaufnahme steuern. Die Erkennung von Sprache und insbesondere deren Abgrenzung zu anderen Muskelbewegungen beispielsweise durch Husten, Lachen, Gähnen etc. war nicht Thema dieser Arbeit.

1.5 Aufbau der Arbeit

Um die Erkennung von kontinuierlicher Sprache zu testen, wurden in dieser Arbeit vier verschiedene Dimensionen gleichzeitig untersucht:

1. Zum einen musste durch eine bessere Vorverarbeitung eine insgesamt robustere Erkennung ermöglicht werden. In vergleichbaren Arbeiten werden unterschiedliche Ansätze gewählt, die entweder im Zeitbereich oder im Frequenzbereich arbeiten. In dieser Arbeit musste geprüft werden, welches Verfahren am besten geeignet ist, um auch kleinere Worteinheiten zu erkennen.

2. Zusätzlich mussten passende Einheiten für die Wortmodellierung gefunden werden, die einerseits die Erkennung von untrainierten Wörter ermöglichen, andererseits optimal auf die Eigenschaften von EMG-Sprache abgestimmt sind. Während bei bisherigen Arbeiten größtenteils mit Modellen für komplette Wörter gearbeitet wurde, wurden in dieser Arbeit zum ersten Mal kleinere Modelleinheiten untersucht, die in mehreren Wörtern vorkommen. Bei der akustischen Spracherkennung bildet ein Phonem eine solche wiederverwertbare Einheit. Da bei Beginn dieser Arbeit jedoch unbekannt war, ob EMG-Sprache sich insbesondere im Bezug auf die Bedeutung des Kontextes mit akustischer Sprache vergleichen lässt,

wurden mit Silben zunächst größere und der EMG-Sprache eher angenäherte Einheiten getestet.

3. Da in der akustischen Spracherkennung durch kontextabhängige Wortmodelle große Verbesserungen erzielt werden konnten und vermutet werden kann, dass aufgrund der längeren Zeitdauer für eine Muskelkontraktion im Vergleich zu den kürzeren Zeitfenstern für die akustische Sprache der Kontext für EMG-Sprache deutlich wichtiger ist, wurden auch Versuche mit verschiedenen kontextabhängigen Einheiten durchgeführt.

4. Als vierte Dimension wurde der Aufnahmemodus betrachtet: Zu Beginn wurde mit isoliert aufgenommenen Wörtern experimentiert; im zweiten Teil wurden Experimente mit kontinuierlich ausgesprochenen Sätzen durchgeführt. Zunächst wurde dabei versucht, durch Modelle, die durch kontinuierlich ausgesprochene Sätze trainiert wurden, die isoliert aufgenommenen Einzelwörter zu erkennen, um zu prüfen, ob Training und insbesondere Segmentierung erfolgreich arbeiten. Im folgenden Schritt sollte dann auch versucht werden, kontinuierliche Sätze zu erkennen.

2 Vergleichbare Arbeiten

Im diesem Kapitel werden verschiedene Arbeiten vorgestellt, die sich mit Elektromyografie im Allgemeinen und deren Anwendung für die Spracherkennung im Speziellen beschäftigen, um die Ausgangsbasis zu Beginn dieser Diplomarbeit darzustellen.

2.1 Zeitliche Einordnung

Bereits in den 60er Jahren wurde versucht, bei myoelektrischen Signalen Muster zu erkennen [MGW91]. In den 70er Jahren wurde untersucht, wie man durch Bewegungs- und Vektorerkennung Prothesen steuern kann. Anfang der 80er Jahren begannen schließlich Untersuchungen zur Spracherkennung durch Elektromyografie. Insbesondere in den letzten fünf Jahren gab es in diesem Bereich verstärkte Forschungen.

2.2 Sensorik

Genauso wie bei der akustischen Spracherkennung die Charakteristik des verwendeten Mikrofons (Headset oder Raummikrofon) über Erfolg oder Misserfolg eines Spracherkennungssystems entscheiden kann, ist auch für EMG-Spracherkennungssysteme entscheidend, welche Muskeln auf welche Weise abgetastet werden.

Hinsichtlich der Aufnahme von Signalen hat sich bei allen vergleichbaren Arbeiten aus Gründen der Benutzerfreundlichkeit die Oberflächen-Elektromyografie durchgesetzt. Eingesetzt werden Silber/Silber-Chlorid-Elektroden (Ag-AgCl), teilweise in Kombination mit einem Gel, das den Kontakt zwischen Haut und Elektrode verbessern soll.

Im Gegensatz zu anderen Arbeiten, bei denen die Elektroden auf der Haut fixiert sind, wählt Manabe [MHS03] einen neuen Ansatz: Um den Benutzerkomfort zu verbessern und einen tatsächlichen Einsatz im Alltag überhaupt erst möglich zu machen, werden in dieser Arbeit keine fixierten Elektroden verwendet, sondern Elektroden, die auf die Finger gesteckt und gegen festgelegte Bereiche im Gesicht gedrückt werden müssen. Dabei ist zu berücksichtigen, dass insbesondere bei der angestrebten Nutzung durch technisch unerfahrene Anwender eine konstante Positionierung weder zwischen verschiedenen Sitzungen noch innerhalb derselben Sitzung garantiert ist; auch durch unterschiedlichen Druck können sich Schwankungen bei der Leitfähigkeit ergeben. Weiterhin ist durch

die Verwendung von trockenen Elektroden ohne Kontaktgel ein insgesamt schlechterer Kontakt gegeben.

Bei der Arbeit von Chan et al. [CKHL02] werden die Elektroden nicht auf der Haut befestigt, sondern stattdessen in die Sauerstoffmaske eines Pilotenhelms integriert, um Kommandos eines Düsenjetpiloten zu erkennen.

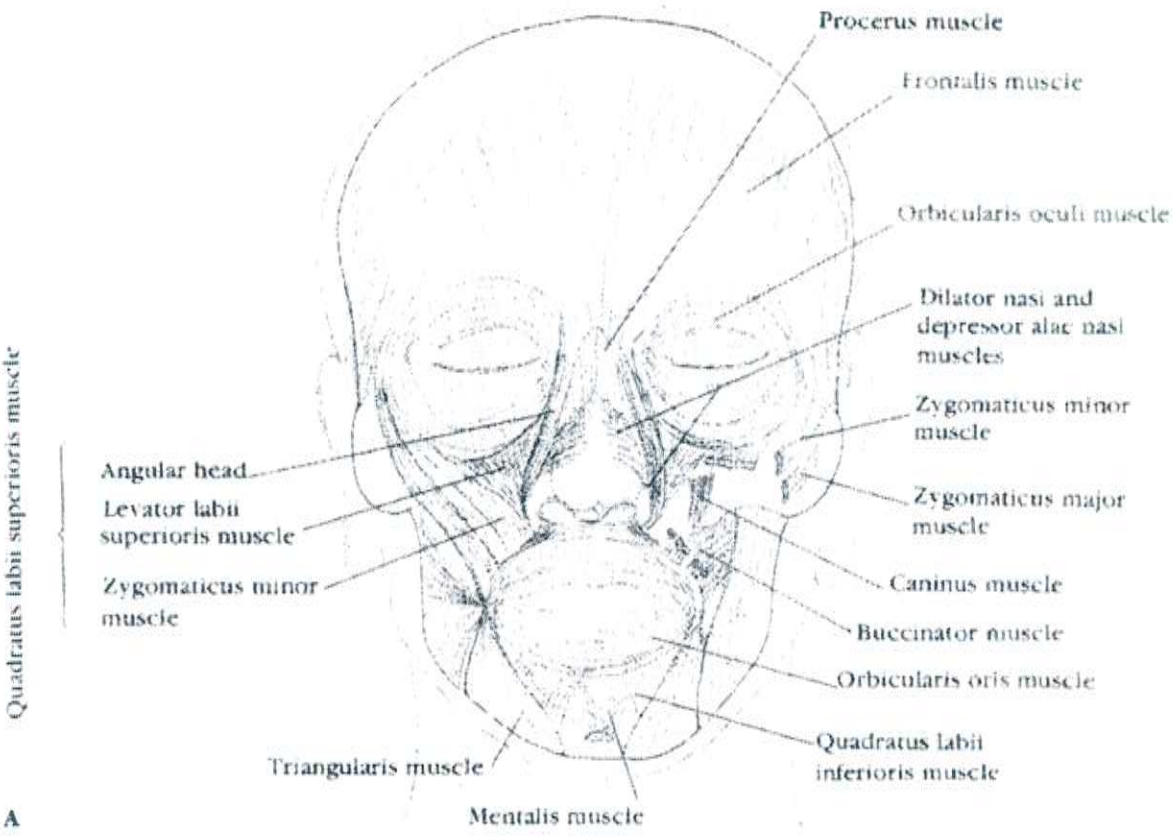

Abbildung 2.1: *Anatomische Darstellung der Gesichtsmuskulatur nach [DMD82]*

Unterschiede gibt es bei der Anzahl der untersuchten Muskeln: Chan und Maier-Hein [MHMSW05] messen insgesamt 5 Muskeln im Gesichtsbereich (levator anguli oris, zygomaticus major, platysma, depressor anguli oris, anterior belly of the digastric, vgl. Abbildung 2.1), bei Chan mit insgesamt fünf Elektrodenpaaren, während bei Maier-Hein drei bipolare und vier unipolare Elektroden an insgesamt zwölf Messpunkten sieben Signale ergeben. Morse dagegen tastet insgesamt nur vier Muskeln ab. Kumar [KKAB04] und Manabe kommen mit drei Muskeln aus, die jeweils durch ein Elektrodenpaar gemessen werden - bei Kumar sind dies mentalis, depressor anguli oris, massetter, und bei Manabe digastricus, zygomaticus major, orbiculas oris.

Jorgensen und Binsted [JB05] haben herausgefunden, dass für die Erkennung von wenigen isoliert aufgenommenen Wörtern insgesamt vier Elektroden ausreichen, die diagonal an der Spalte zwischen Kinn und Kehlkopf aufgeklebt sind.

Als Abtastfrequenz werden Werte zwischen 250 Hz (Kumar) und 10 kHz (Chan) verwendet. Zuvor wird das Signal jedoch durch Hochpass- und Tiefpassfilter auf Frequenzen zwischen 8 Hz und 79 Hz (Kumar) bzw. 20 bis 500 Hz begrenzt.

Da das Ziel seiner Arbeit die Verbesserung der akustischen Spracherkennung ist, erfasst Chan zusätzlich zu den EMG-Signalen das akustische Signal.

2.3 Vorverarbeitung

Größere Unterschiede zwischen den Arbeiten gibt es bei der gewählten Vorverarbeitung: Es werden Vorverarbeitungen im Zeitbereich, im Frequenzbereich und durch die Bildung von Featurevektoren mit Kombinationen aus Zeit- und Frequenzbereich verwendet. In einigen Arbeiten wird die Dimension des Merkmalsraums durch Verfahren wie die Hauptkomponentenanalyse (engl. Principal Component Analysis, PCA)[1] reduziert.

Als Vorverarbeitung im Zeitbereich verwenden Manabe und Kumar die Standardabweichung (engl. Root Mean Square, RMS)[2] für jeweils 400 ms große Zeitfenster.

Für die Vorverarbeitung im Frequenzbereich kommen Verfahren wie „Short Time Fourier Transformation" (STFT)[3] und Wavelets[4] zum Einsatz. Jorgensen und Binsted verwenden bei ihrer Arbeit ausschließlich Vorverarbeitungen im Frequenzbereich, nämlich STFT und Dual Tree Wavelets. Manabe kombiniert in seiner Arbeit insgesamt vier verschiedene Vorverarbeitungen (melskalierte[5] Filterbänke, LPC[6], Mel-cepstral-Koeffizien-

[1] Bei der Hauptkomponentenanalyse versucht man, über den mathematischen Weg der Hauptachsentransformation aus Variablen mit vielen Eigenschaften einige wenige latente Faktoren zu extrahieren, die für diese Eigenschaften bestimmend sind [Pea01].

[2] Die Standardabweichung ist in der Stochastik ein Maß für die Streuung der Werte einer Zufallsvariable um ihren Mittelwert. Sie ist für eine Zufallsvariable X definiert als die positive Quadratwurzel aus deren Varianz und wird als $\sigma_x = \sqrt{Var(X)}$ notiert.

[3] Die Short-Time-Fourier-Transformation behebt ein Problem der Fouriertransformation, die keine Aussage über das zeitliche Auftreten einzelner Frequenzen oder Frequenzbereiche macht. Dazu werden die im Signal enthaltende Frequenzen abschnittsweise mittels klassischer Fourier-Analyse berechnet.

[4] Die Wavelet-Transformation ist eine Form der Frequenz-Transformation. Der große Vorteil gegenüber der Fourier-Transformation ist die zeitliche Lokalität der Basisfunktionen und die geringere Komplexität.

[5] Mel ist die Maßeinheit für die psychoakustische Größe Tonheit und beschreibt die wahrgenommene Tonhöhe.

[6] Linear Predictive Coding ist in der Datenübertragung eine Codierungsart, mit der eine Datenreduktion durch ein Prädiktionsverfahren erreicht wird. Dabei wird die im Sprachsignal enthaltene Redundanz, d.h. die Abhängigkeit aufeinander folgender Abtastwerte voneinander, ausgenutzt.

ten und LPC Cepstrum[7]), die jeweils auf einem 100 ms großen Zeitfenster mit 80 ms Überlappung berechnet werden.

Chan et al. und Maier-Hein verwenden in ihren Arbeiten ein System, das sowohl eine Vorverarbeitung im Zeitbereich als auch im Frequenzbereich verwendet. Chan verwendet den RMS-Wert und 12 Mel-cepstral-Koeffizienten von einem 128 ms großen Fenster. Maier-Hein dagegen berechnet über ein 54 ms großes Fenster den Mittelwert aus dem Zeitbereich und 17 Delta-Koeffizenten aus der STFT, aus denen ein gemeinsamer Feature-Vektor erstellt wird. Für das akustische Signal berechnet Chan über jeweils 25,6 ms große Fenster die Mel-cepstral-Koeffizenten.

2.4 Klassifizierung

Für die Klassifizierung der Merkmale werden bei bisherigen Arbeiten entweder Neuronale Netze oder Hidden-Markov-Modelle (HMM) verwendet. Morse und Kumar verwenden mehrschichtige Neuronale Netze, Chan und Maier-Hein dagegen HMMs. Mit mehr als einem Klassifikator beschäftigt sich Manabe, der sowohl HMMs als auch Neuronale Netze verwendet. Jorgensen vergleicht Neuronale Netze mit einer Support Vector Maschine.

Morse und Kumar verwenden zur Klassifizierung ein auf Back-Propagation basierendes neurales Netzwerk; bei Kumar werden insgesamt zwei versteckte Schichten verwendet.

Für die HMMs kombiniert Maier-Hein die 18 Koeffizienten pro Kanal zu einem Zustandsvektor, dem anschließend mit Hilfe des *Expectation-Maximization* (EM) Algorithmus fünf HMM-Zuständen pro Wort zugeordnet werden. Chan kombiniert die EMG-Feature mit insgesamt 12 MFCC-Koeffizienten und teilt jede Ziffer auf insgesamt 12 HMM-Zustände auf.

Um verschiedene Vorverarbeitungen zu kombinieren, arbeiten Manabe und Zhang in einer weiteren Arbeit [MZ04] mit insgesamt 44 parallelen Multi-stream-HMMs, die anschließend entsprechend den Trainingsergebnissen gewichtet werden. Jedes Wort wird auf fünf Zustände aufgeteilt. Da er HMMs für die temporären und nicht-stationären EMG-Signale für nicht geeignet hält, vergleicht Jorgensen neuronale Netze und Support Vector Maschines. Als neuronales Netz verwendet er skalierte konjugierte Gradienten-Netzwerke entsprechend der Levenberg-Marquardt Implementierung.

[7]Das Cepstrum ist die inverse Fouriertransformation des logarithmierten, durch Division mit einer Bezugsgröße G0 dimensionslos gemachten, einseitigen Autoleistungsspektrums.

2.5 Korpus

Insgesamt spielt sich die EMG-Spracherkennung bislang in sämtlichen vergleichbaren Arbeiten noch auf Einzelwortebene ab. So können beispielsweise die zehn Ziffern von null bis neun erkannt werden.

Teilweise wird auch die Erkennung von speziellen Kommandos untersucht. In der Arbeit von Jorgensen geht es um die Steuerung eines Webbrowsers, dafür werden Steuerungswörtern wie „stop", „go", „left" trainiert.

In der Arbeit von Jorgensen und Binsted sind erste Versuche beschrieben, eine Spracherkennung auf Phonemebene durchzuführen; so experimentieren sie mit 17 einsilbigen Wörtern und erreichen eine Erkennungsgenauigkeit von 50%. Sie stellen u.a. fest, dass zahlreiche Phone nur über die Stimmhaftigkeit zu unterscheiden sind, z.B. 'd' und 't'. Weiterhin konstatieren sie, dass die Forschung in diesem Bereich insgesamt noch in den Kinderschuhen steckt.

Weitere Forschungen zur EMG-Spracherkennung auf Phonemebene finden sich in der Diplomarbeit von Maier-Hein [MH05]. Dabei zeigt sich zum einen, dass der Kontext auf Wortebene für die Erkennung äußerst wichtig ist. Sie dokumentiert weiterhin Experimente, bei denen zehn zuvor auf Wortebene getestete Wörter nun auf kontextunabhängiger Phonemebene erkannt werden sollen. Insbesondere bei stimmloser Sprache erreicht die Erkennung auf Phonemebene jedoch deutlich schlechtere Erkennungsquoten als die Erkennung auf Wortebene; Maier-Hein schließt daraus, dass die Verwendung von Phonemmodellen möglicherweise nicht die beste Möglichkeit zur Erkennung stimmloser Sprache ist.

2.6 Detaillierter Aufbau eines EMG-Spracherkennungssystems

Als ein Beispiel für ein Spracherkennungssystem auf Basis der Elektromyografie wird im Folgenden das von Lena Maier-Hein am Institut für theoretische Informatik entwickelte System [MH05] vorgestellt, das für diese Arbeit als Baseline-System fungiert. Es erreicht bei der Erkennung von 10 Ziffern (*zero, one, two, three, four, five, six, seven, eight, nine*) Erkennungsquoten von bis zu 98%.

In der Arbeit werden 7 Eingangskanäle, die sich aus 4 bipolare Elektrodenpaaren und 3 unipolaren Elektroden ergeben, mit einer Frequenz von 600 Hz abgetastet. Anschließend werden die Kanäle einzeln vorverarbeitet. Alle 4 ms werden jeweils über ein 54 ms

großes Fenster pro Kanal insgesamt 17 STFT-Koeffizienten berechnet. In einem nächsten Schritt werden von diesen Koeffizienten die Delta-Werte berechnet, d.h. die Differenzen zu vorhergehenden Werten. In den folgenden Schritten werden ausschließlich die Delta-Werte berücksichtigt. Zusätzlich wird der Mittelwert im Zeitbereich innerhalb dieses Fensters berechnet. Die 18 Koeffizienten der 7 Kanäle werden anschließend zu einem Feature-Vektor mit insgesamt 126 Werten zusammengefügt.

Für die Klassifikation werden *Left-to-right* Hidden Markov Modelle (HMMs) verwendet. Für die Initialisierung wird die Aufnahme linear auf insgesamt 7 HMM-Zustände aufgeteilt: 1 Silence-Zustand zu Beginn, 5 Zustände für das ausgesprochene Wort und 1 Silence-Zustand am Ende. Mithilfe des KMeans-Algorithmus werden pro HMM-Zustand 14 Gaussians initialisiert und während 4 Durchläufen des Expectation-Maximization-Algorithmus trainiert.

2.7 Andere Nutzung von EMG-Signalen

Neben dem Nutzen für die Spracherkennung beschäftigen sich andere Arbeiten mit der Erkennung von Muskelbewegungen im Arm oder in der Hand. Diese versuchen, die Erkennung durch eine bessere Vorverarbeitung zu optimieren. Einige Arbeiten verwenden dabei die „Independent Component Analysis“ (ICA) (vgl. Kapitel 3.5.4) in Kombination mit einer Wavelet-Transformation des Eingangssignals. Azzerboni et al. beispielsweise versuchen in ihrer Arbeit [AFI+02] die Aktivität der Armmuskulatur zu erkennen, und verwendet dazu 16 auf Brust, Oberarm und Unterarm verteilte Elektroden. Auf deren Signal wendet er zunächst eine diskrete Wavelet-Transformation (dwt) an, um die Ergebnis-Werte anschließend durch eine Hauptkomponentenanalyse (vgl. Kapitel 3.5.4) zu reduzieren. Mit der ICA sollen anschließend EKG-Artefakte entfernt und einzelne Muskeln identifiziert werden. Mit einer Wavelet-Analyse soll anschließend die Aktivität der einzelnen Muskeln erkannt werden.

In der vorliegenden Arbeit wurde untersucht, ob einige dieser Verfahren auch für die EMG-Spracherkennung genutzt werden können.

3 Methodik

In diesem Kapitel werden die theoretischen Grundlagen und Vorgehensweisen bei dieser Arbeit erläutert, angefangen mit den physiologischen Grundlagen der Elektromyografie. Aufbauend auf diesen Grundlagen werden verschiedene Vorverarbeitungen aufgeführt, die den Besonderheiten der Elektromyografie Rechnung tragen. Weiterhin werden unterschiedlichen Einheiten für die Wortmodellierung vorgestellt und insbesondere Grundlagen der silbenbasierten Spracherkennung aufgeführt. Anschließend werden die Verfahren zur Berücksichtigung der Kontextabhängigkeit beschrieben und diskutiert. Zum Schluss werden prinzipielle Unterschiede zwischen isoliert aufgenommenen Einzelwörtern und kontinuierlicher Sprache erwähnt.

3.1 Grundlagen der Elektromyografie

In der menschlichen Anatomie unterscheidet man drei Typen von Muskeln: Glatte Muskulatur, Herzmuskulatur und Skelettmuskulatur. Willentlich steuerbar ist lediglich die quergestreifte Skelettmuskulatur, die mit über 600 Muskeln beim Mann 40% und bei der Frau 23% der Gesamtkörpermasse ausmacht [JCKH98].

Die willentliche Steuerung verläuft durch elektrische Impulse, die vom Gehirn oder Rückenmark ausgesandt und über die Nerven weitergeleitet werden. Diese Impulse lassen sich in eine Ruhephase vor der Erregung, eine Depolarisationsphase zu Beginn der Muskelanspannung, eine anschließende Repolarisationsphase, eine Hyperpolarisationsphase und die Rückkehr zur Ruhephase einteilen (vgl. Abbildung 3.1). Während der Ruhephase liegt in den Nervenzellen ein Spannungspotenzial von -70 bis -90 mV an, das in der Depolarisationsphase auf 20 bis 30 mV ansteigt.

Dieses Aktionspotenzial kann durch spezielle Elektroden gemessen werden. Die Elektroden können entweder in Form von Nadeln direkt in unmittelbare Nähe der Muskeln injiziert werden oder auf der Hautoberfläche über den Muskeln aufgeklebt werden. Im ersten Fall erreicht man eine größere Genauigkeit, allerdings zu Lasten der Benutzerfreundlichkeit. Zum einen ist es durchaus schmerzhaft, Nadeln in Muskelnähe zu injizieren, zum anderen ist eine derartige Anwendung für den Alltagsgebrauch durch normale Menschen wenig realistisch. Deshalb hat sich zumindest im Bereich der Spracherkennung die Messung durch Oberflächenelektroden durchgesetzt. Dabei muss jedoch beachtet werden, dass sich bei derartigen Messungen die Signale von verschiedenen Muskeln vermischen können.

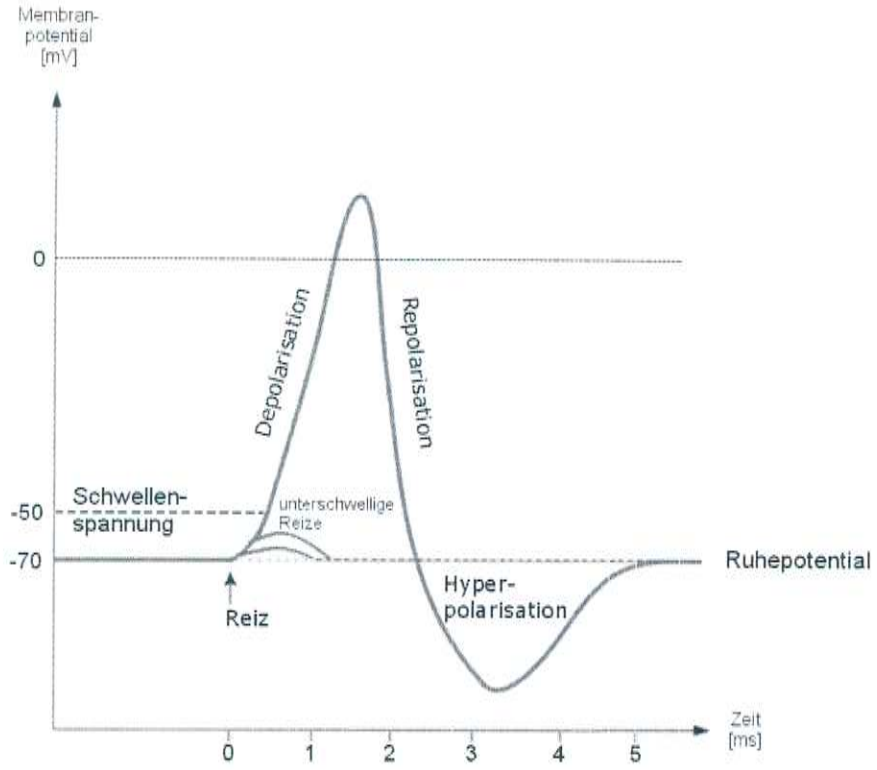

Abbildung 3.1: *Aktionspotential eines Muskels nach [Wika]*

3.2 Menschliche Sprache aus physiologischer Sicht

Bereits bei der akustischen Spracherkennung hat sich die Erkenntnis durchgesetzt, dass es für das Verstehen der Signale, die der Artikulationsapparat erzeugt, durchaus hilfreich ist, die Anatomie der Organe des Menschen, die für die Produktion von Sprache verantwortlich sind, zu begreifen [Rog03]. Tatsächlich gibt es einige Bestrebungen, die Sprache nicht nur als Beobachtung akustischer Phänomene zu betrachten, sondern als akustische Manifestation von Zuständen und Zustandsfolgen der Artikulatoren. Dies ist insbesondere deshalb sinnvoll, weil allein schon aus physikalischen und biologischen Gründen nicht jeder Zustand des Artikulationsapparates jedem anderen unmittelbar folgen kann. Die Artikulatoren müssen stetige Bewegungen durchführen und können keine Sprünge machen. Die Ausprägung einzelner Laute hängt also davon ab, welche Laute davor und danach artikuliert werden. Teilweise wird der Schluss eines Lautes noch von einem Teil der Artikulatoren gesprochen während ein anderer Teil schon den Anfang der folgenden Laute produziert (vgl. Abbildung 3.2).

Für die menschliche Sprache ist neben Lunge und Stimmbändern auch der Vokaltrakt verantwortlich. Dort bestimmen Kiefer, Lippen, Mund, Zunge und Velum die Abstrahlung der modulierten Pulse durch Mundöffnung und Nasenlöcher bzw. ermöglichen durch den Verschluss des Vokaltraktes eine Unterbrechung des Luftstroms [Rog03] (vgl. Abbildung 3.3).

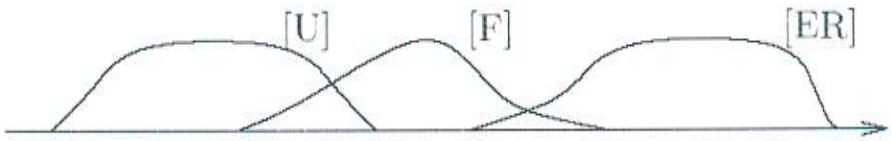

Abbildung 3.2: *Überlagerung und Übergänge von Lauten nach [Rog03]*

3.3 Besonderheiten der Spracherkennung auf Basis der Elektromyografie

Um bei der EMG-Spracherkennung die Sprache möglichst gut zu erkennen, müssen möglichst alle Muskeln abgetastet werden, die die Sprachlaute eindeutig voneinander unterscheiden. Aus diesem Grund werden bei EMG-Spracherkennung für gewöhnlich mehrere Kanäle (bei unserem System insgesamt 7 Kanäle) gleichzeitig abgetastet, während bei akustischer Sprache bei Verwendung eines Headsets nur ein Mono-Signal zur Verfügung steht.

Ein anderer Unterschied zwischen akustischer Spracherkennung und EMG-Spracherkennung betrifft die Abtastrate: Bei akustischen Systemen wird mit 16.000 Hz abgetastet, da die höchsten relevanten Frequenzen im Bereich bis 8 kHz liegen. Bei der EMG-Spracherkennung dagegen wird mit deutlich geringeren Frequenzen im Bereich bis zu 1000 Hz abgetastet.

Wichtig ist auch die unterschiedliche Erfassung der Signale: Während bei akustischer Spracherkennung die von Lunge und Vokaltrakt geformten Impulse durch das Trägermedium Luft übertragen werden, werden bei der Elektromyografie die von den Nerven zu den Muskeln übertragenen Signale erfasst und elektrisch gemessen. Dies hat zur Folge, dass bei der Elektromyografie auch tieffrequente Signale erfasst werden können, zum Beispiel das von Anspannung bis Entspannung eines Muskels erzeugte Signal.

Durch eine genauere Analyse des EMG-Signals im Zeitbereich ist es deshalb möglich, die Anspannung und Entspannung der Muskeln zu messen. Daraus ergibt sich eine wichtige Konsequenz: Bei der EMG-Spracherkennung spielen Veränderungen des Signals eine deutlich wichtigere Rolle als bei der akustischen Spracherkennung. Bei der akustischen Spracherkennung geht es darum, die Laute durch kurzzeitige charakteristische Mischungen von Signalen verschiedener Frequenzen zu erkennen. Dagegen muss man bei der EMG-Spracherkennung versuchen, anhand der Signale die Muskelbewegungen zu erkennen, die den Vokaltrakt in den Zustand verändern, um diese Laute formen zu können. Bei der EMG-Spracherkennung muss also mehr die Dynamik und die Veränderung eines Signals innerhalb eines relativ großen Zeitabschnitts im Vordergrund

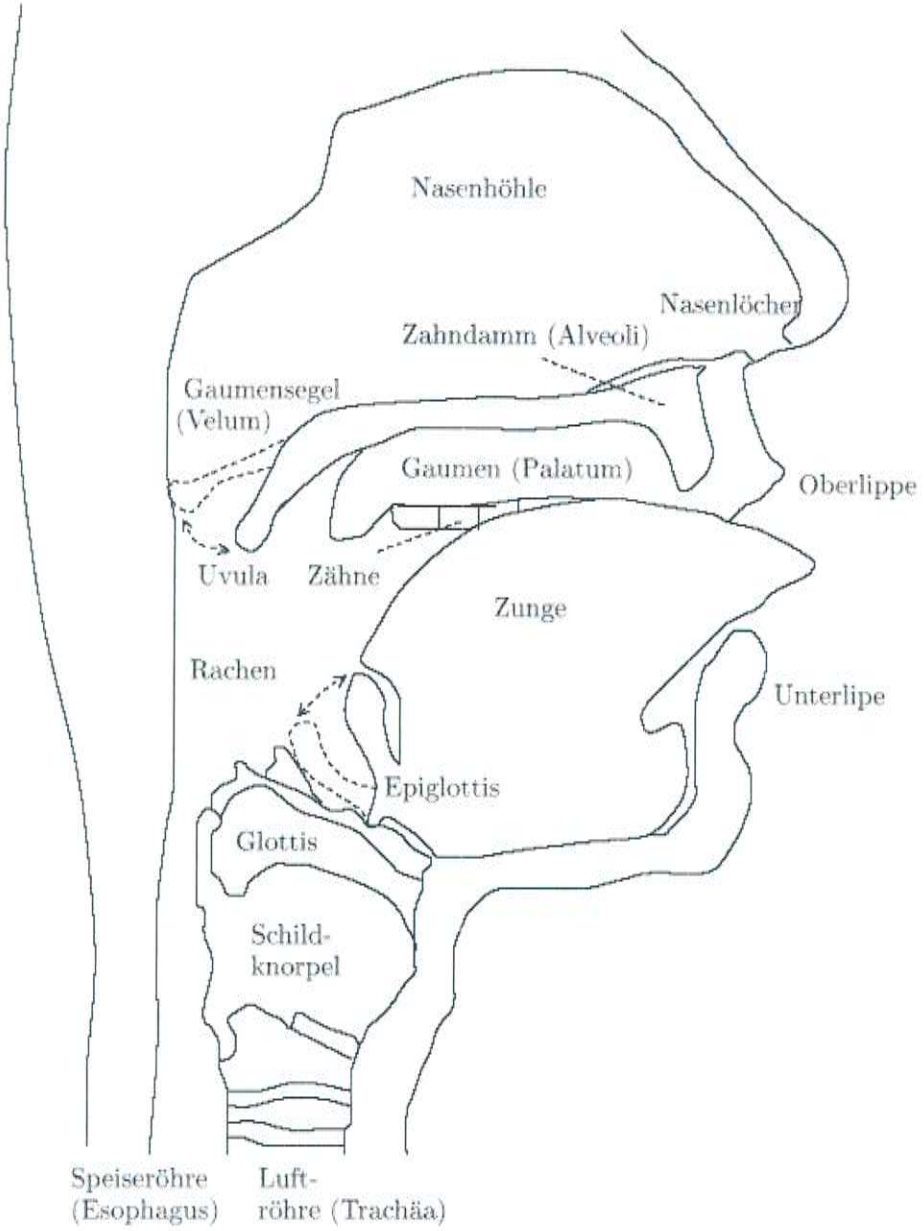

Abbildung 3.3: *Anatomie des Artikulationsapparates nach [Rog03]*

stehen. Quasi-stationäre Signale, die zwar einen Laut, aber keine Muskelbewegungen erzeugen, können dagegen nicht berücksichtigt werden, da es ohne Muskelbewegungen auch keine Veränderung der EMG-Signale gibt.

Diese Muskelbewegungen sind allerdings nicht auf die Phonemgrenzen beschränkt, sondern können auch über längeren Zeitabschnitte verlaufen. Sie beginnen teilweise bis zu 50 ms vor der hörbaren Sprache [JMHSW06]. Dies muss bei der Wahl der Wortuntereinheiten berücksichtigt werden.

Ebenfalls berücksichtigt werden muss, dass die zur Formung eines bestimmten Lautes notwendigen Muskelbewegungen stärker von der Position des Lautes innerhalb des Wortes und der vorherigen und nachfolgenden Laute abhängen. Soll beispielsweise ein

Vokal am Anfang eines Wortes ausgesprochen werden, so muss zunächst der Mund geöffnet werden. Wird der Vokal dagegen in der Mitte oder am Ende eines Wortes ausgesprochen, ist der Mund bereits offen und muss anschließend ggf. geschlossen werden. Das bedeutet also, dass einige für einen Vokal notwendigen Muskelbewegungen bereits einige Zeit vor diesem Vokal stattfinden können. Für die praktische Anwendung bedeutet das, dass sich die Laute je nach Kontext stärker unterscheiden, als es bei der akustischen Spracherkennung der Fall ist, für diesen Kontext jedoch nicht nur die direkten Nachbarn berücksichtigt werden müssen.

3.4 Automatische Segmentierungskorrektur

Da bei den Aufnahmen trotz Verwendung eines vom Sprecher zu betätigenden Knopfes der tatsächliche Sprechakt niemals millisekundengenau direkt nach dem Start beginnen kann, muss der stille Vorlauf erkannt und abgeschnitten werden, um das Training der Phoneme nicht zu beeinflussen.

3.4.1 Stille-Modell

Bereits Maier-Hein [MHMSW05] hat damit experimentiert, zu Beginn und Ende des Wortes ein „Stille"-Modell zu modellieren. Dieses Modell wurde durch ein „Stille"-Pseudo-Wort trainiert, und der Spracherkenner sollte dieses Modell vor und nach dem Wort erkennen und dadurch bestimmen, wie lange der Vor- und Nachlauf ist.

3.4.2 Eigener Spracherkenner zur Sprachdetektion

Bei der akustischen Spracherkennung wird für die Erkennung, ob ein bestimmtes Signal „Sprache" enthält, teilweise auch ein eigener Erkenner verwendet, der nur die Wörter „Sprache" und „Nicht-Sprache" erkennen kann. Ein derartiger Erkenner verwendet typischerweise eine einfachere Vorverarbeitung und verbraucht insgesamt weniger Rechenzeit, um im Echtzeitbetrieb den eigentlichen rechenzeitintensiven Spracherkenner nur dann zu starten, wenn tatsächlich Sprache aufgenommen wird [Rog03].

Auch in dieser Arbeit wurde ein derartiger Sprachdetektor getestet. Da sich bei einer visuellen Analyse der aufgenommenen Signale der Vor- und Nachlauf durch insgesamt geringere Amplituden repräsentieren, wird als Vorverarbeitung statt der Fourier-Transformation die Signalenergie berechnet. Zusätzlich wird der Mittelwert im Zeitbereich berechnet. Für die Initialisierung wurde der Flatstart-Algorithmus verwendet. Dafür wird zunächst der Mittelwert über alle Aufnahmen berechnet und anschließend

alle Modelle mit diesem Mittelwert initialisiert. Nach insgesamt vier Trainingsdurchläufen mit dem Expectation-Maximization-Algorithmus werden mit den derart trainierten Modelle alle Aufnahmen in einen „Nicht-Sprache", einen „Sprache" und einen „Nicht-Sprache"-Teil aufgeteilt und diese Aufteilung abgespeichert.

3.5 Vorverarbeitungen

3.5.1 Fenstergröße

Für die Bildung von Feature-Vektoren ist es notwendig, das Eingangssignal in definierte Abschnitte (Fenster) zu unterteilen, die separat betrachtet werden können. Diese Fenster müssen einerseits groß genug sein, um genügend Daten zu enthalten. Insbesondere bei Phonem-basierten Spracherkennern dürfen diese Fenster aber andererseits auch nicht zu groß sein, damit die einzelnen Phoneme noch getrennt betrachtet werden können. Da die kürzesten Phoneme im Schnitt 40 ms lang sind, sollte das Fenster kleiner als 40 ms sein, um diese Phoneme getrennt erfassen zu können.

Da bei der EMG-Spracherkennung zunächst die Erkennung von einzelnen Wörtern im Vordergrund stand, war es hier zunächst wichtiger, genug Fourier-Koeffizienten zu gewinnen. Deshalb hat sich bei vergleichbaren Arbeiten eine Fenstergröße von 54 ms bis 64 ms durchgesetzt.

In dieser Arbeit wurde untersucht, ob für die Erkennung von Phonemen eine kleinere Fenstergröße sinnvoller sein könnte. Untersucht wurden Größen von 16ms bis 54ms.

3.5.2 Fourier-Transformation

Ähnlich wie bei der akustischer Spracherkennung werden auch bei der EMG-Spracherkennung Vorverarbeitungen verwendet, die den Frequenzbereich betrachten, beispielsweise die Fourier-Transformation bzw. die Short Time Fourier Transformation (STFT). Da für die EMG-Spracherkennung insbesondere die durch die Muskelbewegungen verursachten Veränderungen erkannt werden sollen, werden nur die Veränderungen der STFT-Werte als delta-Werte betrachtet.

Da die Anzahl der Koeffizienten bei der Fourier-Transformation direkt von der Zahl der betrachteten Datenwerte abhängt, spielt die Fenstergröße eine entscheidende Rolle. Je größer das Fenster ist, desto mehr Datenwerte werden gemeinsam betrachtet und desto mehr Koeffizienten können berechnet werden.

3.5.3 Mittelwert im Zeitbereich

Da der Zeitbereich für die EMG-Spracherkennung relevante Informationen enthält, wird er bei vergleichbaren Arbeiten berücksichtigt. Bei der Arbeit von Maier-Hein beispielsweise wird der Mittelwert des Signals im Zeitbereich innerhalb des Betrachtungsfensters ermittelt. Da jedoch nicht nur der aktuelle Wert von Interesse ist, sondern insbesondere dessen Veränderung Rückschlüsse auf die An- oder Abspannung zulässt, wurde in dieser Arbeit ein sogenanntes „Time Domain Context" (TDC) Feature untersucht, bei dem der Mittelwert innerhalb des aktuelles Fensters mit dem Mittelwert des Fensters 40 ms vor und 40 ms nach dem aktuellen Fenster verglichen wurde.

3.5.4 Independent Component Analysis

Die in dieser Arbeit verwendete Oberflächen-Elektromyografie hat gegenüber der Einzelfaserelektromyografie den Nachteil, dass nicht die Aktivität einzelner motorischer Einheiten erfasst werden kann [Wikc], sondern nur ein Gemisch von verschiedenen Muskelfasern oder teilweise sogar verschiedener Muskeln. Da gerade im Gesichtsbereich zahlreiche verschiedene Muskeln in unmittelbarer Nähe verlaufen, ist es also nicht möglich, einzelne Muskeln zu erfassen, an jedem Messpunkt liegt also nur ein Gemisch von verschiedenen Muskeln an.

Aus diesem Grund wurde versucht, durch die Anwendung der Independent Component Analysis (ICA) auf die Signale im Zeitbereich aus diesem Signalgemisch auf mathematischem Weg einzelne Muskelsignale zu isolieren. Bei der ICA als Spezialfall der *Blind source separation* kann ein multivariantes Signal in einzelne statistisch unabhängige Einzelsignale separiert werden, deren Verteilung allerdings nicht der Gaußkurve entsprechen darf [Com94].

Für die Anwendung der ICA wurde in dieser Arbeit zunächst der globale Mittelwert über alle Trainingsaufnahmen einer Sitzung berechnet, und anschließend durch die Hauptkomponentenanalyse [Pea01] die Dimension der Vektoren bestimmt, auf die die 7 Eingangssignale reduziert werden können, während ein möglichst großer Anteil der ursprünglichen Information entsprechend der Formel

$$I(k) = \frac{\sum_{i=1}^{k} D(i,i)}{\sum_{j=1}^{n} D(j,j)} \tag{3.1}$$

abgedeckt wird [Hyv99]. D stellt dabei die Diagonalmatrix der sortierten Eigenwerte dar, k steht für die Anzahl der beibehaltenen Basisvektoren und n entspricht der ursprünglichen Anzahl der Basisvektoren. Anschließend wurde die ICA-Matrix berechnet und damit alle Signale multipliziert, bevor im nächsten Schritt die normale Vorverarbeitung auf der nun reduzierten Kanal-Anzahl ausgeführt wurde.

3.5.5 Lineare Diskriminanzanalyse

Um die Zahl der Werte innerhalb eines Feature-Vektors der Vorverarbeitung für die anschließende Klassifizierung zu verringern und so bei möglichst gleichbleibendem Informationsgehalt eine bessere Diskriminierung zu ermöglichen, hat sich in der akustischem Spracherkennung die lineare Diskriminanzanalyse (engl. Linear Description Analysis, LDA) durchgesetzt. Auch in dieser Arbeit wurden Experimente mit LDA durchgeführt, bei der nach einem ersten regulären Trainingsdurchlauf mit den aus den trainierten Modellen berechneten Labels die LDA-Matrizen für die einzelnen Modelle berechnet wurden.

3.6 Einheiten zur Wortmodellierung

3.6.1 Einzelwörter

Die einfachste Form der Wortmodellierung besteht darin, jedes Einzelwort als eigenes Modell zu betrachten. In diesem Fall braucht man kein kompliziertes und möglicherweise fehlerhaftes Aussprachewörterbuch und hat bei dem hier durchgeführten sprecherabhängigen Training keine Probleme mit Akzenten und Dialekten. Es ist ebenfalls irrelevant, wie sich die Muskelbewegungen im Verhältnis zu den Phonemen oder Silben verhalten, da diese Einheiten eben nicht berücksichtigt werden. Allerdings ist die logische Konsequenz, dass nur trainierte Wörter erkannt werden können. Wenn man zusätzliche Wörter erkennen möchte, müssen diese auch trainiert werden. Deshalb ist diese Form der Wortmodellierung für die praktische Anwendung ungeeignet, da man zu viele Trainingsdaten bräuchte, um alle Wörter eines typischen Spracherkennungs-Wörterbuchen mit 40.000 Wörtern erkennen zu können.

3.6.2 Silben

Bei der akustischen Spracherkennung wurde bereits mit Silben als Einheit experimentiert: Ganapathiraju erzielte mit einem silbenbasierten Erkenner bei der Standard-SWITCHBOARD-Evaluation deutlich bessere Ergebnisse als mit einem vergleichbaren auf Triphonen basierenden Erkenner [GHP+01].

Diese Verbesserung erscheint realistisch, da eine Silbe eher als ein Phonem als akustische Einheit betrachtet werden kann. Eine Silbe entspricht einem eindeutigen Bewegungsablauf innerhalb des Vokaltraktes und deshalb kann insbesondere auch bei der EMG-Spracherkennung eine Verbesserung gegenüber Phonemen erwartet werden.

Problematisch sind aber die benötigten Trainingsdaten: Um 70.000 englische Wörter abzudecken, müssen 9.023 Silben trainiert werden. Für die EMG-Spracherkennung ist es unrealistisch, mit aktuellen sitzungsabhängigen Systemen entsprechend viele Trainingsdaten zu generieren.

3.6.3 Phoneme

Phoneme haben gegenüber den oben diskutierten Modellen den Vorteil, dass sie nur relativ wenig Trainingsdaten benötigen. Im einfachsten Fall werden lediglich Trainingsdaten für 45 Phoneme benötigt. Da Wörter typischerweise aus mehreren Phonemen bestehen, können pro Wort auch mehrere Phoneme trainiert werden.

Aus diesem Grund haben sich die Phoneme als Einheit zur Wortmodellierung in der akustischen Spracherkennung durchgesetzt. Insbesondere im Bezug auf EMG-Spracherkennung muss jedoch geprüft werden, ob Phoneme möglicherweise zu kleine Einheiten sind, um die teilweise längerfristigen Muskelbewegungen korrekt zu erfassen. Eine Lösung dieses Problems kommt auch aus der akustischen Spracherkennung: Indem man kontextabhängige Phoneme verwendet, kann man berücksichtigen, dass ein Phonem je nach vorherigem und nachfolgendem Phonem anders klingt beziehungsweise durch andere Muskelbewegungen erzeugt wird. Dadurch entsteht zwar zunächst ein Bedarf nach mehr Trainingsdaten, diesem kann man jedoch durch Clustern von ähnlichen Phonemmodellen entgegenwirken.

3.7 Kontextabhängigkeit

Bereits bei der akustischen Spracherkennung spielt der Kontext eine wichtige Rolle, da die Laute abhängig vom Vorgänger/Nachfolger unterschiedlich ausgesprochen werden. Bei der EMG-Spracherkennung ist Kontext noch wichtiger, da die Muskelbewegung nicht unbedingt mit den benachbarten Phonem-/Silbengrenzen korrelieren, sondern sich teilweise über einen längeren Zeitraum erstrecken. Insbesondere kann der Spannungsabfall nach der Entspannung in nachfolgende Lauteinheiten einfliessen.

Werden isoliert aufgenommene Einzelwortmodelle betrachtet, erübrigen sich derartige Überlegungen jedoch, da in diesem Fall keine Vorgänger und Nachfolger existieren. Deshalb braucht man in diesem Fall auch keine Kontextabhängigkeit zu betrachten.

Bei Silben- bzw. Phonemmodellen dagegen existieren direkte Vorgänger und Nachfolger, aus denen sich die verwendeten Kontexteinheiten ergeben. Für die Verwendung von Modellen, die den Kontext über einen größeren Zeitraum als bei Triphonen betrachten

(also beispielsweise Quintphone) standen in dieser Arbeit jedoch zuwenig Trainingsdaten zur Verfügung. In dem für diese Arbeit verwendeten Erkenner werden für jedes Triphon eigene Modell-Gewichte berechnet; der Codebuch-Eintrag bleibt jedoch identisch.

Damit diese neuen Modelle auch mit einer begrenzten Menge an Trainingsdaten trainiert werden können, werden die Phonemmodelle zusätzlich noch geclustert. Dabei wird ein bei der akustischen Spracherkennung übliches Frageset verwendet (vgl. Kapitel C.4); für die neuen Modelle werden jeweils eigene neue Codebuch- und Gewichtseinträge erzeugt.

Auf das Clustern der Silbenmodelle wurde verzichtet, da dafür zu wenig Trainingsdaten zur Verfügung standen.

3.8 Unterschiede zwischen isoliert aufgenommenen Einzelwörtern und kontinuierlicher Sprache

Aus Sicht eines Spracherkenners dürfte der Hauptunterschied zwischen isoliert aufgenommenen Einzelwörtern und kontinuierlicher Sprache die Tatsache sein, dass bei isolierten Wörtern die Zahl der zu erkennenden Wörter exakt feststeht. Da stets exakt ein Wort erkannt werden soll, können die Modelle folglich über die gesamte Aufnahme gelegt werden. Bei kontinuierlicher Sprache dagegen steht eben nicht fest, wie viele Wörter in der Aufnahme enthalten sind, und die Segmentierung gestaltet sich entsprechend schwerer. Trotz konzentrierter Aussprache sind außerdem Pausen zwischen den Wörtern nicht immer zu vermeiden. Es kann ebenso passieren, dass Laute oder Silben verschluckt werden, da bei kontinuierlicher Sprache insgesamt weniger präzise gesprochen wird als bei der Aussprache einzelner Wörter.

Ein anderer Unterschied betrifft die Kontextabhängigkeit, die nun auch über Wortgrenzen hinweg berücksichtigt werden muss. Zum einen vergrößert sich dadurch die Zahl der Modelle, andererseits muss die Wortgrenze als spezielles Ereignis innerhalb der Muskelbewegungssequenz berücksichtigt werden.

4 Datenauswahl, -sammlung und -verarbeitung

Um die verschiedenen Ansätze und Ideen aus dem vorangegangenen Kapitel zu testen, wurden innerhalb zahlreicher Sitzungen mit Testpersonen Daten aufgenommen, die einerseits innerhalb definierter Grenzen liegen, andererseits aber so realistisch wie möglich einer zukünftigen Anwendung nahe kommen sollten.

In diesem Kapitel wird der im Wesentlichen auf der Arbeit von Maier-Hein [MH05] aufbauende Versuchsaufbau beschrieben, mit dem die Daten für die in den nächsten Kapiteln aufgeführten Experimente aufgenommen wurden. Dabei werden auch Rahmenbedingungen diskutiert, die bei der Interpretation der Versuchsergebnisse zu beachten sind. Des Weiteren werden die Wörter- und Satzlisten vorgestellt und diskutiert, die die Basis für die Experimente bildeten.

4.1 Sensorik

Um vergleichbare Ergebnisse zu bekommen, ist es wichtig, die Elektroden stets an den gleichen Positionen im Gesicht aufzukleben. In der Arbeit von Maier-Hein wurde die Idee vorgestellt, eine Gipsmaske zu verwenden, um diese gleiche Positionierung zu erreichen. Dieses Verfahren wurde getestet und angewendet und wird im Folgenden kurz dargestellt.

Um die Gesichtsmuskeln levator anguli oris, zygomaticus major, platysma, depressor anguli oris und anterior belly of the digastric (vgl. Abbildung 2.1) zu erfassen, werden 11 Elektroden wie in Abbildung 4.3 zu sehen positioniert.

Um eine vergleichbare Positionierung zu garantieren, wird eine für jeden Sprecher angefertigte Gipsmaske (vgl. Abb. 4.1) verwendet. Durch Löcher in dieser Maske wird die Position der Elektroden mit einem Stift auf der Haut markiert (vgl. Abb. 4.2), anschließend wird die Position mit Alkohol gereinigt und die mit Gel versehene Elektrode aufgeklebt.

Die Signale werden anschließend verstärkt und mit einer Abtastrate von 600 Hz mit Hilfe des Varioport-Systems digitalisiert. Das System ist über eine galvanische Trennung an den seriellen Port eines PCs angeschlossen, der die Daten aufnimmt und verarbeitet.

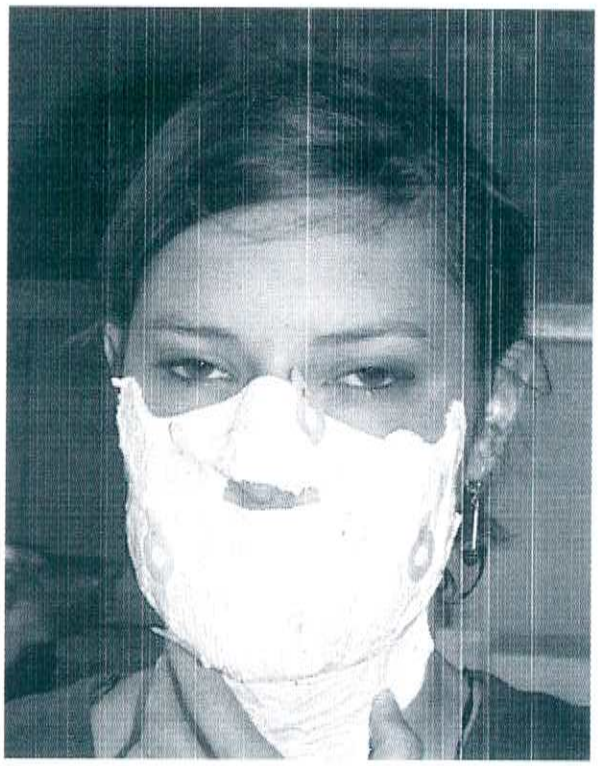

Abbildung 4.1: *Abbildung der Gipsmaske für die Markierung der Elektroden-Positionen*

Zu Beginn wird die in 2.6 beschriebene Vorverarbeitung verwendet, die in den nachfolgend beschriebenen Experimenten verbessert wurde.

4.2 Klassifikation

Da in dieser Arbeit auch Wortuntereinheiten untersucht werden sollten, wurde eine von Maier-Heins Arbeit abweichende Klassifikation verwendet. Um die Ergebnisse der verschiedenen Wortuntereinheiten vergleichen zu können, wurden pro Phonem 3 HMM-Zustände verwendet. Bei der Verwendung von Silben- oder Wortmodellen wurde dementsprechend die Anzahl der Phoneme berechnet und anschließend dreimal so viele HMM-Zustände zur Modellierung des betreffenden Modells verwendet.

4.3 Testpersonen

Zwei Testpersonen stellten sich für die Aufnahme der Daten zur Verfügung: Ein männlicher Sprecher (S1) und eine weibliche Sprecherin (S2). Beide Personen besaßen keine bekannten Sprachfehler. Während S2 Englisch flüssig sprach, hatte S1 einen ungarischen Akzent.

Abbildung 4.2: *Abbildung der Markierungen der Elektroden-Positionen*

4.4 Versuchsaufbau

4.4.1 Positionierung der Elektroden

Bei der momentanen Konfiguration ist die Positionierung der Elektroden durchaus kritisch, da bei falscher Positionierung Störsignale (z.B. der Herzschlag durch die Schlagadern) auftreten können. Deshalb wurde die korrekte Positionierung vor jedem Versuch durch eine Analyse der Signale über das Monitor-Programm überprüft.

4.4.2 Wortlisten

Da ein Ziel dieser Arbeit die Erkennung von unbekannten Wörtern sowohl auf Phonem- als auch auf Silbenbasis ist, musste eine spezielle Liste englischer Ausdrücke erstellt werden. Diese Liste sollte in eine Trainings- und eine Testliste geteilt werden können, und in der Testliste durften nur Phoneme oder Silben auftreten, die auch in den Wörtern der Trainingsliste enthalten sind und daher bereits trainiert wurden.

Folgende Liste wurde bei den Experimenten verwendet: *all, alright, also, alter, always, center, early, earning, enter, entertaining, entry, envy, euro, gateways, leaning, li, liter, n, navy, right, rotating, row, sensor, sorted, sorting, so, tree, united, v, watergate, water, ways* (die Wörter des Trainingssets sind unterstrichen).

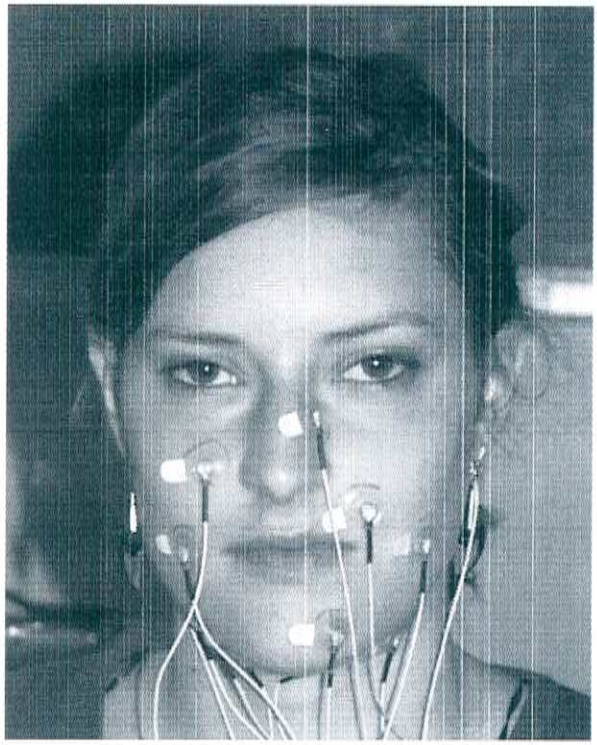

Abbildung 4.3: *Abbildung der Elektroden auf dem Gesicht*

Um die Stille zu Beginn und am Ende der Aufnahme erkennen zu können, wurde ein Pseudo-Wort „Stille" trainiert.

Diese Liste wird im Folgenden als *EINZELWÖRTER* bezeichnet.

Für die Untersuchung des Silben-Korpus müssen die Wörter in Silben aufgeteilt werden. Insgesamt werden 21 Silben verwendet, nämlich: *AOL, EHN, ER, GEYT, LIY, NAY, NIXNG, RAYT, ROW, SAOR, SEHN, SOW, TAXR, TAXD, TEY, TIXNG, TRIY, VIY, WAO, WEYZ, YUH.*

Um die Erkennung von kontinuierlicher Sprache zu untersuchen, wurden zwei weitere Wortlisten erzeugt: Zunächst wurden 38 phonetisch gleichverteilte Sätze aus dem Wall-Street-Korpus ausgewählt, die zum Training verwendet wurden. Anschließend wurde versucht, zum einen die zehn Ziffern von *zero* bis *nine* zu erkennen sowie die Wörter aus dem EINZELWÖRTER-Korpus. Auch hier wurde das Pseudo-Wort „Stille" trainiert. Dieser Korpus wird als *KONT1* bezeichnet (vgl. Kapitel C.2).

Zum Schluss wurde untersucht, inwieweit komplette Sätze erkannt werden können: Dazu wurden die 38 phonetisch gleichverteilten Sätze mit 24 neuen Sätzen aus einem Newspaper-Korpus kombiniert. Diese neuen Sätze wurden verschiedenen Artikeln der Online-Ausgabe der Zeitung New York Times (www.nytimes.com) entnommen. Dieser Korpus wird als *KONT2* bezeichnet (vgl. Kapitel C.3).

4.4.3 Versuchsablauf

Die Wörter aus der Trainingsliste wurden zufällig sortiert und den Testpersonen einzeln am Bildschirm gezeigt. Per Knopfdruck konnten diese die Aufnahme starten und stoppen.

Für das „Stille"-Wort sollten die Testpersonen ca. 2 Sekunden lang keine Gesichtsmuskeln bewegen.

Jede Einheit wurde 20 mal aufgenommen.

4.4.4 Verwendete Programme

Für Training und Erkennung wurde das Spracherkenner-Framework Janus Recognition Toolkit [FGH+97] verwendet.

Das Training wurde folgendermaßen durchgeführt:

1. Kontextunabhängiges Training
   a) Lineare Segmentierung
   b) Initialisierung der Codebücher durch *KMeans*
   c) Vier Trainingsdurchläufe mit *EMTraining*
   d) Ausschreiben der Labels
2. Kontextabhängiges Training
   a) Berechnung der kontextabhängigen Triphone
   b) Initialisierung durch ausgeschriebene Labels
   c) Initialisierung der Codebücher durch *KMeans*
   d) Vier Trainingsdurchläufe mit *EMTraining*
   e) Ausschreiben der Labels
   f) *Clustering-and-Split*
3. Kontextabhängiges Training der geclusterten Modelle
   a) Initialisierung durch ausgeschriebene Labels
   b) Initialisierung der Codebücher durch *KMeans*
   c) Vier Trainingsdurchläufe mit *EMTraining*
   d) Ausschreiben der Labels

Nach dem kontextunabhängigen (context independent, CI) Training, nach dem kontextabhängigen (context dependent, CD) Training und nach dem kontextabhängigen Training der geclusterten (CL) Modelle wurde jeweils die Erkennungsrate gemessen.

5 Training mit isoliert aufgenommener Sprache

In diesem Kapitel werden die Experimente mit isolierten Einzelwörtern vorgestellt. Durch diese Experimente soll das Baseline-System im Bezug auf die Vorverarbeitung, Klassifikation, Modellbildung und Kontextabhängigkeit im Hinblick auf die Erkennung trainierter Wörter verbessert werden. Dabei wird insbesondere auch untersucht, wie sich die Erkennungsraten bei einem über die 10-Ziffern hinausgehenden Vokabular verhalten. Nach diesen Voruntersuchungen soll schwerpunktmäßig analysiert werden, ob es auch möglich ist, untrainierte Wörter zu erkennen. In diesem Zusammenhang wird untersucht, welche Wortmodelle sich dafür am besten eignen.

Für alle folgenden Experimente wurden von beiden Testpersonen innerhalb von jeweils 5 Sitzungen an verschiedenen Tagen und zu verschiedenen Tageszeiten alle 32 Wörter sowie das Pseudo-Wort Stille 20 mal aufgenommen.

5.1 Erkennung von bekannten Wörtern

Für die Erkennung von bekannten Wörtern wurden die Aufnahmen in ein Trainingsset und ein Testset unterteilt, so dass jedes Set 10 Aufnahmen jedes Wortes enthielt. Nach einem kompletten Trainings- und Testdurchlauf wurden beide Sets vertauscht, und ein erneuter Trainings- und Testdurchlauf durchgeführt.

5.1.1 Automatische Segmentierungskorrektur

Zunächst wurde untersucht, ob durch eine automatische Korrektur der manuellen Segmentierung eine Verbesserung der Segmentierung und der Erkennungsleistung insgesamt möglich ist.

Dazu wurde auf Basis der Einzelwortmodelle und des Baseline-Systems insgesamt drei verschiedene Systeme untersucht:

1. ein Erkenner ohne automatische Korrektur
2. ein Erkenner mit automatischer Korrektur
3. ein Erkenner, der zur Segmentierungskorrektur die Labels eines eigenen Erkenners verwendet

Der Versuch wurde anschließend mit einer verkleinerten Fenstergröße wiederholt.

Fenstergröße	ohne Korrektur	mit Korrektur	mit eigenem Erkenner
54 ms	61,7	64,8	54,7
27 ms	76,7	78,1	64,8

Tabelle 5.1: *Wortkorrektheiten (in %) abhängig von der Segmentierungskorrektur.*

Die Ergebnisse in Tabelle 5.1 zeigen, dass die automatische Korrektur zum einen die Erkennung verbessert, während aus der Korrektur durch einen eigenen Erkenner deutlich schlechtere Ergebnisse als beim Baseline-System resultieren. Zusätzlich würde dieser Ansatz auch eine deutlich höhere Komplexität des Systems bewirken, da zwei verschiedene Erkenner nacheinander verwendet werden müssten.

5.1.2 Untersuchung der Vorverarbeitungen

In den folgenden Experimenten soll zunächst ermittelt werden, wie gut die einzelnen Vorverarbeitungen funktionieren. Dazu werden die verschiedenen Verarbeitungen jeweils für das Einzelwortmodell, für das Silbenmodell und das Phonemmodell getestet. Bei dem Silben- und Phonemmodell werden zusätzlich kontextabhängige Modelle getestet. Bei den Phonemmodellen werden die kontextabhängigen Modelle als letzter Schritt geclustert.

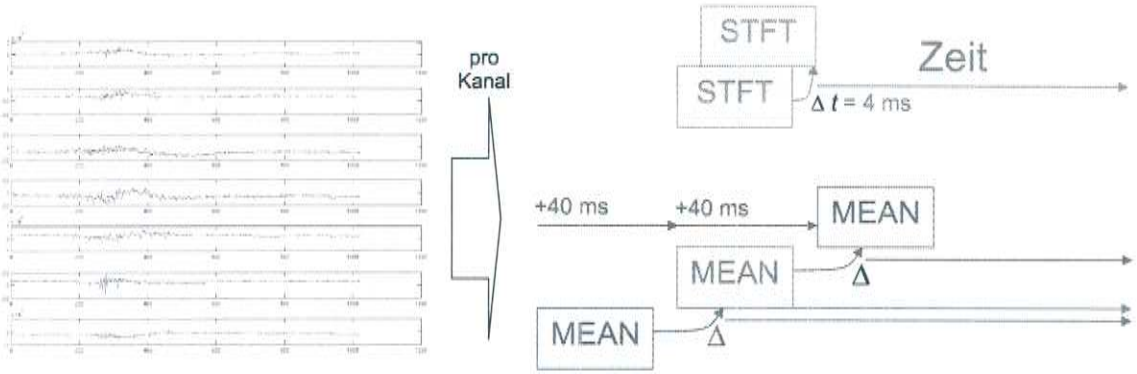

Abbildung 5.1: *Schematische Abbildung der Vorverarbeitungen*

Bewertet wurde dabei die Wortkorrektheit, d.h. der Anteil der korrekt erkannten Wörter im Bezug auf das gesamte Testset.

Im Vergleich mit dem Baseline-System ergibt sich bei dessen Konfiguration (Fenstergröße 54 ms, Wortmodelle, ohne TDC) mit 64,8% ein schlechterer Wert als innerhalb

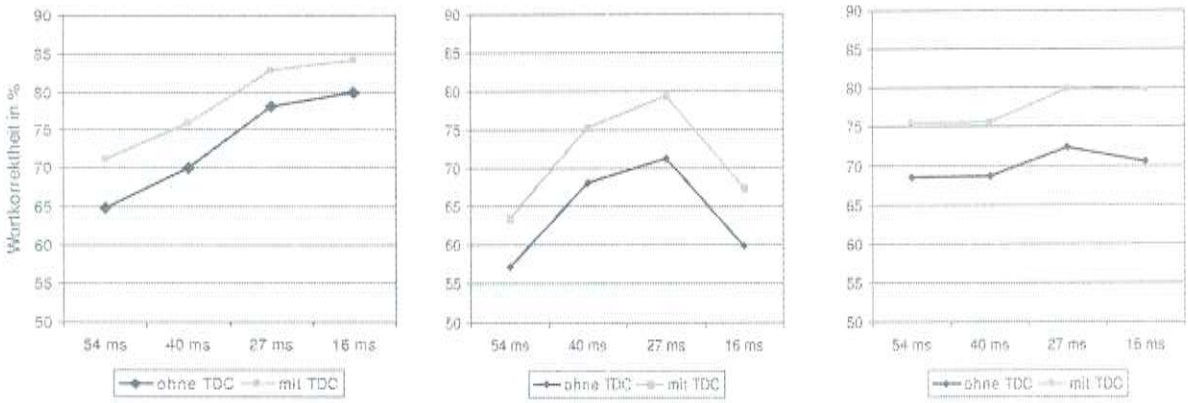

Abbildung 5.2: *Wortkorrektheiten für den Split-Test.*

Vorverarbeitung	Wörter	Silben (CI)	Silben (CD)	Phoneme (CI)	Phoneme (CD)	Phoneme (CL)
Fenstergröße 54 ms	64,8	51,0	57,1	57,3	64,0	68,5
40 ms	69,9	62,0	68,0	55,6	61,5	68,7
27 ms	78,1	64,6	71,2	57,7	62,9	72,4
16 ms	79,9	54,7	59,9	54,1	59,6	70,6
54 ms & TDC 40 ms	69,9	58,0	63,5	65,3	71,7	75,5
40 ms & TDC 40 ms	75,8	69,5	75,2	64,0	69,3	75,7
27 ms & TDC 40 ms	**82,9**	**73,3**	**79,3**	**67,3**	**72,6**	**79,8**
16 ms & TDC 40 ms	84,1	61,0	67,4	63,6	68,9	79,7
27 ms & TDC 27 ms	82,4	72,1	78,2	66,5	71,9	79,4
27 ms & TDC 54 ms	83,0	74,3	79,9	67,1	72,1	80,9
27 ms & TDC 60 ms	82,9	74,5	80,5	67,1	72,4	80,4

Tabelle 5.2: *Wortkorrektheiten (in %) für den Split-Test.*

der ursprünglichen Arbeit ermittelt wurde. Es muss jedoch berücksichtigt werden, dass bei dieser Arbeit mit einem mehr als dreimal so großen Vokabular gearbeitet wird.

Die Ergebnisse zeigen, dass die Experimente mit kleineren Fenstern grundsätzlich bessere Erkennungsraten erreichen als bei dem ursprünglichen 54 ms großen Fenster. Bei den Einzelwortmodellen ist dieser Trend durchgängig und die besten Ergebnisse werden mit den kleinsten 16 ms großen Fenstern erzielt. Nicht so dagegen bei den Silbenmodellen: Dort ist der Trend bis zu 27 ms auch durchgängig, bei 16 ms gibt es allerdings einen starken Abfall weit unter die Ergebnisse des 40 ms Fensters.

Das neue TDC-Feature bewirkt in allen Konstellationen eine weitere Verbesserung von 4,5 bis 10%. Nachdem zunächst mit einem Betrachtungshorizont von 40 ms für das TDC-Feature experimentiert worden ist, wurde anschließend getestet, ob kleinere oder größere Werte besser geeignet wären. Dabei ergab sich ein uneinheitliches Bild: Insbesondere bei größeren Werten gab es teilweise geringfügig bessere Erkennungsraten, in einzelnen Zwischenergebnissen jedoch auch schlechtere Ergebnisse. Aus diesem Grund wurde im Folgenden mit einem Betrachtungshorizont von 40 ms gearbeitet und es wird

vorgeschlagen, in zukünftigen Arbeiten mit einem variablen phonem- und kontextabhängigen Betrachtungshorizont zu arbeiten.

Dieses Experiment zeigt weiterhin, dass die Verwendung von kontextabhängigen Modellen eine deutliche Verbesserung der Erkennungsraten bewirkt. Bei Silbenmodellen ergab es eine Verbesserung von 6 bis 6,5%, bei den Phonemmodellen von ca. 5 bis 6,5%. Durch die Verwendung von geclusterten Phonem-Modellen gibt es dort eine weitere Verbesserung von 4,5 bis 11%. Insbesondere bei kleineren Fenstergrößen sind die Verbesserungen durch die Verwendung von kontextabhängigen Modellen deutlich größer, da jetzt eine präzisere Segmentierung möglich ist und die Übergänge präziser erkannt werden können.

Da die Verkleinerung des Fensters auf 27 ms und die Benutzung des TDC-Features eine deutliche Verbesserung bewirken, werden sämtliche folgenden Experimente auf Basis dieser verbesserten Vorverarbeitung durchgeführt.

5.1.3 Lineare Diskriminanzanalyse

Mithilfe der LDA wurde versucht, die Zahl der Koeffizienten zu reduzieren. Dabei wurde mit einer unterschiedlichen Anzahl von Ausgangskoeffizienten experimentiert. Allerdings stellte sich heraus, dass nach Anwendung der LDA und dem regulären Trainingsdurchlauf keine robuste Erkennung mehr möglich war. Es wird vermutet, dass das TDC-Feature insbesondere durch die feste Größe des Betrachtungshorizonts keine stabile LDA-Matrix für alle Aufnahmen eines Phonemes erlaubt, da die tatsächliche Kontextlänge je nach Position und Kontext variiert.

5.1.4 Independent Component Analysis

Bei der Anwendung der ICA wurde zunächst mit der PCA die Dimension berechnet, die notwendig ist, um einen möglichst großen Anteil der Information des 7-kanäligen Signals mit einer geringeren Dimension abzudecken. Dabei ergaben sich die in Tabelle 5.3 aufgeführten Werte.

Es stellte sich jedoch heraus, dass die Berechnung der ICA-Matrix nur für wenige Sitzungen möglich war, bei dem Großteil der Sitzungen war dagegen keine Konvergenz erreichbar. Als Grund dafür wird vermutet, dass sich innerhalb der Sitzungen die Signalzusammensetzungen an den einzelnen Messpunkten durch unterschiedliche Hauttemperatur und -feuchtigkeit ändert.

Informationsabdeckung	Dimension
>50%	1
>60%	2
>70%	3
>80%	3
>85%	4
>90%	5
>95%	5
>96%	5
>97%	6
>98%	6
>99%	6
>99,9%	7

Tabelle 5.3: *Dimension in Abhängigkeit von der Abdeckung der Information des Signals im Zeitbereich*

5.2 Erkennung von unbekannten Wörtern

Um die Erkennung von unbekannten Wörtern zu testen, wurde die Wörterliste in zwei Hälften aufgeteilt. Um auch mit Silbenmodellen arbeiten zu können, wurde sie derart geteilt, dass jede Silbe mindestens einmal im Trainingsset und einmal im Testset enthalten war. Da jetzt nur noch 16 Wörter zum Training und 16 Wörter für die Erkennung zur Verfügung standen, sank einerseits die Zahl der kontextabhängigen Trainingsinstanzen, gleichzeitig jedoch auch die Perplexität, da nun entsprechend weniger Wörter unterschieden werden mussten. Die Zahl der Trainingsaufnahmen insgesamt bleibt jedoch gleich, da jetzt 20 statt bisher 10 Aufnahmen fürs Training verwendet wurden.

Wenn man untersuchen möchte, ob für diese Aufgabe Silben- oder Phonemmodelle besser geeignet sind, muss bei diesem Vergleich berücksichtigt werden, dass Phoneme in den Trainingswörtern insgesamt häufiger vorkommen, da sie im Vergleich zu Silben eine kleinere Einheit sind. So waren die meisten Silben nur in einem Trainingswort enthalten, und wenige andere Silben traten insgesamt zweimal auf. Als Durchschnittswert ergaben sich 1,19 Trainingsbeispiele pro Silbenmodell.

Die Phoneme traten dagegen bis zu sechsmal in den Trainingswörtern auf, hier ergab sich ein Durchschnittswert von 2,78 Trainingsbeispielen pro Phonemmodell. Um trotzdem einen fairen Vergleich zu ermöglichen, wurde die Anzahl der Trainingsaufnahmen für die Phonem-Modelle reduziert, statt 20 Aufnahmen wurden nur noch 8 Aufnahmen jeden Wortes verwendet.

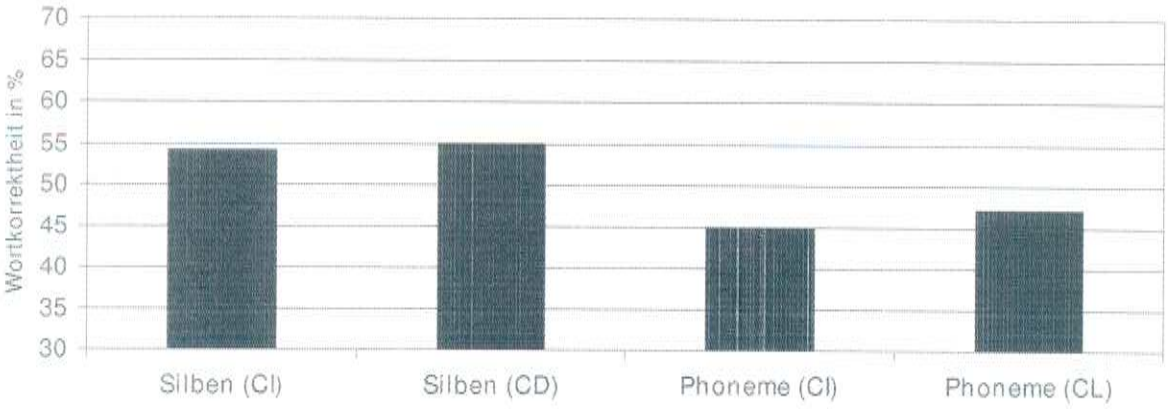

Abbildung 5.3: *Durchschnittliche Wortkorrektheiten für den Unbekannte-Wörter-Test bei gleicher Anzahl der Trainingsbeispiele pro Modell.*

	cbs	dss	S1	S2
Syllables (CI)	154	154	53,6	54,6
Syllables (CD)	154	493	55,6	54,6
Phonemes (CI)	137	137	52,8	37,1
Phonemes (CD)	137	441	52,5	39,9
Phonemes (CL)	246	246	52,8	41,5

Tabelle 5.4: *Durchschnittliche Anzahl der Codebuch-Einträge (cbs) und Gewichts-Einträge (dss) und Wortkorrektheit (in %) für den Unbekannte-Wörter-Test bei gleicher Anzahl der Trainingsbeispiele pro Modell.*

Dieses Experiment zeigt zum ersten Mal, dass auch bei Spracherkennung mittels Elektromyografie unbekannte Wörter durch die Verwendung von Wortuntereinheiten erkannt werden können. Die Erkennungsraten von 55,1% bei den Silbenmodellen und 47,1% bei den Phonemmodellen liegen deutlich über den Ergebnissen einer zufälligen Erkennung von 6,3%. Es zeigt weiterhin, dass Silben unter vergleichbaren Trainingsbedingungen wie erwartet besser geeignet sind, um die Muskelbewegungen zu modellieren. Die Erkennungsraten bei Silbenmodellen liegen 8% über den Erkennungsraten bei Phonemmodellen.

Ein weiteres Ergebnis dieses Versuchs: Die Gewinne durch Verwendung von kontextabhängigen Modellen liegen mit 1 bis 2% deutlich unter den Werten, die sich beim vorherigen Versuch mit den bekannten Wörtern ergaben. Hinsichtlich der Silbenmodelle muss dabei berücksichtigt werden, dass sich jetzt Trainings- und Testwörter unterscheiden. Da die Silbenmodelle allerdings größtenteils jeweils nur in einem Wort auftreten, kommen nun die in den Testwörtern auftretenden Tri-Silben nicht mehr in

	alright	also	alter	always	center	early	enter	entertaining	entry	envy	euro	leaning	liter	navy	sorted	watergate	ERGEBNIS
alright	**60,5**	0	0	**38,0**	0	0	0	0	0	0	0	0	0	0,5	0,5	0,5	**60,5**
also	6,5	**75,5**	5,5	3,0	0	1,0	0	0	2,0	0,5	4,0	0	0,5	0,5	0,5	0,5	**75,5**
alter	4,5	5,0	**55,0**	2,0	5,5	6,0	9,0	2,0	1,0	0	0	0,5	3,0	0,5	2,0	4,0	**55,0**
always	0,5	0	0	**98,5**	0	0	0	0	0	0	0	0,5	0	0,5	0	0	**98,5**
center	3,0	1,5	1,0	5,0	**62,0**	1,0	1,5	2,0	1,0	0	1,0	9,0	3,5	0	8,0	0,5	**62,0**
early	6,5	3,5	**22,5**	13,5	0	**32,5**	0,5	4,5	4,5	0	1,0	3,5	2,5	0,5	1,5	3,0	**32,5**
enter	0,5	3,5	5,0	7,0	16,0	1,0	**41,5**	3,5	7,0	2,5	0,5	4,5	3,0	0,5	4,0	0	**41,5**
entertaining	2,0	0,5	0	1,0	0,5	0	1,5	**67,5**	1,0	0	0	**21,5**	0,5	0	0	4,0	**67,5**
entry	1,5	0,5	0	1,0	0,5	0,5	0	3,5	**86,5**	2,5	0,5	1,0	1,0	0,5	0	0,5	**86,5**
envy	2,5	0,5	0	8,0	0	0,5	0	0	9,5	**70,0**	1,0	0	0	7,0	0	1,0	**70,0**
euro	1,0	3,5	0	2,0	0	0	0	0,5	1,5	0	**89,0**	0	0	0	2,5	0	**89,0**
leaning	0,5	1,0	0	4,0	0	0,5	0	9,0	0,5	0	0	**81,0**	3,0	0,5	0	0	**81,0**
liter	4,5	5,5	15,5	2,0	6,0	3,0	0	0,5	3,5	1,5	4,0	18,5	**29,0**	0	6,5	0	**29,0**
navy	1,0	1,0	0,5	11,5	0	1,5	0	1,5	10,5	**35,5**	0,5	1,5	0,5	**32,0**	1,0	1,5	**32,0**
sorted	5,0	0	0,5	9,0	3,5	0,5	0	1,0	3,0	0	0	3,0	0,5	0,5	**72,0**	1,5	**72,0**
watergate	11,5	0	0	**39,0**	0	0	0	0,5	0	0	0,5	2,5	0	0	0,5	**45,5**	**45,5**
TOTAL	**7,0**	**6,3**	**6,6**	**15,3**	**5,9**	**3,0**	**3,4**	**6,0**	**8,2**	**7,0**	**6,4**	**9,2**	**2,9**	**2,7**	**6,2**	**3,9**	**62,4**

Abbildung 5.4: *Die kombinierten Ergebnisse des Unbekannte-Wörter-Tests bei gleicher Anzahl an Trainingsaufnahmen, angegeben in Wortkorrektheit (in %). In der linken Spalte stehen die Referenzen, in der obersten Zeile die Hypothesen. Die unterste Zeile führt die normalisierte Auftrittshäufigkeit auf (in %).*

den Trainingswörtern vor - deshalb kann die Kontextabhängigkeit faktisch nicht mehr funktionieren.

Bei den Phonemmodellen führt die mehr als halbierte Trainingsdatenmenge dazu, dass die Modelle weniger repräsentativ sind als beim Test zuvor und folglich auch die Veränderungen im Kontext nicht mehr so gut erkannt und berücksichtigt werden können.

Für die praktische Anwendung muss allerdings bedacht werden, dass Silben eine weit größere Menge an Trainingsdaten erfordern als Phoneme (vgl. Kapitel 3.6.2). Da die Menge an Daten allerdings noch durch das Problem der Sitzungsabhängigkeit (vgl. Kapitel 1.4) eingeschränkt wird, erscheint es realistischer, das erhöhte Auftreten der Phonemmodelle im Vergleich zu den Silbenmodellen als Folge der begrenzten Trainingsdaten in Kauf zu nehmen.

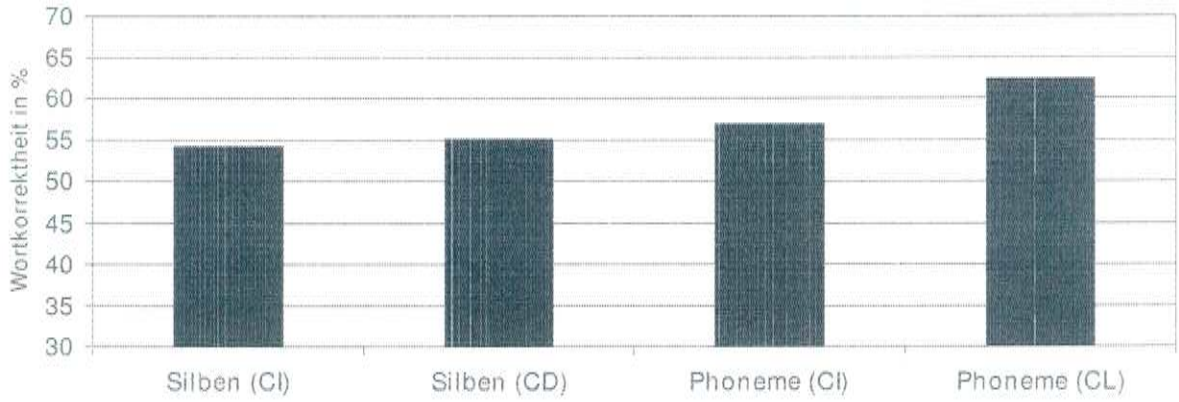

Abbildung 5.5: *Durchschnittliche Wortkorrektheiten für den Unbekannte-Wörter-Test bei gleicher Anzahl an Trainingsaufnahmen.*

	cbs	dss	S1	S2
Syllables (CI)	154	154	53,6	54,6
Syllables (CD)	154	493	55,6	54,6
Phonemes (CI)	137	137	63,3	50,5
Phonemes (CD)	137	441	63,6	53,9
Phonemes (CL)	246	246	66,0	58,8

Tabelle 5.5: *Durchschnittliche Anzahl der Codebuch-Einträge (cbs) und Gewichts-Einträge (dss) und Wortkorrektheit (in %) für den Unbekannte-Wörter-Test bei gleicher Anzahl an Trainingsaufnahmen.*

Die Ergebnisse in Tabelle 5.5 zeigen, dass bei diesem mehr realistischen Test die Phonemmodelle aufgrund der stärkeren Anzahl an Trainingsbeispielen deutlich bessere Erkennungsraten bewirken als die Silbenmodelle. Es zeigt sich weiterhin, dass die Erhöhung der Anzahl an Trainingsbeispielen von 8 auf 20 die Erkennungsraten um 15% verbessert. Die Erhöhung der Trainingsdatenmenge führt also zu einer besseren Modellbildung. Dies wirkt sich auch auf die Berücksichtigung der Kontextabhängigkeit aus. Im Vergleich zur Differenz von 2,1% im vorangegangenen Test beträgt die Verbesserung nun 5,5%.

In der Verwechslungsmatrix 5.4 zeigt sich, dass es große Unterschiede bei der Wortkorrektheit bezüglich der einzelnen Wörter gibt. Einige Wörter wie *liter* werden mit 29% vergleichweise schlecht erkannt, während andere Wörter wie *always* mit 98,5% dagegen sehr gut erkannt werden. Diese Werte korrespondieren mit der Auftrittshäufigkeit für die Hypothesen: Das Wort *liter* wird nur in 2,9% aller Fälle erkannt; zu erwarten wären bei gleichmäßiger Verteilung aber 6,25%. Das Wort *always* wird dagegen in 15,3% der Fälle erkannt, da die Modelle, aus denen es zusammengesetzt ist, offenbar auch anderen Modellen wie beispielsweise im Wort *alright* ähneln. Die Modelle des Wortes *liter* dagegen scheinen weiter differenziert zu sein. Insgesamt muss jedoch berücksichtigt werden, dass eine optimale Modellbildung auch daran scheitert, dass viele Triphone im Trainingsset nicht auftauchen und die Berücksichtigung der Kontextabhängigkeit deshalb noch nicht ihre komplette Wirkung entfalten konnte. Um dem entgegen zu wirken, wurde im Anschluss mit kontinuierlich aufgenommenen phonetisch ausbalancierten Sätzen trainiert, durch die eine wesentlich größere Zahl an Triphonen trainiert werden konnte

6 Training mit kontinuierlicher Sprache

Nachdem zuvor nur isoliert ausgesprochene Einzelwörter untersucht worden sind, wird in diesem Kapitel die Verarbeitung von kontinuierlicher Sprache im Hinblick auf deren Erkennung analysiert. Dazu wird zunächst untersucht, durch welche Verfahren die Modelle am besten initialisiert werden. Anschließend werden kontinuierlich ausgesprochene Sätze zum Training verwendet und zunächst die Erkennung von Einzelwörtern geprüft. Als nächster Schritt wird die Erkennung von kontinuierlichen Sätzen analysiert.

Als Bewertungsmaß für die Erkennung der Einzelwörter wird weiterhin die Wortkorrektheitsrate (word accuracy) verwendet, d.h. der Anteil der Wörter aus dem Testset, der korrekt erkannt wurde.

6.1 Erkennung von Einzelwörtern

Um zunächst zu testen, ob das Training auf der Basis von kontinuierlich ausgesprochenen Sätzen überhaupt möglich ist, wurden innerhalb einer Aufnahmesitzung zunächst die 38 Trainingssätze aufgenommen. Innerhalb der gleichen Sitzung wurden außerdem die 16 Wörter aus der Einzelwörter-Liste aufgenommen, die bereits im vorherigen Kapitel als Testwörter für die Erkennung untrainierter Wörter verwendet wurden. Anschließend wurden die 38 Trainingssätze über das bereits im vorherigen Kapitel vorgestellte System zum Training der Phonemmodelle verwendet.

	cbs	dss	S1	S2	Avg.
Phonemes (CI)	295	295	26	22	24
Phonemes (CD)	295	441	32	30	31
Phonemes (CL)	2332	2332	38	31	35

Tabelle 6.1: *Durchschnittliche Anzahl der Codebuch-Einträge (cbs) und Gewichts-Einträge (dss) und Wortkorrektheiten (in %) für den Unbekannte-Wörter-Test mit den über den kontinuierlichen Sätzen trainierten Modelle.*

Die Verwechslungsmatrix in Abbildung 6.1 zeigt ähnlich wie die Matrix für das in Kapitel 5.2 vorgestellte Experiment große Unterschiede zwischen den Erkennungsraten für die einzelnen Wörter. Im Unterschied zu diesem Experiment gibt es nun jedoch

	alright	also	alter	always	center	early	emer	entertaining	entry	envy	euro	leaning	liter	navy	sorted	watergate	ERGEBNIS
alright	**50**	0	0	5	0	0	10	**25**	10	0	0	0	0	0	0	0	**50**
also	10	**55**	0	10	0	0	0	5	0	0	15	0	5	0	0	0	**55**
alter	15	**35**	**20**	0	10	5	5	10	0	0	0	0	0	0	0	0	**20**
always	10	5	0	**50**	0	5	5	5	0	10	0	0	10	0	0	0	**50**
center	5	0	0	0	**30**	0	5	**40**	0	0	0	0	5	0	10	5	**30**
early	15	5	10	5	0	**20**	0	5	10	0	5	10	0	0	5	10	**20**
emer	5	0	5	0	**20**	5	**15**	15	0	0	0	10	**20**	0	5	0	**15**
entertaining	0	0	0	0	5	5	0	**45**	0	5	0	5	5	0	**25**	5	**45**
entry	10	0	0	10	0	0	10	15	**45**	5	0	0	0	0	5	0	**45**
envy	15	0	0	**50**	5	0	5	0	5	**20**	0	0	0	0	0	0	**20**
euro	5	0	0	0	0	0	0	5	0	15	**70**	0	5	0	0	0	**70**
leaning	5	0	0	0	0	10	5	**25**	10	10	0	**15**	5	10	5	0	**15**
liter	0	0	0	0	**20**	5	5	**30**	10	0	0	0	**20**	0	10	0	**20**
navy	**30**	0	5	0	0	0	5	10	5	**20**	0	0	5	**15**	5	0	**15**
sorted	**20**	0	5	0	0	**20**	5	0	**20**	0	0	5	0	5	**15**	5	**15**
watergate	5	0	0	10	0	5	10	5	0	0	0	0	0	0	5	**60**	**60**
TOTAL	12,5	6,2	2,8	8,8	5,6	5	5,3	15,0	7,2	5,3	5,6	2,8	5,0	1,9	5,6	5,3	35

Abbildung 6.1: *Die kombinierten Ergebnisse des Unbekannte-Wörter-Tests mit den über den kontinuierlichen Sätzen trainierten Modellen, angegeben in Wortkorrektheit (in %). In der linken Spalte stehen die Referenzen, in der obersten Zeile die Hypothesen. Die unterste Zeile führt die normalisierte Auftrittshäufigkeit auf (in %).*

keine eindeutige Korrelation zwischen hohen Auftrittshäufigkeiten und hohen Erkennungsquoten; das Wort *entertaining* beispielsweise taucht zwar in 15% aller Fälle auf, erreicht mit 45% aber nur eine knapp überdurchschnittliche Erkennungsquote. Das Wort *euro*, das mit 70% die beste Erkennungsquote erreicht, taucht mit 5,6% insgesamt jedoch unterdurchschnittlich oft auf. Die bei diesem Experiment deutlich größere Zahl an trainierten Triphonen erlaubt also offenbar bei manchen Modellen eine bessere Erkennung und ebenso eine Abgrenzung von anderen Modellen, gleichzeitig werden jedoch bedingt durch das schwierigere Aligmnent beim Training andere Modelle deutlich schlechter gebildet.

Da die Ergebnisse deutlich unter den Ergebnissen des vorherigen Tests lagen, wurde getestet, ob durch ein aus Sicht der akustischen Spracherkennung fortschrittlicheren Trainingssystems bessere Werte erzielt werden können. Dazu wurde alternativ zu *KMeans* auch *Merge-and-Split* eingesetzt und insgesamt mehr Trainingsdurchläufe angewendet. Folgender Ablauf wurde durchgeführt:

1. Kontextunabhängiges Training
   a) Lineare Segmentierung
   b) Initialisierung der Codebücher durch *KMeans*
   c) Vier Trainingsdurchläufe mit *EMTraining*
   d) Ausschreiben der Labels
2. *Merge-and-Split* Training (3 Durchläufe)
   a) Initialisierung durch ausgeschriebene Labels

b) *Merge-and-Split*

c) Vier Trainingsdurchläufe mit *EMTraining*

d) Ausschreiben der Labels

3. Kontextabhängiges Training

a) Berechnung der kontextabhängigen Triphone

b) Initialisierung durch ausgeschriebene Labels

c) *Merge-and-Split*

d) Vier Trainingsdurchläufe mit *EMTraining*

e) *Clustering-and-Split*

4. *Merge-and-Split* Training (3 Durchläufe)

a) Initialisierung durch ausgeschriebene Labels

b) *Merge-and-Split*

c) Vier Trainingsdurchläufe mit *EMTraining*

d) Ausschreiben der Labels

Schritt	S1	S2	Avg.
Kontextunabhängige Training	26	22	24
Merge-and-Split Training 1	25	31	28
Merge-and-Split Training 2	24	36	30
Merge-and-Split Training 3	23	39	31
Kontextabhängiges Training	24	41	28
Merge-and-Split Training 1	28	39	34
Merge-and-Split Training 2	26	41	34
Merge-and-Split Training 3	32	38	35

Tabelle 6.2: *Wortkorrektheiten (in %) für den Unbekannte-Wörter-Test mit den über den kontinuierlichen Sätzen durch den Merge-and-Split Algorithmus trainierten Modellen.*

Jeweils am Ende eines Trainingszyklus wurde die Erkennungsgenauigkeit gemessen. Die Ergebnisse sind in Tabelle 6.2 festgehalten.

Insgesamt ergibt sich ein uneinheitliches Bild: Während bei S1 im Gegensatz zum vorherigen Test jetzt deutlich schlechtere Werte gemessen werden können, sind die Ergebnisse von S2 deutlich besser als zuvor. Auffällig ist jedoch, dass im Laufe der Trainingsiterationen die Ergebnisse teilweise nicht besser, sondern schlechter werden. Offenbar können

	alright	also	alter	always	center	early	enter	entertaining	entry	envy	euro	leaning	liter	navy	sorted	watergate	ERGEBNIS
alright	**30**	0	5	**30**	0	0	0	10	**20**	5	0	0	0	0	0	0	**30**
also	10	**45**	**25**	15	5	0	0	0	0	0	0	0	0	0	0	0	**45**
alter	5	**40**	**35**	0	0	0	10	0	0	5	0	0	0	0	5	0	**35**
always	**25**	0	10	**45**	0	0	0	0	5	0	10	0	0	0	0	5	**45**
center	0	0	**20**	0	**25**	0	15	10	0	0	5	0	10	5	10	0	**25**
early	10	5	**25**	10	5	**20**	0	10	0	0	5	0	0	0	0	10	**20**
enter	5	0	**20**	0	5	0	**15**	5	0	0	10	5	**20**	5	5	5	**15**
entertaining	0	0	0	0	10	0	0	**50**	5	0	0	**20**	5	5	5	0	**50**
entry	10	0	10	5	0	0	0	15	**50**	5	5	0	0	0	0	0	**50**
envy	10	0	0	0	0	0	0	0	10	**20**	**25**	0	0	**20**	5	10	**20**
euro	15	0	10	**35**	0	5	0	0	0	0	**35**	0	0	0	0	0	**35**
leaning	0	5	0	0	0	10	0	**30**	15	5	0	**30**	0	5	0	0	**30**
liter	0	5	**40**	0	5	10	0	5	10	0	0	10	**10**	0	5	0	**10**
navy	0	5	5	0	0	0	0	0	**35**	10	0	0	5	**40**	0	0	**40**
sorted	10	0	10	0	0	0	0	5	**25**	0	0	15	0	0	**30**	5	**30**
watergate	0	0	0	15	0	0	0	0	0	0	0	0	5	0	5	**75**	**75**
TOTAL	8,1	6,6	13,4	9,7	3,4	2,8	2,5	8,8	10,9	3,1	5,9	5,0	3,4	5,0	4,4	6,9	**35**

Abbildung 6.2: *Die kombinierten Ergebnisse des Unbekannte-Wörter-Tests mit den über den kontinuierlichen Sätzen durch den Merge-and-Split Algorithmus trainierten Modellen, angegeben in Wortkorrektheit (in %). In der linken Spalte stehen die Referenzen, in der obersten Zeile die Hypothesen. Die unterste Zeile führt die normalisierte Auftrittshäufigkeit auf (in %).*

auch mit den verbesserten Trainings- und Codebuch-Initialisierungsalgorithmen unter den gegebenen Bedingungen keine besseren Modelle gebaut werden. Diese These wird auch von der Verwechslungsmatrix in Abbildung 6.2 bestätigt, die keine deutlich Verbesserung oder Veränderung im Vergleich zum voherigen Experiment erkennen lässt.

In einem weiteren Experiment wurde bei den *Merge-and-Split*-Trainingsiterationen zusätzlich eine lineare Diskriminanzanalyse durchgeführt. Dabei zeigte sich aber wie bereits in Kapitel 5.1.3, dass nach Anwendung der LDA keine robuste Erkennung mehr möglich war.

Ebenso wie bei den Experimenten mit Einzelwörtern wurde auch bei den Experimenten mit kontinuierlicher Sprache eine lineare Initialisierung für die Modelle verwendet. Da jetzt aber nicht nur ein Wort pro Aufnahme aufgenommen wurde, sondern ein kompletter Satz, ist es aufgrund der unterschiedlichen Phonem-Längen deutlich schwieriger, mit der linearen Initialisierung ein gutes Alignment für die Phoneme zu finden. Das bedeutet, dass die Modelle deutlich schlechter initialisiert werden als zuvor.

6.2 Erkennung von kontinuierlichen Sätzen

Da die Erkennung von Einzelwörtern zumindest grundlegend funktionierte, wurde im Folgenden untersucht, ob auch die Erkennung von kontinuierlicher Sprache möglich ist. Dazu wurde ein neues Test-System entwickelt, das ähnlich wie ein akustischer Spracherkenner beliebige kontinuierliche Sprache erkennen kann. Da die Erkennung kontinuier-

licher Sprache wesentlich komplexer ist als die Erkennung von Einzelwörtern, wird in der akustischen Spracherkennung typischerweise ein Sprachmodell für kontinuierliche Sprache verwendet. Ebenso wurde hier im kontinuierlichen EMG-Spracherkenner ein Sprachmodell verwendet. Dieses deckte die Test-Domäne ab. Weiterhin wurde das Vokabular auf die 144 Wörter beschränkt, die in den Test-Sätzen vorkamen. Es ergaben sich die in Tabelle 6.3 aufgeführten Werte.

	S1	S2	Avg.
#Korrekt	16,4	8,8	12,6
#Ersetzungen	40,0	38,0	39,0
#Auslassungen	43,7	53,3	48,5
#Einfügungen	3,0	0,8	1,9
Wortkorrektheit	13,3	7,9	10,6

Tabelle 6.3: *Erkennungswerte für die kontinuierlichen Testsätze.*

Diese Versuche wurden mit der Standard-Konfiguration des IBIS-Spracherkenners durchgeführt; durch Veränderung der Parameter sowohl für das Sprachmodell als auch für die Wörterzahl konnten die Erkennungswerte insgesamt nicht verbessert werden.

Die im Vergleich zur akustischen Spracherkennung niedrigen Erkennungsraten sind darauf zurückzuführen, dass die Vorverarbeitung bei der EMG-Spracherkennung trotz bedeutender Verbesserungen noch weiter optimiert werden muss. Mit der aktuell verwendeten Vorverarbeitung und Modellbildung ist es offenbar noch nicht möglich, eine robuste Erkennung von kontinuierlicher Sprache zu ermöglichen.

7 Zusammenfassung, Schlussfolgerung und Ausblick

In diesem Kapitel werden die Ergebnisse der durchgeführten Experimente zusammengefasst und Schlussfolgerungen für die weitere wissenschaftliche Forschung erläutert. Weiterhin werden Ansätze für zukünftige Untersuchungen vorgestellt.

7.1 Unterschiede zwischen akustischer Sprache und EMG-Sprache

Folgende theoretische Vorüberlegungen konnten in der Diplomarbeit bestätigt werden: Im Unterschied zur akustischen Sprache hat der Zeitbereich bei der EMG-Sprache eine deutlich größere Bedeutung. Während bei der akustischen Sprache durch eine immer genauere Analyse des Frequenzbereiches mit Filterbänken, Cepstral-Transformationen etc. deutliche Verbesserungen in der Erkennungsgenauigkeit erzielt werden, wurden in der EMG-Spracherkennung durch Untersuchungen des Zeitbereiches über vergleichsweise längere Zeitfenster (bis zu 120 ms) deutliche Verbesserungen erzielt.

Dies bestätigt die These, dass in der akustischen Spracherkennung eher stationäre Signale innerhalb kurzer Zeitfenster betrachtet werden müssen, während in der EMG-Spracherkennung eher längerfristige Schwankungen des Signals im Zeitbereich Aufschluss auf das gesprochene Wort ermöglichen.

Es hat sich herausgestellt, dass für die Segmentierung weiterhin eine manuelle Vorarbeit notwendig ist, da noch nicht zwischen Sprache und Nicht-Sprache (z.B. Lachen) unterschieden werden kann.

7.2 Erkennung von trainierten Wörtern

Die Erkennung von trainierten Wörtern gelingt auch bei einer Vokabulargröße von 32 Äußerungen. Mit der verbesserten Vorverarbeitung und Modellierung von Kontextabhängigkeit sind Erkennungsraten von ca. 80% erreichbar. Insbesondere durch das von 54 ms auf 27 ms verkleinerte Fenster war eine durchschnittliche Steigerung um 4% bei der Worterkennungsrate bei den geclusterten Phonemen, bei der Verwendung von Einzelwörtern oder Silben als Wortuntereinheiten von 13% möglich. Eine weitere

Verbesserung von 6-7% bei allen Tests konnte durch das neue *Time Domain Context*-Feature erreicht werden, bei dem der Verlauf des Signals innerhalb des Zeitbereiches betrachtet wurde.

Insgesamt ergeben sich bei der Erkennung auf Basis von geclusterten Phonemen vergleichbare Erkennungsraten wie die Erkennung auf Basis von Einzelwortmodellen - allerdings mit dem Vorteil, dass mit diesen geclusterten Phonemmodelle auch die Erkennung von untrainierten Wörtern möglich ist.

7.3 Erkennung von untrainierten Wörtern

Mit dieser Diplomarbeit wurde zum ersten Mal der Beweis erbracht, dass auch die Erkennung von untrainierten Wörtern bei der Spracherkennung mittels EMG möglich ist. Damit ist eine wichtige Grundlage für die weitere Entwicklung der EMG-Spracherkennung gelegt. Insgesamt ergab sich eine durchschnittliche Erkennungsrate von 62,4% bei der Erkennung von 16 verschiedenen Wörtern mittels kontextabhängiger Phonemmodelle und der verbesserten Vorverarbeitung.

Bei dem Vergleich zwischen verschiedenen Wortuntereinheiten stellte sich heraus, dass bei vergleichbaren Bedingungen (also einer gleichen Anzahl von Trainingsbeispielen) Silbenmodelle bessere Erkennungsraten ermöglichen als Phonemmodelle, da sie abgeschlossene artikulatorische Muskelbewegungen modellieren. Unter realistischen Bedingungen und insbesondere bei Berücksichtigung der Sitzungsabhängigkeit können mit Phonemmodellen allerdings insbesondere bei Berücksichtigung der Kontextabhängigkeit bessere Erkennungsraten erzielt werden. Damit erscheint klar, dass auch bei der EMG-Spracherkennung durch die Verwendung von Phonem-Modellen eine ähnliche Entwicklung wie bei der akustischen Spracherkennung möglich ist. Jedoch sollte berücksichtigt werden, dass in dieser Arbeit nicht alle Phoneme einzeln untersucht wurden; es ist also nicht bekannt, ob durch EMG-Spracherkennung beliebige Wörter unterschieden werden können. Allerdings gibt es auch bei der akustischen Spracherkennung Phoneme, die schwierig, aber dennoch zu unterscheiden sind.

7.4 Kontinuierliches Training

In dieser Arbeit wurde gezeigt, dass es möglich ist, auch kontinuierliche Sprache zum Trainieren der EMG-Spracherkennung zu verwenden. Die getestete Erkennung von Einzelwörtern auf Basis der kontinuierlichen Sprache funktioniert mit Erkennungsraten von

35%. Dies zeigt, dass auch kontinuierliche EMG-Sprache wenigstens prinzipiell verarbeitet werden kann. Allerdings ist die Erkennungsquote vergleichsweise schlecht. Dies resultiert aus den noch nicht optimal gebildeten Modellen und zeigt, dass insbesondere die Vorverarbeitung und Kontextmodellierung weiter verbessert werden muss, um auch bei kontinuierlichen Sätzen ein gutes Alignment für die Phoneme zu finden.

Vor allem die Versuche zur Erkennung von kontinuierlicher Sprache haben gezeigt, dass es mit der aktuell verwendeten Vorverarbeitung offenbar noch nicht möglich ist, Modelle zu bilden, die eine robuste Erkennung von kontinuierlicher Sprache ermöglichen.

7.5 Zukünftige Arbeiten

Insbesondere die Experimente mit kontinuierlicher Sprache haben gezeigt, dass die Vorverarbeitung und Kontextmodellierung weiter verbessert werden müssen. Hierfür bieten sich eine Vielzahl von Möglichkeiten an, mit denen in vergleichbaren Arbeiten beachtliche Erfolge erzielt wurden, insbesondere die Verwendung von Wavelets erscheint interessant.

Außerdem erscheint es naheliegend die fixe Kontextlänge für das *Time Domain Context*-Feature durch einen variablen, von Phonem, Kontext und Position abhängigen Wert zu ersetzen.

Nachdem bei dieser Arbeit nur vorgelesene Sprache betrachtet wurde, wäre es längerfristig sehr interessant, Experimente mit spontaner Sprache durchzuführen. Während das Gehirn nämlich bei vorgelesener Sprache bereits zu Anfang der Äußerung einen kompletten Ablaufplan für die betroffenen Muskeln ausarbeiten kann, muss dies bei spontaner Sprache sehr viel kurzfristiger und ungenauer passieren.

Ebenfalls interessant wären Experimente mit der Mimik. Bei dieser Arbeit wurde die Mimik komplett ausgeblendet - dabei dürfte es deutlich leichter sein, die grundlegenden Mimiken (hoch- oder heruntergezogene Mundwinkel) zu erkennen als Sprache. Langfristig wäre es für die Spracherkennung notwendig, Mimiken herausfiltern zu können, und dadurch nicht die Spracherkennung beeinflussen zu lassen.

Genauso wichtig wäre es für die Alltagstauglichkeit, andere Muskelbewegungen (beispielsweise Lachen und Husten) zu erkennen und von der Sprache zu unterscheiden.

A Verzeichnisse

A.1 Abbildungsverzeichnis

A.2 Tabellenverzeichnis

B Auflistung aller aufgenommenen Sitzungen

B.1 Summe aller Aufnahmen

Sprecher	S1	S2
Summe EINZELWÖRTER	5 nicht-hörbar, 1 hörbar	5 nicht-hörbar, 1 hörbar
Summe KONT1	1 nicht-hörbar, 2 hörbar	1 nicht-hörbar, 1 hörbar
Summe KONT2	2 nicht-hörbar	2 nicht-hörbar

Tabelle B.1: *Summe der Aufnahme-Sitzungen pro Sprecher und pro Korpus*

B.2 Aufnahmen von Sprecher S1

Sitzung Nr.	Korpus	Hörbar
076	EINZELWÖRTER	nein
077	KONT1	nein
081	KONT1	ja
081b	EINZELWÖRTER	ja
082	KONT2	nein
086	EINZELWÖRTER	nein
087	EINZELWÖRTER	nein
088	KONT2	nein

Tabelle B.2: *Aufnahmesitzungen von Sprecher S1*

B.3 Aufnahmen von Sprecher S2

Sitzung Nr.	Korpus	Hörbar
011	EINZELWÖRTER	nein
012	EINZELWÖRTER	ja
015	KONT1	ja
016	KONT2	nein
017	EINZELWÖRTER	nein
018	EINZELWÖRTER	nein
019	KONT1	nein
020	EINZELWÖRTER	nein
021	EINZELWÖRTER	nein
022	KONT2	nein

Tabelle B.3: *Aufnahmesitzungen von Sprecher S2*

C Wort- und Satzlisten

C.1 EINZELWÖRTER

- all
- alright
- also
- alter
- always
- center
- early
- earning
- enter
- entertaining
- entry
- envy
- euro
- gateways
- leaning
- li
- liter
- n
- navy
- right
- rotating
- row
- sensor
- sorted
- sorting
- so
- tree
- united
- v
- watergate
- water
- ways

C.2 KONT1

- The female produces a litter of two to four young in November and December.
- Numerous works of art are based on the story of the sacrifice of Isaac.
- Their solution requires development of the human capacity for social interest.
- His most significant scientific publications were studies of birds and animals.
- In recent years she has primarily appeared in television films such as Little Gloria.
- The process by which the lens focuses on external objects is called accommodation.
- Two narrow gauge railroads from China enter the city from the northeast and northwest.
- Some maps use bands of color to indicate different intervals of value.
- Origins or causes of spontaneous mutation are not yet completely clear.
- Unusually high levels of radiation were detected in many European countries.
- Both petroleum and natural gas deposits are scattered through eastern Ohio.
- For the first time in years the Republicans also captured both houses of Congress.
- A tanker is a ship designed to carry large volumes of oil or other liquid cargo.
- The enormous amounts of carbon dioxide in the atmosphere cause this high pressure.
- The population lives by herding goats and sheep or by trading.
- This allows the shaft to change its length and direction as the car wheels move up and down.
- Bismarck serves as a trade and shipping point for an area of large mechanized farms.
- He is a major figure among writers who break away from the American tradition of realism.
- Modern electronics has become highly dependent on inorganic chemistry.
- They began to build boats with the same materials they used for portable shelters.
- Much of the ground beef consumed in the United States comes from dairy cows.
- Eclipses of the sun and moon have long made a deep impression on humankind.
- Philosophers of education often differ in their views on the nature of knowledge.
- During the following years he tried unsuccessfully to get it into production.
- Private free schools were formed both in poor neighborhoods and in middle-class communities.
- In most cases only a few artifacts and the foundations of buildings can be seen.
- It is one of the earliest agricultural villages yet discovered in Southwest Asia.
- The hot fluid is circulated through a tube located in the lower tank of the radiator.
- Military policy was to keep the travel routes open and protect the settled areas.
- Several environmental factors also have an effect on average life expectancy.
- They can also show how the shape and size of continents and oceans have changed over time.
- Almost all students who are accepted into medical schools obtain a medical degree.
- Microbiology is the study of organisms that cannot be seen by the naked eye.
- He introduced the writing of English into a curriculum that had emphasized Latin.
- The Statue of Liberty and Ellis Island are within the New Jersey waters of New York Bay.
- The United States undertook to defend Western Europe against Soviet attack.
- He never obtained a secure academic position or permanent employment.
- They established royal commissions to recover illegally held church lands.

C.3 KONT2

- The explanation once commonly dispensed in textbooks turns out to be wrong
- Scientists expect to discover still more variations in the coming years
- The notion that ice has an intrinsic liquid layer is not a new concept
- Because the layer is so thin it was hard for scientists to see
- In his view friction is the primary reason ice is slippery
- Ice two does not occur naturally on Earth
- With pressure high enough the temperature need not even be cold for ice to form
- At higher pressures the usual hexagonal structure breaks down
- Science is not cold and remote in this setting
- The topics are not always so funny
- But even these critics consider these alternatives a bit of a stretch
- The threat is here and now
- The robots are computers and such
- Or at least until I pull the plug
- These findings are causing a big debate and getting lots of attention
- I already have some important findings
- It just needed a name
- Practice makes perfect
- The perfect is the enemy of the good
- Today is the tomorrow you worried about yesterday
- Size is not everything
- Russia and China say sanctions would only make things worse
- Investigators have yet to find the murder weapon
- Dozens of states are debating the issue

C.4 Frageset

```
; ------------------------------------------------------
;  Name            : ps
;  Type            : PhonesSet
;  Number of Items : 2
;  Date            :
;  Remarks: DX->T, add XL/XM/XN
;    removed DX from SONORANT & VOICED, added X-LMN class
; ------------------------------------------------------
PHONES PAD IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z ZH V F TH S SH HH XL XM XN SIL
HUMANSND IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z ZH V F TH S SH HH XL XM XN
VOLATILE AO EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z ZH V F TH S SH HH XL XM XN
SILENCES SIL
CONSONANT P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG R Y W L ER AXR XL XM XN
CONSONANTAL P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG XL XM XN
OBSTRUENT P B F V TH DH T D S Z SH ZH CH JH K G
SONORANT M N NG R Y W L ER AXR XL XM XN
SYLLABIC AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX ER AXR XL XM XN
VOWEL AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX
DIPHTHONG AY OY EY AW OW
CARDVOWEL IY IH EH AE AA AH AO UH UW IX AX
VOICED B D G JH V DH Z ZH M N NG W R Y L ER AY OY EY IY AW OW EH IH AO AE AA AH UW UH AXR IX AX XL XM XN
UNVOICED P F TH T S SH CH K
CONTINUANT F TH S SH V DH Z ZH W R Y L ER XL
DEL-REL CH JH
LATERAL L XL
ANTERIOR P T B D F TH S SH V DH Z ZH M N W Y L XM XN
CORONAL T D CH JH TH S SH DH Z ZH N L R XL XN
APICAL T D N
HIGH-CONS K G NG W Y
BACK-CONS K G NG W
LABIALIZED R W ER AXR
STRIDENT CH JH F S SH V Z ZH
SIBILANT S SH Z ZH CH JH
BILABIAL P B M W
LABIODENTAL F V
LABIAL P B M W F V
INTERDENTAL TH DH
ALVEOLAR-RIDGE T D N S Z L
ALVEOPALATAL SH ZH CH JH
ALVEOLAR T D N S Z L SH ZH CH JH
RETROFLEX R ER AXR
PALATAL Y
VELAR K G NG W
GLOTTAL HH
ASPIRATED HH
STOP P B T D K G M N NG
PLOSIVE P B T D K G
NASAL M N NG XM XN
FRICATIVE F V TH DH S Z SH ZH HH
AFFRICATE CH JH
APPROXIMANT R L Y W
LAB-PL P B
ALV-PL T D
VEL-PL K G
VLS-PL P T K
VCD-PL B D G
LAB-FR F V
DNT-FR TH DH
ALV-FR SH ZH
VLS-FR F TH SH
VCD-FR V DH ZH
ROUND AO OW UH UW OY AW OW
HIGH-VOW IY IH UH UW IX
MID-VOW EH AH AX
LOW-VOW  AA AE AO
FRONT-VOW IY IH EH AE
CENTRAL-VOW AH AX IX
BACK-VOW AA AO UH UW
TENSE-VOW IY UW AE
LAX-VOW IH AA EH AH UH
ROUND-VOW AO UH UW
REDUCED-VOW IX AX
REDUCED-CON AXR
REDUCED IX AX AXR
LH-DIP AY AW
```

```
MB-DIP OY OW EY
BF-DIP AY OY AW OW
Y-DIP AY OY EY
W-DIP AW OW
ROUND-DIP OY AW OW
LIQUID-GLIDE    L R W Y
W-GLIDE         UW AW OW W
LIQUID          L R
LW              L W
Y-GLIDE         IY AY EY OY Y
LQGL-BACK       L R W
X-LMN           XL XM XN
```

D Verwendete Daten und Programme

Sämtliche verwendete Programme und Daten befinden sich innerhalb des Projektverzeichnisses /project/emg2.

D.1 Daten

Die mit Sprecher S1 und S2 aufgenommenen Daten befinden sich innerhalb des Ordners /project/emg2/session.

D.2 Programme

Die verwendeten Programme befinden sich innerhalb des Ordners /project/emg2/versuche. Im Einzelnen wurden folgende Programme für die entsprechenden Experimente verwendet:

Experiment	Programmordner
Ohne automatische Segmentierungskorrektur Kapitel 5.1.1	versuche/134
Mit automatischer Segmentierungskorrektur Kapitel 5.1.1	versuche/131
Segmentierungskorrektur mit Sprachdetektor Kapitel 5.1.1	versuche/135
Vorverarbeitung Kapitel 5.1	versuche/131
Lineare Diskriminanzanalyse Kapitel 5.1.3	versuche/64
Independent Component Analysis Kapitel 5.1.4	versuche/127
Erkennung unbekannter Wörter Kapitel 5.2	versuche/125
Training mit kontinuierlicher Sprache Kapitel 6.1	versuche/105
Training mit *Merge-and-Split* Kapitel 6.1	versuche/128
Training mit *Merge-and-Split* und LDA Kapitel 6.1	versuche/136
Erkennung von kontinuierlicher Sprache Kapitel 6.2	versuche/105

E Literaturverzeichnis

[AFI+02] AZZERBONI, B., G. FINOCCHIO, M. IPSALE, F. LA FORESTA und F. C. MORABITO: *A New Approach to Detection of Muscle Activation by Independent Component Analysis and Wavelet Transform.* 2486/2002:109–116, 2002. 22

[Bun] BUND, SCHWERHÖRIGEN: *Informationen zur Schwerhörigkeit, Ertaubung und Kommunikation, Stand: 27.6.2006.* `http://www.schwerhoerigen-netz.de/RATGEBER/KOMMUNIKATION/`. 11

[CKHL02] CHAN, A.D.C., K.ENGLEHART, B. HUDGINS und D.F. LOVELY: *Hidden Markov Model Classification of Myolectric Signals in Speech.* Engineering in Medicine and Biology Magazine, IEEE, 21:143–146, 9 2002. 18

[Com94] COMON, P.: *Independent component analysis, A new concept?* Signal Processing, 36:287–314, 1994. 29

[DMD82] DICKSON, D. R. und W. MAUE-DICKSON: *Anatomical & Physiological Bases of Speech.* Lippincott Williams & Wilkins, 1982. 18, 57

[DMW94] DUCHNOWSKI, P., U. MEIER und A. WAIBEL: *See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading.* Proceedings ICSLP, Seiten 547–550, 1994. 11

[FGH+97] FINKE, M., P. GEUTNER, H. HILD, T. KEMP, K. RIES und M. WESTPHAL: *The Karlsruhe Verbmobil Speech Recognition Engine.* Proceedings ICASSP 97, München; Germany, (4), 1997. 37

[GHP+01] GANAPATHIRAJU, A., J. HAMAKER, J. PICONE, M. ORDOWSKI und G. R. DODDINGTON: *Syllable-Based Large Vocabulary Continuous Speech Recognition.* IEEE Transactions on speech and audio processing, 9(4), May 2001. 30

[Hyv99] HYVARINEN: *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis.* IEEE Transactions on Neural Networks, 10(3), May 1999. 29

[JB05] JORGENSEN, C. und K. BINSTED: *Web Browser Control Using EMG Based Sub Vocal Speech Recognition.* Proceedings of the 38th Hawaii International Conference on System Sciences, 2005. 18

[JCKH98] JEFFREY, R., PH.D. CRAM, GLENN S. KASMAN und JONATHAN HOLTZ: *Introduction to Surface Electromyography.* Jones & Bartlett Publishers, 1998. 23

[JMHSW06] JOU, S.-C., L. MAIER-HEIN, T. SCHULTZ und A. WAIBEL: *Articulatory feature classification using surface electromyographie.* Proceedings ICASSP 06, Toulouse; France, 2006. 26

[KKAB04] KUMAR, S., D. K. KUMAR, M. ALEMU und M. BURRY: *EMG based voice recognition.* Proceedings of the Intelligent Sensors, Sensor Networks and Information Processing Conference, Seiten 593–597, Dec 2004. 18

[MGW91] MORSE, M. S., Y. N. GOPALAN und M. WRIGHT: *Speech Recognition Using Myoelectric Signals With Neural Networks.* Engineering in Medicine and Biology Society, 13, 1991. 17

[MH05] MAIER-HEIN, L.: *Speech Recognition Using Surface Electromyography.* Jul 2005. 21, 33

[MHMSW05] MAIER-HEIN, L., F. METZE, T. SCHULTZ und A. WAIBEL: *Session independent non-audible speech recognition using surface electromyography.* Proceedings ASRU, Cancun, Mexico, Nov 2005. 18, 27

[MHS03] MANABE, H., A. HIRAIWA und T. SUGIMURA: *Unvoiced Speech Recognition using EMG - Mime Speech Recognition -.* In: *Proceedings of the 2003 Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA*, Seiten 794–795, 4 2003. 17

[MM76] MCGURK, H. und J. MACDONALD: *Hearing lips and seeing voices.* Nature, 264:746–748, 1976. 11

[MZ04] MANABE, H. und Z. ZHANG: *Multi-stream HMM for EMG-Based Speech Recognition.* In: *Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA*, Seiten 4389–4392, Sep 2004. 20

[Pea01] PEARSON, K.: *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 2:559–572, 1901. 19, 29

[PSF+05] PAULIK, M., S. STÜKER, C. FÜGEN, T. SCHULTZ, T. SCHAAF und A. WAIBEL: *Speech Translation Enhanced Automatic Speech Recognition.* Proceedings of the Automatic Speech Recognition and Understanding Workshop, Cancun, Mexico, 2005-12, Nov 2005. 13

[Rog03] ROGINA, I.: *Sprachliche Mensch-Maschine-Kommunikation. Entwurf einer Habilitationsschrift.* 2003. 13, 24, 25, 26, 27, 57

[Wika] WIKIPEDIA: *Aktionspotenzial, Stand: 18.6.2006.* http://de.wikipedia.org/wiki/Aktionspotenzial. 24, 57

[Wikb] WIKIPEDIA: *Electromoygraphie, Stand: 18.6.2006.* http://en.wikipedia.org/wiki/Electromyographie. 12

[Wikc] WIKIPEDIA: *Elektromoygraphie, Stand: 18.6.2006.* http://de.wikipedia.org/wiki/Elektromyographie. 12, 29