Universität Karlsruhe
Fakultät für Informatik
Institut für Logik, Komplexität und Deduktionssysteme
Prof. Dr. A. Waibel

Diplomarbeit

Audio-Visuelle Spracherkennung auf großem Vokabular

Jan Kratt

31. März 2004

Betreuer:
Prof. Dr. A. Waibel
Dipl.-Phys. F. Metze
Dr.-Ing. R. Stiefelhagen

Hiermit erkläre ich, die vorliegende Arbeit selbstständig erstellt und keine anderen als die angegebenen Quellen verwendet zu haben.

Karlsruhe, den 31.03.2004 .....

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

1.1 Motivation

Die Art der Bedienung von technischen Geräten und speziell Computern hat sich in den vergangenen Jahrzehnten deutlich vereinfacht, z.B. durch die Einführung von grafischen Benutzeroberflächen und Mäusen beim PC. Auch andere Geräte, wie Videorekorder lassen sich heute viel einfacher bedienen, da sie wichtige Informationen in Menüs auf dem TV-Bildschirm darstellen können und nicht mehr über kryptische Zahlenkürzel zu programmieren sind.

Durch einen ständig steigenden Funktionsumfang und die wachsende Komplexität wird aber ein großer Teil dieser Fortschritte wieder zunichte gemacht. Aus diesem Grund ist es erforderlich, andere, besser geeignete Bedienungsformen zu finden. Hierbei bietet sich die akustische Informationsübermittlung, das Sprechen, an, da es die natürliche Kommunikationsform des Menschen ist.

Für Menschen ist es einfach, gesprochene Sprache zu verstehen, für den Computer ist die *automatische Spracherkennung* (ASR, **A**utomatic **S**peech **R**ecognition) eine hohe Herausforderung, die trotz jahrelanger Forschung noch nicht perfekt gelöst ist.

Im Rahmen dieser Arbeit geht es nicht um das Verstehen des Inhalts, das Erfassen des Sinns der Äußerung, sondern um das Erkennen der gesagten Wörter, das Erstellen einer Transkription des gesprochenen Textes. Für die Lösung dieser Aufgabe stehen die so genannten speech-to-text-Systeme zur Verfügung.

Automatische Spracherkennungssysteme sind mittlerweile hinreichend gut, um sie in kommerziellen Produkten einsetzen zu können. Die im Handel befindlichen ASR-Systeme werden mit Erkennungsraten von bis zu 98% beworben [1], aber diese hervorragenden Erkennungsleistungen werden nur auf einer eingeschränkten Domäne erreicht, die nicht verlassen werden darf. Außerdem muss häufig auch ein spezielles Mikrofon verwendet werden, um eine optimale Erkennung zu erreichen.

Des Weiteren ist diese hohe Erkennungsrate nur unter idealsten Bedingungen zu erreichen, d.h. ein Sprecher, der langsam und deutlich spricht. Das System muss auf den Sprecher trainiert, eingestellt und optimiert werden. Zusätzlich müssen die Audiodaten noch in optimaler Form vorliegen. In dem

Raum, in dem die Aufnahmen gemacht werden, muss es ansonsten sehr ruhig sein, der Sprecher muss sich nahe am Mikrophon befinden und der Abstand sollte sich nicht verändern. Daraus ergibt sich, dass der Sprecher idealerweise ein hochwertiges Nahbesprechungsmikrophon benutzen sollte, welches in der Nähe des Mundes befestigt werden muss.

In der Praxis gibt es aber viele Anwendungsgebiete, bei denen eine Datenerfassung unter diesen idealisierten Bedingungen nicht möglich ist. Bereits heute wird die Sprachsteuerung zur Bedienung von Navigationssystemen bzw. Bordcomputern [2, 3] in Oberklassefahrzeugen eingesetzt. Auch wenn diese Fahrzeuge heute eine sehr gute Geräuschdämmung besitzen, so gibt es noch immer deutlich mehr Nebengeräusche, wie im Labor. Außerdem ist es nicht akzeptabel, die ganze Zeit ein Mikrophon zu tragen, wenn nur hin und wieder der Radiosender gewechselt werden soll. Der Sicherheitsgewinn der Sprachsteuerung geht aber verloren, wenn während der Fahrt erst das Mikrophon in die richtige Position gebracht werden muss, in diesem Fall kann man das Radio auch direkt bedienen.

Der Einfluss der Störgeräusche - beim Auto wären dies vorwiegend Motor-, Wind- und Abrollgeräusche der Reifen - kann durch geeignete Maßnahmen, z.B. durch Cepstrale Mittelwertsubtraktion [4, 5], verringert werden. Es ist aber nicht möglich, durch den Einsatz von Filtern jegliche Störungen des Signals auszugleichen. Wenn sie zu stark werden, dann sinkt die Erkennungsleistung. Der Mensch kann sich unter diesen Bedingungen aber im Normalfall noch gut unterhalten. Also muss es auch möglich sein, die automatische Spracherkennung so zu erweitern, dass sie unter diesen Bedingungen bessere Ergebnisse erzielen kann.

Der Mensch bedient sich im Allgemeinen auch nicht nur der akustischen Daten für das Verstehen von Sprache. Er ist, z.B. beim Telefonieren, zwar dazu im der Lage, hat er aber die Möglichkeit, so bedient er sich zusätzlich visueller Informationen. Jeder Mensch nutzt unbewusst seine Fähigkeit zum Lippenlesen, während er kommuniziert und ist dadurch in der Lage sein Gegenüber besser zu verstehen [6].

Vor allem unter schlechten akustischen Bedingungen wird der Einfluss des Lippenlesens immer größer, da viele Informationen in den akustischen Daten verloren gehen. Ein großer Vorteil der visuellen Informationen liegt in der Unabhängigkeit der Signalqualität zu den akustischen Informationen. Eine laute Umgebung muss kein schlechtes Videosignal liefern und Dunkelheit beeinflusst die Qualität der Audiodaten nicht.

Darum wurde in dieser Arbeit die Einführung einer zusätzlichen Modalität, dem Lippenlesen, zur Verbesserung der Erkennungsleistung eines Spracherkennungssystems gewählt, um weitere Informationen über die zu transkribierenden Äußerungen zu erhalten. Es wird also ein *gewöhnliches* automatisches Spracherkennungssystem zu einem *audiovisuellen* Spracherkennungssystem erweitert. Durch diese Maßnahme ist man nicht mehr ausschließlich darauf angewiesen, Informationen aus den Audiodaten zu gewinnen, sondern kann, bei verschlechterter Signalqualität, verstärkt auf die Videodaten zurückgreifen, wie es auch der Mensch macht [7].

Auch unter guten akustischen Bedingungen darf man auf eine leichte Verbesserung der Erkennungsleistung hoffen, da es Laute gibt, die akustisch nur sehr schlecht, visuell aber gut zu unterscheiden sind. Es stehen also zusätzliche Informationen zur Verfügung, die dem normalen, rein akustischen Spracherkenner fehlen.

1.2 Zielsetzung

Um eine robustere Erkennung unter schwierigen Bedingungen zu erreichen, soll der an der Universität Karlsruhe vorhandene Spracherkenner Janus [8, 9] um Fähigkeiten zum Verarbeiten von visuellen Informationen erweitert werden. Dieses System soll als Basis für zukünftige Forschungen in der audiovisuellen Spracherkennung dienen, um möglichst schnell ein state-of-the-art-System zu erhalten. Ein solches System ist in [10] beschrieben.

An diesem System hat eine Gruppe von Experten mit mehrjähriger Erfahrung in der audiovisuellen Spracherkennung mitgewirkt. Es sollte versucht werden, mit dem für diese Arbeit erzeugten audiovisuellen Spracherkenner möglichst nah an das System aus [10] heran zu kommen, um eine gute Basis für weitere Forschungen mit audiovisuellen Spracherkennern zu erhalten. Es wurde versucht, einige der dort beschriebenen Experimente auf den Janus Spracherkenner zu übertragen, um die Ergebnisse vergleichen zu können.

Insgesamt wurden drei unterschiedliche Ansätze zur Kopplung von Audio- und Videodaten zu einem audiovisuellen Spracherkenner verfolgt. Der erste und einfachste Ansatz war eine einfache Verknüpfung der Daten, im nächsten Versuchsaufbau wurde getestet, ob es möglich ist, die relevanten Informationen mittels LDA vor der Verknüpfung der Daten hervorzuheben. Der letzte Ansatz für die Verknüpfung von Audio- und Videodaten war eine Multi-Stream-Architektur.

Es sollte die Eignung der unterschiedlichen Verfahren für das Aufgabengebiet der audiovisuellen Spracherkennung getestet werden, um herauszufinden, welcher dieser Ansätze das meiste Potential für weitere Optimierungen bietet.

1.3 Überblick

Eine Übersicht über andere Arbeiten zur audiovisuellen Spracherkennung wird in Kapitel 2 gegeben, beginnend mit den ersten Versuchen, die bereits Jahrzehnte zurück liegen, bis zu aktuellen Arbeiten.

In Kapitel 3 wird zuerst ein Überblick über die herkömmliche, rein akustische Spracherkennung gegeben. Es wird auch gezeigt, dass visuelle Informationen zu einer Verbesserung der Erkennungsleistung beitragen können [11, 12]. Anschließend wird darauf eingegangen, welche Änderungen für die visuelle Spracherkennung, das Lippenlesen, notwendig sind. Am Ende dieses Kapitels wird der aktuelle Stand der Technik in der audiovisuellen Spracherkennung beschrieben.

Kapitel 4 beginnt mit einer Beschreibung der verwendeten Datenbasis. Darauf folgt eine Darstellung der Verarbeitungsschritte, die für eine erfolgreiche automatische Spracherkennung notwendig sind. Zuerst wird die Verarbeitung der Audiodaten beschrieben, anschließend die Umformungen, welche an den Videodaten vorgenommen werden.

Danach werden in Kapitel 5 die durchgeführten Experimente ausführlich beschrieben und die Ergebnisse dieser Experimente aufgeführt. Hierzu wird zunächst ein rein akustisches Referenzsystem beschrieben, mit welchem die Gewinne oder Verluste der audiovisuellen Spracherkennungssysteme gemessen werden konnten. Anschließend wird der rein visuelle Spracherkenner beschrieben und die verschiedenen Ansätze der audiovisuellen Spracherkennung.

Kapitel 6 beschreibt weitere mögliche Optimierungsmöglichkeiten, die in diese Arbeit - auf Grund des zeitlich begrenzten Horizonts - nicht mehr eingeflossen sind, die aber eine weitere Steigerung der Erkennungsleistung ermöglichen sollten.

2 Verwandte Arbeiten

Die ersten Ansätze der audiovisuellen Spracherkennung sind bereits fast 40 Jahre alt, schon im Jahre 1965 erhielt IBM ein Patent für eine Entwicklung von Ernie Nassimbene. Hierbei wurde, auf Grund der zu dieser Zeit noch sehr eingeschränkten Rechenleistung, lediglich berücksichtigt, ob die Zähne des Sprechers sichtbar sind oder nicht.

Danach gab es für lange Zeit keine großen Fortschritte in der audiovisuellen Spracherkennung, erst im Jahr 1984 erschien eine Arbeit von Eric Petajan [13], in der bereits komplexere Merkmale zum Einsatz gekommen sind. Die audiovisuelle Spracherkennung wurde zu diesem Zeitpunkt aber nur auf Ziffernfolgen oder buchstabierten Sequenzen betrieben.

Diese frühen Arbeiten haben bereits gezeigt, dass durch das Hinzufügen von Informationen aus Videodaten, die Erkennungsleistung gesteigert werden konnte. Dieser Effekt konnte später ebenfalls festgestellt werden [11, 14, 15], obwohl sich die Erkennungsleistung der rein akustischen Spracherkennungssysteme deutlich verbessert hatte.

Auch heute beschäftigen sich noch viele Arbeiten mit diesem kleinen Aufgabengebiet, in [16, 17] geht es um das Erkennen von Buchstaben und in der Arbeit [18] müssen lediglich Ziffern erkannt werden.

Im Bereich der audiovisuellen Spracherkennung auf kontinuierlich gesprochener Sprache mit großen Vokabular ist die Arbeitsgruppe um Chalapathy Neti und Gerasimos Potamianos Momentan führend. Die von dieser Gruppe für den Workshop 2000 an der Johns Hopkins University [10] genutzte Datenbasis wurde uns freundlicherweise zur Verfügung gestellt und dient als Basis für diese Arbeit.

Wie man an der Arbeit [10] erkennen kann, ist der erzielbare Gewinn durch eine Unterstützung der klassischen, rein akustischen, Spracherkennung sehr stark abhängig von der Qualität der akustischen Daten. Wenn die Audioaufnahmen in hinreichend guter Qualität vorliegen, fällt die mögliche Steigerung der Erkennungsleistung deutlich geringer aus. Die Verbesserung ist sowohl absolut, als auch relativ geringer als in den Fällen mit verrauschten Audiodaten.

Mit guten, unverrauschten Audiodaten wurde in [10] im besten Fall eine Verringerung der Wortfehlerrate (**W**ord **E**rror **R**ate (WER), siehe Kapitel 3.1) von 14,44% auf 13,47% mithilfe einer Multi-Stream-Architektur erreicht. Dies

entspricht absolut einer Verbesserung um ca. 1% und einer relativen Verbesserung von etwa 7%. Dass es nicht einfach ist, unter diesen Bedingungen eine Verbesserung zu erzielen, zeigt sich an der Tatsache, dass in insgesamt sechs Experimenten mit audiovisueller Spracherkennung nur in drei Fällen eine Verbesserung für den Fall *clean audio* erreicht werden konnte.

Die Aufgabe, die Erkennungsleistung bei verrauschten Audiodaten zu verbessern ist deutlich einfacher, da der Abstand zu einer perfekten Erkennung viel größer ist. In [10] wurde z.B. ein weiteres Sprachsignal als Störrauschen zu dem eigentlichen Sprachsignal dazugemischt, dieses zweite Signal besaß einen Signal-Rausch-Abstand von 10 db zum ersten.

Unter diesen Bedingungen trat im rein akustischen Fall, wie zu erwarten, eine deutliche Verschlechterung ein: die WER stieg auf 48,10%. Der Einfluss des Lippenlesens ist nun deutlich höher, als bei den unverrauschten akustischen Daten; alle sechs der in [10] durchgeführten Experimente führten zu einer deutlichen Verbesserung. Die besten Ergebnisse wurden mit dem gleichen Experiment erzielt, wie im ersten Fall, mit den nicht verrauschten Audiodaten. Die WER konnte auf 35,21% reduziert werden. Dies entspricht absolut einer Verbesserung von knapp 13% und einer relativen Verbesserung von ungefähr 27%. Die einzelnen Ergebnisse sind in Tabelle 1 nachzulesen.

Aus den Ergebnissen dieser Experimente lässt sich deutlich erkennen, dass der Einsatz von audiovisueller Spracherkennung gegenüber normaler, rein akustischer Spracherkennung vor allem eine deutliche Verbesserung bringt, wenn die Erkennungsleistung auf Grund von Störgeräuschen beeinträchtigt ist.

	Clean Audio	Noisy Audio
Audio-only	14,44%	48,10%
AV-Concat	16,00%	40,00%
AV-HiLDA	13,84%	36,99%
AV-MS-1	14,62%	36,61%
AV-MS-2	14,92%	38,38%
AV-MS-UTTER	13,47%	35,27%
AV-PROD	14,19%	35,21%

Tabelle 1: Wortfehlerraten in Prozent bei den Experimenten des Johns Hopkins Workshop 2000 [10].

3 Grundlagen

3.1 Fehlerarten und Optimierungskriterium

Bei text-to-speech-Systemen können drei verschiedene Arten von Fehlern auftreten. Diese möglichen Fehlerklassen sind die Ersetzungs-, Auslassungs- und die Einfügefehler.

Die erste Klasse von denkbaren Fehlern eines automatischen Spracherkenners sind die Ersetzungsfehler. Ein Ersetzungsfehler tritt auf, wenn ein Wort falsch erkannt wird, das System erkennt z.B. *und* statt *Hund*.

Die zweite Fehlerklasse sind die Auslassungsfehler, hier wird ein Wort als Störgeräusch interpretiert, oder zwei einzelne Wörter als eines erkannt, so dass in der Hypothese ein Wort fehlt.

Die dritte Klasse von möglichen Fehlern sind die Einfügefehler. In diesem Fall wird ein Wort in die Hypothese eingefügt, welches in der Äußerung gar nicht vorhanden war. Dies kann z.B. bei langen Wörtern geschehen, die fälschlicher Weise als zwei kürzere interpretiert werden.

`Ref.:` Spracherkennung ist eine schwierige Aufgabe.
`Hyp.:` Sprache Kennung ist schwierige Aufgabe.

In diesem Beispiel sind alle drei Fehlerklassen vorhanden. Zuerst tritt ein Ersetzungsfehler auf, das Wort *Spracherkennung* wird als *Sprache* erkannt. Da das Wort *Spracherkennung* nicht wirklich als *Sprache* erkannt wurde, sondern in zwei Wörter aufgetrennt worden ist, gibt es gleich im Anschluss einen Einfügefehler des Wortes *Kennung*. Der Auslassungsfehler tritt etwas später auf, da das Wort *eine* in der Hypothese nicht erkannt worden ist.

Nachdem geklärt ist, welche Arten von Fehlern bei der automatischen Spracherkennung auftreten können, muss nun ein geeignetes Maß gefunden werden, um die Qualität der Erkennung zu bewerten.

Ein einfaches Maß für die Erkennungsleistung ist z.B. die Worterkennungsrate. Diese berechnet sich wie folgt:

$$\frac{\#\ erkannte\ Wörter}{\#\ gesagte\ Wörter} * 100 = Erkennungsrate\ in\ Prozent$$

In dem obigen Beispiel wird eine Worterkennungsrate von 60% erzielt, da drei der fünf Wörter aus der Referenz korrekt erkannt worden sind.

Dieses Maß lässt sich aber überlisten, indem man zu jedem potentiellen Wort alle bekannten Wörter ausgibt. So macht man zwar sehr viele Einfügefehler, die in diesem Maß aber nicht berücksichtigt sind:

`Ref.`: Hier ist alles richtig.
`Hyp.`: Hier und heute ist fast alles komplett richtig.

Würde man die Worterkennungsrate als Maß für die Güte heranziehen, so hat man in dem zweiten Beispiel eine perfekte Erkennung. Die Worterkennungsrate beträgt 100%. Es ist zu sehen, dass die Worterkennungsrate kein geeignetes Optimierungskriterium für einen Spracherkenner darstellt.

Ein anderes Maß zum Vergleichen unterschiedlicher ASR-Systeme ist die Wortfehlerrate oder auf Englisch: **W**ord **E**rror **R**ate (WER). Dieses Maß berechnet sich wie folgt:

$$\frac{\#\ Ersetzungen + \#\ Auslassungen + \#\ Einfügungen}{\#\ Wörter} * 100 = WER$$

Da bei diesem Maß die Fehler und nicht die erkannten Wörter gezählt werden, kann die WER nur null sein, wenn wirklich keine Fehler gemacht worden sind. Andererseits kann die WER aber sehr wohl auf über 100% ansteigen, wenn von dem System zu viele Einfügefehler gemacht werden.

Die anderen Fehlerklassen können nicht zu einer WER von über 100% führen, da pro Wort in der gesagten Äußerung nur jeweils ein Fehler aus diesen Klassen möglich ist. Die Zahl der Einfügefehler ist dagegen nicht durch die Länge der gemachten Äußerung beschränkt.

Für das erste Beispiel errechnet man eine WER von 60%: bei einer Äußerung aus fünf Wörtern wurden drei Fehler gemacht. Im zweiten Beispiel beträgt die WER 100%, da zu den vier korrekt erkannten Wörtern noch vier weitere Wörter eingefügt worden sind.

Neben der Wortfehlerrate kann auch die Wortakkuratheit als Optimierungskriterium eines automatischen Spracherkennungssystems verwendet werden. Die Wortakkuratheit lässt sich einfach aus der Wortfehlerrate errechnen:

$$100 - Wortfehlerrate = Wortakkuratheit$$

Für die beiden in diesem Abschnitt besprochenen Beispiele ergibt sich eine Wortakkuratheit von 40% bzw. 0%.

3.2 Automatische Spracherkennung

3.2.1 Herausforderung ASR

Dem Menschen fällt es im Allgemeinen sehr leicht, gesprochene Sprache zu verstehen. Man könnte also meinen, dass die automatische Spracherkennung eine relativ einfache Aufgabe sein müsste, dies ist aber nicht der Fall. Warum fällt es Computern so schwer, menschliche Äußerungen zu verstehen?

Ein direkter Mustervergleich zwischen einer Äußerung und einer gespeicherten Referenzäußerung führt zu keinem zufrieden stellenden Ergebnis. Auf Signalebene haben zwei unterschiedliche, aber in der gleichen Lautstärke gesprochene Wörter mehr Ähnlichkeit miteinander, als ein Wort, welches einmal leise und einmal Laut ausgesprochen worden ist.

Verschiedene Aufnahmen ein und desselben Wortes/Satzes werden sich immer unterscheiden. In Abbildung 1 sieht man zwei Fouriertransformierte des Satzes *good luck comma Arnie period*, es ist zu erkennen, dass beide Äußerungen aus gleich vielen Wörtern bestehen, die jeweiligen Worte unterscheiden sich jedoch deutlich voneinander. Die Fouriertransformation ist bereits ein Schritt zur Vereinheitlichung der Signale, ein direkter Vergleich der Signalstärken zu gegebenen Zeitpunkten bringt noch weniger Übereinstimmungen. Darum müssen andere Verfahren für die Spracherkennung eingesetzt werden. Die häufigsten sind neuronale Netze [19] und stochastische Ansätze [20] zur Mustererkennung.

Wenn man es schafft, das Problem der sehr großen Varianz der Sprache in den Griff zu bekommen, so stellen sich noch weitere Probleme. Beim Sprechen kommt es häufig zu abgebrochenen Sätzen, wenn den Sprechern etwas Neues einfällt. Die Grammatikregeln werden weit weniger eingehalten, als dies in der Schriftsprache der Fall ist. Für die Erkennung verwendete Grammatikregeln können also nicht einfach aus dem Duden übernommen werden.

Auch die Trennung einzelner Wörter ist in gesprochener Sprache oft sehr schlecht. Im Deutschen ist dieses Problem nicht so ganz stark ausgeprägt,

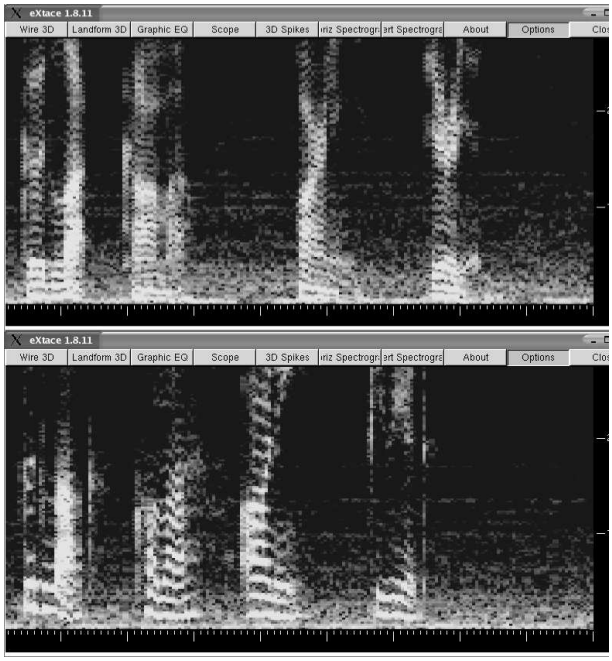

Abbildung 1: In beiden hier gesagten Äußerungen wurde der Text: *good luck comma Arnie period* gesagt.

im Englischen und Französischen kommen solche Verschleifungen aber sehr häufig vor.

Durch diese Reihe von Widrigkeiten fällt es auch dem Menschen sehr schwer, jedes einzelne Wort zu erkennen. Menschen sind jedoch auf Grund ihrer Erfahrung in der Lage, die fehlenden Teile aus dem Zusammenhang sehr gut zu rekonstruieren. Somit fällt es ihnen bei einer normalen Konversation nicht auf, wenn man nur einen Teil versteht.

Dem Computer fehlen diese Erfahrungen, er weiß im Allgemeinen nichts über den Inhalt der gesagten Äußerungen und versucht nur über neuronale Netze [19] oder Statistiken [20] passende Wörter zu den erkannten Lauten zu finden.

Auch innerhalb der speech-to-text-Systeme gibt es noch unterschiedliche Anwendungsgebiete, welche die Komplexität der Aufgabe der automatischen Spracherkennung bestimmen [21]. Einen Überblick über die verschiedenen Anforderungen gibt Tabelle 2.

	leichter	schwerer
kontinuierliche Sprache	nein	ja
Vokabular	klein	groß
Sprechart	vorgelesen	frei
OOV Worte erlaubt	nein	ja

Tabelle 2: Unterschiedliche Komplexität bei speech-to-text-Systeme.

Es ist möglich, dass lediglich Einzelwörter [22] erkannt werden sollen oder aber ganze Sätze [7]. Hierbei ist die Erkennung von Sätzen die deutlich schwierigere Aufgabe, da bei der Einzelworterkennung im Allgemeinen die kurzen, leicht zu verwechselnden Wörter (*a* vs. *the*) nicht enthalten sind. In der audiovisuellen Spracherkennung wird häufig noch mit Eingaben in buchstabierter Form [16, 17] oder mit Ziffern [18] gearbeitet.

Ein weiterer Punkt der die Komplexität der Aufgabe von speech-to-text-Systemen beeinflusst ist die Frage, wie gesprochen wird. Handelt es sich um vorbereiteten, vorgelesenen Text [7, 10] oder um spontane Sprache [23]? Hier ist klar, dass die spontane Sprache eine viel größere Herausforderung darstellt. Außerdem spielt die Größe des verwendeten Vokabulars auch eine Rolle dabei, wie aufwändig es ist, die Aufgabe zu lösen: In einem kleinen Vokabular sind wesentlich weniger Einträge vorhanden, die ähnlich klingen.

Ein kleines Vokabular kann von lediglich zwei Einträgen (Ja/Nein) über Ziffernfolgen (10 Einträge), buchstabieren (26 Einträge) bis zu einem recht eingeschränkten Wortschatz mit einigen hundert Einträgen reichen. Von einem großen Vokabular spricht man bei einem zur Verfügung stehenden Wortschatz von mehreren tausend Wörtern. Das für diese Arbeit verwendete Wörterbuch enthält ca. 10.500 Einträge.

Auch stellt sich die Frage, ob alle gesprochenen Wörter bereits beim Training bekannt sind, bzw. der Sprecher darauf beschränkt ist, nur bekannte Wörter zu verwenden. Ist es ihm erlaubt, aus allen Wörtern frei zu wählen, so wird die Aufgabe für den Spracherkenner viel schwieriger: es ist möglich, Wörter zu verwenden, die im Vokabular nicht vorkommen, so genannte ***O**ut **O**f **V**okabulary* Wörter (OOV-Wörter). Da solche Wörter nicht erkannt werden können und auf etwas anderes abgebildet werden, ziehen sie meist noch Folgefehler nach sich.

3.2.2 Mathematische Problemstellung

Das Ziel von automatischer Spracherkennung ist es, eine Wortfolge W' zu finden, die am besten zu der gesprochenen Äußerung X passt. Gesucht wird also die Wortfolge, welche folgende Bedingung erfüllt:

$$W' = argmax_W P(W|X)$$

Nun kann $P(W|X)$ nicht direkt berechnet werden, weil nur das Sprachmodell und das akustische Modell zur Verfügung stehen. Das Sprachmodell gibt für jede mögliche Wortfolge eine Wahrscheinlichkeit an, kann also als $P(W)$ geschrieben werden.

Das akustische Modell gibt an, wie gut die aufgezeichnete Äußerung zu einer gegebenen Phonemfolge passt. Es lässt sich mathematisch als $P(X|W)$ beschreiben. Das akustische Modell könnte bereits als Spracherkenner dienen, die Erkennungsgenauigkeit von Phonemen ist jedoch nicht hoch genug, um eine gute Hypothese zu erzeugen.

Zur Lösung des Problems, die Wahrscheinlichkeit $P(W|X)$ zu berechnen, kann nun der Satz von Bayes heran gezogen werden:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)}$$

Müsste man den Wert der Wahrscheinlichkeit von $P(W|X)$ bestimmen, so hätte man das Problem, dass nicht bekannt ist, mit welcher Wahrscheinlichkeit die Äußerung X auftritt. Der Wert von $P(X)$ ist aber für den hier vorliegenden Fall unerheblich, da lediglich die Äußerung mit der größten Wahrscheinlichkeit bestimmt werden soll, nicht die Wahrscheinlichkeit selber. Dadurch darf der Nenner weggelassen werden und es sind alle Elemente bekannt, die benötigt werden, die wahrscheinlichste Wortfolge zu berechnen:

$$W' = argmax_W P(W|X) =$$
$$argmax_W \frac{P(X|W)P(W)}{P(X)} = argmax_W P(X|W)P(W)$$

3.2.3 Einheiten in der Spracherkennung

Um einen automatischen Spracherkenner gut trainieren zu können ist es notwendig, dass von jeder zu lernenden Einheit ausreichend Trainingsdaten vorhanden sind. Komplette Wörter direkt zu erkennen ist demnach eine sehr schwere Aufgabe, da sehr große Mengen an Trainingsdaten benötigt werden. Zusätzlich stellt sich noch das Problem, dass ein und dasselbe Wort auf sehr unterschiedliche Weisen ausgesprochen werden kann.

Aus diesem Grund versucht man kleinere Einheiten, so genannte (Sub-)Phoneme, zu erkennen und aus diesen dann mögliche Wörter zu bilden [24]. Jedes Phonem ist ein eindeutiger Laut. Aus diesen Lauten werden dann die Wörter aufgebaut. Da ein Laut nicht einem Buchstaben in der geschriebenen Sprache entspricht, benötigt ein ASR-System ein Aussprachelexikon. In diesem sind die Phonemfolgen enthalten, aus denen Wörter zusammengesetzt werden können. Hier sind drei Aussprachevarianten des Wortes *hello* zu sehen, wie sie in dem, für diese Arbeit verwendeten, Aussprachelexikon vorkommen:

{ { HH WB} EH L { OW WB}}
{ { HH WB} AX L { OW WB}}
{ { HH WB} L { OW WB}}

Es gibt zwar mehr Phoneme als es Buchstaben gibt, aber deutlich weniger als Wörter, nämlich nur 50 bis 100 Stück in einer Sprache. Dadurch kommen die einzelnen Phoneme wesentlich häufiger in den Trainingsdaten vor als ganze Wörter und können somit besser trainiert und anschließend erkannt werden.

Die *International Phonetic Association* hat ein Phonemalphabet standardisiert, welches unter [25] zu finden ist. Ein Beispiel, in dem diese Lautschrift des Phonemalphabets genutzt worden ist, zeigt Abbildung 2.

Abbildung 2: Ein Text *International Phonetic Association* im IPA-Alphabet.

3.2.4 Neuronale Netze

Eine Möglichkeit die Trainingsdaten zu erlernen und anschließend für die automatische Spracherkennung zu nutzen, sind neuronale Netze. Neuronale Netze sind der Versuch, ein Gehirn - in sehr stark vereinfachter Form - im Computer nachzubilden. Sie bestehen aus einzelnen, miteinander verbundenen Neuronen, von denen einige für die Eingabe von Informationen dienen und einige andere für die Ausgabe, der aus der Eingabe gewonnenen Schlussfolgerung.

Neuronale Netze sind vor allem früher bei kleinen Problemstellungen, z.B. der Erkennung von buchstabierten Sequenzen, in der automatischen Spracherkennung zum Einsatz gekommen. An der Universität Karlsruhe gab es vor etwa zehn Jahren bereits eine Diplomarbeit über audiovisuelle Spracherkennung [26], in welcher erfolgreich neuronale Netze zum Einsatz gekommen sind.

Unter diesen Bedingungen haben sie durchaus gute Ergebnisse geliefert, aber in dem Aufgabengebiet, kontinuierliche Sprache mit großem Vokabular zu erkennen, hat sich gezeigt, dass die neuronalen Netze den stochastischen Mustererkennern unterlegen sind. Aus diesem Grund konzentriert sich diese Arbeit auf die Aspekte der stochastischen Ansätze.

3.2.5 Stochastische Mustererkenner / Hidden-Markov-Modelle

Heute werden neuronale Netze nur noch selten in der automatischen Spracherkennung verwendet, wesentlich gebräuchlicher sind stochastische Mustererkenner. Diese bestimmen Wahrscheinlichkeiten für das Auftreten eines bestimmten (Sub-)Phonems bei einer gegebenen Eingabe.

Um (Sub-)Phoneme zu erkennen, wird der kontinuierliche Datenstrom in kleine Teilstücke von üblicherweise 20 ms Länge und 10 ms Versatz zwischen den einzelnen Teilen zerlegt. Für jedes dieser Stücke erhält man Wahrscheinlichkeiten für die Zugehörigkeiten zu den (Sub-)Phonemen. Anschließend errechnet man aus den wahrscheinlichsten Phonemfolgen mögliche Wörter mithilfe des Aussprachelexikons. Mit Unterstützung des Sprachmodells [27] wird dann die wahrscheinlichste Wortfolge berechnet und diese als Hypothese ausgegeben. In Abbildung 3 sieht man die Zuordnung der Hypothese eines Wortes zu einem möglichen Kandidaten aus dem Lexikon. Solche Zuordnungen werden für alle wahrscheinlichen Wörter durchgeführt und aus diesen wird dann versucht, Hypothesen für ganze Sätze zu bilden.

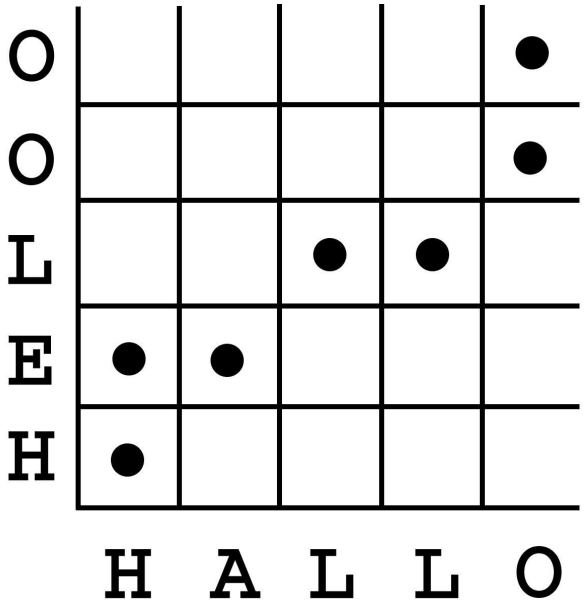

Abbildung 3: Berechnung der Distanz zweier Äußerungen. Diagonal: perfekte Zuordnung, Horizontal: einfügen, Vertikal: auslassen.

Zur Bestimmung der Wahrscheinlichkeiten der (Sub-)Phoneme werden üblicherweise Hidden-Markov-Modelle (HMM) [28] verwendet.

Ein Hidden-Markov-Modell besteht zum Einen aus einer Markovkette, ein Beispiel zeigt Abbildung 4. Sie repräsentiert das Wort *can*, bestehend aus den Phonemen */k/ /ae/ /n/*. Jedes dieser Phoneme besteht wiederum aus drei Subphonemen, einem Anfangs-, einem Mittel- und einem Endzustand.

Zusätzlich zur Markovkette besitzt ein Markov-Modell noch Übergangswahrscheinlichkeiten zwischen den einzelnen Zuständen der Kette. Ebenso gibt es zu jedem Zustand eine Emissionswahrscheinlichkeit mit der eine Ausgabe erzeugt wird. Das *Hidden* des Hidden-Markov-Modells kommt daher, dass ein Beobachter nur die Ausgaben des Systems zu sehen bekommt, er erhält jedoch keine Informationen über die Zustandfolge, die durchlaufen worden ist. Ein Hidden-Markov-Modell ist in Abbildung 5 zu sehen. Die Wahrscheinlichkeit eines Zustands, ein Anfangszustand zu sein ist hier immer null, außer für den ersten Zustand.

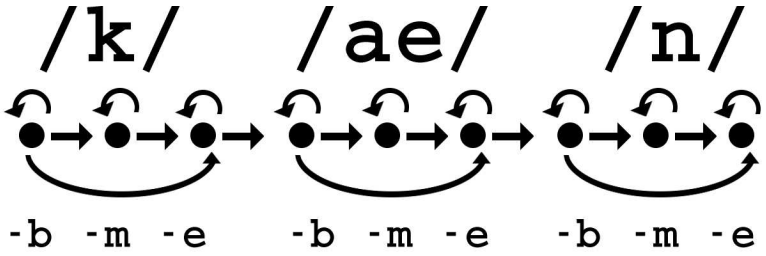

Abbildung 4: Eine Markovkette für das Wort *can* bestehend aus drei Subphonemen je Phonem.

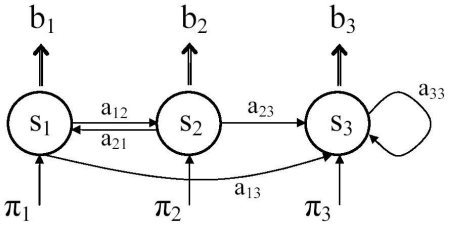

Abbildung 5: Zu sehen ist ein Hidden-Markov-Modell mit den Zuständen s_i, den Übergangswahrscheinlichkeiten a_ij und den Emissionswahrscheinlichkeiten $b_i(k)$. Die Wahrscheinlichkeit, dass s_i ein Anfangszustand ist beträgt π_i.

Das Training eines Spracherkenners besteht aus der Optimierung der Wahrscheinlichkeiten, so dass die emittierte Ausgabefolge der korrekten Phonem-

folge möglichst gut entspricht. Eine Beschreibung des HMM-Trainings gibt es z.B. unter [29], eine deutlich ausführlichere Beschreibung von Hidden-Markov-Modellen ist unter [28] zu finden. Bevor das Training beginnen kann, müssen die Parameter initialisiert werden. Hierfür werden im Allgemeinen die Parameter eines bereits existierenden Spracherkenners verwendet.

Wenn nun Anfangswerte vorhanden sind, kann die erste Iteration des Trainingsprozesses starten. Für das Training wird der Forward-Backward- oder der Viterbi-Algorithmus eingesetzt. Anschließend werden die Parameter des HMM mithilfe des Baum-Welch-Verfahrens neu geschätzt. Erklärt wird der Algorithmus in [30] Seite 136ff. Dieses Vorgehen wird nun iterativ wiederholt, bis für andere, neue Daten, die nicht in den Trainingsdaten vorhanden sind, ein Optimum erreicht worden ist.

3.3 Audiovisuelle Spracherkennung

3.3.1 Menschliche Fähigkeit des Lippenlesen

Auch der Mensch bedient sich beim Erkennen von Sprache mehrerer Modalitäten gleichzeitig [6]. Er beschränkt sich im Allgemeinen nicht nur auf Eingaben in akustischer Form, auch wenn er dazu in der Lage ist, z.B. beim Telefonieren.

Wenn die Akustik sich aber verschlechtert, bedient sich der Mensch verstärkt der Möglichkeit des Lippenlesens, um sein Gegenüber besser verstehen zu können. Dieses Phänomen ist bekannt unter dem Namen Cocktailparty-Effekt oder Barty-Effekt. Menschen sind in der Lage, einem Gespräch auch dann noch zu folgen, wenn sich mehrere Personen in einem Raum mit ähnlicher Lautstärke unterhalten [31]. Dazu ist es von Vorteil, dass der Mensch zwei Ohren besitzt und somit orten kann, aus welcher Richtung Schall zu ihm dringt. Außerdem schauen sich die Menschen intensiver auf den Mund, wenn die Umgebungsgeräusche zunehmen. Im Normalfall schaut man sich häufig im Raum um, wenn es lauter wird blickt man aber wesentlich länger auf das Gesicht seines Gesprächspartners.

Selbst wenn viele Menschen gleichzeitig sprechen und man auf einer reinen Tonaufnahme keinen der einzelnen Sprecher mehr heraushören kann, so ist der Mensch dennoch in der Lage, in einer echten Gesprächssituation sein Gegenüber zu verstehen. Diese Fähigkeit erlangt der Mensch zum einen daraus, dass er orten kann, aus welcher Richtung die für ihn relevante Stimme kommt

und andere Geräusche im Gehirn ausgefiltert werden können, zum anderen aber auch daraus, dass die relevanten akustischen Informationen von visuellen Informationen unterstützt werden.

In vielen Fällen lässt sich das Ohr sogar vom Auge überzeugen etwas anderes zu erkennen, als wirklich gesprochen wurde. Dieses Phänomen nennt man den McGurk-Effekt [32]. McGurk machte einen Versuch, bei dem zu dem selben akustischen Signal unterschiedliche *visuelle Laute* gezeigt wurden. Das Video zeigte einen Sprecher, der die Lippenbewegungen der Laute */ba/* und */ga/* macht. Die dazu abgespielte Tonspur enthielt jedoch immer denselben Laut. Die Versuchspersonen haben meist die *visuellen Laute* wahrgenommen. Daran kann man erkennen, dass in der Lippenbewegung durchaus für das Sprachverstehen relevante Informationen vorhanden sind. Ziel ist es nun, diese zusätzlichen Informationen für die automatische Spracherkennung zu nutzen, um die Erkennungsleistung zu verbessern.

3.3.2 Einheiten in der visuellen Spracherkennung

Genau wie ein akustischer Spracherkenner die Wörter aus eindeutigen Lauten zusammensetzt, gibt es für den visuellen Fall Klassen von optisch unterscheidbaren *visuellen Lauten*. Diese Klassen nennen sich Viseme, und sie werden, genau wie die Phoneme, als Grundbausteine für den Aufbau der Wörter genutzt.

Im Normalfall werden Phoneme, die sich visuell (fast) nicht unterscheiden lassen zu einem Visem zusammengefasst. Ob ein Laut stimmhaft oder stimmlos ist, lässt sich nicht an der Stellung der Lippen ablesen, z.B. die Phoneme */f/* und */v/* werden als ein Visem behandelt. Somit hat man im Allgemeinen deutlich weniger Viseme als Phoneme.

In dieser Arbeit wurden dreizehn verschiedene Viseme verwendet, die der Erkenner erkennen musste, und zwar die in [10] aufgeführten und in Tabelle 3 zu sehenden. Bis auf das *silence*-Visem besitzen sie jeweils einem Anfangs-, Mittel- und Endzustand.

Dies ergibt insgesamt 36 verschiedene Visem-Zustände, die Laute repräsentieren, und einen 37. Visem-Zustand für den Ruhezustand, das Visem SIL-m.

Unter diesen dreizehn Visemen gibt es Laute, die sich sehr ähnlich anhören, aber sehr unterschiedlich aussehen (*/b/* vs. */d/*). Ganz besonders bei solchen Lauten können zusätzliche visuelle Informationen die Erkennungsgenauigkeit steigern.

silence	/sil/
V01	/ao/, /ah/, /aa/, /er/, /oy/, /aw/, /hh/
V02	/uw/, /uh/, /ow/
V03	/ae/, /eh/, /ey/, /ay/
V04	/ih/, /iy/, /ax/
V05	/l/, /el/, /r/, /y/
V06	/s/, /z/
V07	/t/, /d/, /n/, /en/
V08	/sh/, /zh/, /ch/, /jh/
V09	/p/, /b/, /m/
V10	/th/, /dh/
V11	/f/, /v/
V12	/ng/, /k/, /g/, /w/

Tabelle 3: Liste der verwendeten Viseme.

3.3.3 Verbindung zur akustischen ASR

Zur visuellen Spracherkennung, dem Lippenlesen, können dieselben Verfahren verwendet werden, wie bei der akustischen Spracherkennung. Da es sich um lernende Verfahren handelt, können Daten beliebiger Herkunft zum Trainieren verwendet werden. Lernenden Verfahren ist es egal, ob sie mit Audiodaten oder mit Videodaten trainiert werden, solange in den verwendeten Trainingsmerkmalen ausreichend Information vorhanden ist, um etwas zu lernen.

Der Unterschied in den Systemen liegt in der Vorverarbeitung der Daten: bei der visuellen Spracherkennung wird z.B. bevorzugt die Cosinustransformation verwendet, bei der akustischen die Fouriertransformation. Die genauen Methoden der Vorverarbeitung von Audio- und Videodaten werden im nächsten Kapitel beschrieben.

Ein weiterer Unterschied besteht darin, dass in der Videoverarbeitung die einzelnen Wörter nicht aus Phonemen sondern aus Visemen zusammengesetzt werden.

Der Trainings- und Erkennungsprozess läuft ansonsten genau so ab, wie beim rein akustischen Spracherkenner. Für eine erfolgreiche audiovisuelle Spracherkennung muss allerdings noch eine geeignete Methode zur Fusionierung von Audio- und Videodaten implementiert werden.

4 Datenbasis und -verarbeitung

4.1 Datenbasis

Die Experimente dieser Arbeit wurden auf der Datenbasis des Workshops 2000 der Johns Hopkins University durchgeführt [10]. Es war ein großer Vorteil, dass bereits eine umfangreiche Datenbasis zur Verfügung stand und diese nicht erst erhoben werden musste. Das Sammeln von Daten für die Spracherkennung ist, bereits für den rein akustischen Fall, eine sehr aufwändige und kostspielige Angelegenheit. Für audiovisuelle Daten ist die Erhebung nochmal aufwändiger, aus diesem Grund sind Weltweit nur wenige große Datenbasen für dieses Anwendungsgebiet vorhanden.

Die in dieser Arbeit verwendete Datenbasis besteht aus Videoaufnahmen, in welchen die Gesichter der Sprecher frontal von vorne in Großaufnahme zu sehen sind. Die Videos liegen im mpeg-Format vor und besitzen eine Auflösung von 704 x 480 Pixel bei einer Frequenz von 30 Hz. Abbildung 6 zeigt Bilder der Personen aus einigen willkürlich ausgewählten Videosequenzen. Es handelt sich um Farbaufnahmen und der Ton ist als mp3-Stream in die Videos eingebettet. Aufgenommen wurden die Daten mit einer Samplingfrequenz von 16 kHz, bei der Konvertierung ins mp3-Format wurden sie jedoch mit einer Samplingfrequenz von 22 kHz abgespeichert. Ein Teil der Daten besitzt lediglich einen Ton-Kanal, der andere Teil der Videos liegt mit Stereoton vor. Zu allen Äußerungen sind außerdem Transkriptionen der gesprochenen Wörter vorhanden.

Insgesamt haben die aufgezeichneten Äußerungen eine Gesamtlänge von ca. 40 Stunden, diese teilen sich in etwa 35 Stunden Trainingsdaten und knapp fünf Stunden Testdaten auf. Für das Training stehen 17.111 Äußerungen von 261 Sprechern zur Verfügung, für den Test gibt es 1.893 Äußerungen von 26 Sprechern. Die Einteilung in Sprecher für Trainings- bzw. Testdaten, zu sehen in Tabelle 4, erfolgte wie in [10].

	Äußerungen	Dauer	Sprecher
Training	17.111	34,9h	261
Test	1.893	4,6h	26

Tabelle 4: Anzahl und Dauer der Äußerungen in der Datenbasis für das Training bzw. Test.

Abbildung 6: Beispielbilder aus zufällig ausgewählten Videosequenzen.

4.2 Trennung von Audio- und Videodaten

Die vorhandenen Videodaten, die für das Training des Spracherkenners genutzt werden sollen, liegen im Dateityp mpeg-Video vor, d.h. Video- und Audiodaten sind in einer gemeinsamen Datei vorhanden und müssen im ersten Schritt voneinander getrennt werden. Außerdem sind beide Datentypen nur in komprimierter Form vorhanden, auf der einen Seite der mpeg-Videostream für die visuellen Informationen, auf der anderen Seite ein in das Video eingebetteter mp3-Stream, der die akustischen Informationen enthält.

Zur Trennung von Audio- und Videodaten in zwei separate Dateien wurde das Programm *transcode* [33] verwendet und diese Dateien werden dann wie im Folgenden beschrieben weiterverarbeitet.

Nachdem akustische und visuelle Daten voneinander getrennt worden sind, müssen sie noch so aufbereitet werden, dass der Spracherkenner sie verarbeiten kann.

Ziel dieser Vorverarbeitung ist es, die Audiodaten in der üblichen Form, d.h. als wav-Datei bzw. adc-Datei, vorliegen zu haben. Die Videodaten sollten in einer Matrix mit einer Zeile für die Mundregion jedes Bildes abgelegt werden.

4.3 Audioverarbeitung

Nach der Trennung von akustischen und visuellen Merkmalen liegen die Audiodaten nun im mp3-Format mit einer Samplingfrequenz von 22 kHz vor. Als Eingabe für den Spracherkenner werden die Daten aber im wav- oder adc-Format mit einer Samplingfrequenz von 16 kHz benötigt.

Um diese Konvertierungen durchzuführen, wird auf das Programm *lame* [34] zurückgegriffen. In einem ersten Schritt wird die Samplingfrequenz von 22 kHz auf 16 kHz reduziert. Anschließend werden, wieder mithilfe des Programms *lame*, die mp3-Dateien in wav-Dateien konvertiert. Da wav-Dateien sehr groß sind und sich relativ gut komprimieren lassen, werden sie am Ende noch mit dem Programm *shorten* gepackt, wodurch sich der benötigte Festplattenbedarf für die Audiodaten in etwa auf die Hälfte reduziert.

Bei *shorten* handelt es sich im Gegensatz zu mp3 um verlustlose Kompression, womit sich die Qualität gegenüber den verlustbehafteten mp3-Dateien nicht steigern lässt, weil bei der Konvertierung keine neuen Informationen gewonnen werden können. Im Gegensatz zu mp3 schadet es den Audiodaten in

verlustlos komprimierten Dateiformaten aber nicht, wenn auf diese mehrere Verarbeitungsschritte angewandt werden.

Obwohl die Daten ursprünglich in einem Datenformat mit verlustbehafteter Kompression vorgelegen haben, ist das für die Spracherkenner kein großes Problem, da nur eine recht schwache Kompression vorgenommen wurde und somit noch sehr viele charakteristische Merkmale der Sprache vorhanden sind. Außerdem arbeitet das mp3-Format so, dass die Veränderungen möglichst nicht vom menschlichen Ohr wahrgenommen werden können. Da auch der Mensch Spracherkennung betreibt, müssen die für diesen Zweck relevanten Informationen weiterhin in dem Signal vorhanden sein.

Da bis jetzt nur Konvertierungen vorgenommen worden sind, die dazu dienen, dass der hier verwendete Spracherkenner die Daten verarbeiten kann, so beginnt nun die eigentliche Audioverarbeitung zur Spracherkennung, bei der möglichst relevante Merkmale des Datenstroms hervorgehoben und extrahiert werden sollen. Es ist nämlich nicht möglich, den Erkenner direkt auf den wav-Dateien zu trainieren, da zum einen die Datenmengen zu groß und zum anderen die relevanten Informationen noch zwischen vielen unbedeutenden, bzw. sogar irreführenden Informationen verborgen ist. Die Schritte der Audioverarbeitung sind in den meisten ASR-Systemen sehr ähnlich, auch in dieser Arbeit wurde eine Standardvorverarbeitung durchgeführt, wie sie in Abbildung 7 zu sehen ist.

Im ersten Schritt wird das Signal geglättet, kurze Ausschläge wie z.B. Knackser werden, wie in Abbildung 8 dargestellt, herausgefiltert. Dies geschieht, da sonst in der nächsten Stufe diese Knackser sehr stark zur Geltung kämen, auch wenn sie für das Sprachverstehen keinen Beitrag liefern.

Nach dem Entfernen der Störungen werden Kurzzeit-Fouriertransformierte des Signals berechnet, d.h. ein kurzes Stück von 20 ms wird als sich ständig wiederholendes, periodisches Signal betrachtet und auf diesem Signal wird eine Fouriertransformation berechnet. Anschließend wird das Fenster um 10 ms verschoben und der Vorgang wiederholt. So erhält man alle 10 ms einen Satz von 160 Fourierkoeffizienten. Diese geben an, wie stark die unterschiedlichen Frequenzen in dem untersuchten Teilstück der Audiodaten vorhanden sind. In Abbildung 9 ist die Fouriertransformierte einer Äußerung zu sehen; die hellen Bereiche sind stark im Signal vertretene Frequenzen, die dunklen Bereiche enthalten wenig Signalenergie.

Auf diesen Fourierkoeffizienten wird als Nächstes eine Normalisierung der Vokaltraktlängen berechnet. (Englisch: **V**ocal **T**ract **L**ength **N**ormalisation, kurz VTLN)

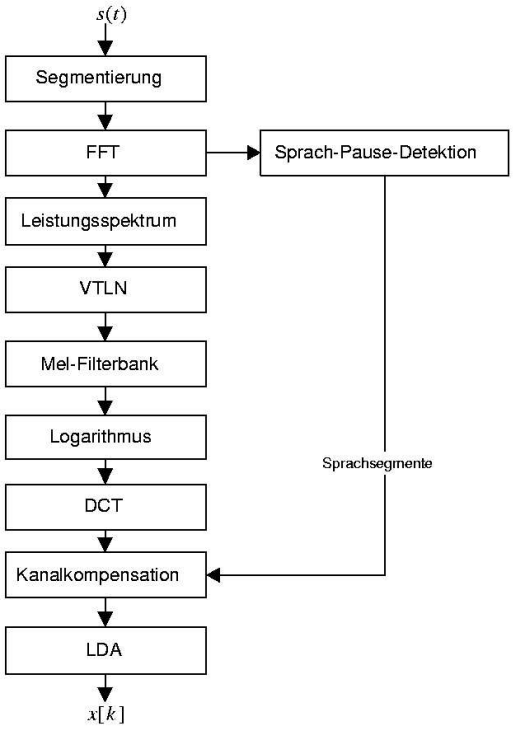

Abbildung 7: Eine Übersicht über die Standard Audioverarbeitung, wie sie in automatischen Spracherkennungssystemen durchgeführt wird.

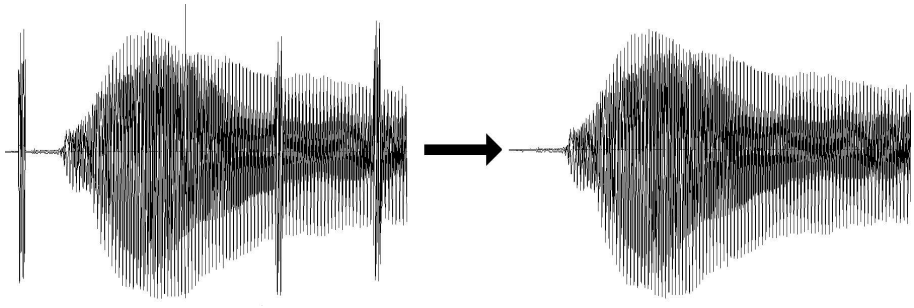

Abbildung 8: Entfernung von Knacksern, um möglichst gute Ausgangsbasis für Fouriertransformation zu erzeugen.

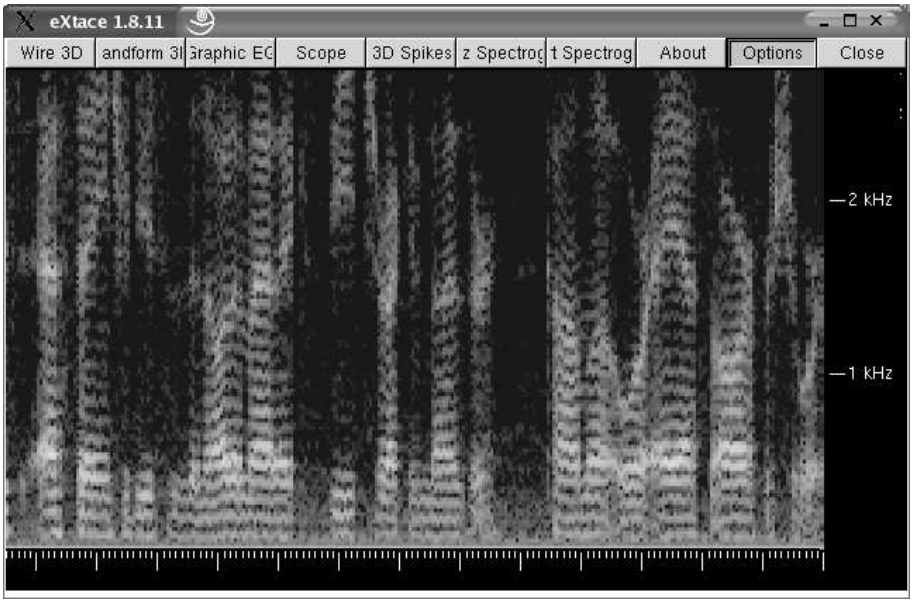

Abbildung 9: Die Fouriertransformierte einer Äußerung. Helle Bereiche enthalten viel, dunkle wenig Energie.

Durch die Länge des Vokaltraktes wird die Höhe der Stimme bestimmt, da durch seine Größe die Resonanzfrequenz des Sprechers festgelegt ist. Frequenzen, die einen sehr starken Anteil in dem Sprachsignal besitzen, sind Vielfache dieser Resonanzfrequenz. Durch die VTLN wird nun versucht, Fourierkoeffizienten so zu verschieben, dass die Resonanzfrequenz möglichst nicht mehr vom Sprecher abhängt, sondern nur von dem gesprochenen Laut (Abbildung 10).

Da der Mensch tiefe Frequenzen deutlich besser unterscheiden kann als hohe, versucht man dieses Verhalten auch bei der Spracherkennung zu nutzen. Kleine Unterschiede können bei tiefen Frequenzen bereits einen relevanten Unterschied für die Bedeutung ausmachen, während ein deutlich größerer Unterschied in einem höreren Frequenzbereich noch irrelevant sein kann. Um diesem Verhalten Rechnung zu tragen, werden die Daten als Nächstes in die Mel-Skala transformiert.

In der Mel-Skala entsprechen die Werte dem Hörempfinden der Menschen. Die Daten werden so skaliert, dass ein gleich großer Unterschied im Zahlenwert auch gleich wahrgenommen wird, unabhängig von dem absoluten Wert. Die Mel-Skala gleicht also Unterschiede im Auflösungsvermögen des menschlichen Ohres aus. Durch die Mel-Skalierung hat man erreicht, dass die Relevanz der Informationen linear verläuft, d.h. es können für die wei-

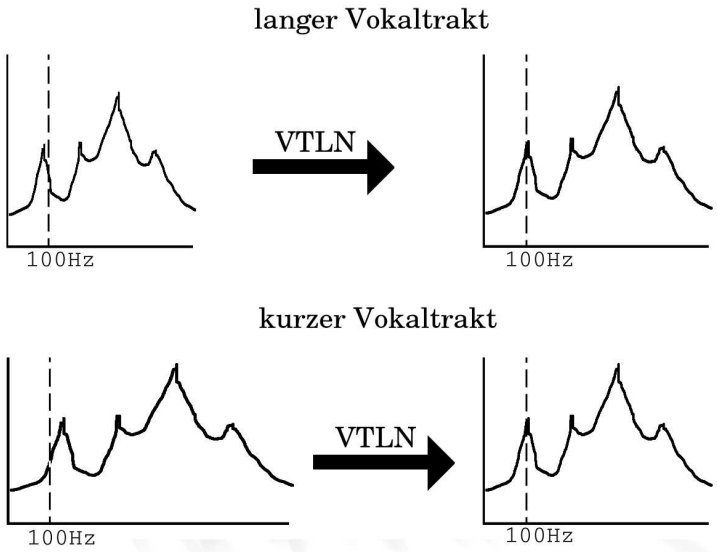

Abbildung 10: Durch die Längennormalisierung des Vokaltraktes werden die Spektren so gestreckt, dass die Formanten übereinstimmen.

tere Verarbeitung Filterbänke konstanter Breite verwendet werden, was die Verarbeitung vereinfacht.

$$f_{mel} = 1125 * log(0,0016 * f + 1)$$

Durch die Mel-Skalierung werden aus den Fourierkoeffizienten die Cepstralkoeffizienten berechnet. Das Cepstrum besitzt die Eigenschaft, dass die für die Spracherkennung relevanten Daten in den unteren Cepstralkoeffizienten stecken, man also zur Datenreduktion einfach die obersten Koeffizienten abschneiden kann. Bei dem hier verwendeten Spracherkenner werden nur die ersten 13 Cepstralkoeffizienten verwendet. Als Nächstes wird eine cepstrale Mittelwertsubtraktion durchgeführt, dieser Schritt dient der Lautstärkenormalisierung der gesprochenen Äußerung.

Anschließend werden die Cepstralkoeffizienten des gerade zu erkennenden Zeitpunktes noch mit einigen benachbarten verknüpft, um etwas zeitlichen Kontext zu erhalten. Wenn dieses geschehen ist, werden mithilfe einer LDA die relevanten Daten hervorgehoben und unwichtige oder sogar verwirrende Daten abgeschwächt [35, 36]. Mit diesen LDA-Koeffizienten wird nun der akustische Teil des Spracherkenners trainiert.

4.4 Videoverarbeitung

Zur Analyse der Videodaten und zur Extraktion der Mundregion wurde ein Programm von Rainer Stiefelhagen [37, 38] verwendet, auf welches ich im nächsten Abschnitt noch genauer eingehen werde.

Da dieses Programm keine mpeg-Videos verarbeiten kann, sondern eine leicht abgewandelte Form des ppm-Dateiformats, müssen die Videos zuerst in dieses Format umgewandelt werden.

Das ppm-Dateiformat ist eigentlich ein sehr einfaches Format für Bilddaten. Eine Beschreibung ist unter [39] zu finden. Dieses Format wurde für die Gesichtsanalyse zu einem sehr einfachen Videoformat erweitert. Der Header ist unverändert zum Originalformat, dann folgen die eigentlichen Bilddaten. Die Bilddaten der einzelnen Bilder werden hintereinander gespeichert, wie in Abbildung 11 zu sehen ist.

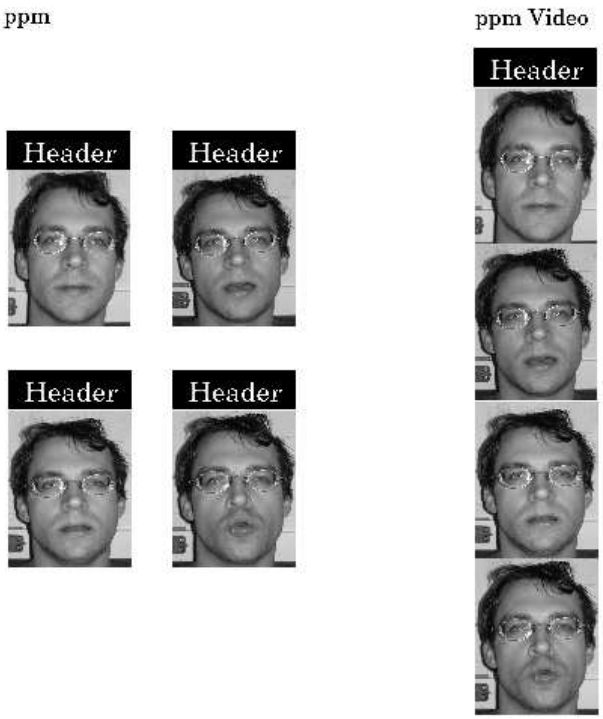

Abbildung 11: Erweiterung des ppm-Dateiformates zu einem einfachen Videoformat.

Um Dateien in diesem Videoformat zu erzeugen, wurde das ppm-output-Plugin von *transcode* [33] so verändert, dass eine Datei pro Video erzeugt wird und nicht mehr eine Datei pro Frame, wie es dies im Original macht.

Diese Datei wird als Eingabe für das in [37] beschriebene Programm genutzt. Dieses sucht in jedem Bild die Augen, die Nasenlöcher und die Mundwinkel des abgebildeten Gesichtes. Im Rahmen dieser Arbeit wurde das Programm so erweitert, dass mithilfe der gefundenen Koordinaten der Mundwinkel, zu sehen in Abbildung 12, die Mundregion automatisch ausgeschnitten werden kann. Dieser Bereich wird, unabhängig von der Größe in den Originalvideodaten, auf eine Größe von 64 x 64 Pixel skaliert, um unterschiedliche Abstände der Sprecher zur Kamera in den Aufnahmen ausgleichen zu können.

Abbildung 12: Bildausschnitt um den Mund des Sprechers mit gefundenen Mundwinkeln und Nasenlöchern.

Für die akustische Spracherkennung liegen zwischen zwei betrachteten Zeitfenstern jeweils 10 ms. Also gibt es insgesamt 100 Fenster pro Sekunde, und es wäre ideal, wenn die Videodaten ebenfalls eine Frequenz von 100 Hz aufweisen würden. Da die Videos aber nur mit 30 Hz aufgenommen sind, muss das Videomaterial interpoliert werden. Um von 30 Hz auf 100 Hz zu kommen bietet es sich an, zwei Bilder jeweils dreimal zu wiederholen und das darauf folgende Bild viermal. Hiermit kommt man dann ebenfalls auf eine Bildfrequenz von 100 Hz, mit einfachem Wiederholen der Bilder. Durch aufwändigere Verfahren der Interpolation sollte an diesem Punkt noch die Möglichkeit für Verbesserungen bestehen, die für dieses initiale System aber noch nicht ausgeschöpft worden sind.

Da die Lippenbewegungen der Tonerzeugung immer einen Moment vorauseilen [40, 41], ist es für eine ideale Synchronisation zwischen Audio- und Videodaten notwendig, die Videosignale etwas zu verzögern. Hierzu werden die

Videodaten um 60 ms verschoben. Die Größe von 60 ms ist ein experimentell ermittelter Wert, bei dem der Erkenner die besten Ergebnisse liefert.

Anschließend wird auf der ausgeschnittenen und skalierten Mundregion eine Histogrammnormalisierung durchgeführt, um Unterschiede in der Beleuchtung bzw. Hautfarbe ausgleichen zu können. Die Wirkungsweise dieser Maßnahme ist in Abbildung 13 zu erkennen. Die so vorverarbeiteten Bilddaten der Mundregionen werden dann in einer Matrix gespeichert. Jede Zeile der Matrix entspricht dabei einem Bild des Videos, bestehend aus 4096 Grauwerten für die 64 x 64 Pixel.

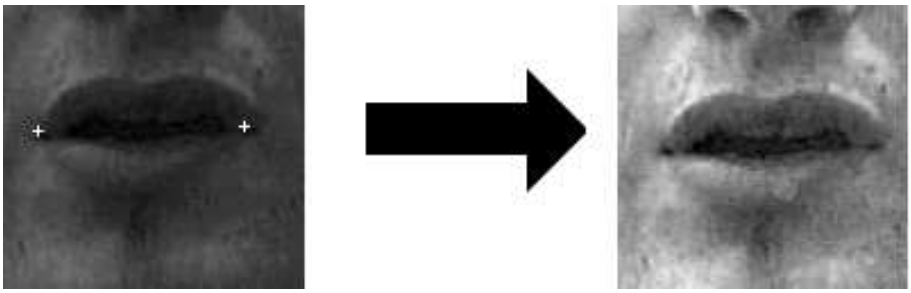

Abbildung 13: Mundregion eines Sprechers vor und nach der Histogrammnormalisierung.

Im nächsten Verarbeitungsschritt wird, wie in [10], für jeden Frame die Cosinustransformierte berechnet. Anhand der Cosinustransformierten kann man erkennen, welche Frequenzen in welcher Richtung im Bild besonders stark vorhanden sind und in welcher Richtung diese verlaufen.

Da der Mund in den Aufnahmen deutlich dunkler ist als das umgebende Gesicht, entspricht die erste tiefe Frequenz, die in vertikaler Richtung deutlich höher als erwartet ist, der Höhe des Mundes. In horizontaler Richtung ist sie dementsprechend ein Indiz für die Breite des Mundes. Das Ergebnis einer Cosinustransformation eines 8 x 8 Pixel großen Bildes ist in Abbildung 14 zu sehen. Eine allgemeine Eigenschaft der Cosinustransformierten ist, dass die Elemente in der oberen linken Ecke sehr hohe Werte besitzen, die vom Betrag her nach unten links abnehmen.

Die Matrizen, bestehend aus jeweils einem cosinustransformierten Bild einer Mundregion pro Zeile, stellen das Ende der Videovorverarbeitung außerhalb des Janus Spracherkenners dar. Die Daten liegen jetzt in einer Form vor, die von dem Spracherkenner verarbeitet werden kann und werden, um den Platzbedarf zu reduzieren, ebenfalls gepackt.

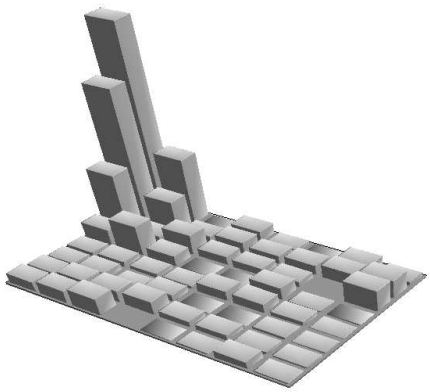

Abbildung 14: Es ist der prinzipielle Aufbau einer Cosinustransformierten zu sehen, mit großen Werten in der linken oberen Ecke, die nach unten rechts abnehmen.

Nicht alle Koeffizienten der Cosinustransformation haben den selben Informationsgehalt bezüglich der Spracherkennung. Frequenzen, die im Bild der Mundregion eine besondere Bedeutung besitzen, haben in der Cosinustransformierten hohe Werte. Es ist somit möglich, die für die Spracherkennung relevanten Koeffizienten, durch aufsummieren der einzelnen Koeffizienten über alle Bilder, zu erhalten. Hierdurch ist es nicht notwendig, alle 4096 Cosinuskoeffizienten zu betrachten, die Verwendung der 64 stärksten führt zu einem fast identischen Ergebnis.

Diejenigen Koeffizienten, bei denen eine ihrer Koordinaten null ist, werden hierbei nicht berücksichtigt. Dies entspricht in Abbildung 15 der ersten Zeile und Spalte. Diese werden ignoriert, da sie die konstanten Teile des Bildes enthalten und nicht für Frequenzen im Bild stehen. Diese Teile nicht zu berücksichtigen entspricht in etwa der Mittelwertsubtraktion im akustischen Fall.

Anschließend werden, analog zum akustischen Fall, Koeffizienten zeitlich benachbarter Frames verknüpft, um einen gewissen Kontext zu erzeugen. Da im visuellen Fall bis jetzt einfach Frames wiederholt werden, um die benötigten 100 Hz zu erreichen, kann man nicht die nächsten bzw. vorigen 5 Frames benutzen. Es wird vielmehr nur jedes dritte Bild genutzt, damit nicht immer dieselben Daten im Kontext stecken. Diese würden nur den Rechenaufwand vergrößern, nicht aber die zur Verfügung stehenden Informationen.

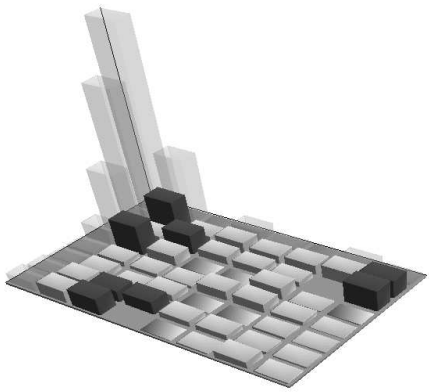

Abbildung 15: Die wichtigsten Komponenten der Cosinustransformierten sind farblich hervorgehoben. Die erste Spalte und Zeile sind von dieser Suche ausgenommen.

Bei gleicher Breite wie im akustischen Fall reicht der visuelle Kontext zeitlich also dreimal so weit in die Vergangenheit und Zukunft wie der akustische. Dies ist kein Problem, da die visuellen Informationen sich auch langsamer verändern als die akustischen.

Zum Abschluss der Videoverarbeitung wird, wieder analog zum akustischen Fall, eine LDA [35, 36] für die zur Verfügung stehenden Daten berechnet. Hierdurch werden die relevanten Daten wieder hervorgehoben und das Training bzw. die Erkennung verbessert und erleichtert.

4.5 Merkmalsextraktion

Zur Analyse der Gesichter und zur Extraktion der hierdurch gewonnenen Mundregion der Sprecher wird ein Programm von Dr. Rainer Stiefelhagen [37] verwendet, welches ursprünglich für einen anderen Zweck, der Bestimmung der Blickrichtung, entwickelt wurde. Dieses Programm sucht in einem Bild nach dem Gesicht des Sprechers und markiert in diesem die Augen, Nasenlöcher und Mundwinkel, wie in Abbildung 16 zu sehen.

Da dieses Programm bereits oft sehr zuverlässig die Mundwinkel eines Sprechers findet, wurde es aus Zeitgründen bei dieser Arbeit als Grundlage der visuellen Merkmalextraktion verwendet und nicht extra ein neues, auf dieses Aufgabengebiet spezialisiertes Programm entwickelt. Das Programm funktioniert mit den meisten Sprechern zuverlässig genug für die Erkennung, nur

Abbildung 16: In jedem Bild wird das Gesicht des Sprechers gesucht, in der Gesichtsregion werden anschließend die Augen, Nasenlöcher und Mundwinkel bestimmt.

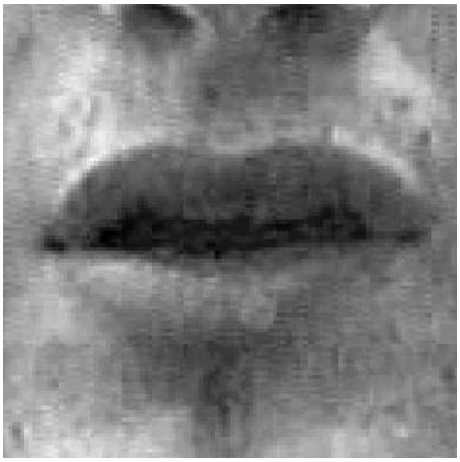

Abbildung 17: Eine extrahierte Mundregion, dargestellt nach der Histogrammnormalisierung.

wenn die Sprecher eine sehr dunkle Hautfarbe haben, wird es häufig verwirrt. Auch bei Brillenträgern oder Menschen mit einem Bart kann das Programm häufig nicht sein volles Potential entfalten.

Zum Auffinden der Mundregion wird zunächst versucht, das Gesicht des Sprechers zu identifizieren. Hierzu wird die größte zusammenhängende hautfarbene Region im Bild gesucht. Ausgehend von der Geometrie des Gesichtes werden die Gesichtsmerkmale gesucht.

Als Erstes wird in der oberen Hälfte des Gesichtes nach zwei dunklen Punkten, den Augen, gesucht. Sind diese identifiziert, so wird aus der Augenposition ein Bereich des Gesichtes bestimmt, in welchen der Mund vermutet wird. Sind die Mundwinkel erfolgreich gefunden, wird noch versucht, zwischen Mund und Augen noch die Nasenlöcher zu finden. Eine ausführliche Beschreibung der Detektion der Gesichtsmerkmale ist unter [38] zu finden.

Bis zu diesem Punkt kann das Programm für die automatische audiovisuelle Spracherkennung unverändert eingesetzt werden, nur im Anschluss an die Detektion der Gesichtsmerkmale ist es noch um die Fähigkeit zur Ausgabe von Bildern der Mundregion in einem geeigneten Datenformat erweitert worden.

Da es hin und wieder vorgekommen ist, dass die korrekten Positionen der Mundwinkel für einen kurzen Moment nicht gefunden werden konnten, wird für die Mundwinkel der Median über die letzten 10 Bilder verwendet. Diese geglätteten Positionen der Mundwinkel werden zur Berechnung der Mundregion herangezogen.

Die Mundregion ist quadratisch und hat die durch die Mundwinkel vorgegebene Breite. Sie wird so gelegt, dass die Mundwinkel in vertikaler Richtung in der Mitte des ausgewählten Bereiches liegen 17. Diese Region wird nun noch auf die einheitliche Größe von 64 x 64 Pixel skaliert, bevor die Bilddaten auf die Festplatte geschrieben werden.

Um die Präzision bei der Extraktion der Gesichtsmerkmale zu bestimmen, wurde die Veränderung der Position des linken Mundwinkels in X-Richtung und auch in Y-Richtung innerhalb einer Äußerung gemessen. Zusätzlich wurde noch die Veränderung der Breite des Mundes gemessen. Für den Einsatz mit dem audiovisuellen Spracherkenner als noch ausreichend befunden wurden alle Sprecher, bei denen es keine Aufzeichnung gibt, bei der einer der drei Werte über 125 Pixel schwankt. Hierdurch standen für das Training 120 der 261 Sprecher zur Verfügung und für die Testmenge konnten 17 der 26 Sprecher verwendet werden.

Diesen Wert als Grenze zu wählen führt dazu, dass die maximalen Schwankungen des linken Mundwinkels, sowohl in horizontaler als auch in vertikaler Richtung, bei den 9 aussortierten Sprechern der Trainingsmenge viermal so groß ist, wie bei den 17 ausgewählten Sprechern. Dieser Unterschied ist bei der Veränderung der Breite des Mundes nicht ganz so gravierend, aber ein Faktor von zwei gibt es auch hier. Die genauen Werte für die Testdaten finden sich in Tabelle 5, für die Trainingsdaten sind die Ergebnisse ähnlich.

	gewählte Sprecher (17 beste)	aussortierte Sprecher (9 schlechteste)
X-Achse	67,50 Pixel	259,00 Pixel
Y-Achse	63,75 Pixel	260,75 Pixel
X-Achse	75,50 Pixel	176,50 Pixel

Tabelle 5: Zeigt die Unterschiede in der Präzision der Gesichtserkennung zwischen den für den Test gewählten Sprechern und den aussortierten.

5 Szenarien und Ergebnisse

5.1 Überblick

Für diese Arbeit wurden mehrere verschiedene Szenarien der audiovisuellen Spracherkennung und der rein akustischen bzw. visuellen Spracherkennung durchgespielt. Zum einen wurde ein rein akustischer Erkenner trainiert, der als Referenzsystem genutzt wurde, um mögliche Verbesserungen durch den zusätzlichen Einsatz der Videodaten erkennen zu können. Des Weiteren wurde ein rein visueller Erkenner trainiert, dieser kam aber nie auf wirklich brauchbare Worterkennungsraten. Es wurde aber auch nicht erwartet, den rein visuellen Spracherkenner zur Erkennung einsetzen zu können. Dieses System wurde aufgesetzt, um zu verifizieren, dass mithilfe der visuellen Daten etwas über die Sprache gelernt werden konnte, dass die Videovorverarbeitung funktioniert.

Außerdem wurden noch drei verschiedene Systeme von audiovisuellen Spracherkennern trainiert. Im ersten Fall wurden die Merkmale einfach verknüpft, und darauf der gleiche Erkennungsprozess angewendet, wie im rein akustischen Fall. Die zweite Form des audiovisuellen Erkenners besaß eine mehrstufige LDA und der letzte Fall war ein Spracherkenner mit Multi-Stream-Architektur, in dem ein akustischer und ein optischer Erkenner gekoppelt werden.

Unter schlechten akustischen Bedingungen zeigten alle drei audiovisuellen Erkenner ein vergleichbares Verhalten. Es war eine mehr oder weniger starke Verbesserung der Erkennungsleistung festzustellen. Waren die akustischen Bedingungen jedoch gut, so konnten die Ergebnisse durchaus unterschiedlich ausfallen. Auf die genauen Ergebnisse werde ich in den nächsten Abschnitten noch genau eingehen.

5.2 Akustischer Spracherkenner

Um Verbesserungen, durch den Einsatz eines audiovisuellen Spracherkenners feststellen zu können, ist es notwendig zu wissen, welche Erkennungsleistung ein rein akustisches System besitzt. Zu diesem Zweck wurde zum Beginn der Arbeit ein rein akustisches Referenzsystem aufgebaut, an welchen sich die audiovisuellen Ansätze messen lassen mussten.

Für dieses Referenzsystem wurden zum Training die akustischen Informationen aus den Videodateien verwendet. Auch wenn die Gesichtserkennung nicht bei allen zur Verfügung stehenden Sprechern funktioniert, so sind die Audiodateien in allen Fällen verwendbar. Darum werden für das Referenzsystem auch bis zu 261 Sprecher der Trainingsmenge verwendet, während für die visuelle Verarbeitung nur 120 Sprecher zur Verfügung standen. Die Ergebnisse sind den Tabellen 6 und 7 zu entnehmen.

	26 Testsprecher	17 Testsprecher
unverrauschte Audiodaten	80,2%	74,7%
verrauschte Audiodaten	-	46,1%

Tabelle 6: Die Ergebnisse der rein akustischen Tests mit 261 Trainingssprechern.

	5 Testsprecher
30 Trainingssprecher	60,7%
14 Trainingssprecher	51,7%

Tabelle 7: Die Ergebnisse der rein akustischen Tests mit kleiner Trainingsmenge auf nicht verrauschten Audiodaten.

Es gibt neben dem großen Audioreferenzsystem, welches mit allen 261 Sprechern trainiert worden ist, auch noch zwei kleine mit 14 bzw. 30 Sprechern trainierte Systeme. Diese sind erzeugt worden, da die ersten Tests der audiovisuellen Ansätze ebenfalls nur mit diesen kleinen Trainingmengen trainiert worden sind, um schnell den optimalsten Ansatz auswählen zu können.

Wie bei den Trainingsdaten gab es auch bei den Testdaten Sprecher, bei denen die Extraktion der Gesichtsmerkmale nicht gut funktioniert hat. Aus diesem Grund musste, neben dem Referenzsystem mit 26 Testsprechern, noch eines mit lediglich 17 Testsprechern berechnet werden.

In diesem zweiten Fall mit 17 der 26 Testsprecher wurde eine Wortakkuratheit von 74,7% , gegenüber von 80,2% mit allen Testsprechern, erreicht. Der Rest des Systems ist identisch, nur die für den Test benutzen Sprecher machen den Unterschied aus.

Ein zweiter Test wurde noch mit verrauschten Audiodaten durchgeführt. Hierzu wurde den Audiodaten weißes Rauschen mit einem Signal-Rausch-Abstand von 15 db beigemischt. Hierdurch verschlechterte sich die Signalqua-

lität und die Erkennungsleistung ist auch deutlich abgesunken. Der Audio-Referenzerkenner mit 17 Testpersonen erreichte in diesem Versuch eine Wortakkuratheit von 46,1%.

Für die kleinen Trainingsmengen mit maximal 30 Sprechern wurde die Testmenge aus Zeitgründen noch einmal auf 5 Sprecher reduziert.

5.3 Visueller Spracherkenner

Nachdem das Audioreferenzsystem erstellt war, stand als nächste Aufgabe an, eine funktionierende Videovorverarbeitung zu erzeugen und einen ersten Spracherkenner zu implementieren, der mit visuellen Daten arbeiten kann. Hierzu wurde ein rein visuelles System gewählt, da in diesem sichergestellt werden kann, dass die Lernerfolge ausschließlich mithilfe der Videodaten zu Stande gekommen sind.

Mithilfe dieses Systems konnte man nachweisen, dass die Videovorverarbeitung prinzipiell funktioniert und sinnvolle Merkmale für das Training liefert, mit deren Hilfe der Erkenner in der Lage ist, etwas zu lernen. Dieses System diente also nicht wirklich der Spracherkennung, sondern ist vielmehr ein Schritt auf dem Weg zu einem funktionierenden audiovisuellen Spracherkenner. In der Arbeit von [10] wurden zwar bessere Ergebnisse des rein visuellen Spracherkenners erzielt, aber auch diese sind im Vergleich zu akustischen oder audiovisuellen Erkennern noch recht gering.

Dieser Abstand ist darauf zurückzuführen, dass es im Rahmen dieser Arbeit nicht möglich war, eine so aufwändige Videovorverarbeitung wie in [10] zu implementieren. Dies bedeutet aber gleichzeitig auch, dass die anderen, audiovisuellen, Experimente unter dem gleichen Problem leiden. Trotzdem ist in diesen Fällen bereits eine gute Verbesserung gegenüber dem rein akustischen Erkenner erreicht worden.

Auch wenn die Worterkennungsrate nur minimal war, so konnte man zeigen, dass Viseme relativ zuverlässig erkannt wurden. In einem untrainierten System müssten die erkannten Viseme ungefähr gleich verteilt sein, d.h. es würde geraten, was für Viseme auf den Bildern zu erkennen sind. Funktioniert die Videovorverarbeitung nicht korrekt und das System erhält für die Spracherkennung unbedeutende Merkmale, so wird ein ähnliches Ergebnis auftreten, wie im untrainierten Fall.

In Abbildung 18 ist die Präzision der Erkennungsleistung auf Visembasis zu erkennen. Insgesamt gibt es in dem System 37 verschiedene Visem-Zustände.

Auch beim Audioerkenner werden nicht sehr viele dieser Visem-Zustände ganz korrekt erkannt, aber fast immer wird zu einem Zeitpunkt der korrekte Visem-Zustand als einer der wahrscheinlichsten gefunden. Aus diesem Grund ist hier aufgeführt, wie häufig der richtige Visem-Zustand als einer der besten zehn erkannt worden ist. Auf der X-Achse sind die Äußerungen der Sprecher aufgeführt und auf der Y-Achse die Erkennungsleistungen, wie groß der Anteil der Viseme ist, die innerhalb der besten zehn lagen.

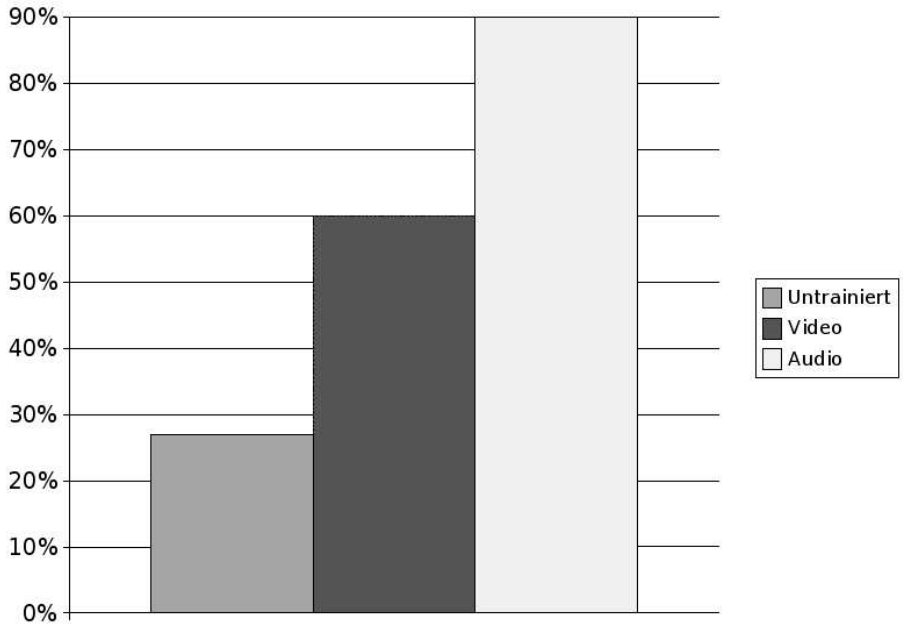

Abbildung 18: Mithilfe von Videos kann etwas über die Sprache gelernt werden, jedoch nicht so zuverlässig wie aus akustischen Informationen.

Bei einem untrainierten Erkenner müsste dieser Wert bei ca. 27% liegen. Die Abbildungen 18 und 19 zeigen jedoch einen deutlichen Lernerfolg für den visuellen Fall, bei diesem wurde in ca. 60% der Fälle der richtige Visem-Zustand unter den zehn Wahrscheinlichsten gefunden. Es ist aber auch zu erkennen, dass doch eine Diskrepanz zwischen dem visuellen und dem akustischen Fall liegt. Der akustische Erkenner erreicht in diesem Versuch eine Quote von 89%.

Um einen Trainingeffekt nachzuweisen, ist es nicht nötig gewesen, mit dem gesamten Trainingsset zu arbeiten. Es genügte, und war aus Zeitgründen vorteilhaft, mit einem kleinen Teil des Trainingssets zu arbeiten. Hierzu sind 11 Sprecher für das Training und 6 Sprecher für den Test genutzt worden.

Noch eine weitere Eigenschaft des Videoerkenners lässt sich aus Abbildungen 19 ablesen. Im Vergleich zwischen den Ergebnissen mit verrauschten Daten und dem visuellen Erkenner ist zu sehen, dass mit Bilddaten arbeitende Systeme eine höhere Variabilität zwischen einzelnen Sprechern besitzen. Die Erkennungsleistung zwischen einzelnen Sprechern schwankt im visuellen Fall deutlich stärker als im akustischen. Dies ist ein Indiz dafür, dass bei der Normalisierung der Videodaten noch Potential für weitere Verbesserungen vorhanden ist.

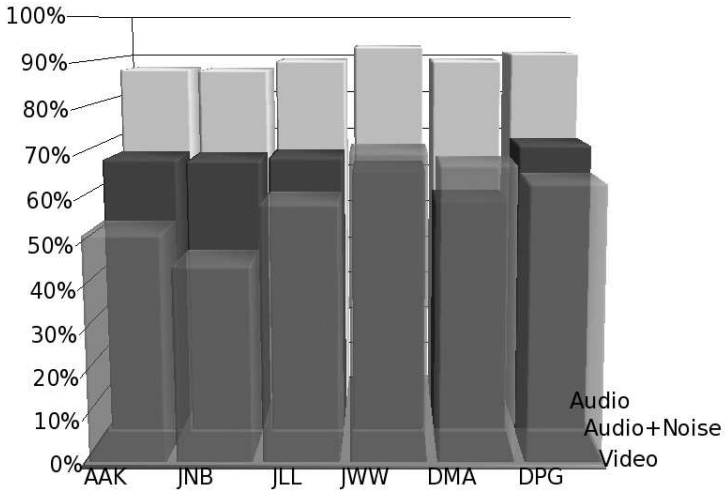

Abbildung 19: Vergleich der Visemerkennung zwischen Audio- und Videoerkenner.

Aber das Ziel ist kein rein visueller Erkenner, der alle seine Informationen aus dem Videosignal gewinnen muss, sondern ein audiovisueller. Dieser bekommt akustische Informationen und kann durch die zusätzlichen visuellen Informationen unterstützt werden. Ob in den visuellen Daten wirklich Informationen stecken, die in den akustischen Signalen nicht vorhanden sind, werden die nächsten Experimente zeigen.

Das Videosignal alleine reicht für die automatische Spracherkennung nicht aus, aber in diesem Signal können durchaus Informationen enthalten sein, die in den akustischen Daten nicht vorhanden sind. Um dieses zu überprüfen, wurde der rein visuelle Spracherkenner herangezogen. Dieser musste für zufällig ausgewählte Äußerungen verschiedene Hypothesen bewerten. Zum einen war dies die korrekte Wortfolge als Referenz und zum anderen die Hypothese, die der rein akustische Erkenner erzeugt hat. Wenn die visu-

ellen Merkmale ausreichend zusätzliche Informationen besitzen, dann sollte die korrekte Äußerung vor der akustischen Hypothese bevorzugt werden.

Selbstverständlich durften für diesen Test nur Äußerungen verwendet werden, die nicht zu 100% korrekt erkannt worden sind, da sonst keine Unterschiede auftreten können.

Bei neun der zehn getesteten Äußerungen hat der Videoerkenner die korrekte Aussage besser bewertet, als die Hypothese des Audioerkenners. In Tabelle 8 sind die einzelnen Ergebnisse aufgelistet, kleinere Zahlenwerte bedeuten eine bessere Bewertung. In der Kombination aus beiden Bewertungen schneidet somit die in Wirklichkeit gemachte Aussage gegenüber der akustischen Hypothese besser ab als bei einem System, welches auf die Zusatzinformationen der Videodaten verzichtet.

Äußerung Nr.	Referenz	Audio-Hypothese
1	**9,899873e+04**	9,904535e+04
2	**9,320744e+04**	9,321290e+04
3	**7,836189e+04**	7,849109e+04
4	**8,674477e+04**	8,678766e+04
5	**6,118579e+04**	6,122995e+04
6	**3,551762e+04**	3,563007e+04
7	**2,929482e+04**	2,937038e+04
8	4,160565e+04	**4,156348e+04**
9	**3,706526e+04**	3,715476e+04
10	**3,532131e+04**	3,532855e+04

Tabelle 8: Die korrekte Äußerung wird vom Videoerkenner meistens besser bewertet als die Hypothese des Audioerkenners.

5.4 Audiovisuelle Spracherkenner

Das Experiment mit dem rein visuellen Spracherkenner hat gezeigt, dass die Vorverarbeitung funktioniert und das System in der Lage ist, etwas zu lernen. Nun stellt sich die Frage, ob in den Videodaten genug neue Informationen vorhanden sind, die es in den akustischen Daten noch nicht gibt. Nur in diesem Fall ist es möglich, den akustischen Spracherkenner zu verbessern, im Idealfall sogar für nicht verrauschte Audioaufnahmen. Um dieser Frage auf

den Grund zu gehen sind die folgenden Experimente durchgeführt worden, die in ähnlicher Form auch in [10] beschrieben worden sind.

Im ersten Experiment wird in Abschnitt 5.4.1 die Konkatenation von Audio- und Videodaten zu einem neuen audiovisuellen Merkmal beschrieben.

In Abschnitt 5.4.2 wird eine hierarchische LDA aufgebaut, um die Dimension des Merkmalraumes verringern zu können, ohne Informationen zu verlieren.

Das letzte Experiment in Abschnitt 5.4.3 beschreibt eine Multi-Stream-Architektur, in der Audio- und Videoerkenner unabhängig voneinander trainiert werden können. Eine Verknüpfung findet erst beim Erkennen statt.

Nachdem alle Experimente vorgestellt worden sind, werden in Abschnitt 5.4.4 die Ergebnisse der audiovisuellen Spracherkennungssysteme präsentiert.

5.4.1 Konkatenation von Audio- und Videofeatures

Das erste Experiment zur audiovisuellen Spracherkennung ist auch zugleich das einfachste Modell. Die zur Verfügung stehenden Daten werden einfach gekoppelt, d.h. für jedes Zeitfenster werden die aus der Vorverarbeitung gewonnenen Merkmale verbunden. Im Fall der akustischen Daten handelt es sich um die 13 Cepstralkoeffizienten und für die visuelle Erkennung werden die 64 Koeffizienten der Cosinustransformierten der Mundregion verwendet, welche die meiste Information enthalten.

Daraus resultieren für jedes Zeitfenster 77 Koeffizienten. Durch Erzeugung eines Kontextes von 5 Frames vor und nach dem aktuellen Zeitpunkt erhöht sich die Zahl auf 847. Damit ist das für dieses Experiment verwendete Merkmal größer, als die der Folgenden. Auch wenn das Modell logisch das einfachste ist, so ist es gleichzeitig das rechenintensivste, da hierbei mit den größten Matrizen gerechnet werden muss.

Als Feature für die Weiterverarbeitung wird also die neu erzeugte Matrix mit 77 Einträgen pro Zeile (Zeitintervall) verwendet anstelle der Cepstralkoeffizienten im akustischen Fall. Ansonsten läuft der Trainings- und Erkennungsprozess aber genau so ab, wie dies im rein akustischen System der Fall ist.

Dies ist das einfachste System, es sind nur sehr wenig Änderungen in einen bestehenden akustischen Erkenner zu implementieren. Es ist lediglich nötig,

zusätzlich die Videodaten zu laden und sie mit den Audiodaten zu verknüpfen. Training und Erkennung können unverändert bleiben. Auf der anderen Seite bietet es aber auch die geringste Flexibilität aller hier durchgeführten Experimente. Dementsprechend sind für diesen Fall auch die schlechtesten Ergebnisse zu erwarten.

Im einfacheren Fall, dem Test auf verrauschten Audiodaten, war bereits mit diesem System eine Verbesserung zu erkennen. Im anderen Fall, dem Test auf nicht verrauschten Daten, haben die nicht so perfekten visuellen Informationen sogar die Erkennungsleistung verschlechtert. Dies konnte passieren, da die visuellen Daten mengenmäßig in etwa sechsmal so stark vertreten sind, wie die akustischen. Auch die anschließend durchgegeführte LDA-Transformation kann die überflüssigen oder verwirrenden Teile nicht komplett beseitigen.

Obwohl dieses Experiment bereits vielversprechende Ergebnisse für den Testfall mit verrauschten Audiodaten liefert, so sollte man doch noch andere Experimente durchführen, um zu sehen, ob man nicht doch im Fall von unverrauschten Audiodaten eine Verbesserung erreichen kann.

5.4.2 Hierarchische LDA

In diesem Experiment ist eine hierarchische Struktur von LDA-Transformationen, wie in [10], zum Einsatz gekommen. Hierdurch ist es möglich, die Dimension des Merkmalraumes zu verringern, ohne dabei für die Spracherkennung wichtige Informationen zu verlieren. Eventuell ist es auch möglich, verwirrende Informationen durch die mehrfache LDA-Transformation zu eliminieren und dadurch die Erkennungsleistung gegenüber dem ersten Fall zu verbessern.

Für dieses System wurde ein Ansatz mit zweischichtiger LDA verwendet. In der ersten Schicht wird sowohl für die Videodaten als auch für die Audiodaten getrennt eine LDA-Matrix berechnet. Die Datenmatrizen werden mit der jeweiligen LDA-Matrix multipliziert, die hieraus resultierenden neuen Datenmatrizen werden verknüpft und aus der resultierenden Matrix wird erneut eine LDA berechnet.

Dieser Ansatz bietet Vorteile gegenüber dem ersten Ansatz, dem einfachen Verknüpfen der Daten. Als Erstes sind die benötigten LDA-Matrizen deutlich kleiner. Es müssen zwar drei Multiplikationen mit LDA-Matrizen berechnet werden, da aber der Aufwand bei Matrixmultiplikationen $O(n^3)$ mit n als

Kantenlänge der Matrix beträgt, ist dieser Ansatz schneller, obwohl mehr Matrixmultiplikationen durchgeführt werden müssen. Insgesamt müssen jedoch weniger einzelne Multiplikationen berechnet werden.

Als weiteren Vorteil gibt einem dieser Aufbau bessere Möglichkeiten zur Gewichtung von den akustischen zu den visuellen Daten. Im Normalfall sind die visuellen Daten wesentlich umfangreicher als die akustischen; in dem hier eingesetzten System bestanden die akustischen Daten aus 143 Koeffizienten für jeweils 10 ms und für das visuelle System werden im gleichen Zeitraum 704 Koeffizienten benötigt. Im Fall des Systems aus Kapitel 5.4.1 muss man einen Kompromiss finden. Werden zu viele visuelle Koeffizienten eingesetzt, so besteht trotz Berechnung einer LDA die Gefahr, dass ein zu großer Teil der akustischen Daten nicht berücksichtigt werden kann. Dies ist nicht wünschenswert, da die akustischen Koeffizienten mehr Informationen über die gesprochenen Worte enthalten.

Auf der anderen Seite darf man auch nicht zu wenige visuelle Koeffizienten verwenden, da diese durchaus Zusatzinformationen beinhalten, die in den akustischen Daten nicht vorhanden sind. Diese möchte man gerne nutzen, ohne zu viele Informationen aus den akustischen Daten zu verlieren.

In diesem Punkt liegt der Vorteil der mehrschichtigen LDA gegenüber der einfachen Verknüpfung der Signale. Durch die beiden LDA-Transformationen auf der ersten Schicht sind, sowohl aus den Video- als auch aus den Audiodaten, die wichtigsten Koeffizienten an die vordersten Stellen verschoben worden. Soll nun die Gewichtung zu Gunsten der akustischen Daten bzw. der visuellen Daten verschoben werden, so werden nur die unwichtigeren Koeffizienten abgeschnitten, evtl. werden sogar Koeffizienten entfernt, die eher verwirrende als hilfreiche Informationen enthalten.

Dadurch ist es jetzt möglich, den visuellen Daten einen so großen Einfluss einzuräumen, dass die in ihnen vorhandenen Zusatzinformationen genutzt werden können, ohne wichtige Informationen aus den akustischen Daten zu verlieren.

Werden von beiden LDA-Transformierten aus der ersten Ebene, also den LDA-transformierten Cepstralkoeffizienten im akustischen und den LDA-transformierten Koeffizienten der Cosinustransformierten, gleich viele Komponenten verwendet, so verhält sich dieses System ähnlich wie das vorherige. Im Fall von unverrauschten akustischen Daten ist keine Verbesserung im Vergleich zu dem Erkennungssystem zu verzeichnen, welches nur mit akustischen Daten arbeitet.

In diesem System bietet sich die Möglichkeit, von der akustischen LDA eine größere Anzahl von Koeffizienten in die zweite Schicht einfließen zu lassen, als von den Visuellen. Bei einem Verhältnis von 4:1 - d.h. 80% der Komponenten auf welchen die LDA der zweiten Schicht berechnet wird stammen aus den akustischen Daten - ist eine Verbesserung der Erkennungsleistung um ca. 1% absolut gegenüber dem rein akustischen System, für nicht verrauschte Audiodaten, zu verzeichnen. Damit ist dies das erste System, bei welchem auch unter guten akustischen Bedingenden keine Verschlechterung eingetreten ist, sondern sogar eine leichte Verbesserung zu verzeichnen ist.

5.4.3 Multi-Stream-Architektur

Das letzte System, welches im Rahmen dieser Arbeit implementiert wurde, war ein audiovisueller Spracherkenner mit Multi-Stream-Architektur. Dieses System besteht im Prinzip aus beliebig vielen unabhängigen Spracherkennern, die einzeln trainiert werden können und erst, wenn Sprache erkannt werden soll, zusammengeschaltet werden, um ihre Fähigkeiten zu kombinieren.

In dem hier vorliegenden Fall eines audiovisuellen Spracherkenners besteht das gesamte System nur aus zwei Einheiten. Einem rein akustischen Spracherkenner und einem rein visuellem Spracherkenner. Es stellt also eine Kombination aus den in Kapitel 5.2 und Kapitel 5.3 beschriebenen Systemen dar.

Das Training läuft für jeden Datenstrom separat ab, es ist also möglich, für den akustischen Teil eine andere Datenbasis zu verwenden als für den visuellen. Es ist einfacher, Daten für den akustischen Erkenner zu sammeln bzw. es sind bereits große Datenbasen vorhanden, für die es keine korrespondierenden Videodaten gibt. Diese Datenbasen müssen nun nicht weggeschmissen werden, sondern können für den audiovisuellen Erkenner weiterverwendet werden, obwohl sie keine visuellen Informationen enthalten.

Auch in dieser Arbeit sind für den akustischen und den visuellen Teil des Spracherkenners mit Multi-Stream-Architektur unterschiedliche Datenbasen verwendet worden. Für die akustische Komponente standen alle 261 Sprecher der Trainingsmenge zur Verfügung, im visuellen Fall wurden nur 120 Sprecher verwendet.

Das Training läuft bei diesem System für die einzelnen Datenströme unabhängig ab. Erst wenn mit dem System etwas erkannt werden soll, müssen die beiden Systeme kombiniert werden. Hierzu wird eine Suche in einem

gemeinsamen Wahrscheinlichkeitsraum durchgeführt, damit für den akustischen und den visuellen Fall die gleichen Hypothesen betrachtet werden. Anschließend wird die Hypothese, die am wahrscheinlichsten gesprochene Äußerung, ausgegeben, die in der momentan gewählten Kombination der Streamgewichte die beste Gesamtbewertung erhalten hat.

Die Bewertung - der *score* - eines Datenstroms ist eine Summe der logarithmierten Wahrscheinlichkeiten über alle Zeitintervalle. Die einzelnen Bewertungen der Datenströme werden dann, entsprechend der Streamgewichte, kombiniert. Für den hier bertachteten Fall eines audiovisuellen Erkenners mit jeweils einem Datenstrom für die akustischen und einem für die visuellen Informationen, berechnet sich die Gesamtbewertung einer Hypothese, wenn eine Gewichtung von 70% für den Audiostream und 30% für den Videostream angenommen wird, wie folgt:

$$0,7 * akustik + 0,3 * visuell = score$$

Wobei *akustik* die Bewertung des akustischen Teils des audiovisuellen Spracherkenners darstellt und *visuell* für den Wert der Hypothese des visuellen Teils des Erkenners steht.

Darin liegt auch der große Vorteil dieses Systems gegenüber den bisher vorgestellten: Die Gewichtung des Einflusses von akustischen und visuellen Daten auf das Erkennungsergebnis wird nur bei der Erkennung festgelegt. Damit ist es möglich, die Gewichtungen zu verändern, ohne das ganze System neu trainieren zu müssen, wie es bei den beiden anderen Ansätzen, der einfachen Verknüpfung der Daten und der mehrschichtigen LDA, der Fall war.

Man kann dieses System unter verschiedenen Bedingungen einsetzen, ohne es neu trainieren zu müssen. Man kann es auch adaptiv einsetzen, wenn man zusätzliche Informationen über die momentan vorhandenen Störgeräusche, z.B. Hintergrundrauschen, hat. Wenn es in einem Moment wenig Störeinflüsse gibt, so können die akustischen Daten höher bewertet werden und den visuellen kann ein geringer Einfluss eingeräumt werden. Analog dazu wird der visuelle Einfluss erhöht, wenn die Störgeräusche zunehmen (Abbildung 20).

Dadurch erreicht man mit der Multi-Stream-Architektur eine wesentlich höhere Flexibilität als mit den anderen Systemen. Man kann **einen** Spracherkenner trainieren, der unter vielen verschiedenen Bedingungen sehr gut funktionieren kann. Es ist möglich ein Spracherkennungssystem zu bauen, welches

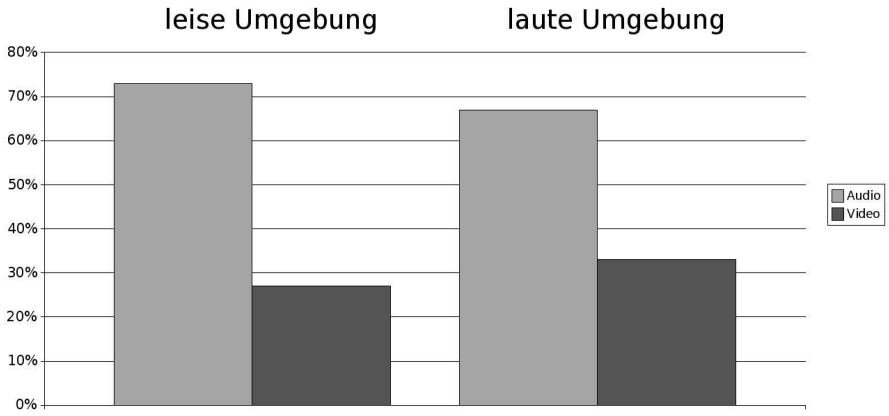

Abbildung 20: Wenn die normale Audioerkennung auf Grund von Störungen nicht gut gelingt, wird der Einfluss der Videomerkmale erhöht.

sowohl unter guten akustischen Bedingungen hervorragend funktioniert, als auch unter schlechten noch eine gute Leistung bringen kann. Darüber hinaus kann der Einfluss der Videodaten auch sehr einfach reduziert werden, wenn diese schlechter werden, zum Beispiel durch eine Verschlechterung der Beleuchtung.

Bei allen vorher betrachteten Systemen müsste man für jeden dieser Fälle einen eigenen Spracherkenner trainieren. Dies würde einen deutlichen Mehraufwand bei der Entwicklung und Wartung der Systeme bedeuten. Außerdem sind selbst bei einem vielfachen Trainingsaufwand nur die Extremfälle abgedeckt, nicht jedoch Zwischenstufen, wie z.B. leichtes Rauschen. Die Multi-Stream-Architektur ist aber auch für diese Übergangsfälle gut gerüstet, ohne neu trainiert werden müssen.

5.4.4 Audiovisuelle Ergebnisse

In diesem Abschnitt werden die Ergebnisse, der drei unterschiedlichen audiovisuellen Spracherkennungsansätze, welche in dieser Arbeit implementiert wurden, dargestellt. Die detailliertesten Ergebnisse gibt es zu der Multi-Stream-Architektur, da diese am universellsten einsetzbar ist, einfacher und schneller zu trainieren ist, als die anderen und zusätzlich auch noch die besten Ergebnisse geliefert hat. Darum sind die meisten Versuche mit diesem System durchgeführt worden und bei den anderen Versuchen aus Zeitgründen nur Teilmengen für Training und Test eingesetzt worden.

Der erste Versuch, die einfache Verknüpfung der akustischen und visuellen Informationen, ist der rechenintensivste. Aus diesem Grund wurden die Versuche auch auf dem kleinsten Trainingsset durchgeführt, wodurch die geringen Werte in der Erkennungsgenauigkeit zu erklären sind. Für die jetzt beschriebenen Ergebnisse wurden die Systeme auf lediglich vierzehn der insgesamt 261 Sprecher trainiert und fünf der 26 Sprecher für den Test verwendet. Im Vergleich mit den beiden anderen audiovisuellen Experimenten ist in Tabelle 9 deutlich zu erkennen, dass dieses System am schlechtesten abschneidet.

	Wortakkuratheit	Veränderung
Audio	51,72%	
Konkatenation	48,64%	-3,08
hierarchische LDA	50,84%	-0,88
Stream	50,72%	-1,00

Tabelle 9: Ergebnisse der audiovisuellen Spracherkennung auf unverrauschten Audiodaten, trainiert mit 14 Sprechern.

Bei diesem Test haben alle drei audiovisuellen Erkenner gegenüber dem akustischen Referenzsystem verloren. Diese Verluste kommen von der größeren Varianz der visuellen Daten, durch diese sind größere Trainingsmengen notwendig für visuelle Daten als für akustische. Wie sich später zeigt, verschiebt sich das Ergebnis immer weiter zu Gunsten der audiovisuellen Erkenner, je größer die Trainingsmenge wird.

In diesem Fall hat das erste System gegenüber dem akustischen Spracherkenner etwa drei Prozentpunkte verloren. Die anderen beiden Systeme, die mehrschichtige LDA und die Multi-Stream-Architektur zeigen in etwa gleichwertige Ergebnisse und verlieren einen Prozentpunkt auf das Audiosystem.

Da sich der einfache Ansatz mit Verknüpfen der akustischen und visuellen Merkmale - wie erwartet - als der schwächste herausgestellt hat, werden die weiteren Tests, mit größeren Trainingsmengen, nur mit den beiden anderen Ansätzen durchgeführt. Es ist nun interessant, ob die Experimente mit der hierarchischen LDA und der Multi-Stream-Architektur auch bei größeren Trainingsmengen eine fast identische Erkennungsleistung erbringen.

Es muss auch noch gezeigt werden, ob die audiovisuelle Erkennung von größeren Trainingsmengen profitiert, so dass die Erkennungsleistung höher liegt, als im rein akustischen Fall. Ist dies der Fall, so bietet es sich an, den Spracherkenner mit Multi-Stream-Architektur auf Grund seiner höheren Flexibilität zu verwenden. Kann die Erkennungsleistung jedoch nicht über das

Niveau des akustischen Referenzsystems gesteigert werden, so ist der Aufwand der audiovisuellen Spracherkennung nicht notwendig und man kann einen gewöhnlichen, rein akustischen Spracherkenner verwenden.

Für die zweiten Versuche sind 30 Sprecher für das Training verwendet worden. Der Audioerkenner erreicht in diesem Fall 60,71% Wortakkuratheit, für den Erkenner mit mehrschichtiger LDA wurde eine Wortakkuratheit von 61,52% ermittelt. Damit ist das erste System gefunden, welches bereits auf nicht verrauschten Audiodaten ein besseres Ergebnis liefert, als der normale, rein akustische Spracherkenner. Der Spracherkenner mit Multi-Stream-Architektur bewegt sich auf gleichem Niveau, ist sogar minimal besser, er erreicht eine Wortakkuratheit von 61,76%. Eine Übersicht über die Ergebnisse findet sich in Tabelle 10.

	Wortakkuratheit	Veränderung
Audio	60,71%	
hierarchische LDA	61,52%	+0,81
Stream	61,76%	+1,05

Tabelle 10: Ergebnisse der audiovisuellen Spracherkennung auf unverrauschten Audiodaten, trainiert mit 30 Sprechern.

Die beiden Versuche, mit 14 bzw. 30 Sprechern in der Trainingsmenge, haben gezeigt, dass der Ansatz mit hierarchischer LDA und die Multi-Stream-Architektur gleichwertige Ergebnisse liefern. Aus diesem Grund wird die Multi-Stream-Architektur, auf Grund ihrer Vorzüge, für die weiteren Experimente ausgewählt.

Zunächst ist interessant, ob es bei steigender Größe der Trainingsmenge zu einer weiteren Verschiebung zu Gunsten des audiovisuellen Erkenners kommt, weil er weiterhin von der steigenden Menge an Trainingsdaten mehr profitieren kann, als der akustische Erkenner. Andererseits kann auch der Fall eintreten, dass der akustische Erkenner durch die größere Menge an Trainingsdaten so gut wird, dass mit den zusätzlichen visuellen Informationen kein so deutlicher Gewinn mehr erzielt werden kann, wie bei dem kleinen System.

Um dieses in Erfahrung zu bringen ist folgender Streamerkenner trainiert worden: Für den Audiostream sind alle 261 Sprecher der Trainingsmenge verwendet worden, um einen möglichst guten akustischen Erkenner zu bilden. Der Videostream ist mit den 120 Personen trainiert worden, bei denen

die Merkmalsextraktion am zuverlässigsten funktioniert. Außerdem ist die Testmenge von 5 Sprechern auf 17 vergrößert worden.

An dem absoluten Unterschied zwischen dem Referenzsystem mit 74,74% und dem audiovisuellen Spracherkenner mit Multi-Stream-Architektur mit 75,90% hat sich durch die Vergrößerung der Trainingsmenge nichts geändert. Es gibt immer noch einen Gewinn von etwa einem Prozentpunkt. Bei der relativen Veränderung der Wortfehlerrate (WER) schneidet dieses System deutlich besser ab, da es eine höhere Wortakkuratheit besitzt.

In diesem Fall konnte die WER um ca. 4,5% gesenkt werden, für das Experiment mit 30 Sprechern in der Trainingsmenge beträgt die Verbesserung lediglich 2,6%. Eine deutlich höhere Verbesserung durch den Einsatz eines audiovisuellen Erkenners gegenüber einem akustischen Spracherkenner, sowohl absolut als auch relativ, kann bei verrauschten Audiodaten beobachtet werden. Hierzu wurde den Audiodaten weißes Rauschen untergemischt mit einem Signal-Rausch-Abstand von ca. 15 db. Die einzelnen Ergebnisse sind in Tabelle 11 aufgeführt.

Streamgewichte Audio:Video	nicht verrauschte Audiodaten	verrauschte Audiodaten
100:0	74,74%	46,06%
90:10	75,22%	48,81%
80:20	75,70%	51,47%
70:30	**75,90**%	**52,63**%
60:40	75,71%	51,95%
50:50	74,48%	47,28%

Tabelle 11: Ergebnisse der audiovisuellen Spracherkennung trainiert mit 120 Sprechern für Video und 261 für Audio.

Die Wortakkuratheit stieg in diesem Fall von 46,06% ohne Unterstützung durch Videodaten auf 52,63% für den audiovisuellen Spracherkenner mit Multi-Stream-Architektur. Das entspricht einem absoluten Anstieg um sechseinhalb Prozentpunkte, die WER konnte um 12,5% gesenkt werden.

Wie man an Tabelle 12 deutlich erkennen kann nimmt mit steigender Größe der Datenmenge, welche für das Training eingesetzt wird, die Homogenität in der Erkennungsleistung zu. Bei dem großen Erkennungssystem beträgt die maximale Verbesserung der Wortakkuratheit lediglich 2,44% gegenüber 4,07% bei dem kleinen System. Die Werte bezeichnen jeweils absolute Veränderungen in der Wortakkuratheit.

Sprecher	30 Sprecher	120 Sprecher
AV3GMF01	+4,07%	+0,55%
AV1JFM01	+3,59%	+2,11%
AV3DLN01	+2,86%	+0,00%
...	...	...
AV3PJB01	-9,70%	+0,91%
AV3JXP01	-10,87%	+1,51%
AV1MDP01	-11,50%	+0,00%

Tabelle 12: Schwankungen einzelner Sprecher zwischen rein akustischer Spracherkennung und audiovisueller Sprecherkennung, trainiert mit 30 bzw. 120 Sprechern der Trainingsmenge.

Dafür werden bei dem großen System für fast alle Sprecher Verbesserungen erzielt, nur bei einem einzigen Sprecher ist eine leichte Verschlechterung um 0,42% zu verzeichnen. Ein ganz anderes Ergebnis zeigt das kleine System: hier gibt es Verschlechterungen für acht der siebzehn Sprecher. Die Verschlechterungen fallen auch deutlich höher aus, als bei dem großen System; einzelne Sprecher verlieren über 10%. Dies zeigt, dass die einzelnen Sprecher sich noch recht stark voneinander unterscheiden. Aus diesem Grund werden für den visuellen Teil eines Spracherkenners mehr Trainingsdaten benötigt, bis ein stabiles Ergebnis erreicht wird.

Außerdem muss man noch überprüfen, ob die optimale Gewichtung der einzelnen Datenströme von der Qualität der Audiodaten abhängig ist. Es ist zu erwarten, dass bei verrauschten Audiodaten der Videostream einen größeren Einfluss bekommen muss, um ein optimales Ergebnis zu erzielen. Hierzu wurde für die 17 Testsprecher die optimale Gewichtung der Datenströme bestimmt, wobei jeweils von nur Audio, d.h. Streamgewicht 100:0 (=100% Audiostream + 0% Videostream), bis Audio- und Videostream gleichrangig, also Streamgewicht 50:50, in 10%-Abständen getestet worden ist.

Bei der in Tabelle 11 vorgenommenen, groben Unterteilung, hat sowohl der Fall mit nicht verrauschten Audiodaten, als auch der Fall mit verrauschten Audiodaten, das beste Ergebnis bei einer Verteilung der Streamgewichte von 70:30. Die Durchschnittswerte über die 17 Sprecher liegen für nicht verrauschte Audiodaten leicht über 70% Einfluss für den Audiostream und bei Verrauschten ein wenig unterhalb. Insgesamt ist eine Verschiebung in die vorhergesagte Richtung um ca. 4,5% zu erkennen, siehe auch Abbildung 20.

Wesentlich stärker als von der Audioqualität, wird die optimale Gewichtung jedoch von der Größe der Trainingsdaten beeinflusst. Hierbei war festzustellen, je größer die Datenmenge, um so größer war auch der optimale Einfluss der Videodaten, siehe Tabelle 13. Bei dem System mit 30 Sprechern in der Trainingsmenge war die ideale Gewichtung 90:10, bei nur 14 Sprechern sogar 100:0, d.h. jeder Einfluss der Videodaten hat das Ergebnis verschlechtert.

In Abbildung 21 wird zum Abschluss dieses Kapitels noch einmal die Verbesserung veranschaulicht, die durch den audiovisuellen Spracherkenner mit Multi-Stream-Architektur auf der großen Trainingsmenge erzielt worden ist.

Verteilung der Streamgewichte	
	Audio : Video
14 Sprecher	100:0
30 Sprecher	90:10
120 Sprecher	70:30

Tabelle 13: Hier ist die Verschiebung in Richtung visuelle Informationen bei steigender Größe der Trainingsmenge zu sehen.

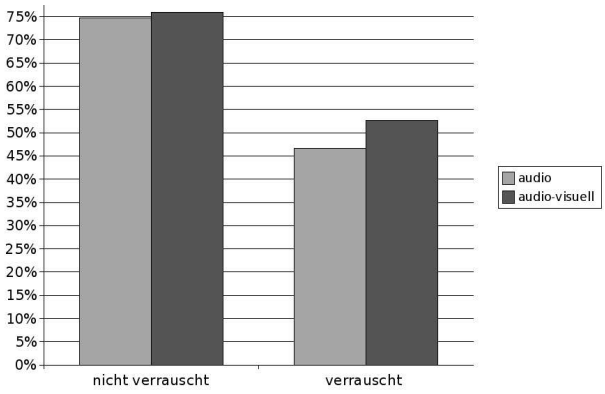

Abbildung 21: Durch den Einsatz eines audiovisuellen Spracherkenners anstelle eines rein akustischen konnte die Erkennungsleistung gesteigert werden.

6 Optimierungspotential

6.1 Merkmalsextraktion

Der hier erstellte audiovisuelle Spracherkenner stellt einen ersten Schritt dar. Es ist bereits eine Verbesserung gegenüber dem rein akustischen Erkenner zu verzeichnen, aber es besteht noch ein Abstand zu den in [10] erzielten Ergebnissen. Die an diesem System beteiligten Experten haben aber bereits eine jahrelange Erfahrung in der audiovisuellen Spracherkennung und haben viel Know How und viel Zeit investiert.

In allen Stufen des Erkennungsprozesses gibt es noch Verbesserungsmöglichkeiten, mit denen eine weitere Steigerung der Erkennungsleistung erreicht werden kann.

Als erstes arbeitet die automatische Detektion der Mundregion noch nicht robust genug für die visuelle Spracherkennung. Dies macht sich gleich in mehreren Punkten bemerkbar:

Es ist für die visuelle Spracherkennung sehr wichtig, die Mundwinkel sehr genau zu finden. Wie Heckmann, Berthommier, Savariaux und Kroschel in [42] gezeigt haben, verschlechtert sich die Erkennungsleistung bereits dramatisch, wenn die Mundregion nur um wenige Pixel variiert. Zum anderen kommt es hin und wieder vor, dass der Algorithmus zum Auffinden der Gesichtsmerkmale [37] mal die Mundregion verliert. Im Normalfall findet er sie bereits wenige Frames später wieder, um aber diese Ausreißer auszugleichen, wird die Mundposition über mehrere Bilder geglättet. Dadurch kann eine pixelgenaue Detektion der Mundwinkel nicht mehr gewährleistet werden. Außerdem hat der Algorithmus häufig Probleme, beim Finden der Merkmale von Personen, die eine Brille oder einen Bart tragen. Auch eine sehr dunkle Hautfarbe kann das Programm verwirren. In diesen Fällen ist es oft nicht möglich, die Mundregion automatisch erkennen zu lassen. Deshalb wurden die Experimente auch nur auf einer Teilmenge der in [10] genutzten Datenbasis durchgeführt.

Es wurden zuerst die Sprecher gewählt, bei denen die gefundene Mundregion sich im Laufe der Aufzeichnung möglichst wenig bewegt hat. Da fast alle Sprecher während der Aufnahme relativ ruhig gesessen haben, ist dies in diesem Fall ein gutes Kriterium für die Qualität der Detektion der Mundregion.

Aber auch nach der Detektion der Mundregion gibt es noch zahlreiche Bereiche, in denen auf Grund der beschränkten Zeit das Optimum noch nicht erreicht ist und noch Verbesserungen vorgenommen werden können.

6.2 Videoverarbeitung

Da der Rahmen dieser Arbeit zeitlich natürlich beschränkt war, gibt es in dem Bereich der Videoverarbeitung noch etliche Stellen, an denen das Optimum noch nicht erreicht ist. Es ist zwar eine Verbesserung der Spracherkennung erreicht worden, sogar im Fall von guten, nicht verrauschten Audiodaten. Aber es ist klar, dass nicht alle Optimierungsmöglichkeiten genutzt werden können, für die andere Gruppen mehrere Jahre Entwicklungszeit investiert haben.

Um die bestmöglichen Erkennungsergebnisse zu erreichen, ist es notwendig, die genutzten Bilder soweit wie möglich zu normieren. Unterschiede in der Beleuchtung können durch eine Histogrammnormalisierung in gewissen Grenzen ausgeglichen werden. Ist die Aufnahme allerdings zu stark über- oder unterbelichtet, so sind die Informationen verloren und können auch nicht mehr durch die Histogrammnormalisierung oder Anpassung des Gamma-Wertes rekonstruiert werden. Die Grenzen der Histogrammnormalisierung zeigt Abbildung 22.

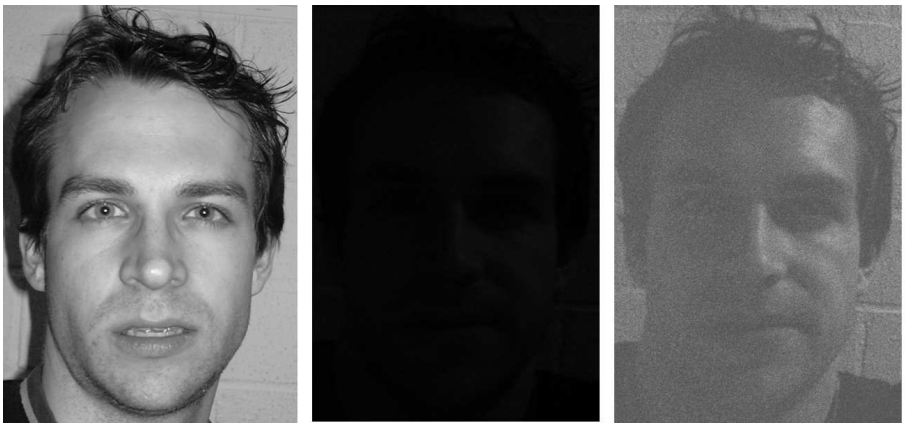

Abbildung 22: Durch die Belichtungskorrektur können Fehler nur bis zu einem gewissen Grad ausgeglichen werden.
1. korrekte Belichtung, 2. unterbelichtet, 3. korrigiert.

Verschiedene Abstände des Sprechers zur Kamera werden bereits recht gut ausgeglichen, indem die Mundwinkel des Probanden gesucht werden, um diese herum ein Bereich ausgeschnitten wird und auf eine einheitliche Größe skaliert wird. Eine Verbesserung in diesem Bereich ist möglich mithilfe einer verbesserten Detektion der Mundwinkel.

Nachdem der Abstand zwischen Sprecher und Kamera vereinheitlicht ist, gibt es aber immer noch Unterschiede im Bild. Es wäre zum Beispiel auch noch wünschenswert, die Kopfhaltung des Sprechers auszugleichen, wie in Abbildung 23 zu sehen. Hierzu müsste man anhand der selektierten Mundwinkel den Neigungswinkel des Kopfes berechnen und das Bild dementsprechend rotieren.

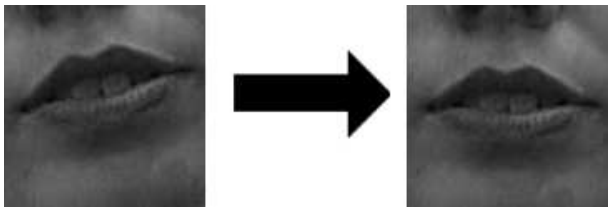

Abbildung 23: Durch Rotation des Bildes kann ein schief halten des Kopfes ausgeglichen werden.

Dieses Problem tritt bei der hier verwendeten Datenbasis nur sehr selten in extremer Form auf, da die meisten Sprecher den Kopf relativ aufrecht halten und nicht stark verdrehen. Trotzdem sollte eine Normierung dieses Faktors zu einer Verbesserung beitragen können.

Bis zu diesem Zeitpunkt werden auch nur Normierungen der Vorverarbeitung vorgenommen. Es sind noch keinerlei Verfahren zur Sprecheradaption in das System eingeflossen. Mithilfe dieser dürfte es vor allen Dingen möglich sein, die Erkennung für diejenigen Sprecher zu verbessern, die bis jetzt weit unterdurchschnittlich abschneiden.

6.3 Usability

In den letzten beiden Abschnitten wurde beschrieben, wie die Erkennungsleistung noch weiter gesteigert werden kann. Eine möglichst gut Erkennungsrate ist natürlich sehr wichtig für das Produkt Spracherkenner, mindestens genauso wichtig ist aber eine komfortable Bedienung. Wenn der Benutzer es nicht schafft, mit dem Produkt umzugehen, wird er es auch nicht benutzen, egal wie hervorragend die Erkennungsleistung ist.

Es gibt bei diesem System auch in dem Bereich Komfort noch Optimierungsmöglichkeiten. Noch muss die momentane Qualität der Audio- und Videosignale manuell bestimmt werden. Außerdem muss eine, bei diesen Bedin-

gungen optimal arbeitende, Abstimmung der Stream-Gewichte experimentell ermittelt werden.

Wünschenswert wäre ein System, welches die Qualität der Eingangsdaten automatisch bestimmen kann und daraus, ebenfalls ohne Unterstützung des Benutzers, die Parameter optimal einstellt. So wäre es möglich, auch unter wechselnden Bedingungen, immer eine optimale Erkennung zu erreichen.

Es gibt auch Sprecher, die für die visuelle Spracherkennung weniger geeignet sind, als andere, weil sie z.B. den Mund beim Reden nur sehr wenig bewegen oder dieser ganz oder teilweise verdeckt ist, z.B. durch einen Bart. Es wäre auch vorteilhaft, wenn das System einen solchen Fall selbstständig erkennen könnte und den Einfluss der visuellen Merkmale automatisch verringern würde. Auch wenn die audiovisuelle Erkennung im Mittel besser funktioniert als die rein akustische, ist es in einem solchen Fall besser auf diese zurück zu greifen, als das Ergebnis von unbrauchbaren visuellen Informationen negativ beeinflussen zu lassen.

7 Zusammenfassung

Die Zielsetzung dieser Arbeit war es, einen audiovisuellen Spracherkenner auf der Basis des Janus Spracherkennungssystems zu entwerfen, welches als Basis für weitere Forschung dienen kann. Es sollte versucht werden, eine Verbesserung der Erkennungsleistung des Janus Spracherkenners - durch den zusätzlichen Einsatz von visuellen Informationen - zu erreichen.

Für das Training des audiovisuellen Spracherkenners stand die Datenbasis des Workshops 2000 der Johns Hopkins University zur Verfügung. Diese Datenbasis besteht aus ca. 35 Stunden Videomaterial von insgesamt 261 Sprechern für das Training und 4,6 Stunden von 26 Sprechern für den Test.

Die Experimente dieser Arbeit wurden jedoch nur auf einer Teilmenge dieser Daten durchgeführt. Zu diesem Kompromiss war man gezwungen, weil die Extraktion der Gesichtsmerkmale nicht bei allen Sprechern zuverlässig genug funktioniert hat.

Mit den zur Verfügung stehenden Daten wurden Experimente mit drei Ansätzen zur Kombination von Audio- und Videodaten in einem audiovisuellen Spracherkenner durchgeführt. Der erste Ansatz bestand aus einem simplen Verknüpfen von vorverarbeiteten Audio- und Videodaten, im zweiten Ansatz kam eine hierarchische LDA zum Einsatz und als drittes wurde eine Multi-Stream-Architektur implementiert.

Zuerst wurden alle drei Ansätze mit einer kleinen Teilmenge der Daten, bestehend aus 14 bzw. 30 Sprechern, trainiert, um die unterschiedlichen Verfahren miteinander zu vergleichen. Hierbei zeigte sich, dass die einfache Konkatenation der Daten die geringste Erkennungsleistung besitzt. Die Ergebnisse der hierarchischen LDA und der Multi-Stream-Architektur liegen auf einem vergleichbaren Niveau.

Der Multi-Stream-Erkenner wurde für die weiteren Experimente auf einem größeren Trainingsset ausgewählt, da dieser mehr Flexibilität bei gleicher Erkennungsleistung bietet, als der Ansatz mit der hierarchischen LDA.

Die besten Ergebnisse wurden erzielt mit einem System, bei welchem der Audiostream mit allen 261 Sprechern trainiert wurde und der Videostream mit den 120 Sprechern, bei denen die Merkmalsextraktion am präzisesten funktioniert hat. Für den Test dieses Erkenners kamen 17 der 26 Sprecher zum Einsatz. Die relative Wortfehlerrate konnte für die Originalaudiodaten um 4,5% gesenkt werden und bei verrauschten Audiodaten sogar um 12,5%.

Wie im vorhergehenden Kapitel beschrieben, ist mit dieser Verbesserung, noch mal veranschaulicht in Abbildung 24, noch nicht das Optimum erreicht. Es gibt noch viele Stellen in dem gesamten Erkennungsprozess, an denen Optimierungen vorgenommen werden können, um die bereits erzielte Verbesserung noch weiter zu steigern. Sowohl die Merkmalsextraktion kann noch verbessert werden, damit die gesamten Trainingsdaten verwendet werden können, als auch bei der Videoverarbeitung gibt es noch Optimierungsmöglichkeiten, z.B. eine Rotation der Bilder um eine nicht ganz gerade Kopfhaltung auszugleichen.

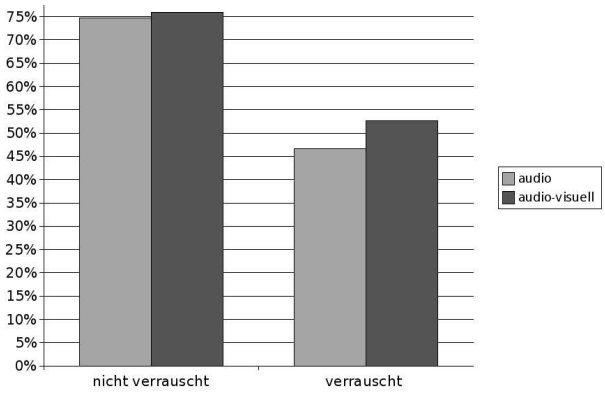

Abbildung 24: Durch den Einsatz eines audiovisuellen Spracherkenners anstelle eines rein akustischen konnte die Erkennungsleistung gesteigert werden.

Literatur

[1] S. Bennett, J. Hewitt, D. Kraithman, C. Britton. Making Chalk and Talk Accessible. Proc. of the 2003 Conference of Universal Usability, 2003

[2] Petra Geutner, Matthias Denecke, Uwe Meier, Martin Westphal and Alex Waibel. Conversational Speech Systems For On-board Car Navigation And Assistance ICSLP '98,Sydney, Australia, 1998

[3] Westphal, Martin. Robuste kontinuierliche Spracherkennung für mobile Informationssysteme. Dissertation an der Universität Karlsruhe (TH). Shaker Verlag, 2001

[4] G. Iliev, N. Kasabov. Adaptive Filtering with Averaging in Noise Cancellation fot Voices and Speech Recognition. In Future Directions for Intelligent Systems and Information Science. International Conference on Neural Information, 1999

[5] A. Singh. Adaptive Noise Cancellation. Central Elektronica Engineering Research Institute, University of Dehli, 2001

[6] Barbara Dodd and Ruth Campbell. Hearing by Eye: The Psychology of Lip-Reading. Psychology Pr., 1987

[7] G. Potamianos, C. Neti, G. Iyengar, Eric Helmuth. Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans, Proc. Eurospeech, 2001

[8] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal "The Karlsruhe-VERBMOBIL Speech Recognition Engine", in Proceedings of ICASSP, Munich, Germany, 1997.

[9] H. Soltau, F. Metze, C. Fügen, A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", in Proc. of ASRU, Trento, Italy, 2001.

[10] C. Neti, G. Potamianos et al. Audio-Visual Speech Recognition - Workshop 2000 Final Report. Center for Language and Speech Processing, The Johns Hopkins University, 2000

[11] G. Potamianos, C. Neti, S. Deligne. Joint Audio-Visual Speech Processing for Recognition and Enhancement. Proceedings of AVSP 2003, 2003

[12] R. Goecke, G. Potamianos, C. Neti. Noisy Audio Feature Enhancement using Audio-Visual Speech Data. ICASSP 02, 2002

[13] Eric D. Petajan, Automatic lipreading to enhance speech recognition, Proceedings of the IEEE Communication Society Global Telecommunications Conference, Atlanta, Georgia, 1984.

[14] A.J. Goldschen, O.N. Gracia, E. Petajan. Continous optical automatic speech recognition by lipreading. 28th Annual Asimolar conference on Signal speech and Computers.

[15] P. Duchnowski, U. Meier, A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. Internation Conference on Spoken Language Processing, ICSLP, pages 547-550, 1994

[16] Uwe Meier, Rainer Stiefelhagen, Jie Yang, Alex Waibel. Towards Unrestricted Lipreading. International Journal of pattern Recognition and Artificial Intelligence, Vol. 14, No. 5, pp. 571-785, 2000, Second International Conference on Multimodal Interfaces (ICMI99), 1999.

[17] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. Proc. IEEE Intl. Conf. Acous. Speech Sig. Process, pp. 669-672, 1994

[18] J. Huang, G. Potamianos, C. Neti. Improving Audio-Visual Speech Recognition with an Infrared Headset. Proceedings of AVSP 2003, 2003

[19] R. de Córdoba, X. Menéndez Pidal, J. Macías-Guarasa, A. Gallardo and J.M. Pardo. Development and Improvement of a Real-Time ASR System for Isolated Digits in Spanish Over the Telephone Line. Proceedings of the 4th European Conference on Speech Communication and Technology 1995 (EUROSPEECH'95), pp. 1537-1540, 1995

[20] E.G. Schukat-Talamazzini. Statistische Spracherkennung. Künstliche Intelligenz, 3:7-9, 1995.

[21] C.C. Chibelushi, F. Deravi, J.S.D. Mason. A Review of Speech-Based Bimodal Recognition. IEEE Transaction on Multimedia, Vol. 4, No. 1, 2002

[22] A. Ogihara, S. Asao. An isolated word speech recognition based on fusion of visual and auditory information using 30-frames/s and 24-bit color image. IEICE Trans. Fund. Electron., Commun. Comput. Sci., vol. E80A, no 8, pp. 1417-1422, 1997

[23] Zhipeng Zhang, Sadaoki Furui. MDL-Based Cluster Number Decisions Methods for Speaker Clustering and MLLR Adaptation. Tokyo Institute of Technologie, 2001

[24] E.G. Schukat-Talamazzini. Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen. Seiten 165 - 197. Vieweg Verlag, Braunschweig, 1995.

[25] The INTERNATIONAL PHONETIC ASSOCIATION
http://www.arts.gla.ac.uk/IPA/ipa.html

[26] Christoph Bregler. Lippenlesen als Unterstützung zur robusten automatischen Spracherkennung. Karlsruhe, Univ., FZI, Dipl.-Arb., 1993

[27] E.G. Schukat-Talamazzini. Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen. Seiten 199 - 230. Vieweg Verlag, Braunschweig, 1995.

[28] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE, 77(2), Seiten 257-286

[29] Kai Nickel. Erkennung von Zeigegesten basierend auf 3D-Tracking von Kopf und Händen. Karlsruhe, Univ., Institut für Logik, Komplexität und Deduktionssysteme, Dipl.-Arb., 2003

[30] E.G. Schukat-Talamazzini. Automatische Spracherkennung - Grundlagen, statistische Modelle und effiziente Algorithmen. Seiten 121 - 163. Vieweg Verlag, Braunschweig, 1995.

[31] Lisa J. Stifelman. The Cocktail Party Effect in Auditory Interfaces: A Study of Simultaneous Presentation. MIT Madia Laboratory, 1994

[32] H. McGurk and J. MacDonald. Hearing lips and seeing voices. Nature 1976

[33] Thomas Östreich
http://zebra.fh-weingarten.de/~transcode/

[34] The LAME Project
http://lame.sourceforge.net/

[35] Keinosuke Fukunaga. Introduction to statistical pattern recognition. New York : Acad. Pr., 1972

[36] S. Balakrishnama, A. Ganapathiraju. Linear Discriminant Analysis - A Brief Tutorial. Institute for Signal and Information Processing, Mississippi State University, 1998

[37] Rainer Stiefelhagen and Jie Yang. Gaze Tracking for Multimodal Human-Computer Interaction. Proc. of the International Conference on Acoustics, Speech and Signal Processing: ICASSP'97, Munich, Germany, April 1997.

[38] R. Stiefelhagen, J. Yang, A. Waibel. A Model-Based Gaze Tracking System. Int. j. of artif. intell. tools 6 (1997) H. 2 S. 193-209. Universität Karlsruhe; Institut für Logik, Komplexität und Deduktionssysteme, 1997.

[39] Bescheibung des Dateifomates z.B. unter:
http://www-lehre.informatik.uni-osnabrueck.de/
~cg/2000/skript/10_2_PBM_PGM_.html
http://www-lehre.inf.uos.de/~cg/2002/Pdf/skript09.pdf

[40] G. Gravier, G. Potamianos and C. Neti. Asynchrony modeling for audio-visual speech recognition. Proc. Human Language Technology Conference, 2002

[41] Martin Heckmann. Adaptive Datenfusion für die audio-visuelle Spracherkennung. Shaker Verlag, 2003

[42] M. Heckmann, F. Berthommier, C. Savariaux, K. Kroschel. Effects of Image Distortions on Audio-Visual Speech Recognition