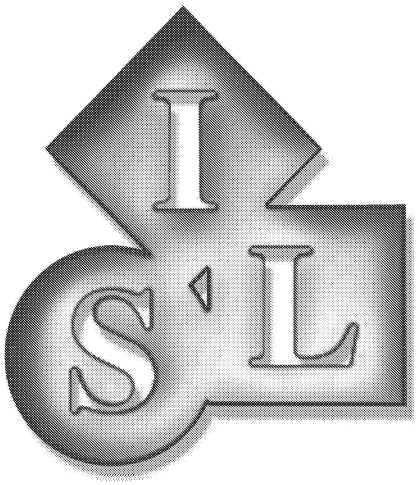
Flexible Ballungsverfahren für Graphembasierte Spracherkennung



Interactive Systems Lab

Universität Karlsruhe (TH)

Studienarbeit

Borislava Mimer

Betreuer:
Dipl.-Inf. Sebastian Stüker
Prof. Dr. Alex Waibel

Juni 2004

Danksagung

Ich möchte mich herzlich bei meinem Betreuer Sebastian Stüker bedanken, der für mich immer Zeit hatte, der mich stets mit Tat und Rat unterstützte und von dem ich sehr viel lernen konnte. Für die Möglichkeit, diese Studienarbeit an seinem Institut durchführen zu können, bedanke ich mich bei Herrn Prof. Dr. Alex Waibel. Mein Dank geht an Dr. Tanja Schultz für ihre Interesse und nützliche Hinweise. Bei Matthias Wölfel, dem Zimmerkollege von Sebastian Stüker, bedanke ich mich, weil er mutig die Unterbrechungen durch meines Erscheinen ertrug. Ich danke allen, die mir während dieser Studienarbeit Halt gaben.

Zusammenfassung

In modernen Spracherkennungssystemen mit großem Vokabular ist das Aussprachewörterbuch eine der teuersten Komponenten. Es zu erstellen, erfordert häufig linguistisches Wissen. Existierende, automatische Verfahren zur Erzeugung von phonembasierten Wörterbüchern erfordern oft eine manuelle Nachbearbeitung durch Experten. Diese manuelle Nachbearbeitung ist teuer und zeitaufwendig.

Aus diesem Grund wurde in letzter Zeit vermehrt die Verwendung von graphembasierten Spracherkennern erforscht. Hierbei werden Wörter in ihrer Grapheme als Unterworteinheiten zerlegt, an Stelle von Phonemen. Grapheme als Modellierungseinheiten haben den Vorteil gegenüber Phonemen, dass sie die Erzeugung des Aussprachewörterbuchs trivialisieren.

Wegen der im Vergleich zu Phonemen loseren Relation zwischen Aussprache und Graphemen sind die kontextabhängige Modellierung der Einheiten und die gemeinsame Nutzung von Parametern von zentraler Bedeutung. Ein graphembasiertes Wörterbuch ist ein Einzel-Aussprache Wörterbuch in Reinform. Aussprachevarianten müssen implizit modelliert werden.

In dieser Studienarbeit präsentiere ich die Anwendung des flexiblen Ballungsverfahrens von Hua et. al. auf graphembasierte Erkennungssysteme in den Sprachen Deutsch und Englisch. Populäre Ballungsverfahren, z. B. divisive, basieren auf Entscheidungsbäumen. Bei der kontextabhängigen Modellierung mit divisiven Vorgängen benutzt man Kontextentscheidungsbäume, die ein Bestandteil der allgemeine Entscheidungsbäume (classifikation and regression trees CART) sind. In jedem Knoten des Ballungsbaums wird entschieden, wie die Knoten aufgeteilt werden müssen. In der Regel werden mehrere Kontextentscheidungsbäume gebaut, z. B. für jeden Unterzustand (Beginn/Mitte/End) des Phonems ein eigener Baum. Bei dem in dieser Studienarbeit angewendeten flexiblen Ballungsverfahren wird nur ein Entscheidungsbaum für jeden Unterzustand für alle Grapheme konstruiert. Mehrere Polygrapheme/Unterzustände können jeden Knoten gemeinsam benutzen. Durch Aufhebung dieser Restriktionen wird eine größerer Suchraum beim Finden einer optimalen Parameterkopplung abgesucht.

Im Laufe der Studienarbeit habe ich graphembasierte Spracherkenner für zwei Sprachen - Deutsch und Englisch- trainiert und die Wortakkuratheit bei Verwendung des bisherigen Ballungsverfahren mit der Akkuratheit bei Anwendung des flexiblen Ballungsverfahren verglichen.

Inhaltsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Über die letzten Jahrzehnte gab es einen rasanten Fortschritt in der Wissenschaft der automatischen Spracherkennung. Spracherkennungssysteme werden heute in verschiedenen Formen in der realen Welt eingesetzt, z. B. eingebettet in Geräte, wie Mobiltelefone oder Navigationssysteme, oder als Einzelanwendungen auf Heimrechnern, etwa in Form von Diktiersystemen. Mit dieser steigenden ökonomischen Bedeutung ist es immer wichtiger geworden, in der Lage zu sein, Spracherkennungssysteme schnell zu erweitern, etwa um neue Worte, oder schnell auf neue, unbekannte Sprachen oder Domänen zu portieren. Geschwindigkeit und Kosteneffizienz sind hier von zentraler Bedeutung.

Eine der teuersten Komponenten eines Spracherkennungssystems ist das Aussprachewörterbuch. In dem Aussprachewörterbuch sind alle zu erkennenden Wörter durch sprachliche Untereinheiten repräsentiert. Man braucht oft linguistisches Wissen und Hilfe von Expertern, um es zu erzeugen. Deshalb wird in letzter Zeit vermehrt nach Einheiten zur Modellierung von Sprache geforscht, deren Anwendung bei der Erstellung des Aussprachewörterbuchs kein spezielles Wissen erfordern und somit Zeit und Geld sparen.

1.1 Motivation

In modernen Spracherkennern werden heutzutage Unterworteinheiten als Modellierungseinheiten eingesetzt. Unterworteinheiten haben eine geringere zeitliche Ausdehnung als Wörter. Außerdem nehmen Untereinheiten einen kleineren Bereich des Merkmalraums ein, was sehr hilfsreich sein kann, wenn die Merkmale mancher Wörter überhaupt nicht im Training vorkommen. Das Training mit Unterworteinheiten ist robuster, weil sie häufiger als Wörter vorkommen.

Traditionell verwendet man Phoneme als Unterworteinheiten bei der Aussprachewörterbucherzeugung. Das Phonem ist die kleinste, bedeutungsunterscheidende sprachliche Einheit. Die Anzahl von Phonemen ist geringer als die von Silben, und Phoneme sind besser zu trainieren als Silben. In diesem Fall, wenn ein neu-

es Wort in das Erkennervokabular aufgenommen werden soll, braucht man kein Trainingsmuster für dieses Wort.

In dieser Arbeit werden Grapheme als Unterworteinheiten benutzt. Grapheme haben den Vorteil gegenüber Phonemen, dass sie die Erstellung des Aussprachewörterbuchs viel einfacher machen. Die Anwendung von Graphemen hilft, die Kosten bei der Erzeugung des Aussprachewörterbuchs zu senken. Die Unterteilung eines Graphems in Subgrapheme (Beginn/Mitte/End) repräsentiert die Dynamik innerhalb eines Graphems.

1.2 Grapheme

[Gal85] definiert als Grapheme die kleinsten, schreibsprachlichen Struktureinheiten, die sich sowohl formal wie auch funktional definieren lassen.

Eine der formal definierten Graphemklassen heißt Grapheme (im engeren Sinn). Sie lassen sich in selbständige und unselbständige unterteilen. Als selbständige werden Buchstaben, Hilfszeichen, Leerzeichen, Ziffern und Sonderzeichen verstanden. Unselbständige Grapheme (im engeren Sinn) sind diakritische Zeichen.

Als Buchstaben gelten die Grapheme, die in einer festen Reihe, Alphabet genannt, angeordnet sind. Z. B. besteht das Alphabet der englischen Sprache aus 26 Buchstaben:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Als Variation derselben Grapheme werden Kleinbuchstaben und Großbuchstaben angesehen.

Das Graphem <ß>, Eszett, das in der deutschen Schrift verwendet wird, kann man als ein weiteres Grundgraphem oder als die allographische Realisierung der Graphemgruppe <ss>betrachten.

Unselbständige Grapheme können nur in Kombination mit selbständigen vorkommen. Im Deutschen kommen für diese Graphemklasse im Wesentlichen nur die Umlautpunkte in Frage. Es gibt verschiedene Möglichkeiten die Umlaute <ä>, <ö> und <ü> anzuordnen: als eigenständige Grapheme, als allographische Realisierung der Graphemgruppen <ae>,<oe>, <ue> oder als Verbindungen von <a>, <o>, <u> mit dem diakritischen Zeichen „Umlautpunkt“. Im dieser Studienarbeit wird die erste Version verwendet, nur statt dem diakritischen Zeichen „Umlautpunkt“, wird das wellenförmigen Zeichen „Tilde“ (∼) vor dem Buchstaben eingegeben.

Für diese Studienarbeit ist eine der funktional definierte Graphemklassen relevant - Grundgrapheme. Grundgrapheme sind ausschließlich bedeutungsunterscheidende Zeichen, die sich zu bedeutungstragenden Zeichengruppen zusammenschließen. Im allgemein fungieren Buchstaben und diakritische Zeichen als Grundgrapheme. Sie bilden ein Analogon zu den Phonemen der gesprochenen Sprache. Oft wird diese Subklasse der Grundgrapheme als Grapheme bezeichnet,

zum Beispiel in [D+84]. Wir gehen davon aus, dass die geschriebene Standardsprache ein Subsystem bildet, das parallel zur gesprochenen Standardsprache Teil des Gesamtsystems der deutsche Sprache (bzw. englischen Sprache) ist. Zwischen Grundgraphemen (bzw. Grundgraphemgruppen) einerseits und Phonemen (bzw. Phonemengruppen) anderseits bestehen Äquivalenzrelationen. Mit dem Terminus Graphem werden allerdings nur graphische Elemente, nicht aber auch die Relation dieser Elemente mit den Elementen eines anderen Subsystems der Sprache bezeichnet.

Die Phonem-Graphem-Beziehung (phonematisches Prinzip) ist zweifellos sehr wichtig, insbesondere für die Spracherkennung. Wenn jedem Phonem nur ein Graphemzeichen entspräche, wäre die Schreibung optimal geregelt. Diese 1:1-Beziehung ist aber nicht immer gegeben. Im Deutschen entsprechen einige Phoneme zwei oder drei Buchstaben, z. B. entspricht /x/ den beiden Graphemen *ch*, wie in *suchen*. Die Phonemfolgen /k/+/s/ und /t/+/s/ werden unter bestimmten Bedienungen durch eine einzigen Buchstaben wiedergegeben: x, z, zum Beispiel in Hexe, Brezel. Außerdem gibt es zahlreiche Ausnahme von den Regeln.

In diese Studienarbeit benutzen wir der Einfachheit halber nur den Terminus Graphem für die Einheiten des Aussprachewörterbuches und des Fragenkatalogs (auch, wenn es in einigen Fällen, wie z. B. für die Stille und den Geräuschen, um Graphemgruppen geht). Als Graphembestand werden unter anderen folgende Einheiten verwendet: Graphem @ für das Füllzeichen, ein Graphem für die Stille, für die Englische Sprache - die Großbuchstaben des aktuellen lateinischen Alphabets und zwei Graphemen für Geräusche, für die Deutsche Sprache - die Großbuchstaben des aktuellen lateinischen Alphabets, Umlaute und ein Graphem für das Geräusch.

1.3 Schriftsysteme

Sprache und Schrift sind auf das engste miteinander verbunden. Der Begriff „Schriftsystem“ definiert die Beziehung einer Menge von Schriftzeichen zum System einer einzelnen Sprache. Die Entwicklung der heutigen Schriften hat einen langen Weg vom nicht-phonologischen System zum phonologischen hinter sich. Unter phonologischen versteht man Schriftsysteme, bei denen es eine klare Beziehung zwischen den Symbolen und den Lauten der Sprache gibt, und unter nicht-phonologischen Systemen solche, bei denen dies nicht der Fall ist.

Piktographische und ideographische Systeme sowie die Keilschrift, ägyptische Hieroglyphen und die logographische Schriften gehören zu den nicht-phonologischen Systemen [Cry93].

In einer Wortschrift (Logographie) stellen die Grapheme ganze Wörte dar. Solche Symbole nennt man Logogramme. Auch heute finden Logogramme in modernen Schriftsystemen ihren Platz. Derartige Systeme sind mehrdeutig und haben mehrere Nachteile. Zum Beispiel ist es nicht möglich, mit Hilfe von Logogrammen

Eigennamen zu repräsentieren.

Aus Piktogrammen sind sämtliche Schriftsysteme der Welt hervorgegangen. Piktogramme sind stilisierte Abbildungen. So steht zum Beispiel die Zeichnung der Sonne für das gesprochene Wort Sonne. Man muß die Symbole kennen, um die Schrift zu verstehen. Mehrdeutigkeit, unzureichende Ausdrucksmöglichkeit und fehlender Kontext der piktographischen Schriften führte zu weiteren Entwicklungen der Schriftsysteme.

In Sumer haben die Schreiber etwa um 3000 die Keilschrift erfunden. Dies war die erste Schrift in der Geschichte, die auch an andere, völlig fremde Sprachen angepasst werden konnte. Sumerische, ägyptische und hethitische Schriften sind eine Mischform aus piktographischen, ideographischen und sprachlichen Einheiten. In Ideographien bedeutet ein Zeichen auf der Grundlage einer Konvention nicht mehr den Gegenstand, den es darstellt, sondern einen (abstrakten) Sachverhalt. Die phonologische Abbildung der Realität ist in diesem Fall nicht mehr zu erkennen.

Ohne Phonetisierung wäre der entscheidende Durchbruch zu einer „richtigen Sprache“ nicht erfolgt. Phonologische Schriften unterteilen sich in die Silbenschriften und die Alphabetschriften. Bei einer Silbenschrift wird jedes Zeichen (Graphem) zur Darstellung einer gesprochene Silbe (in der Regel eines Konsonant-Vokal-Paars), verwendet. Mit weniger als 100 Zeichen kann man schon bei einer reinen Silbenschrift auskommen, selten werden mehr als 200 gebraucht. Der Nachteil der Silbenschrift ist, dass man normalerweise mehr Grapheme benötigt, um eine gegebene Äußerung zu schreiben, als bei einer Logographie.

Die Entwicklung von eine Silbenschrift zu einer vollständigen Alphabetschrift (Buchstabenschrift) ist durch die lautliche Trennung der Konsonanten von den Vokalen und ihre getrennte Darstellung in der Schrift geschehen. Dieser Schritt war aber nicht zwangsläufig. Viele Schriften, wie etwa die japanische Kana-Schrift, haben die Entwicklung zur Buchstabenschrift nicht vollzogen. Ein solcher Schritt war aber notwendig bei den Sprachen, die viele einzelne Konsonanten aufweisen, wie bei den indogermanischen Sprachen.

Ein direkter Zusammenhang zwischen Graphem und Phonem besteht bei den Alphabet-Schriften. Ein Alphabet versucht im Idealfall, jeden Einzellaut durch ein Graphem darzustellen. Solche völlig regelmäßigen Systeme sind aber selten. Als Beispiel kann man das koreanische Alphabet (das vollkommenste phonetische System, das man kennt) nennen. Im Allgemeinen sind die heute verwendeten Alphabete unregelmäßig. Die Gründe dafür sind, dass entweder mit Aussprache-veränderungen nicht Schritt gehalten wurde, oder, dass die Sprache ein für sie ursprünglich nicht entwickeltes Alphabet benutzt.

Die meisten Alphabete verfügen über etwa 20 bis 30 Symbole. Dagegen hat das Rotokas-Alphabet, das auf den Salomon-Inseln verwendet wird, nur elf Buchstaben, während das Khmer-Alphabet 74 Buchstaben hat und somit das längste Alphabet der Welt ist.

Das erste bekannte Alphabet ist das nordsemitische. Dieses Alphabet bestand

ausschließlich aus 22 Konsonanten, weil die semitische Sprache kaum Vokale hat, und diese auch kaum hörbar ausgesprochen werden. Vokale wurden als bekannt vorausgesetzt und sind beim Lesen zu ergänzen. Das nordsemitische Alphabet wurde für das hebräische, das arabische und das phönizische Alphabet zum Vorbild. Im modernen hebräischen und arabischen Alphabet sind immer noch ausschließlich Konsonanten vertreten, wobei das erstere 22 und das letztere 28 Zeichen besitzt. Zur Darstellung langer Vokale können einige Konsonanten benutzt werden. Vokale können durch Vokalpunkte und -striche (diakritischer Zeichen) unterhalb, oberhalb oder neben den Konsonanten gekennzeichnet werden. Man schreibt von rechts nach links. Aus dem semitischen Alphabet haben sich mehrere Zweige herausgebildet.

Später übernahmen die Griechen die phönizische Variante des semitischen Alphabets. Die Leistung der Griechen bestand darin, dass sie für die Aufzeichnung des Griechischen nicht benötigte Konsonantenzeichen umformten und ihnen Vokalwerte zuwiesen. Somit war die erste voll phonetisierte Schrift entstanden.

Auf dem von den Slawen besiedelten Gebiet verbreitete sich im 10. Jahrhundert das kyrillische Alphabet. Es stammt von der griechischen Unzialschrift ab und bestand aus 43 Buchstaben. Zur Darstellung der slawischen Laute wurden zusätzliche Zeichen entworfen, die es nicht im Griechischen gab. Das kyrillische Alphabet verdrängte das bis dahin gebräuchliche glagolitisches Alphabet (Glagoliza). Später wurde die kyrillische Schrift durch die Schriftreform Peters des Großen vereinfacht. In verschiedenen Ausprägungen wird heute das kyrillische Alphabet im Russischen, Ukrainischen, Bulgarischen und anderen Schriften verwendet. Allerdings hat das moderne kyrillische Alphabet einige Buchstaben verloren. Das heutige russische Alphabet existiert seit 1918 und besitzt 33 Buchstaben.

Die griechische Schrift verbreitete sich und wurde zu verschiedenen Formen modifiziert: dem etruskischen, oskischen und umbrischen Alphabet [C$^+$98]. Früher meinten Wissenschaftler, dass die Römer die griechische Schrift von den Griechen übernommen hätten. Heute sind sie der Meinung, dass die Römer die Buchstabenschrift von den Etruskern gelernt haben. Die römischen Eroberungen führten zu der Verbreitung der lateinischen Sprache und so wurde das römische Alphabet zu Grundlage aller Sprachen Westeuropas und Teilen Osteuropas (bei den Polen, Tschechen und Slowaken).

Seit der römischen Zeit hat das lateinische Alphabet relativ geringe Änderungen erfahren. Beim schriftlichen Fixieren der Aussprache eines Wortes kommt es manchmal zu Schwierigkeiten. Die 26 Buchstaben, die im aktuellen lateinischen Alphabet zur Verfügung stehen, reichen nicht aus, um die Laute schriftlich auszudrücken, die allein in den europäischen Sprachen vorkommen. Manche Sprachen haben zusätzliche Laute und verwenden Akzente und sonstige Sonderzeichen. Allein in Europa gibt es 85 Buchstaben, die auf der lateinischen Schrift basieren, zum Beispiel á,ç. Im Deutschen sind Buchstabenkombinationen notwendig, um Zischlaute (-sch-) und andere Laute auszudrücken. Umlaute wurden durch die

Kombination von Buchstaben und diakritische Punkten (das Trema) erzeugt.

Es gibt große Unterschiede hinsichtlich der Graphem-Phonem-Beziehung zwischen den Sprachen. Bei einigen, wie etwa dem Spanischen und dem Finnischen, entsprechen die Grapheme den Phonemen sehr genau. Bei anderen, wie etwa dem Englischen, das viele Unregelmäßigkeiten aufweist [Bod76], ist das nicht der Fall. Der Grund dafür ist, dass die Modifikationen der Schrift nicht mit der Entwicklung der Aussprache einhergingen. Nur bei wenigen Sprachen wurde ein neues Schriftsystem entwickelt oder ein bestehendes so angepasst, dass Phoneme und Grapheme perfekt einander entsprechen. Ein Beispiel für ein solches Alphabet ist das Türkische. Beim Erlernen von Sprachen, bei denen Grapheme und Phoneme einander kaum entsprechen, stößt man auf eine große Anzahl Rechtschreibregeln. Bei der Spracherkennung ist aber eine enge Graphem-Phonem-Beziehung sehr erwünscht, denn je enger sie ist, desto einfacher lässt sich die Aussprache eines Wortes aus seiner Schreibweise generieren.

1.4 Automatische Spracherkennung

Automatische Spracherkennung beschäftigt sich mit der Untersuchung und Entwicklung von Verfahren, deren Verwendung es erlaubt, mit Hilfe von Automaten, insbesondere Computern, gesprochene Sprache zu erkennen und zu verarbeiten. Sprache zu erkennen bedeutet, für eine gesprochene Äußerung, die in der Regel eine zeitliche Folge von Merkmalen darstellt, diejenige Wortsequenz zu finden, für die die Wahrscheinlichkeit gemäß des erkennerinternen Modells am größten ist [Rog98].

In der Spracherkennung wird die getrennte Modellierung der Akustik und der Linguistik praktiziert. Diese Trennung veranschaulicht die Gleichung:

$$P(W|X) = \frac{p(X|W) \cdot P(W)}{p(X)} \tag{1.1}$$

Die Gleichung (1.1), die durch Umformung mit Hilfe der Bayes-Regel gewonnen wurde, ist als die *Fundamentalformel der Spracherkennung* bekannt. P(W|X) ist die Wahrscheinlichkeit, dass die Wortsequenz W gesprochen wurde, unter der Bedingung, dass eine gesprochene Äußerung X beobachtet wird. p(X) ist die a-priopi Wahrscheinlichkeit dafür, dass der Merkmalvektor X überhaupt beobachtet werden kann. p(X|W) ist die bedingte Wahrscheinlichkeit, das Signal X zu beobachten, wenn die Wortfolge W gesprochen wurde, und P(W) ist die a-priori Wahrscheinlichkeit dafür, dass die Wortsequenz W gesprochen wird.

Das Problem der automatischen Spracherkennung besteht darin, für ein gegebenes Signal X diejenige Wortfolge $\bar{W}$ zu finden, die bei Beobachtung von X am

wahrscheinlichsten produziert wurde:

$$\begin{aligned} \bar{W} &= \underset{W}{argmax}\ P(W|X) \\ &= \underset{W}{argmax}\ \frac{p(X|W) \cdot P(W)}{p(X)} \\ &= \underset{W}{argmax}\ p(X|W) \cdot P(W) \end{aligned} \tag{1.2}$$

Wie die Umformung der Gleichung (1.2) zeigt, ist p(X) für das Maximieren unerheblich. Und so bleiben zwei Faktoren von Bedeutung: das *Akustische Modell* p(X|W) und *Sprachmodell* P(W).

Ein monolinguales Spracherkennungssystem besteht im Allgemeinen aus einer Komponente zur Merkmalextraktion, einem Akustische Modell, einem Aussprachewörterbuch, in dem alle zu erkennenden Wörter durch sprachliche Untereinheiten dargestellt sind, einem Sprachmodell und einem Dekoder. Die Aufgabe des Dekoders ist es, die beste Wortfolge $\bar{W}$ unter Ausnutzung aller oben genannten Komponenten zu finden.

1.5 Ziel der Arbeit

Die Aufgabe der Studienarbeit bestand darin, das flexibles Ballungsverfahren aus [YW03],[YS03] auf graphembasierte Spracherkennungsysteme umzusetzen. Es mussten monolinguale Spracherkenner für zwei Sprachen - Deutsch und Englisch - trainiert werden. Bei der Anwendung des flexiblen Ballungsverfahren sollte die Anzahl der Teilbäume und Gaussiens in zahlreichen Experimenten variiert werden. Drei Fragenmengen für den Entscheidungsbaum musste ich untersuchen.

Die Ergebnisse der Experimente machen es möglich, die Wortakkuratheit bei Anwendung des flexiblen Ballungsverfahren mit der Akkuratheit bei der Verwendung des bisherigen Ballungsverfahren zu vergleichen.

Kapitel 2

Verwandte Arbeiten

2.1 Arbeiten über Grapheme

In letzter Zeit wurde vermehrt die Verwendung von Graphemen für Spracherkennungssysteme untersucht. Die Arbeiten aus diesem Gebiet sind in der Studienarbeit nachgeforscht worden.

Schillo, Fink und Kummert haben zwei Ansätze zum Bau eines graphembasierten Spracherkennungssystems präsentiert: Erkenner für kontextunabhängige graphemische Einheiten mit verschiedene statistische Sprachmodellen (engl. graphemic Typewriter) und lexikonbaumbasierte Erkenner mit Trigraphemen [SFK00]. Diese Erkenner sind fähig, mit einem großen Vokabular umzugehen. Die Ergebnisse der Experimenten zeigen, dass, obwohl das Niveau der Standarderkenner nicht erreicht wurde, der Verlust der Akkuratheit teilweise mit anderen Mitteln kompensiert werden kann.

In [KN02] schlagen Kanthak und Ney vor, einen graphembasierten Entscheidungsbaum zusammen mit phonetisch motivierte Fragen zu nutzen. Das ist wichtig für die automatische Erzeugung des Aussprachewörterbuchs für automatische Spracherkennungssysteme. Es wurde gezeigt, dass die automatische Fragengeneration es erlaubt, ohne manuelle Anstrengungen bei der Fragenkatalogerstellung auskommen.

Bei der automatische Konstruktion des Aussprachewörterbuchs wird meistens die graphische oder orthographische Repräsentation von Wörter in eine phonetische umgewandelt.

Kanthak und Ney haben vorgeschlagen den zustandsgebundene Entscheidungsbaum für die orthographische Repräsentation des Wörtes anzuwenden.

Den Fragenkatalog für diesen Entscheidungsbaum kann man manuell oder automatisch generieren. Fragen werden zu den Graphemen gestellt und können leicht aus den phonetischen gewonnen werden. Wenn das Graphem ein Teil eines Phonems ist, dann wird es der phonetischen Frage zugewiesen. Bei der manuellen Erstellung braucht man allerdings in einigen Fällen eine Hilfe der Experten

in Phonetik. Für die automatische Generierung von Fragen wird ein Bottom-up Ballungsalgorithmus verwendet. Die Fragen werden aus akustischen Trainingsdaten gelernt; somit braucht man dafür keine phonetischen Kenntnisse. Die authomatische Erzeugung der Fragen eliminiert die Notwendigkeit, den Fragenkatalog manuell zu kreieren.

Die Experimente wurden auf vier Korpora mit verschiedenen Sprachen (Niederländisch, Deutsch, Italienisch und Englisch) durchgeführt. Die Wortfehlerrate auf drei Korpora (Niederländisch,Deutsch und Italienisch) ist nur bis zu 2 % relativ gewachsen, auf dem Englische Korpus, allerdings, ist sie um 20 % relativ gestiegen.

In [KN03] haben Kanthak und Ney multilinguale akustische Modellierung mittels Graphemen realisiert. In diesem Fall wurden für die Experimente Korpora in folgenden vier Sprachen benutzt: Niederländisch, Französisch, Italienisch und Deutsch.

Die Hauptidee ist die Anwendung der automatische Fragengeneration für den zustandsgebundenen Entscheidungsbaum zur multilingualen, akustischen Modellierung, was die vollständige Elimination von manuelle Anstrengungen erlaubt. Im Gegensatz zu den phonembasierten Unterworteinheiten ist es nicht notwendig, eine gemeinsame Menge akustischer Unterworteinheiten zu finden.

Eine Arbeit im Bereich der graphembasierten Spracherkennung wurde von Killer, Stüker und Schultz in [KSS03] präsentiert. Der Unterschied des graphembasierte Spracherkenners zum phonembasierten liegt in den Unterworteinheiten, im Aussprachewörterbuch und im Fragenkatalog für die kontextabhängige Modellierung. Es wurden graphembasierte Spracherkenner in drei Sprachen gebaut und trainiert. Dann wurden die Ergebnisse mit denen von äquivalenten, phonembasierten Erkennern verglichen.

Große Aufmerksamkeit wurde der Fragenkatalogerstellung für den Entscheidungsbaum gewidmet. Es wurden vier Methoden untersucht:

1. Phonem-Graphem Fragen: Der für den phonembasierten Erkennern verwendeter Fragenkatalog wurde mit Hilfe sprachabhängiger Phonem-zu-Graphem Abbildungen umgewandelt und für die Erzeugung kontextabhängiger Modellen benutzt.

2. Bottom-up Entropy Fragen: Die Menge von Monographemen wurde mittels eines agglomerativen (bottom-up) Ballungsverfahren solange vereinigt, bis nur eine Gruppe bleibt. Als Distanzmaß wurde die Entropie-Distanz eingesetzt. Die Knoten in dem resultierenden Baum sind dann als Fragen anzusehen.

3. Hybrid Entropy Fragen: Das Ballungsverfahren stellt sich als Hybrid von top-down (divisive) und bottom-up Verfahren dar. Die engsten Grapheme aus der Menge von Monographemen werden beim bottom-up Verfahren

zusammengelegt, bis der Raum möglicher Aufteilungen vollständig abgesucht werden kann. Das erlaubt den Vergleich von allen möglichen Gruppierungsmöglichkeiten, was zur Erstellung von zwei maximal separaten Klasse führen soll. Am Anfang des nachfolgenden Rekursionsschritts wird der beste Anteil ausgewählt. Die Gruppierungsschritte wiederholen sich an jeder resultierende Teilmenge mit nachfolgender ausführlicher Suche. Die Teilmengen bilden ein Baum für das top-down Verfahren, wenn sie in jeder Rekursionsschritt gespeichert werden. Die Zwischenknoten dienen als Fragen.

4. Singleton: Man fragt, welche Art von Graphemen im linken oder rechten Kontext sind. Jede Frage besteht aus einem Graphem.

Die Singleton Fragen liefern niedrigere Wortfehlerrate als Phoneme-Grapheme Fragen in Englisch und in Deutsch, aber nicht in Spanisch. Die Bottom-up Fragenmengen zeigen generell schlechtere Zahlen als der Hybrid Entropy Fragenkatalog.

Die Experimente führen zur Folgerung, dass es vernünftig ist, graphembasierte Erkenner zu bauen. Den Fragenkatalog für polyphonemische Entscheidungsbäume kann man ohne linguistische Wissen erstellen. Wie es sich herausgestellt hat, zeigt die graphembasierte Modellierung ähnlich gute Resultate, wie die phonembasierte, wenn die Sprache eine enge Graphem-zu-Phonem Beziehung hat. Weil das in Englisch nicht der Fall ist, sind die Ergebnisse des phonembasierten Spracherkenner besser. Außerdem wurde die Teilung der Trainingsdaten über die Sprachen hinweg untersucht, um einen graphembasierten sprachenunabhängigen Erkenner zu bauen. Die Ergebnisse der Experimente erlauben zu sagen, dass sprachenspezifische Grapheme sprachenunabhängige übertreffen.

2.2 Ballung

Die Anzahl der verschiedenen, kontextabhängigen Modelle kann schon bei relativ kleinen Kontextbreiten sehr groß sein. Die Modellzahl kann sogar Millionen erreichen. Um so viele Modelle zu trainieren, reichen die Trainingsdatenmengen im Allgemeinen nicht aus. Eine Zusammenfassung (Ballung) der kontextabhängigen Modelle in eine kleinere Zahl an Modellenklassen kann dieses Problem lösen. Es kann auch passieren, dass manche Modelle im Training gar nicht vorkommen. Sie können aber später in den Testdaten, die sich von den Trainingsdaten unterscheiden können, auftauchen. Ein Ballungsalgorithmus sollte für nicht-trainierte Phänomene eine Lösung anbieten. Das Ballungsverfahren unterstützt die Generalisierungsfähigkeit der kontextabhängigen Modellen und verringert so die Gefahr, dass wegen der spezifischen Kontexte, mit denen die Modelle trainiert werden, diese Modelle auf neue Daten nicht mehr gut passen.

Die Ballungsalgorithmen unterscheiden sich nach folgende Kriterien [Sch00]:

- welche Grundeinheiten zur Ballung verwendet werden,

- ob es ein agglomeratives (bottom-up) oder divisives (top-down) Verfahren ist,
- welches Distanzmaß benutzt wird (Likelihood-Distanz, Entropie-Distanz).

Kai-Fu Lee hat als erster ein agglomeratives Ballungsverfahren durchgeführt [K.F88]. Als Distanzmaß setzte er die Entropie-Distanz ein, als Grundeinheiten verwendete er Triphone. Der große Nachteil bei agglomerativen Verfahren ist die Vokabularabhängigkeit. Nur wenn die Kontexte im Training gesehen wurden, können diese Triphone einer geeigneten Klasse zugeordnet werden. Der andere Nachteil: die Zahl möglicher Ballungen wächst quadratisch mit der Zahl der Modelle. Divisive Verfahren haben auch Probleme, nämlich die Frage, nach welchen Kriterien die Modelle eines Knotens in zwei Unterknoten aufgeteilt werden können. [Rog03] Trotzdem wird es bevorzugt, weil der Einsatz eines divisiven Ballungsalgorithmus die schwerwiegenden Nachteile des agglomerativen Verfahrens behebt.

Bei divisiven Verfahren wird in jedem Knoten des Ballungsbaums entschieden, wie die Knoten aufgeteilt werden müssen. Es wird ein Fragenkatalog erstellt. Die Fragen können z. B. phonetisch motiviert sein und werden über den Kontext von Phonemen gestellt. Eine Antwort auf so eine Frage ist “ja” oder “nein”. Aus dem Fragenkatalog wird in jedem Ballungsschritt die Frage ausgewählt, bei deren Anwendung der größte Gewinn einer Auftrennung vorkommt. Normalerweise werden in der Spracherkennung mehrere Kontextentscheidungsbäume gebaut, zum Beispiel für jeden Unterzustand des Phonems und für jedes Mittelphonem eines Polyphons. Einerseits erleichtert das die Arbeit, weil bei Verwendung kleiner Bäume weniger Entscheidungen getroffen werden müssen. Anderseits gehen einige Freiheiten verloren, akustische Phänomene gemeinsam zu modellieren.

Die Werke [YW03] und [YS03] haben wesentlich diese Studienarbeit beeinflusst. Im Gegensatz zu zustandsgebundenen Entscheidungsbäumen lässt die neue Ballungsmethode die Parameterteilung quer durch die verschiedene Phoneme zu. Die effektive Parameterteilung im akustischen Modell erlaubt die akkurate Aussprachemodellierung. Flexible Ballungsverfahren wurden für die implizite Aussprachemodellierung eingesetzt.

Man hat je einen Entscheidungsbaum für alle Unterzustände (Beginn/Mitte/End) aller Phonemen konstruiert. Der Ballungsalgorithmus startet am Wurzelknoten mit allen Polyphonen. Mehrere Polyphone/Unterzustände können jeden Knoten gemeinsam benutzen. Fragen werden bezüglich der Identität des zentralen Phonems, der Nachbarphoneme und der Unterzustände gestellt. Die beste Frage wird nach dem Prinzip des größten Informationsgewinns in jedem Knoten gewählt, und dann wird der Baum geteilt. Wenn der Baum eine bestimmte Größe erreicht oder einen minimale Schwellwert für die Menge an Trainingsmaterial pro Modell durchkreuzt, wird der Prozess gestoppt.

In [HAH01] wurde gezeigt, dass verschiedene Phoneme gleichen Einfluß auf die Nachbarphoneme ausüben. Darum ist es erstrebenswert, die Parameterteilung zuzulassen, wenn die akustischer Realisation identisch ist. In dem Fall, wenn als

Grundeinheiten im HMM generalisierte Subpolyphone benutzt werden, liefert die unterzustandbasierte Ballung eines Phonems bessere Ergebnisse, als die zustandsbasierte (modellbasierte).

In [Hai02] wurde eine Methode zur konsistenten Reduktion der Anzahl von Aussprachevarianten zu einer Aussprache pro Wort vorgestellt. Die durchgeführten Experimente auf gelesener und gesprochener Sprache bestätigen, dass die Anwendung des Vokabulars mit nur einer Aussprache pro Wort vergleichbare oder bessere Ergebnisse liefert als normale Systeme mit vielen Aussprachevarianten. Die implizite Modellierung der Aussprachevariante ist gleich oder besser als die explizite.

Saraçlar et al. schlagen vor, bei der Aussprachemodellierung Gaussiens über Modellierungseinheiten zu teilen. In [SNK00] beschreiben sie den Einschluss von Aussprachevarianten ins akustischem Trainingsmodell. Es wurde ein neues Aussprachemodell präsentiert - state-level pronunciation model (SLPM). SLPM ermöglicht alternative Realisierungen der Wahrscheinlichkeitsfunktionsfläche eines Phonems und erlaubt den HMM Zuständen, eines Phonemmodells die Gaussiens mit alternativen Realisationen der Modelle zu teilen. Die Hauptidee der SLPM ist die Ausgangsdichte des HMM Zustands eines Phonems so zu ergänzen, dass es möglich ist alternative Wahrscheinlichkeitsfunktionsflächenrealisierungen zu modellieren.

Experimente zeigen, dass diese Methode besonders gut für akustische Modelle, die für die spontane Sprache trainiert wurde, passen. Die Anwendung der neuen impliziten Methode führt zu einer bedeutenden Reduzierung der WER auf dem Switchboard Korpus, was auch auf zwei unabhängige Testmengen nachgewiesen wurde.

Kapitel 3

Systemaufbau

Die in der Studienarbeit trainierten graphembasierten Erkenner haben die gleiche Datenbasis, das gleiche Sprachmodell und die gleiche Signalvorverarbeitung wie die phonembasierten Basiserkenner. Der Unterschied liegt in den Modellierungseinheiten, in der Prozedur der Erstellung des Aussprachewörterbuchs und der Fragenmengen für die kontextabhängige Modellierung.

Die Basiserkenner wurden von T. Schultz für das GlobalPhone Projekt [Sch00] gebaut und trainiert. Die Sammlung der multilingualen Datenbasis GlobalPhone stellt akustische Trainingsmaterial in 15 Sprachen zur Verfügung. Hörbare Effekte spontaner Sprache wie Häsitationen, Wortabbrüche und Stottern wurden in die Verschriftungen eingearbeitet. Die Basiseinheiten der Basisspracherkenner sind die einzelnen Phoneme einer Sprache. Das Basisinventar eines Erkenners muss nicht unbedingt mit dem phonologisch definierten Phoneminventar einer Sprache übereinstimmen. Die intern im System verwendete Bezeichnung eines Phonemmodells besteht aus einem an die IPA-Schreibweise angelehnten Kürzel und einem zwei-Buchstaben-Etikett, das die Sprachenzugehörigkeit des Phonems darstellt.

3.1 Datenbasis

3.1.1 Trainings- und Testdaten

Die Sprecher wurden disjunkt im Verhältnis 80:10:10 in allen Sprachen auf drei Mengen verteilt: Trainings-, Kreuzvalidierungs- und Evaluationsmenge. Kreuzvalidierungsdaten werden auch Entwicklungsdaten genannt (engl. development data) und statt dem Begriff Evaluierungsdaten wird manchmal der Begriff Testdaten verwendet.

Große Aufmerksamkeit wurde bei dieser Aufteilung der Homogenität der Geschlechts- und Altersverteilung (und, sofern vorhanden, der Dialekte) in den einzelnen Sprecheruntermengen gewidmet. In der Trainingsmenge wurden diejenigen Sprecher gelegt, deren Textvorlage mehrmals benutzt worden waren. Die

Sprachdaten der Trainingsmenge wurden für das Training der akustische Modelle verwendet und das Textmaterial der Trainingsdaten wurde zur Sprachmodellierung mitverwendet. Freie Parameter des Spracherkenner (zum Beispiel die Sprachmodellgewichtung) wurden mittels der Kreuzvalidierungsmenge eingestellt. Zur Feststellung der Wortfehlerraten wurde die Evaluierungsmenge angewendet.

3.1.2 Textauswahl und Datenaufnahme

Die Daten für Englisch wurden nicht im Rahmen von GlobalPhone gesammelt, sondern vom Wall Street Journal-Korpus übernommen. Tabelle 3.1 und 3.2 [St"03] geben ein Überblick über die gesammten Datenbasis. 103 Sprecher haben Daten gespendet.

Die nachträgliche Verschriftung des geprochene Textes ist der aufwendigste und teuerste Teil einer Sammlung von Sprachdaten. Die Faktoren Kosten und Zeit spielten eine wesentliche Rolle beim Entwurf des GlobalPhone-Korpus. Um diese Faktoren auf niedrigen Niveau zu halten, wurden abgelesene Sprachdaten gesammelt. Die im Internet verfügbaren, überregionale Tageszeitungen der jeweiligen Länder wurden als Textquellen ausgewählt. Die Themengebiete umfassen Wirtschaftsberichte, internationales und nationales Tagesgeschehen. Die Domäne entspricht so den Erfordernissen eines Spracherkennes für große Wortschätze.

Für die deutschen Erkenner wurden die Textdaten von den Zeitungen „Süddeutsche Zeitung“ (http://www.sueddeutsche.de) und „Frankfurter Allgemeine Zeitung“ (http://www.faz.de) genommen. Die Daten wurden von 77 Sprechern abgelesen. Für einen Überblick über die Datenbasis siehe Tabelle 3.1 und 3.2.

	#Äußerungen (Stunden)		
Sprache	Training	Kreuzvalidierung	Evaluation
EN	7,137 (15.0)	144 (0.4)	152 (0.4)
DE	9,259 (16.9)	199 (0.4)	250 (0.4)

Tabelle 3.1: Statistik über die Äußerungen des GlobalPhone-Korpus

Sprache	#Wörter
EN	9,461
DE	24,000

Tabelle 3.2: Größe des GlobalPhone Wörterbuchs

Mit einem tragbaren DAT-Rekorder SonyTDC-8 und einem Nahsprachmikrofon der Firma Sennheiser HD-440-6 wurden alle Sprachdaten aufgenommen und auf DAT-Bänder mit einer Abtastfrequenz von 48 kHz digital aufgezeichnet. Dann wurden sie mit Hilfe einer Hardware-Karte der Firma MICROWAVE optisch auf einen PC übertragen. Mit der Software WAVE der Firma Turtle Beach wurden die Daten an diesem PC mit 16-bit Auflösung abgespeichert. Danach wurden die Aufnahmen auf eine Abtasrate von 16 kHz transformiert.

3.2 Monolinguale Erkenner

3.2.1 Vorverarbeitung

Nachdem die Abtastrate von 48 kHz auf 16 kHz reduziert worden war, wurde der Mittelwert aller Abtastwerte von den Abtastwerten der Aufnahme substrahiert. Das dient zum Ausgleich eines eventuellen Offset des A/D-Wandlers. Dann wurde eine Kurzzeitanalyse durchgefürt. Dafür wurden aus dem Signal jeweils Zeitsegmente von 16 ms mit Hilfe des Hamming-Fensters ausgeblendet, was 256 Abtastwerten entspricht. Nach Annahme, soll das Signal über diesen Segmenten stationär bleiben. Das Hamming-Fenster wurde mit einem Versatz von 10 ms über das Signal geschoben und so überlappen sich benachbarte Segmente um jeweils 6 ms. Die Anwendung einer diskrete Fouriertransformation auf den je 256 Abtastwerten dient zur Berechnung der 129 Spektralkoeffizienten.

Diese Koeffizienten wurden später mit einer Mel-skalierten Filterbank auf 30 Dimensionen reduziert. *Mel* ist die Abkürzung für *Melody* und ist dadurch motiviert, die Frequenzskala so in Abschnitte zu unterteilen, wie sie ein (ungeübter) Mensch einteilen würde. Dabei werden mehrere Frequenzbänder so zusammengefaßt, dass die Frequenzauflösung in hohen Frequenzbereichen stärker und in niedrigen Frequenzbereichen weniger stark reduziert wird.

Die gewonnenen 30 Mel-skalierte Koeffizienten wurden logarithmiert und anschließend mit Hilfe einer diskrete Cosinus-Transformation in 30 Cepstral-Koeffizienten transformiert. Nur die ersten 13 Koeffizienten werden für die weitere Verarbeitung verwendet. Um die Cepstren mittelwertzubereinigen, wurde eine cepstrale Mittelwertsubstraktion durchgeführt. Die Vektoren, jeweils aus 13 Koeffizienten, werden alle 10 ms berechnet. Weil diese Vektoren stationäre Momentaufnahmen des Sprachsignals sind, wurden sie durch dynamische Merkmale ergänzt, so wie die Approximation der ersten und zweiten Ableitungen der 13 Cepstren. Im Anschluß wurden die resultierenden 43 Merkmale zu einem Vektor zusammengefaßt. Für die Reduktion der Dimension auf 32 wurde eine LDA-Transformation (Lineare Diskriminanzanalyse) durchgeführt.

3.2.2 Training

Zum Training und zur Evaluation beider Erkenner wurde der JANUS-Spracherkenner verwendet. JANUS (oder Janus) ist ein Projekt, das das Ziel hat, einen allgemein-nutzbares Spracherkennungstoolkit für Forschung und Anwendungen zur Verfügung zu stellen. Die Software besteht aus dem JRTk (Janus Recognition Toolkit), das zur Entwicklung von Spracherkennungssystemen benutzt wird und dem Ibis Dekoder. JRTk wurde von Universität Karlsruhe (TH)(Deutschland) und Carnegie Mellon University (USA) entwickelt. JRTk ist in C implementiert und in eine Tcl/Tk-Umgebung eingebettet [JAN02]. Für die monolingualen Erkenner wurde das JRTk-Spracherkennungstoolkit V 5.0 Patch-Level 13 verwendet [JAN03].

Um ein Training mit JANUS zu starten, braucht man eine Datenbasis, bestehend aus digitalisierten Aufnahmen und einer möglichst genauen Transkription des Gesprochenen. Die Datenbasis ist in eine Trainings-, Kreuzvalidierungs- und Evaluationsmenge aufgeteilt. Man braucht auch ein Aussprachewörterbuch, das alle im Training verwendeten und beim Evaluieren zu erkennenden Wörter beinhaltet.

Die allgemeine Trainingsschema sieht wie folgt aus:

1. Erzeugung der verschiedenen Beschreibungsdateien.
   Oft muss man nur schon existierende Dateien für die eigene Ziele anpassen, z.B. die Datei desc.tcl. Diese Datei sagt den Spracherkennern, was für Beschreibungsdateien verwendet werden, wo sie sich befinden und andere notwendige Informationen. Außerdem werden Dateien, die die Architektur des Erkenners beschreiben und andere zusätzliche Information beinhalten, generiert. Es ist auch möglich solche Dateien von anderen, bereits trainierten Systemen zu übernehmen.

2. Erzeugung der Labels.
   Labels geben uns die Information über die zeitliche Zuordnung von Sprachmerkmalsvektoren zu HMM-Zuständen. Es wird ein existierende Spracherkenner genommen und ein Viterbi- oder ein Forward-Backward-Algorithmus auf den Trainingsaufnahmen gerechnet. Die von existierenden Spracherkenner gewonnene Labels haben geringere Qualität, als von Menschen erzeugte. Die berechnete Zeitzuordnungen werden in sogennanten Labels-Dateien abgespeichert, um die zeitaufwendige Labelserstellung nicht im jedem Entwicklungsschritt wiederholen zu müssen. Außerdem, können diese Labels-Daten zur Initialisierung eines neuen Erkenners benutzt werden.

3. Training eines kontextunabhängigen Systems.
   In diesem Fall verwendet der Erkenner nur kontextunabhängige akustische Modelle. Als erste wird eine LDA Matrix berechnet. Als nächstes wurden Beispielvektoren aus der Trainingsdaten extrahiert. Auf diesen Beispielvek-

toren wurde der k-means Algorithmus angewendet und so für jedes Subphonem ein Codebuch und eine Mixturgewichtsverteilung initialisiert. Das Trainiern der akustischer Parameter geschieht in mehrere Trainingsiterationen. In jeder Trainingsiteration werden dem System alle Trainingsdaten präsentiert und für die Optimierung der HMM-Parameter der EM-Algorithmus angewendet. Vier Iterationen haben für das Training dieser monolinguale Erkenner gereicht.

4. Einführung der Subpolyphone und Ballung eines kontextunabhängigen Systems.
   Für diesen Schritt brauchen wir kontinuierliche, kontextunabhängige Systeme. Als Klasse sind in Initiale Systeme Subphoneme benutzt worden. Es wird die Kontextmodellierung durch Subpolyphone vorbereitet. Erst werden für jedes Subpolyphon, das in Trainingsdatenbasis repräsentiert ist, eine eigene Mixturgewichverteilung definiert. Dann wird mit einem EM-Training ein semikontinuierliches System trainiert. In dem resultierenden System teilen sich Subpolyphone desselben Subphonems ein Codebuch. Nach dem Training werden die vielen kontextabhängigen Mixturgewichte mit Hilfe der Ballungsverfahren zu einer definierten Anzahl Klassen zusammengeballt. Die Klassen werden in dem entstehenden Kontextentscheidungsbaum angeordnet.

5. Bau und Training eines kontextabhängigen Systems.
   Das Training erfolgt so, wie beim kontextunabhängigen System. Man erhält ein voll kontinuierliches System, das auf Subpolyphonen als Klassen statt auf Subphonemen aufgebaut ist.

3.2.3 Test

Wenn der Erkenner trainiert ist, ist es wichtig zu wissen, wie gut er funktioniert. Neben dem akustischen Modell wird zusätzlich ein Sprachmodell benötigt und es muss ein Suchvokabular festgelegt werden.

Zur Bewertung eines Erkenners wird eine gesprochene Äußerung dekodiert und die resultierende Wortkette (Hypothese) abgespeichert. Um die Güte der Hypothesen zu bewerten, ist ein Fehlermaß erforderlich. Bei der Erkennung geht es darum, aus der unendlichen Anzahl möglicher Satzhypothesen die richtige rauszusuchen. Man spricht in diesem Zusammenhang auch von Suche. Für die Erkennung wird der One-Stage-Dymanic-Time-Warping-Algorithmus eingesetzt. Bei diesem wird eine Sequenz beliebig koppelbarer Wortmodelle aufgebaut. Innerhalb der einzelnen Wortmodelle ist die Berechnung des besten Pfades nach dem Viterbi-Algorithmus realisiert. Jedem Wort aus dem Erkennungsvokabular wird in diesem Fall eine entsprechende HMM-Zustandsfolge zugeordnet. Vom letzten Zustand jedes Wortes kann in den ersten Zustand jedes Wortes gesprungen werden. Die Wahrscheinlichkeit eines solchen Sprungs in ein neues Wort wird

vom Sprachmodell bestimmt. Somit sind die Sequenzen von Einzelwörten sowohl durch den akustisch wahrscheinlichsten Pfad von einem zum anderen Wort als auch durch den vorausgegangenen textuellen Kontext gesteuert. Die aussichtslosen partiellen Suchpfaden werden nicht bis zu Ende berechnet, sondern frühzeitig abgeschnitten. Man spricht von Pruning.

Die besten Suchpfade werden rückverfolgt und daraus entsteht eine Sequenz von Wörtern - Hypothesen. Wenn die N-besten Suchpfade berechnet werden, wird eine N-besten Liste von Hypothesen vom Erkenner ausgegeben. Diese N-besten Listen können in Form eines Worthypothesengraphen (WHG) repräsentiert werden.

Bei der Dekodierung werden die Emissionswahrscheinlichkeiten des HMM mit den Wortübergangswahrscheinlichkeiten des Sprachmodells kombiniert. Dabei weichen die Mittelwerte und Varianzen der Wahrscheinlichkeiten des akustisches Modells und des Sprachmodells stark ab. Es muss ein Korrektur eingesetzt werden. Die Parameter der Sprachmodels lassen sich manuell durch die Vorgabe von *lz*,*lp*-Paaren feinstellen. Mittels Parameter *lz* (Gewicht des Sprachmodels) wird das Sprachmodell relativ zum akustischen Modell gewichtet. Die Wortübergangsstrafe *lp* hilft die unterschiedlichen Länge betrachteter Wortfolgen W zu normieren, die an keine anderen Stelle betrachtet würde. Die Einstellung der *lz*,*lp*-Paare erfolgt mit Hilfe der Kreuzvalidierungsmenge.

$$P(W|X) = \frac{p(X|W) \cdot P(W)}{p(X)} \xrightarrow{lz,lp} \frac{p(X|W) \cdot P(W)^{lz} \cdot lp^{|W|}}{p(X)} \tag{3.1}$$

Die Dekodierung für die monolingualen Erkenner wird mit Hilfe des Ibis Dekoder durchgefürt.

Der Ibis Dekoder ist Ein-Suchlauf Dekoder. Es ist möglich, den Dekoder für Lattice-Rescoring zu benutzen. Die Hypothesen, die für verschiedene Werte von *lz* und *lp* die beste Bewertung (score) haben, werden aus dem Graphen ausgelesen. Man kann viele *lz*,*lp*-Paare ausprobieren. Dabei kann der Einfluß des Sprachmodells innerhalb des Graphen nachträglich verändert werden, was zu eine neue wahrscheinlichsten Hypothese führen kann. Die Wahrscheinlichkeiten innerhalb des Graphen müssen neue berechnet werden, deswegen bezeichnet man den Vorgang als Rescoring. Der Graph präsentiert aber nur einen Unterbereich des gesamten Suchraumes. Somit kann es passieren, dass die Werte für *lz* und *lp*, bei denen aus diesem Graph ausgewählten Hypothesen die beste WA haben, nicht unbedingt mit den Werten, die sich gut für die eigentliche Erkennung (baseLz und baseLp) eignen, übereinstimmen.

Tests werden normalerweise auf einer Kreuzvalidierungsmenge durchgeführt, um festzustellen, dass sich die Erkennungsrate verbessert und keine Probleme aufgetreten sind.

Kapitel 4

Experimente

4.1 Phonembasierte Basiserkenner

Die Tabelle 4.1 fasst die Ergebnisse der Experimenten für phonembasierte Basiserkenner zusammen.

Sprache	Kreuzvalidierungsdaten
DE	85,56
EN	90,3

Tabelle 4.1: Wortakkuratheit in % für phonembasierte Basiserkenner

4.2 Graphembasierte Basiserkenner

Im Laufe der Studienarbeit wurden graphembasierte Basiserkenner unter Verwendung der bisherigen Entscheidungsbäume für Deutsch und Englisch trainiert. Die Evaluation ist genau so wie bei den phonembasierten Basiserkennern erfolgt (Unterkapitel 3.2.2).

In jedem Fall wurden zuerst kontextunabhängige Erkenner trainiert. Jeder Trainingslauf bestand aus 4-6 Iterationen. Die Initiallabels und die Dateien, die die Architektur des Erkenners beschreiben und andere zusätzliche Information beinhalten, sind von [Kil03] übernommen worden. Als Modellierungseinheiten wurden Subgrapheme eingesetzt.

Nach dem ersten Trainingsschritt wurden neue Labels geschrieben und das System erneuert trainiert.

Nachdem der kontextunabhängige Erkenner fertig war, wurde auf seiner Basis ein kontextabhängiges System gebaut.

Die Kontextmodellierung erfolgte aufgrund der Untereinheiten der Polygrapheme, die in der Datenbasis zu beobachten sind. In dieser Studienarbeit habe ich mit einer Kontextbreite von 1, also mit Trigraphemen, gearbeitet.

Nachdem kontextabhängige Elementarmodelle trainiert wurden, können sie geballt werden. Das Geräuschmodell und das Stillemodell werden nicht kontextabhängig modelliert und nicht in den Ballungsprozeß einbezogen. Der Ballungsalgorithmus sollte nach 3000 Modellen terminieren.

Bevor der Ballungsprozess gestartet hat, haben alle Subpolyphone desselben Subgraphems sich das gleiche Codebuch geteilt. Die Kontexte waren mit verschidene Mixturgewichtverteilungen modelliert. So ein System ist semikontinuierlich. Bei dem Ballungsverfahren wurden Subpolyphone unter Verwendung der Entropie-Kriterien und des Fragenkatalogs zu einer definierte Anzahl Klassen zusammengefasst. Für den Entscheidungsbaum habe ich den Singleton Fragenkatalog benutzt. Die kontextabhängigen Systeme wurde geballt und ein Kontextentscheidungsbaum entstand. Es wurde ein neuer kontextabhängiger Spracherkenner erzeugt, in dem für jede Klasse ein eigenes Codebuch und eine eigene Mixturgewichverteilung verwendet wird. Diese System ist ein voll kontinuierlicher HMM-Erkenner.

Der erste Trainingslauf für das kontextabhängige System fand mit den Labels der kontextunabhängige Systeme statt. Dann wurden neue Labels geschrieben und das System noch mal trainiert.

Die Ergebnisse des Tests sind in Tabelle 4.2 zu sehen:

Sprache	Kreuzvalidierungsdaten
DE	85,28
EN	81,88

Tabelle 4.2: Wortakkuratheit in % für graphembasierte Basiserkenner

4.3 Parameteranpassung

Im Rahmen der Studienarbeit habe ich mehrere Möglichkeiten zum Erreichen einer gute Erkennungsleistung ausprobiert.

Es wurden die Anzahl der Teilbäume und die Anzahl der Gaussiens für die semikontinuierlichen Systeme für die Ballung variiert.

Im folgenden werden die Ergebnisse der Forschung bei Anwendung des flexiblen Ballungsverfahrens detalliert vorgestellt.

4.3.1 Anzahl der Gaussiens

In der vorliegende Arbeit habe ich die verschiedene Anzahl der Gaussiens für die Codebücher des semikontinuierlichen kontextabhängigen Systems zur Ballung ausprobiert und die Resultate verglichen. In früheren Experimenten wurde die Anzahl der Gaussiens im Basiserkenner auf 32 gesetzt. Für das Geräuschmodell und das Stillemodell wurde die Anzahl der Gaussiens (32) nicht geändert.

Tabelle 4.3 gibt Auskunft über Verfahren:

	Deutsch Kreuzvalidierungsdaten	
# Gaussiens für semikont. Erkenner	Baum ohne Unterteilung und Singleton	Baum mit Unterteilung und ergänzte Singleton
256	85,15	85,31
1024	85,35	85,69
1500	-	85,78
2048		85,51

Tabelle 4.3: Wortakkuratheit in % für vollkontinuierlichen graphembasierte Spracherkenner in Deutsch bei Anwendung des flexiblen Ballungsverfahren

Die hier dargestellte Ergebnisse wurden mit Hilfe von Singleton und ergänzte Singleton Fragenmenge gewonnen. Es wurde ein Entscheidungsbaum mit Unterteilung auf Vokale und Konsonanten und ein Entscheidungsbaum ohne diese Unterteilung bei den Experimenten benutzt.

Nachdem festgestellt wurde, dass die Verwendung des Baums mit Unterteilung und ergänzten Singleton Fragenkatalog die besten Erkennungsrate bringt, wurden die Experimente mit 1500 und 2048 Gaussiens nur mit obengenannten Baum und Fragenmenge weitergeführt.

4.3.2 Anzahl der Teilbäume

Die Experimente wurden für zwei Typen von Entscheidungsbäume durchgeführt. In einem Fall wurde ein Baum mit zwei verschiedenen Teilbäumen mit der Unterscheidung zwischen Vokalen und Konsonanten untersucht, im anderen Fall ein Baum ohne diese Aufteilung. In jedem Fall sind drei Baumwurzelknoten dargestellt, da jedes Graphem mit drei HMM-Zuständen modelliert wurde, die als Anfangs-, Mittel-, und Endzustand bezeichnet wurden.

Obwohl es prinzipiell möglich wäre, verschiedenartige Zustände, so wie Anfangs-,

und Endzustände, zusammenzuballen, habe ich auf die Möglichkeit wegen des großen Rechenaufwands verzichtet. Dafür habe ich aber gleichartige Zustände verschiedener Grapheme zusammengeballt, was auch viel Rechenzeit gekostet hat. So wurde ein Teilbaum für jeden Unterzustand der Vokale und Konsonanten gebaut (insgesamt 6 Teilbäume) und 3 Teilbäume sind ohne die genannte Aufteilung entstanden.

Abbildung 4.1 präsentiert eine schematische Darstellung des Entscheidungsbaums mit Unterscheidung zwischen Vokalen und Konsonanten für den Mittelzustand der Grapheme.

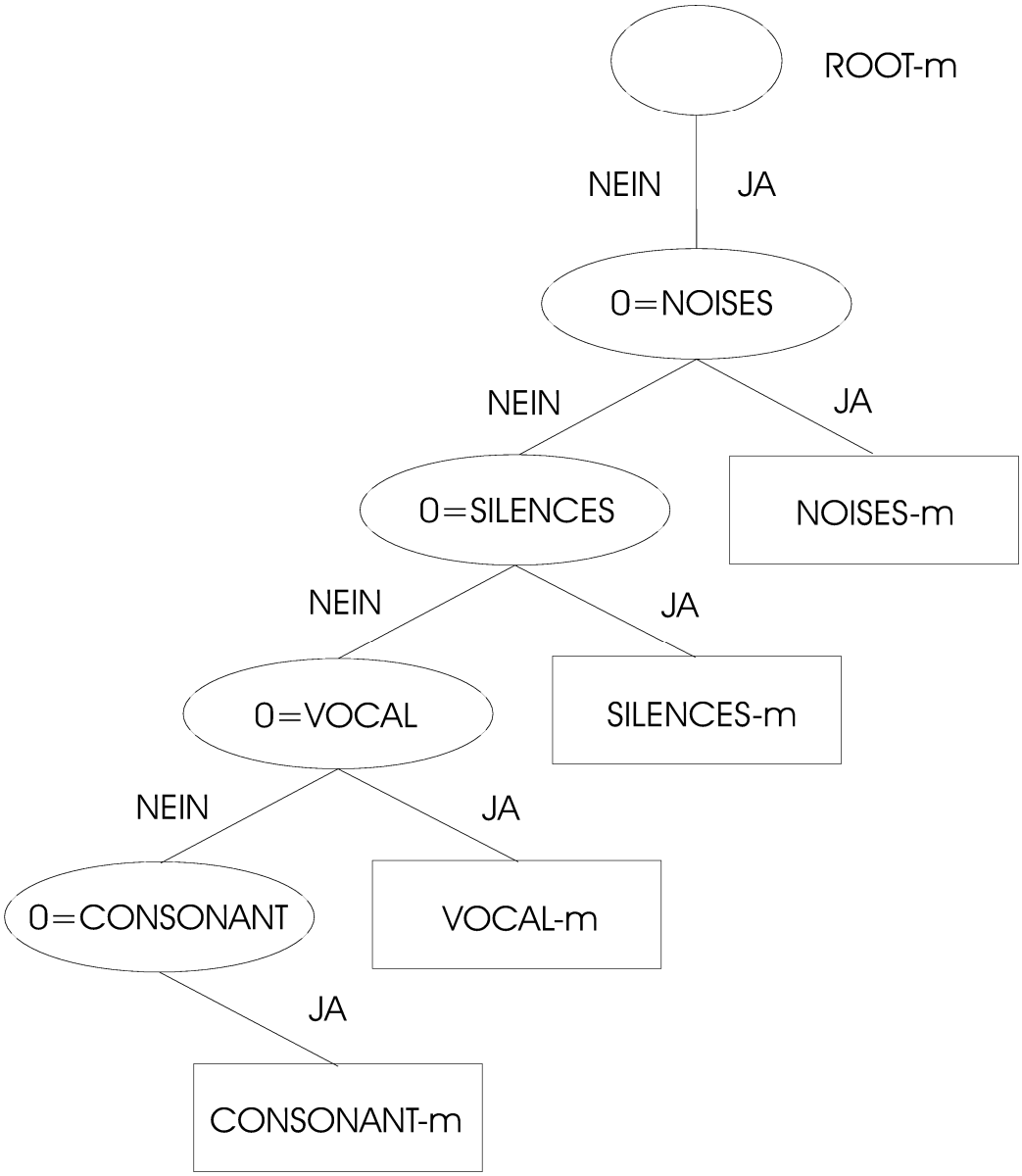


Abbildung 4.1: Darstellung eines Entscheidungsbaums vor der Ballung

Das Geräuschmodell und das Stillemodell wurden als einzelne Knoten betrachtet.

Der Vorgang startet in jedem Wurzelknote des Entscheidungsbaums. In jedem Knoten wird die Antwort auf die Frage des Baums verlangt: zu welcher Klasse gehört das zentrale Subgraphem? Zum Beispiel, ob das Subgraphem ein Geräusch ist (0=NOISES)? Wenn mit NEIN geantwortet wird und der Knoten nicht der letzte ist, kommt der Kontext in den NEIN-Zweig und es wird die nächste Frage gestellt. Wenn die Antwort JA ist, wird zur Modellierung das akustische Modell derjenigen Ballungsknotens verwendet, der sich im Blatt des JA-Zweiges befindet.

Alle gleichartigen Zustände verschiedener Grapheme, die zu einem Knoten verallgemeinert wurden, teilen als gemeinsame Parameter die Gaussmixtur.

4.3.3 Fragenkatalog

Die Menge aller mögliche Polygrapheme für die Standardsprache in dem kontextabhängigen Spracherkenner ist normalerweise sehr groß. Um die Menge von Daten zu reduzieren, wurden die Polygrapheme in Gruppen zusammengefasst.

Die Reduktion der Anzahl der freie Parameter erlaubt es, die übrig gebliebene Parameter robuster zu schätzen. Es ist sehr wichtig, dass die Kontextklassen sich maximal voneinander unterscheiden.

Die Fragenmenge wird für die Ballungsprozedur kreiert. Als Distanzmaß ist hier die Entropie-Distanz genommen worden, die für die Optimierung der Information in Parameterraum des Erkenners benutzt wird.

Eine Elementarfrage ist die Frage nach der Zugehörigkeit eines Graphems zu einer bestimmten Graphemmenge.

Im Laufe der Studienarbeit wurden drei Fragenkatalogen untersucht:

1. Singleton.
2. Ergänzte Singleton.
3. Phonem-Graphem Fragen.

Die Idee von Singleton und Phonem-Graphem Fragen habe ich von [Kil03] genommen, die schon in Unterkapitel 2.1 erklärt wurde. Hier möchte ich nur in Tabelle 4.4 und 4.6 zeigen, wie diese Fragenmengen in meinem Fall aussehen.

Ergänzte Singleton hat zuzüglich zum Singleton Graphemmengen noch zwei weitere: VOCAL und CONSONANT. Tabelle 4.5 präsentiert den ergänzten Singleton Fragenkatalog.

In Englisch wurden zwei Geräuschmodell in die Experimente einbezogen, nämlich M_+QK_EN und M_+hGH_EN, in Deutsch nur eins : M_+QK_DE.

Kontextfrage	Liste der Grapheme
PHONES	A_DE B_DE C_DE D_DE E_DE F_DE G_DE H_DE I_DE J_DE K_DE L_DE M_DE N_DE O_DE P_DE Q_DE R_DE S_DE T_DE U_DE V_DE W_DE X_DE Y_DE Z_DE ∼A_DE ∼O_DE ∼U_DE @ SIL M_+QK_DE
SILENCE	SIL
VOCAL	A_DE E_DE I_DE U_DE O_DE ∼A_DE ∼O_DE ∼U_DE
CONSONANT	B_DE C_DE D_DE F_DE G_DE H_DE J_DE K_DE L_DE M_DE N_DE P_DE Q_DE R_DE S_DE T_DE V_DE W_DE X_DE Y_DE Z_DE
NOISES	M_+QK_DE
AFFRIKATE	A_DE
ALLE_DIPHTONGE	B_DE
BEHAUCHT	C_DE
DENTAL_ALV	D_DE
⋮	

Tabelle 4.4: Beispiel der Phonem-Graphem Fragenmenge für Deutsch

Kontextfrage	Liste der Grapheme
PHONES	A_DE B_DE C_DE D_DE E_DE F_DE G_DE H_DE I_DE J_DE K_DE L_DE M_DE N_DE O_DE P_DE Q_DE R_DE S_DE T_DE U_DE V_DE W_DE X_DE Y_DE Z_DE ∼A_DE ∼O_DE ∼U_DE @ SIL M_+QK_DE
VOCAL	A_DE E_DE I_DE U_DE O_DE ∼A_DE ∼O_DE ∼U_DE
CONSONANT	B_DE C_DE D_DE F_DE G_DE H_DE J_DE K_DE L_DE M_DE N_DE P_DE Q_DE R_DE S_DE T_DE V_DE W_DE X_DE Y_DE Z_DE
NOISES	M_+QK_DE
SILENCES	SIL
A_DE	A_DE
B_DE	B_DE
C_DE	C_DE
⋮	
∼U_DE	∼U_DE

Tabelle 4.5: Beispiel der ergänzte Singleton Fragenmenge für Deutsch

Kontextfrage	Liste der Grapheme
PHONES	A_DE B_DE C_DE D_DE E_DE F_DE G_DE H_DE I_DE J_DE K_DE L_DE M_DE N_DE O_DE P_DE Q_DE R_DE S_DE T_DE U_DE V_DE W_DE X_DE Y_DE Z_DE ∼A_DE ∼O_DE ∼U_DE @ SIL M_+QK_DE
NOISES	M_+QK_DE
SILENCES	SIL
A_DE	A_DE
B_DE	B_DE
C_DE	C_DE
⋮	
∼U_DE	∼U_DE

Tabelle 4.6: Beispiel der Singleton Fragenmenge für Deutsch

Die Ergebnisse der Experimenten mit der Singleton und der ergänzten Singleton Fragenmenge wurden schon in Unterkapitel 4.3.1 gezeigt. In Tabelle 4.7 stelle ich die Resultate für den Phonem-Graphem Fragenkatalog dar.

Fragenmenge	Deutsch Kreuzvalidierungsdaten
Phonem-Graphem	85,79

Tabelle 4.7: Wortakkuratheit in % für graphembasierte Erkenner in Deutsch bei Anwendung 1500 Gaussiens und Entscheidungsbaum mit Aufteilung auf Vokale und Konsonanten

4.4 Deutsche Evaluation

Um die Leistungsbewertungen zu vergleichen, wurden die Experimente auf der Evaluationsmenge durchgeführt. Die Ergebnisse für den Spracherkenner in Deutsch kann man in Tabelle 4.8 beobachten.

Herangehensweise	Deutsch Evaluationsmenge
Phonembasierte Basiserkenner	84,4
Graphembasierte Basiserkenner	86,0
Graphembasierte Erkenner mit ergänzter Singleton Fragenmenge, dem Baum mit Unterteilung auf Vokale und Konsonanten, 1500 Gaussiens	87,3
Graphembasierte Erkenner mit Phonem-Graphem Fragenmenge, dem Baum mit Unterteilung auf Vokale und Konsonanten, 1500 Gaussiens	87,3

Tabelle 4.8: Wortakkuratheit in % für den Spracherkenner in Deutsch für Experimente auf der Evaluationsmenge

4.5 Transfer auf Englisch

Nachdem die Experimente bei der Anwendung des flexiblen Ballungsverfahren für verschiedene Anzahl der Gaussiens und Teilbäume für die Deutsche Sprache durchgeführt wurden, habe ich mit den Parametern, die für die beste Wortakkuratheit gesorgt haben, die Experimente für Englisch wiederholt.

Tabelle 4.9 zeigt die Resultate für den Transfer.

Fragenmenge	Englisch Kreuzvalidierungsdaten
Ergänzte Singleton	83,19
Phonem-Graphem	82,44

Tabelle 4.9: Wortakkuratheit in % für die Englische Sprache bei Anwendung 1500 Gaussiens und des Entscheidungsbaums mit Aufteilung auf Vokale und Konsonanten

Für Englisch wurden ähnlich wie für Deutsch die Experimente auf der Evaluationsmenge gemacht. Die Ergebnisse sind in Tabelle 4.10 dargestellt.

Herangehensweise	Englisch Evaluationsmenge
Phonembasierte Basiserkenner	88,5
Graphembasierte Basiserkenner	80,8
Graphembasierte Erkenner mit ergänzte Singleton Fragenmenge, dem Baum mit Unterteilung auf Vokale und Konsonanten, 1500 Gaussiens	81,4
Graphembasierte Erkenner mit Phonem-Graphem Fragenmenge, dem Baum mit Unterteilung auf Vokale und Konsonanten, 1500 Gaussiens	81,6

Tabelle 4.10: Wortakkuratheit in % für den Spracherkenner in Englisch für Experimente auf Evaluationsmenge

Kapitel 5

Analyse der Ergebnisse

In diesem Kapitel werden die Ergebnisse der Experimente bewertet. Die Effizienz der Anwendung des flexiblen Ballungsverfahren auf graphembasierte Erkennungsystem wird analysiert.

Die Vergrößerung der Anzahl der Gaussiens für die Codebücher des semikontinuierlichen, kontextabhängigen Systems zur Ballung hat zur Steigerung der Wortakkuratheit deutlich beigetragen. 1500 Gaussiens scheint die optimale Anzahl in diesem Fall zu sein. Eine weitere Vergrößerung der Anzahl der Gaussiens konnte die Erkennungsrate nicht mehr verbessern.

Auch die Idee der Anwendung des ergänzten Singleton Fragenkatalogs für die Ballungsprozedur hat sich gelohnt. So eine Fragenmenge ist leicht zu erstellen und man braucht dafür keine linguistischen Kenntnisse. Die Verwendung der Phonem-Graphem Fragenmenge bringt vergleichbare Ergenbnisse, ist aber nicht ohne linguistisches Wissen zu erzeugen.

Bei der Anpassung der Anzahl der Teilbäume hat sich herausgestellt, dass der Baum mit der Unterscheidung zwischen Vokalen und Konsonanten einen Leistungsgewinn gegenüber dem Baum ohne Unterteilung bietet.

Durch die Verwendung des flexiblen Ballungsverfahrens wurde eine deutliche Reduktion der Wortfehlerrate erreicht. Die Wortfehlerrate für den graphembasierte Erkenner in Deutsch ist bei den Experimenten auf der Evaluationsmenge um 9.3 % relativ gesunken im Vergleich zum graphembasierte Basiserkenner. Dabei spielt es keine Rolle, ob die ergänzte Singleton Fragenmenge oder die Phonem-Graphem Fragenmenge eingesetzt wird. Die Anzahl der Gaussiens ist 1500, der Entscheidungsbaum unterliegt der Aufteilung auf Vokale und Konsonanten.

Für den graphembasierten Erkenner in Englisch unterscheiden sich die Wortakkuratheiten bei den Experimenten auf der Evaluationsmenge ein wenig für den Fall, ob die ergänzte Singleton Fragenmenge oder der Phonem-Graphem Fragenkatalog verwendet wird. Die sonstigen Parameter sind die gleichen wie für den deutschen Erkenner. Wenn die Singleton Fragenmenge benutzt wird, sinkt die Wortfehlerrate relativ um 3.1 %, bei der Phonem-Graphem um 4.1 % im Vergleich zum graphembasierten Basiserkenner.

Der Ursprung des Unterschiedes in den Ergebnisse zwischen dem graphembasierte Erkenner in Deutsch und Englisch liegt in dem Unterschied in der Graphem-zu-Phonem-Relation der Sprachen. In Deutsch sind Grapheme und Phoneme besser als in Englisch aufeinander abgestimmt.

Bei der Analyse der Ergebnisse kommt man auf die Untersuchung des Kontextetscheidungsbaums nach dem flexiblen Ballungsverfahren. Am Beispiel des Erkenners in Deutsch werden hier exemplarisch einige Parameter erörtert.

Insgesamt sind in dem Kontextentscheidungsbaum 4760 Knoten. 168 von denen sind Knoten mit der Frage nach dem Mittelgraphem. Die Länge des längsten Pfades im Baum ist 31. Die durchschnittliche Tiefe des Knotens mit einer Frage nach dem Zentergraphem - 13,6. Die durchschnittliche Tiefe aller Konten - 14,7. Somit steht die Frage nach dem Mittelgraphem ziemlich nah zur durchschnittlichen Tiefe aller Knoten. Man kann vermuten, dass die Kontextmodellierung von der Frage nach dem Zentergraphem nicht viel mehr abhängig ist als von anderen Fragen, wenn es um Trigrapheme geht.

Polygrapheme mit unterschiedlichem Mittelgraphem teilen sich das gleiche Modell. Beim alten Ballungsverfahren wäre das nicht möglich gewesen. Die Anzahl der geteilten Modellen liegt bei 68.

Tabelle 5.1 präsentiert Beispiele für die geteilten Modellen und die Mengen der Mittelgrapheme, bei denen es möglich ist, gleiche Modelle zu teilen.

Modelle	Menge der Mittelgrapheme
VOCAL(|)-e (465)	O_DE ∼O_DE U_DE
VOCAL(|)-e (691)	∼O_DE U_DE
CONSONANT(|)-b (1150)	F_DE H_DE J_DE Q_DE X_DE Y_DE
CONSONANT(|)-b (1522)	Q_DE V_DE X_DE Y_DE
CONSONANT(|)-m (837)	J_DE P_DE Q_DE V_DE W_DE X_DE Y_DE
CONSONANT(|)-m (1276)	C_DE F_DE Q_DE X_DE Y_DE
CONSONANT(|)-e (681)	B_DE G_DE J_DE M_DE Q_DE V_DE W_DE X_DE
CONSONANT(|)-e (417)	C_DE P_DE Q_DE X_DE Y_DE

Tabelle 5.1: Die Mengen der Mittelgrapheme und die geteilten Modelle

Wie man aus der Tabelle 5.1 sieht, teilen sich Polygrapheme mit verschiedenem Mittelgraphem das gleiche Modell nur für die Endzustände der Klasse VOCAL. Bei der Klasse CONSONANT ist das nicht der Fall. Die Modelle werden für jeden Unterzustand geteilt.

Die Tatsache, dass einige Modelle von mehreren Polygraphemen mit unter-

schiedlichem Mittelgraphem mitbenutzt werden können, führt zur Bemerkung: die Polygrapheme können gleichen Einfluß auf die unterschiedliche Mittelgrapheme ausüben. Allerdings, bei der Klasse VOCAL, üben die Polygrapheme nur auf den Endzustand der Mittelgraphem einen gleichen Einfluß aus. Somit wurde gezeigt, dass es für Grapheme, genau so wie für Phoneme rational ist, die Parameterteilung zuzulassen, wenn die akustische Realisation identisch ist.

Kapitel 6

Fazit und Ausblick

Die Idee, das flexible Ballungsverfahren auf graphembasierte Spracherkennungssysteme anzuwenden, hat sich als erfolgreich erwiesen.

Frühere Versuche haben gezeigt, dass für bestimmte Sprachen Grapheme als Modellierungseinheiten ähnlich gut geeignet sind wie Phoneme. Dabei ist die Erstellung des Aussprachewörterbuchs mit Graphemen viel einfacher als mit Phonemen. Aber, weil bei der Verwendung von graphembasierten Wörterbüchern keine Aussprachevarianten erhältlich sind, müssen diese impliziert modelliert werden. Für die implizite Aussprachemodelierung wurde das flexible Ballungsverfahren eingesetzt.

Die Anwendung von Graphemen reduziert die Entwicklungskosten. Das ist eine gute Lösung für diejenigen Entwickler, die sich hohe Ausgaben nicht leisten können. Allerdings ist die Wortakkuratheit von phonembasierten Spracherkennern für manche Sprachen noch größer als von graphembasierten. Aber die Forschungen auf dem Gebiet der Verwendung von Graphemen für Spracherkennungssysteme haben erst eine relativ kurze Geschichte. Deswegen kann man hoffen, dass die Erkennungsrate für graphembasierte Erkenner in Zukunft stetig wachsen wird.

Die Verwendung des Baums mit der Unterteilung auf Vokale und Konsonanten bringt bessere Ergebnisse als der Baum ohne so eine Unterteilung. Das kann daran liegen, dass das Entropie-Kriterium für die Ballung nicht gut genug ist. Man könnte probieren, ein anderes Kriterium einzusetzen. Vielleicht wäre es mit dem Likelihood-Distanzmaß möglich gewesen, auf die Unterscheidung zwischen Vokalen und Konsonanten bei dem Baum zu verzichten. Ein anderer Grund, wieso der Baum mit der Unterteilung für bessere Ergebnisse gesorgt hat, kann sein, dass die Distributionsgewichte nicht genug trainiert waren; oder, dass das flexible Ballungsverfahren ein Greedy-Algorithmus ist.

Eine Möglichkeit, die Erkennungsrate der phonembasierten und graphembasierten Basiserkenner zu verbessern, könnte die Vergrößerung der Anzahl der Gaussiens für die Codebücher des semikontinuierlichen kontextabhängigen Systems zur Ballung sein. Allerdings, soll die Anzahl der Gaussiens nur ganz langsam

vergrößert werden, damit die Anzahl der Parameter nicht zu stark zunimmt.

Durch die Verwendung des flexiblen Ballungsverfahrens wurde die Wortfehlerrate der graphembasierten Erkenner in Deutsch um bis zu 9.3% und in Englisch um bis zu 4.1 % relativ gesenkt und so gezeigt, dass für die Sprachen Englisch und Deutsch Grapheme geeignete Einheiten zur impliziten Aussprachemodellierung sind. In Deutsch und Englisch sind aber Grapheme und Phoneme nicht perfekt aufeinander abgestimmt. Es wäre deswegen interessant das flexible Ballungsverfahren auf graphembasierte Erkenner für Türkisch oder Kroatisch anzuwenden, das heißt für die Sprachen, bei denen die Buchstaben den Lauten sehr genau entsprechen.

Literaturverzeichnis

[Bod76] F. Bodmer. *Die Sprachen der Welt.* Köln, Berlin: Kiepenheuer und Witsch, 5 edition, 1976.

[C+98] Bernard Comrie et al., editors. *Bildatlas der Sprachen.* Augsburg: Bechtermünz Verlag, 1998.

[Cry93] D. Crystal. *Die Cambridge Enzyklopädie der Sprache.* Frankfurt am Main/New York : Campus Verlag, 1993.

[D+84] Günther Drosdowski et al., editors. *Duden: Grammatik der deutschen Gegenwartssprache*, volume 4. Dudenverlag, 4 edition, 1984.

[Gal85] Peter Gallmann. *Graphische Elemente der geschriebene Sprache.* Tübingen: Niemeyer, 1985.

[HAH01] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing.* Prentice Hall, 2001.

[Hai02] T. Hain. Implicit Pronunciation Modelling in ASR. In *Pronunciation Modelling and Lexicon Adaption for Spoken Language Technology (PMLA)*, pages 129–134, 2002.

[JAN02] Online JRTk dokumentation. http://isl.ira.uka.de/∼jrtk/janus-doku.html, 2002.

[JAN03] JRTk und JANUS. The Ibis-Gang. http://isl.ira.uka.de/∼jrtk/doc.Janus5/janus-doku/janus-doku.html, August 18, 2003.

[K.F88] K.F.Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* Dissertation, Carnegie Mellon University, 1988.

[Kil03] Mirjam Killer. Graphem Based Speech Recognition. Master's thesis, Carnegie Mellon University,USA and Eidgenssische Technische Hochschule Zürich, Schweiz, 2003.

[KN02] S. Kanthak and H. Ney. Context-Dependent Acoustic Modeling Using Graphemes For Large Vocabulary Speech Recognition. In *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848, Orlando, FL, May 2002.

[KN03] S. Kanthak and H. Ney. Multilingual Acoustic Modeling Using Graphems. In *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1145–1148, Geneva, Switzerland, September 2003.

[KSS03] M. Killer, S. Stüker, and T. Schultz. Grapheme Based Speech Recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-2003)*, Geneva, Switzerland, September 2003.

[Rog98] Ivica Rogina. *Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular.* Dissertation, Universität Karlsruhe (TH), Shaker Verlag, 1998.

[Rog03] Ivica Rogina. Sprachliche Mensch-Maschine-Kommunikation. Skriptum zum Vorlesung "Sprachliche Mensch-Maschine-Kommunikation", Universität Karlsruhe (TH), 2003.

[Sch00] Tanja Schultz. *Multilinguale Spracherkennung - Kombination akustischer Modelle zur Portierung auf neue Sprachen.* Dissertation, Universität Karlsruhe (TH), Shaker Verlag, 2000.

[SFK00] C. Schillo, G. A. Fink, and F. Kummert. Grapheme Based Speech Recognition For Large Vocabularies. In *International Conference on Spoken Language Processing*, volume 4, pages 584–587, Beijing, China, 2000.

[SNK00] M. Saraçlar, H.J. Nock, and S. Khudanpur. Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models. In *Computer Speech and Language*, volume 14, pages 137–160, 2000.

[St"03] Sebastian Stüker. Multilingual Articulatory Features. Master's thesis, Carnegie Mellon University,USA und Universität Karlsruhe (TH),Germany, 2003.

[YS03] H. Yu and T. Schultz. Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-2003)*, Geneva, Switzerland, September 2003.

[YW03] H. Yu and A. Waibel. Flexible Parameter Tying for Conversationl Speech Recognition. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003.