Universität Karlsruhe (TH)
Fakultät für Informatik
Lehrstuhl Prof. Dr.rer.nat. A. Waibel
Institut für Anthropomatik

Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen

Studienarbeit
von

Sebastian Ochs

Dezember 2008

Betreuer: Dipl.-Ing. Matthias Wölfel
Dipl.-Inform. Sebastian Stüker

Sebastian Ochs
Meisenweg 17
76275 Ettlingen

Hiermit erkläre ich an Eides statt, dass ich die von mir vorgelegte Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ettlingen, den 18. Juli 2009

(Unterschrift)

_____

Sebastian Ochs

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einführung

1.1 Motivation, Zielsetzung und Beitrag der Arbeit

Bereits seit einiger Zeit werden an vielen Hochschulen Lehrveranstaltungen in Ton und Bild digital aufgezeichnet. Diese Aufzeichnungen werden gewöhnlich den Studierenden als Dienstleistung zugänglich gemacht, z.B. über das Internet, und können als Ergänzung der persönlichen Vorlesungsmitschrift dienen. Darüber hinaus bietet es sich an, solche Aufzeichnungen auch als Material für die Forschung zu nutzen sowie als Basis für weitere Anwendungen. Allerdings sind Tonaufnahmen relativ unflexibel wenn es darum geht, den Inhalt auf sprachlicher Ebene weiter zu verarbeiten. Es wäre deshalb wünschenswert, eine Transkription, d.h. eine textuelle Repräsentation der in der Aufnahme gesprochenen Sprache, zu haben. Das Vorhandensein solcher Transkriptionen würde die Erschließung vielfältiger, zusätzlicher Dienstleistungen ermöglichen, beispielsweise automatische Untertitelung, Volltextsuche, thematische Indizierung, automatische Zusammenfassung und Übersetzung in andere Sprachen. Insbesondere schwerhörige oder ausländische Studierende könnten von solchen Anwendungen profitieren. Da das manuelle Erstellen von Transkriptionen aber sehr aufwändig ist, möchte man zu diesem Zweck automatische Spracherkennungssysteme einsetzen.

Die automatische Transkription von Vorlesungen stellt jedoch eine Herausforderung für die Spracherkennung dar, denn die Sprache in Vorlesungen ist sehr variabel was Sprechstil und Wortschatz betrifft [Cettolo 04]. Es handelt sich um kontinuierliche Sprache mit großem Vokabular, die sowohl spontane Äußerungen als auch formelle Fachsprache umfasst. Im Hinblick auf den Wortschatz treten insbesondere in technischen Vorlesungen auch fremdsprachliche Wörter auf, z.B. aus dem Englischen. Das Erkennen von Wörtern aus anderen Sprachen ist allerdings ein grundlegendes Problem in der automatischen Spracherkennung, denn die Systeme sind in der Regel auf einzelne Sprachen spezialisiert.

An der Universität Karlsruhe (TH) wird ein Lecture Translation System enwickelt, das Vorlesungen nicht nur in Echtzeit transkribieren sondern auch simultan übersetzen kann [Fügen 06]. In vorangehenden Versuchen mit der Spracherkennungskomponente

dieses Systems wurde festgestellt, dass auf deutsch gehaltene, technische Vorlesungen englische Wörter enthalten und dass diese Wörter überwiegend falsch erkannt werden. Deren korrekte Erkennung ist aber von großer Wichtigkeit, da es sich oft um bedeutungstragende Wörter wie etwa Eigennamen oder Fachbegriffe handelt.

Das Ziel dieser Arbeit bestand darin, den Einfluss der englischen Wörter auf die Gesamterkennungsleistung zu untersuchen und verschiedene Methoden zu entwickeln und anzuwenden, um die Erkennung englischer Wörter zu verbessern. Mit allen Ansätzen, die in dieser Arbeit vorgestellten werden, gelang es die englische Wortfehlerrate auf ungefähr die Hälfte zu reduzieren, wenn auch auf Kosten der deutschen Wortfehlerrate. Insgesamt konnte die Erkennungsleistung um 1,1% absolut verbessert werden.

1.2 Aufbau eines Spracherkennungssystems

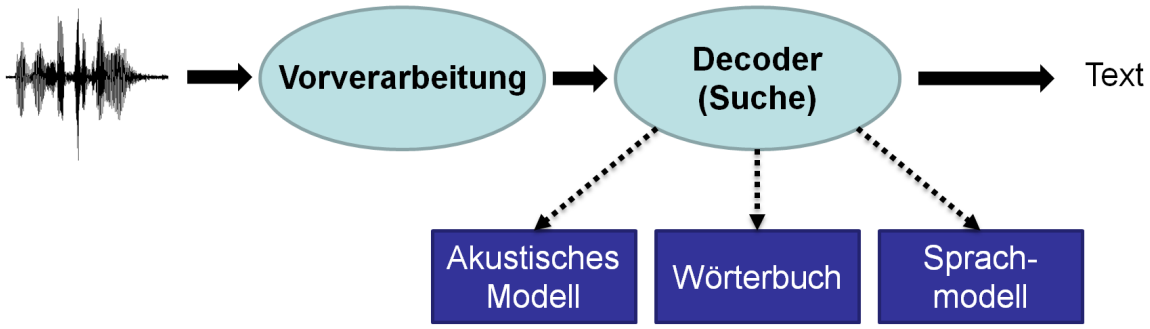

Abbildung 1.1: Hauptkomponenten eines Spracherkennungssystems.

Abbildung 1.1 zeigt den schematischen Aufbau eines Spracherkennungssystems. Am Anfang steht das Audiosignal, aus ihm werden durch eine Vorverarbeitung Merkmalsvektoren gewonnen, die auf kompakte Weise die relevanten Informationen des Audiosignals für einen Zeitabschnitt enthalten. Die Folge von Merkmalsvektoren wird anschließend vom *Decoder* benutzt, um diejenigen Wörter zu suchen, die mit größter Wahrscheinlichkeit dem Audiosignal entsprechen. Der Decoder greift dafür auf drei Informationsquellen zurück, das *Akustische Modell*, das *Aussprachewörterbuch* und das *Sprachmodell*. Das Akustische Modell stellt einen Zusammenhang zwischen Merkmalsvektoren und Aussprachen her, indem es durch Wahrscheinlichkeitsverteilungen beschreibt, wie gut eine Folge von Merkmalsvektoren zu einer bestimmten Aussprache passt. Die Aussprachen, repräsentiert durch phonetische Einheiten, sind dabei durch das Aussprachewörterbuch, auch Lexikon genannt, Wörtern zugeordnet. Das Sprachmodell schließlich beschreibt, wie wahrscheinlich eine bestimmte Folge von Wörtern in der Sprache ist.

Die in dieser Arbeit vorgestellten Methoden greifen in das Wörterbuch und das Akustische Modell ein, die anderen Komponenten des Spracherkenners werden nicht verän-

dert. Die erwähnten phonetischen Einheiten, die im Wörterbuch und im Akustischen Modell verwendet werden, sind in unserem Fall Phoneme, aber auch andere Einheiten wie z.B. Silben sind möglich. Hier ist anzumerken, dass in dieser Arbeit die Begriffe *Phonem* und *Phon* als gleichbedeutend angesehen werden, denn eine strenge Unterscheidung wie sie in der Linguistik üblich ist, ist von unserem technischen Standpunkt aus nicht nötig. Die verwendeten deutschen und englischen Phonembezeichner sind in Anhang A und Anhang B aufgelistet.

Kapitel 2

Frühere Arbeiten

In diesem Kapitel werden frühere Arbeiten vorgestellt, die in Bezug zu dieser Arbeit stehen.

2.1 Lecture Translator

In unserer zunehmend globalisierten Welt wird es immer wichtiger, über Kultur- und Sprachgrenzen hinweg kommunizieren zu können, um Handel zu treiben, Dienstleistungen zu erbringen oder Wissen zu vermitteln. In vielen solcher Situationen sind menschliche Übersetzer zu teuer oder nicht verfügbar. Auch Englisch als Lingua Franca hat seine Grenzen, da es nicht von jedermann in ausreichendem Maße beherrscht wird. Die Technik kann hier möglicherweise eine revolutionäre Lösung bieten. Aus diesen Gedanken heraus, angetrieben von den Fortschritten der Forschung auf dem Gebiet der Sprachtechnologien, wird an der Universität Karlsruhe der Lecture Translator entwickelt [Fügen 06]. Dabei handelt es sich um ein echtzeitfähiges, domänenunbeschränktes Sprach-zu-Sprach-Übersetzungssystem, d.h. ein System, das Ansprachen, Vorträge und Vorlesungen diverser Themengebiete simultan übersetzen kann. Das System läuft auf handelsüblicher PC-Hardware. Die Präsentation der Übersetzung kann je nach Anwendungsfall auf verschiedene Arten erfolgen, so wurden neben Kopfhörern und Untertiteln auch spezielle Richtlautsprecher und Brillen mit Head-Up-Display getestet. Abbildung 2.1 zeigt den typischen Aufbau eines Sprach-zu-Sprach-Übersetzungssystem. Der Spracherkenner (ASR) und der Übersetzer (SMT) sind dabei immer vorhanden. Je nachdem, ob die Übersetzung nur als Text oder auch als Sprache ausgegeben werden soll, ist zusätzlich eine Sprachsynthese-Einheit (TTS) präsent. Ein solches System ist allerdings mehr als die bloße Aneinanderreihung dieser Teile, es erfordert zusätzliche Komponenten wie z.B. einen Segmentierer, der die Ausgabe des Erkenners in für den Übersetzer geeignete Abschnitte zerlegt. Außerdem ist eine engere Bindung zwischen Erkenner und Übersetzer denkbar, weil der Übersetzer von zusätzlichen Informationen des Erkenners profitieren kann. Da beim Lecture Translator alle Komponenten in einem Client-Server-Framework verbunden sind, ist es außerdem möglich, mehrere Übersetzer

gleichzeitig mit den Daten desselben Erkenners zu speisen, man kann also eine Sprache simultan in N andere Sprachen übersetzen lassen. Zunächst wurde jedoch nur das Sprachpaar Englisch nach Spanisch entwickelt und getestet. In einer laufenden Arbeit wird das Sprachpaar Deutsch nach Englisch hinzugefügt. Dabei ergaben sich einige zusätzliche Schwierigkeiten, die es bei Englisch als Eingabesprache nicht gab. Eine davon ist das Auftreten von englischen Wörtern in deutschen Vorlesungen. Diese englischen Wörter führen zu vermehrten Fehlern, nicht nur im Spracherkenner, sondern auch im Übersetzer, da es eine starke Korrelation zwischen Qualität der Spracherkennung und Qualität der Übersetzung gibt. Hier setzt diese Arbeit an, mit der Absicht, der deutschen Spracherkennungskomponente des Lecture Translators englische Wörter besser verständlich zu machen.

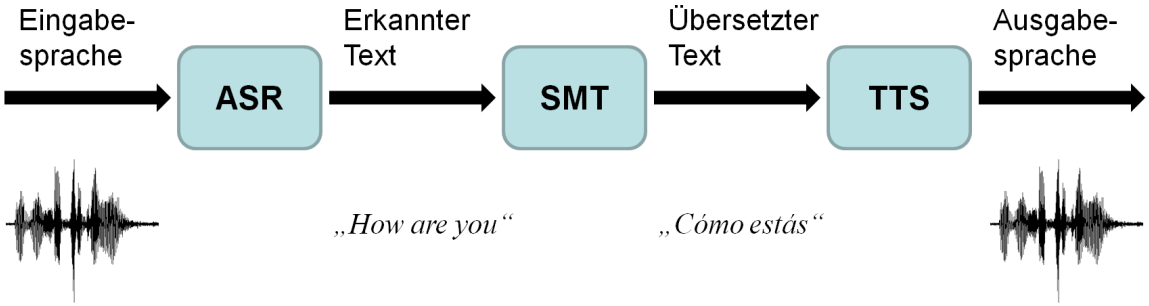

Abbildung 2.1: Hauptkomponenten eines Sprach-zu-Sprach-Übersetzungssystems.

Abbildung 2.2: Der Lecture Translator in Aktion.

2.2 Kombination Akustischer Modelle

Das spontane, gelegentliche Wechseln eines Sprechers von einer Sprache L1 (z.B. Deutsch) in eine andere Sprache L2 (z.B. Englisch) wird als Code-Switching bezeichnet und ist ein bekanntes Problem in der der multilingualen Sprachverarbeitung [Schultz 06]. Wenn dabei für die Aussprache von L2 Phoneme verwendet werden, die L1 fremd sind, so stellt sich die Frage, wie man ein Spracherkennungssystem anpassen muss, um die Akustischen Modelle von L1 und L2 zu kombinieren, so dass beide Sprachen erkannt werden können. Dafür werden in dieser Arbeit drei Methoden untersucht.

Methode 1: man benutzt weiterhin ein monolinguales System das nur L1-Phoneme kennt. Die Phoneme von L2 werden durch Phoneme aus L1 ersetzt, d.h. gesucht ist eine Funktion, die L2-Phoneme auf möglichst ähnliche L1-Phoneme abbildet.

Methode 2: man erzeugt ein bilinguales System, das die kompletten akustischen Modelle von L1 und L2 parallel betreibt.

Methode 3: man vereint die akustischen Modelle von L1 und L2 zu einem gemeinsamen Modell, in dem es sprachunabhängige (d.h. sprachübergreifende) Phoneme und sprachabhängige Phoneme gibt.

Methode 1 ist aufgrund ihrer Einfachheit weit in der Literatur verbreitet. Nach der Art und Weise auf die die Abbildungsfunktion erstellt wird, unterscheidet man wissensbasierte Verfahren, datengetriebene Verfahren oder Kombinationen der beiden. Datengetriebene Verfahren definieren gewöhnlich ein Distanz- oder Ähnlichkeitsmaß für Phoneme, damit die Abbildungsfunktion präzise und automatisch errechnet werden kann. In [Andersen 94] wurde ein datengetriebenes Verfahren vorgestellt, dessen Ähnlichkeitsmaß auf einer Phonem-Konfusionsmatrix basiert. [Ma 98] evaluierten sowohl wissensbasierte und datengetriebene Verfahren als auch eine Mischform, um chinesische Sprache von einem englischen Erkenner verarbeiten zu lassen. Die Mischform schnitt im Test am besten ab, dicht gefolgt vom wissensbasierten System. Als Ähnlichkeitsmaß diente ebenfalls eine Konfusionsmatrix. [Köhler 96] dagegen definierte eine Funktion zur Berechnung der Distanz zwischen zwei Hidden-Markov-Modellen (HMM) um die Ähnlichkeit von Phonemen zu bestimmen. Die meisten Arbeiten auf diesem Gebiet beschränken sich auf kontextunabhängige Modelle, obwohl Spracherkenner im Allgemeinen kontextabhängige Modelle verwenden, da diese bessere Ergebnisse liefern, weil sie die beim Sprechen auftretenden Koartikulationseffekte berücksichtigen. Ein Ähnlichkeitsmaß für kontextabhängige Modelle wurde von [Imperl 99] vorgeschlagen und in einem Clusteringprozess zur Bestimmung multilingualer Triphone verwendet. [Le 06] erweiterte diese Idee auf Cluster von Polyphonen eines Entscheidungsbaums, um ein Spracherkennungssystem schnell auf eine neue Sprache (Vietnamesisch) zu portieren. Hierbei wurden ebenfalls wissensbasierte und datengetriebene Abbildungen verglichen, mit dem Ergebnis, dass das wissensbasierte System leicht geringere Fehlerraten produzierte. In Kapitel 4.1 dieser Arbeit werden wissensbasierte, datengetriebene und kombinierte Abbildungen für kontextunabhängige Phoneme behandelt, wobei drei verschiedene Distanzmaße zum Einsatz kommen.

Methode 2 ist der typische Ausgangspunkt für bilinguale Systeme, birgt jedoch einige Nachteile und ist daher in dieser Form nicht weit verbreitet. So werden für diesen Ansatz zwei vollwertige, monolinguale akustische Modelle benötigt, es werden keine Gemeinsamkeiten der beiden Sprachen ausgenutzt. Somit ist das resultierende akustische Modell doppelt so groß und würde durch das Hinzufügen neuer Sprachen immer weiter wachsen. Dies widerspricht der Annahme, dass sich alle Sprachen ein begrenztes Phoneminventar, wie etwa das *Internationale Phonetische Alphabet* [IPA 05], teilen und macht den Ausbau eines solchen Systems zu einem multilingualen System sehr problematisch. Da sich diese Arbeit jedoch auf Deutsch und Englisch beschränkt, testen wir diese Methode dennoch, wie in Kapitel 4.2.1 beschrieben.

Methode 3 löst die angesprochenen Probleme der Methode 2. Sie basiert auf einer Idee von Dalsgaard und Andersen [Dalsgaard 92], die ihm Rahmen der Sprachidentifizierung die Bezeichnungen *Polyphonem* und *Monophonem* eingeführt haben. Polyphoneme sind hierbei Phoneme, die in zwei oder mehr Sprachen ähnlich genug sind, um als identisch angesehen zu werden, während Monophoneme keine ausreichend ähnlichen Phoneme in anderen Sprachen haben. Somit ist dieser Ansatz eine Art Mittelweg zwischen den Methoden 1 und 2. Statt alle englischen Phoneme durch deutsche zu ersetzen oder alle deutschen und englischen Phoneme beizubehalten, werden gleichartige Phoneme über Sprachen hinweg verschmolzen und nur ungleichartige beibehalten. In früheren Experimenten zeigte diese Methode gute Ergebnisse [Schultz 06; White 08]. In Kapitel 4.2.2 wird unsere Version eines Systems mit Mono- und Polyphonemen vorgestellt.

Kapitel 3

Wörterbuch und Sprachidentifizierung

In diesem Kapitel wird zunächst die Rolle des Aussprachewörterbuchs erläutert. Danach wird eine einfache Methode vorgestellt, mit der die Sprachzugehörigkeit eines geschriebenen Wortes festgestellt werden kann. Diese Methode wird dazu benutzt, alle potentiell englischen Wörter im Aussprachewörterbuch des Spracherkenners zu identifizieren. Außerdem ermöglicht sie es später, bei Testläufen des Spracherkenners die Verteilung der Sprachen in Referenz und Hypothese zu analysieren, sowie die Wortfehlerrate nach Sprache aufzuschlüsseln.

3.1 Rolle des Aussprachewörterbuchs

Das Aussprachewörterbuch spielt eine zentrale Rolle beim Spracherkennungsprozess, es ordnet Wörtern jeweils eine oder mehrere Aussprachen zu. Beim Training legt es somit fest, welche akustischen Merkmale den Phonemmodellen zugewiesen werden und bei Erkennungsläufen bestimmt es zusammen mit dem Sprachmodell diejenigen Wörter, die mögliche Kandidaten für die Hypothesen des Spracherkenners sind. Ein System, das wie in Abbildung 1.1 aufgebaut ist, hat daher gewöhnlich die Einschränkung, dass es nur die Wörter erkennen kann, welche im Aussprachwörterbuch aufgeführt sind. Um sogenannte *Out-of-Vocabulary* (OOV) Fehler während des Erkennungsprozesses zu vermeiden, sollte das Wörterbuch deshalb möglichst alle für das Einsatzgebiet relevanten Wörter enthalten. Es sollte also einerseits nicht zu eingeschränkt sein, andererseits aber auch nicht zu umfassend, denn mit zunehmender Wortanzahl steigt der Suchaufwand für den Decoder und die Wahrscheinlichkeit der Verwechslung bei der Erkennung. Durch den Vergleich des Wörterbuchs mit Referenztexten kann festgestellt werden, ob das Wörterbuch einer Ergänzung bedarf. Falls ja, so werden die fehlenden Wörter meist automatisch hinzugefügt, da eine manuelle Bearbeitung zu aufwändig wäre. Die Aussprachen dieser neuen Wörter werden dabei entweder aus bereits existierenden Lexika

übernommen oder von regelbasierter Software erzeugt. Für letzteres verwenden wir beispielsweise die Sprachsynthesesoftware Festival [Black 97].

Für den Lecture Translator bedeutet das, dass sein Wörterbuch zunächst um diverse englische Begriffe und Namen, die in technischen Vorlesungen auftreten, erweitert wurde. Dies führte zwar zu einer Reduzierung der OOV-Rate, allerdings war in Testläufen zu beobachten, dass die englischen Wörter nicht gut erkannt wurden. Grund dafür war die schlechte Repräsentation der englischen Aussprachen im Wörterbuch, denn sie wurden automatisch nach deutschen Regeln erzeugt. Aber selbst bei manueller Erzeugung gibt es Probleme, da wegen der zum Teil sehr unterschiedlichen Phoneme gewisse Aussprachen nicht korrekt dargestellt werden können (z.B. kann der englische /th/-Laut nur mangelhaft durch deutsche Phoneme nachgebildet werden). Tabelle 3.1 zeigt einen kleinen Auszug aus dem Wörterbuch. Solche unpassenden Aussprachen gilt es zu vermeiden, denn sie führen zu zwei Arten von Fehlern. Erstens verschmutzen sie beim Training die Modelle, da die aus dem Audiosignal extrahierten Merkmale den falschen Phonemmodellen zugewiesen werden und zweitens werden die Wörter bei Erkennungsläufen nicht erkannt, da die modellierte Aussprache zu sehr von der tatsächlichen abweicht.

broadcast	{B R OH A T K A S T}
joke	{J OH K E2}
through	{T R UH K}
wireless	{V IE R EH L E S}

Tabelle 3.1: Beispiele für schlecht modellierte Aussprachen englischer Wörter

3.2 Identifizierung englischer & deutscher Wörter

Um die Repräsentation englischer Wörter in unserem deutschen Aussprachewörterbuch zu verbessern, müssen wir für sie korrekte Aussprachen generieren, die auf englischen Regeln basieren. Hierfür müssen die englischen Wörter zuerst identifiziert werden. Genaugenommen suchen wir nicht einfach Wörter, die allein der englischen Sprache zugehörig sind, sondern Wörter, deren Aussprache durch englische Regeln gebildet wird. Wörter wie beispielsweise „Handy“, „Service“ und „Beamer“ haben sich zwar im Deutschen etabliert, sie werden aber eher englisch als deutsch ausgesprochen. Schwieriger wird es bei Wörtern, die deutsche und englische Morpheme mischen, wie z.B. „tracken“ und „gedownloadet“. Idealerweise müssten sie in ihre Bestandteile zerlegt werden, um dann für jeden Teil die Ausspracheerzeugung separat vorzunehmen. Da solch eine Zerlegung allerdings ein nichttriviales Problem ist, ignorieren wir diese Wörter vorerst und entscheiden uns für eine einfachere Methode, das Nachschlagen in Lexika. Zu diesem Zweck verwenden wir die freie Software Hunspell [Németh 03] mit zwei passenden Lexika, einem deutschen[1] und einem englischen[2]. Hunspell ist eine Rechtschreibprüfung,

[1] de_DE-frami, Stand 2008-03-06

[2] en_US, Stand 2006-02-07

die z.B. von OpenOffice benutzt wird. Sie zeichnet sich dadurch aus, dass sie speziell für Sprachen mit komplexen Wortbildungsregeln entworfen wurde und daher auch Wörter in flektierter und komponierter Form erkennt. Dies ist insbesondere für Deutsch wichtig, mit seinen vielfältigen Fällen und beliebig langen Kettenwörtern.

Zur Bestimmung seiner Sprachzugehörigkeit lassen wir ein Wort von beiden Hunspell-Lexika prüfen. Dadurch ergeben sich vier Fälle. Wenn das Wort nur dem deutschen Lexikon bekannt ist, klassifizieren wir es als *deutsch*. Entsprechend klassifizieren wir es als *englisch*, wenn es nur dem englischen Lexikon bekannt ist. Für den Fall, dass ein Wort beiden Wörterbüchern oder keinem bekannt ist, wird es in die Klassen *ambivalent* bzw. *unbekannt* eingeteilt. Beispiele für ambivalente Wörter wären etwa „bald“, „kind“ und „man“, da sie je nach Kontext deutsch und englisch ausgesprochen werden können. Auch die bereits erwähnten deutschen Wörter mit englischer Aussprache wie „Handy“ usw. fallen darunter.

Diese Methode der Sprachidentifizierung ist natürlich nicht optimal, da sie auf Lexika und Affixregeln beruht, die ihrerseits nicht perfekt sind, aber sie ist unkompliziert und liefert gute Ergebnisse, wie in Kapitel 5 gezeigt wird.

3.3 Neuaufbau des Aussprachewörterbuchs

Wie in Abbildung 3.1 dargestellt, verwenden wir diese Klassifizierungsmethode mit Hunspell, um alle potentiell englischen Wörter (d.h. die Klassen *englisch* und *ambivalent*) in unserem deutschen Aussprachewörterbuch zu identifizieren. Wir benutzen Festival für die Erzeugung der neuen, englischen Aussprachen. Die bestehenden Aussprachen rein deutscher und unbekannter Wörter werden unverändert übernommen. Im Falle der rein englischen Wörter ersetzen wir die alten Aussprachen durch die neuen. Für ambivalente Wörter jedoch behalten wir die alten Aussprachen und fügen die neuen als zusätzliche Varianten hinzu, da wir ohne weitere Informationen nicht entscheiden können, ob diese Wörter deutsch oder englisch ausgesprochen werden. Hier können zukünftige Optimierungen ansetzen, um die Zahl der unnötig hinzugefügten Aussprachen zu reduzieren. So ist es beispielsweise unwahrscheinlich, dass die englische Aussprache des ambivalenten Wortes „nun“ (Nonne) für eine technische Vorlesung benötigt wird, ebensowenig wie die deutsche Aussprache von „Computer“ („Komputer“ statt „Kompjuter“) auftreten wird.

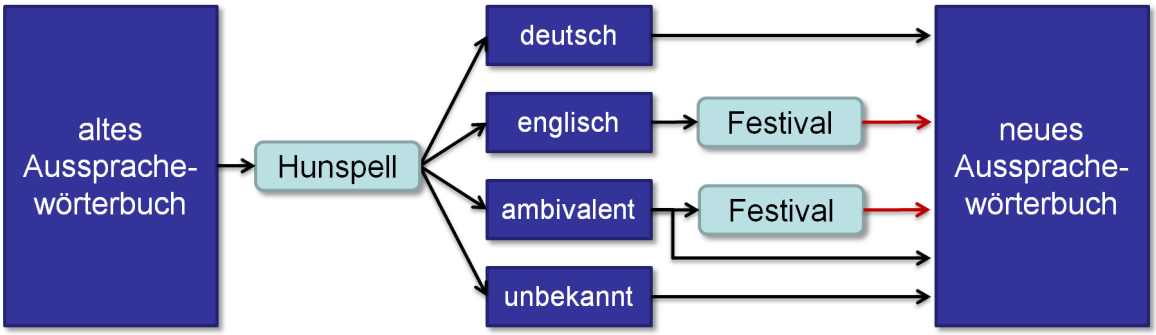

Abbildung 3.1: Zerlegung des Wörterbuchs durch Hunspell und Wiederaufbau mit neuen, englischen Aussprachen von Festival.

Kapitel 4

Modellierung englischer Wörter im deutschen System

Da wir wie in Kapitel 3.3 beschrieben neue Aussprachen, bestehend aus englischen Phonemen, die unserem deutschen Spracherkennungssystem unbekannt sind, in unser Wörterbuch eingeführt haben, müssen wir zusätzliche Anpassungen vornehmen. Entweder wir ersetzen alle englischen Phoneme durch deutsche, oder wir erweitern das akustische Modell des Spracherkenners. Im Folgenden werden beide Ansätze vorgestellt.

4.1 Abbildung von englischen Phonemen auf deutsche Phoneme

In diesem Kapitel werden mehrere Methoden beschrieben, mit denen man Abbildungen erhält, die Phoneme einer Sprache (Englisch) möglichst ähnlichen Phonemen einer anderen Sprache (Deutsch) zuordnen. Dadurch gelingt es, englische Wörter in das Aussprachewörterbuch aufzunehmen und mit dem akustischen Modell für deutsche Sprache abzudecken, ohne dass erneutes Training oder weitere Anpassungen des Spracherkennungssystems nötig wären. Wir verwenden dabei sowohl einen wissenbasierten als auch einen datengetriebene Ansatz sowie eine Kombination der beiden. Für das datengetriebene Verfahren werden drei Distanzmaße benutzt: die euklidische Distanz, die Mahalanobis-Distanz und die Kullback-Leibler-Distanz. In der Literatur wird dieser Ansatz oft benutzt, um Phoneme für den Bau von multilingualen Systemen zu clustern und um ein monolinguales System schnell von einer Sprach auf eine andere zu portieren [Schultz 06].

4.1.1 Wissensbasierter Ansatz

Der wissensbasierte Ansatz beruht auf der Annahme, dass die akustischen Ausprägungen von Phonemen über verschiedene Sprachen hinweg im Allgemeinen so ähnlich sind, dass Phoneme als sprachunabhängige Einheiten angesehen werden können. Dies kann durch die physikalischen Eigenschaften und Grenzen des menschlichen Sprachorgans begründet werden. Als universelles Phoneminventar hat sich das *Internationale Phonetische Alphabet* (IPA) der International Phonetic Association [IPA 05] bewährt. Um die verschiedenen Phonembezeichner unseres Spracherkennungssystems (siehe Anhänge A und B) auf einen Nenner zu bringen, ermitteln wir für sie zunächst die entsprechende IPA-Notation. Für einige Phonembezeichner gibt es dann eine direkte Übereinstimmung, in diesem Fall ist eine Abbildung trivial (z.B. englischer Bezeichner /AX/ = mittlerer Zentralvokal = deutscher Bezeichner /E2/). Falls es jedoch keine direkte Entsprechung für ein englisches Phonem im Deutschen gibt, so wählen wir ein verfügbares aus der Nachbarschaft (z.B. englisch /TH/ = stimmloser dentaler Frikativ, im Deutschen nicht vorhanden, also abbilden auf benachbarte Frikative /S/ oder /F/). Manche englische Phonembezeichner repräsentieren Kombinationen zweier Laute (z.B. Diphthonge und Affrikate), falls es für eine solche Kombination keinen passenden, einzelnen Bezeichner im Deutschen gibt, so behandeln wir die Laute getrennt und bilden dadurch einen englischen Bezeichner auf eine Folge von zwei deutschen ab (z.B. englisch /CH/ = stimmlose postalveolare Affrikate = deutsch /T/ + /SCH/). Tabelle 4.1 zeigt unsere vollständige Abbildungsfunktion.

Der Vorteil dieses wissensbasierten Vorgehens ist seine Einfachheit, neben dem linguistischen Wissen um die Phonembezeichner und das IPA werden keine weiteren Daten benötigt. Der Nachteil ist sowohl der manuelle Aufwand als auch die fehlende Objektivität der Methode, denn die tatsächlichen akustischen und statistischen Eigenschaften der Phonemmodelle des Spracherkennungssystems werden nicht berücksichtigt.

Englisch	AA	AE	AH	AO	AW	AX	AXR	AY	B	CH	D	DH
Deutsch	AH	AEH	A	O	AU	E2	ER	AI	B	T+SCH	D	Z
Englisch	EH	ER	EY	F	G	HH	IH	IX	IY	JH	K	L
Deutsch	E	OE	E+I	F	G	H	I	I	IE	D+SCH	K	L
Englisch	M	N	NG	OW	OY	P	R	S	SH	T	TH	UH
Deutsch	M	N	NG	O+U	EU	P	R	S	SCH	T	S	U
Englisch	UW	V	W	XL	XM	XN	Y	Z	ZH			
Deutsch	UH	V	U	E2+L	E2+M	E2+N	J	Z	SCH			

Tabelle 4.1: Wissensbasierte Abbildungsfunktion.

4.1.2 Datengetriebener Ansatz

Die Grundlage für einen rein datengetriebenen Ansatz zur Bestimmung von guten Abbildungsfunktionen ist ein Ähnlichkeitsmaß für die Phonemmodelle des Spracherken-

nungssystems. Ein solches Ähnlichkeitsmaß kann aus einer Phonem-Konfusionsmatrix konstruiert [Andersen 94; Ma 98] oder durch die Verwendungen von bekannten Distanzmaßen erhalten werden [Sooful 01; Wölfel 09]. Wir wählen letzteres und vergleichen drei verschiedene Distanzmaße: die euklidische Distanz, die (erweiterte) Mahalanobis-Distanz und die Kullback-Leibler-Distanz. Hierfür werden zunächst Merkmale aller deutschen und englischen Phoneme aus entsprechenden, sprecherspezfischen Trainingsdaten extrahiert und danach alle Modelle mit jeweils einer Gaußkurve trainiert. In früheren Experimenten haben wir festgestellt, dass es für ein Phonemtraining zum Zweck der Distanzberechnung keine nennenswerten Vorteile bringt, wenn Gaußmixturen statt nur einer Gaußkurve verwendet werden [Heger 07; Stüker 08]. Im nächsten Schritt berechnen wir für jedes deutsche Phonemmodell die Abstände zu allen englischen Phonemmodellen. Das Ergebnis ist eine Distanzmatrix, die Abbildungen 4.2, 4.3 und 4.4 stellen diese für jedes Distanzmaß grafisch dar. Die Abstandswerte wurden hier zum Zweck einer einheitlichen Darstellung auf den Wertebereich 0 bis 1 normalisiert, wobei 0 für die Identität und 1 für den maximal auftretenden Abstand des jeweiligen Maßes steht. In diesen Grafiken fällt auf, dass die Hauptdiagonale und ihre unmittelbare Umgebung einige niedrige Abstände enthält. Das liegt an der Tatsache, dass die Zeilen und Spalten der Matrizen alphabetisch nach Phonembezeichner sortiert sind und deutsche und englische Bezeichner mit gleichem oder ähnlichem Namen auch oft für denselben Laut stehen. Ebenfalls bemerkenswert ist der große Abstand, den die Reibelaute /CH/, /S/, /SH/ und /ZH/ zu den meisten anderen Phonemen aufweisen.

Um eine Abbildungsfunktion zu erhalten, wird schließlich für jedes englische Phonem das deutsche Phonem mit dem geringsten Abstand ausgewählt. Die Tabellen 4.3, 4.4 und 4.5 zeigen die ermittelten Funktionen. Bei einem direkten Vergleich in Tabelle 4.6 fällt auf, dass alle drei Distanzmaße weitestgehend dasselbe Ergebnis liefern. Sie differieren lediglich in 6 von 45 Phonemen, sind also zu ca. 87% identisch. Dies widerspricht den Experimenten von [Sooful 01], in denen die Mahalanobis-Distanz deutlich schlechter als die euklidische und Kullback-Leibler-Distanz abschnitt.

4.1.3 Mischansatz

Der datengetriebene Ansatz hat den Nachteil, dass er ein englisches Phonem nur durch ein einzelnes deutsches Phonem ersetzen kann. Der englische Phonemsatz enthält allerdings einige Kombinationslaute wie z.B. /CH/, die – wie im wissensbasierten Ansatz geschehen – besser auf eine Folge von zwei deutschen Phonemen abgebildet werden sollten. Außerdem gibt es eine Reihe von englischen Phonemen, die keine gute Entsprechung im Deutschen haben, sondern mehrere Alternativen ähnlichen Abstands, wie etwa das /TH/. Aus diesen Gründen erstellen wir eine gemischte Abbildungsfunktion, welche die wissensbasierte mit den datengetriebenen kombiniert und bei unklaren Fällen englische Phoneme durch mehrere alternative deutsche ersetzt. Im Gegensatz zu den vorigen Funktionen handelt es sich also nicht mehr um eine 1:1 sondern eine 1:n Abbildung. Jede dieser Alternativen führt zu einer eigenen Aussprachevariante, die dem Wörterbuch hinzugefügt wird. Dadurch wächst das Wörterbuch zwar beträcht-

lich, jedoch konnte diese Methode das beste Ergebnis in dieser Arbeit liefern. Tabelle 4.2 zeigt diese gemischte Abbildungsfunktion. Die deutschen Alternativen sind durch Komma getrennt, ihre Reihenfolge hat keine Bedeutung, Kombinationen sind mit „+“ gekennzeichnet.

Die Testergebnisse unserer Experimente mit allen Abbildungsfunktionen werden in Kapitel 5.2 präsentiert.

Englisch	Deutsch	Englisch	Deutsch	Englisch	Deutsch
AA	AH, A	F	F	R	R
AE	AI, E	G	G	S	S
AH	A, ER	HH	H	SH	SCH
AO	O, AU	IH	I	T	T
AW	AU, AH	IX	EH, I	TH	TS, F, Z, S
AX	E2	IY	IE	UH	UH, UE
AXR	R, ER	JH	SCH, T+SCH, D	UW	UH, UE
AY	AI	K	K	V	V
B	B	L	L	W	U
CH	SCH, T+SCH, CH	M	M	XL	U, OH, L
D	D	N	N	XM	N, UH, M
DH	D, V	NG	NG	XN	N
EH	E	OW	OH, O+U	Y	J, IE
ER	R, OE, OEH	OY	E, EU	Z	S, Z
EY	AEH, EH+I	P	P	ZH	SCH

Tabelle 4.2: Gemischte Abbildungsfunktion von englischen auf deutsche Phoneme.

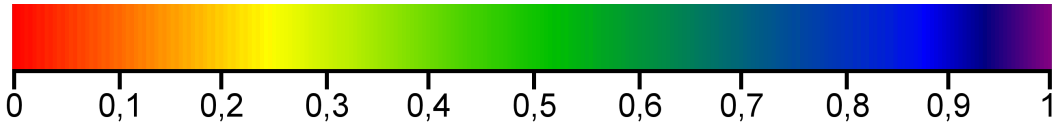

Abbildung 4.1: Skala für die nachfolgenden Distanzmatrizen.

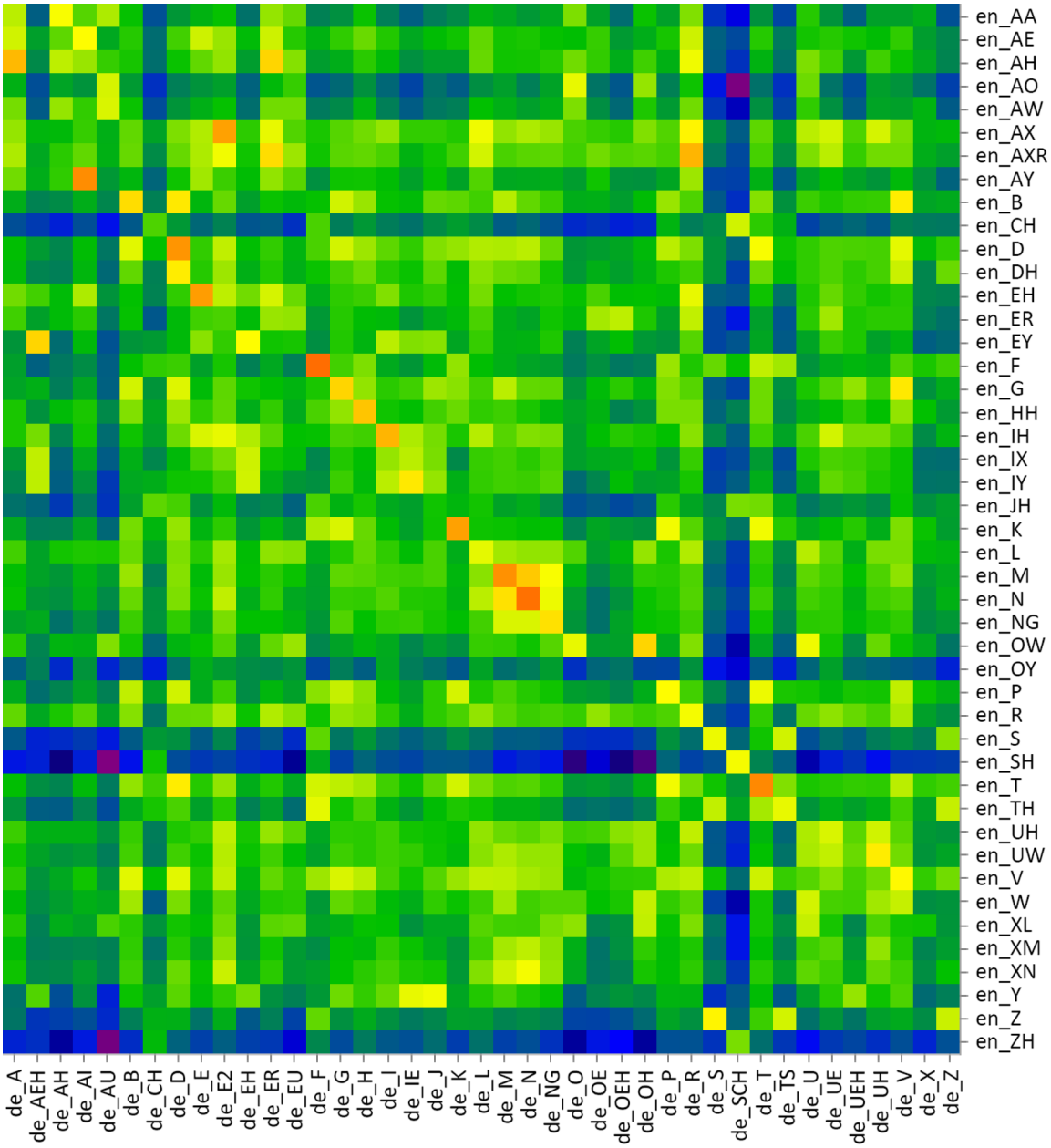

Abbildung 4.2: Phonemdistanzmatrix euklidisch (normalisiert).

Englisch	AA	AE	AH	AO	AW	AX	AXR	AY	B	CH	D	DH	EH	ER	EY
Deutsch	AH	AI	A	O	AU	E2	R	AI	B	SCH	D	D	E	R	AEH
Englisch	F	G	HH	IH	IX	IY	JH	K	L	M	N	NG	OW	OY	P
Deutsch	F	G	H	I	EH	IE	SCH	K	L	M	N	NG	OH	E	P
Englisch	R	S	SH	T	TH	UH	UW	V	W	XL	XM	XN	Y	Z	ZH
Deutsch	R	S	SCH	T	TS	UE	UH	V	U	U	N	N	J	S	SCH

Tabelle 4.3: Aus euklidischem Distanzmaß gewonnene Abbildungsfunktion.

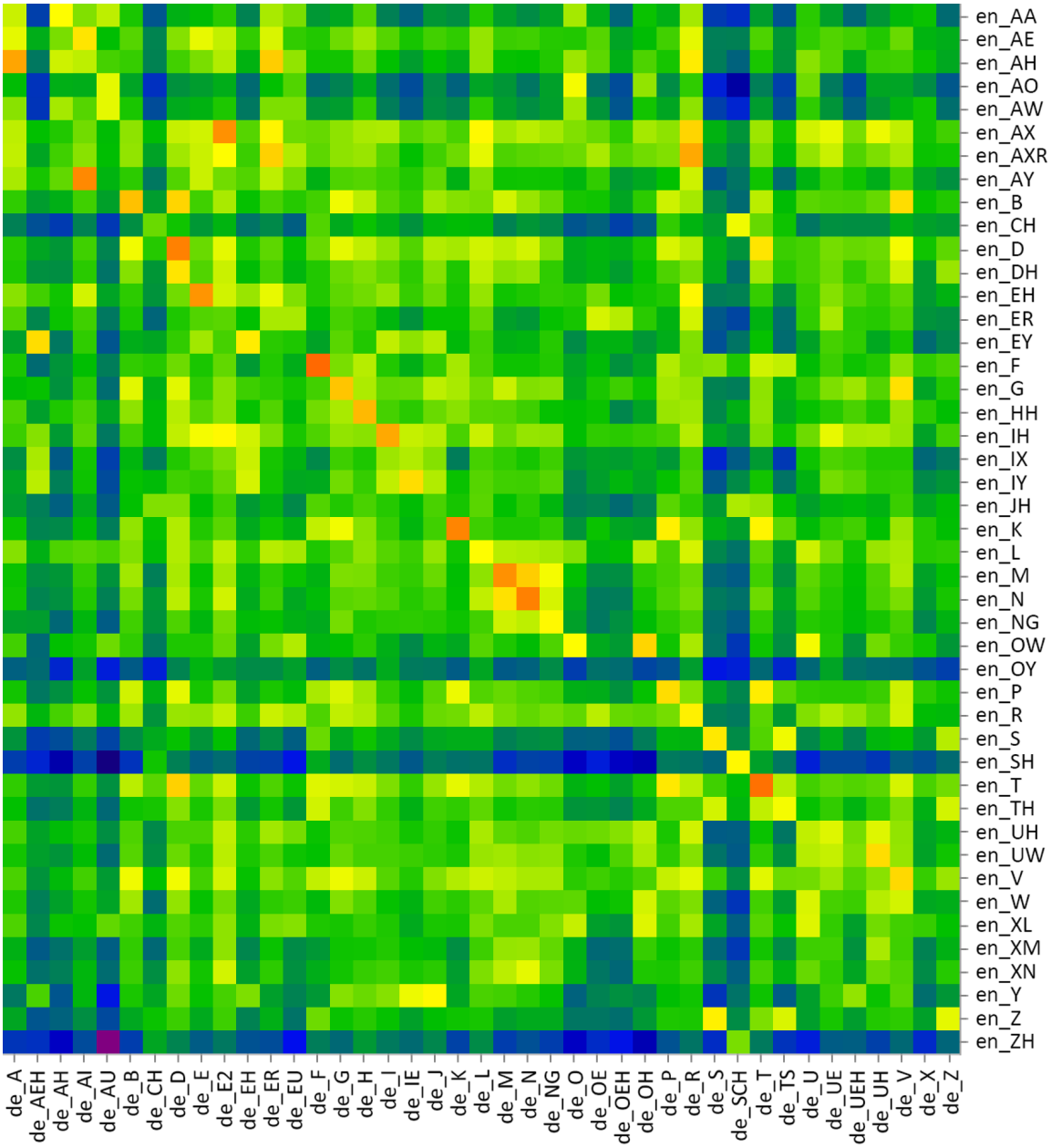

Abbildung 4.3: Phonemdistanzmatrix Mahalanobis (normalisiert).

Englisch	AA	AE	AH	AO	AW	AX	AXR	AY	B	CH	D	DH	EH	ER	EY
Deutsch	AH	AI	A	O	AU	E2	R	AI	B	SCH	D	D	E	OE	AEH
Englisch	F	G	HH	IH	IX	IY	JH	K	L	M	N	NG	OW	OY	P
Deutsch	F	G	H	I	EH	IE	SCH	K	L	M	N	NG	OH	E	P
Englisch	R	S	SH	T	TH	UH	UW	V	W	XL	XM	XN	Y	Z	ZH
Deutsch	R	S	SCH	T	TS	UH	UH	V	U	OH	UH	N	J	S	SCH

Tabelle 4.4: Aus Mahalanobis-Distanzmaß gewonnene Abbildungsfunktion.

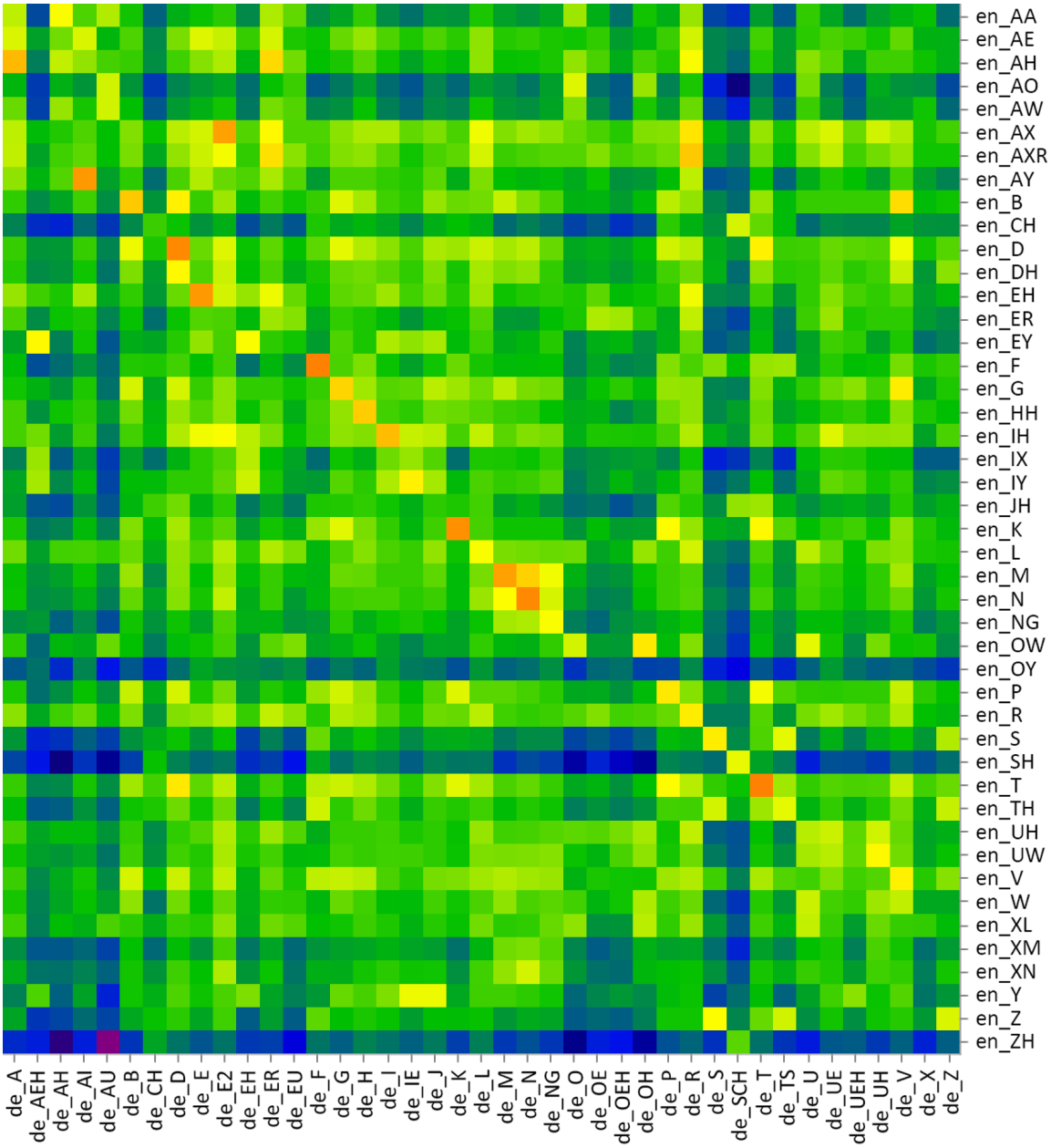

Abbildung 4.4: Phonemdistanzmatrix Kullback-Leibler (normalisiert).

Englisch	AA	AE	AH	AO	AW	AX	AXR	AY	B	CH	D	DH	EH	ER	EY
Deutsch	AH	E	A	O	AU	E2	R	AI	B	SCH	D	D	E	R	AEH
Englisch	F	G	HH	IH	IX	IY	JH	K	L	M	N	NG	OW	OY	P
Deutsch	F	G	H	I	EH	IE	T	K	L	M	N	NG	OH	E	P
Englisch	R	S	SH	T	TH	UH	UW	V	W	XL	XM	XN	Y	Z	ZH
Deutsch	R	S	SCH	T	TS	UH	UH	V	U	U	N	N	IE	S	SCH

Tabelle 4.5: Aus Kullback-Leibler-Distanzmaß gewonnene Abbildungsfunktion.

Englisch	Deutsch			
	IPA-basiert	Euklidisch	Mahalanobis	Kullback-Leibler
AA	AH	AH	AH	AH
AE	AEH	AI	AI	E
AH	A	A	A	A
AO	O	O	O	O
AW	AU	AU	AU	AU
AX	E2	E2	E2	E2
AXR	ER	R	R	R
AY	AI	AI	AI	AI
B	B	B	B	B
CH	T+SCH	SCH	SCH	SCH
D	D	D	D	D
DH	Z	D	D	D
EH	E	E	E	E
ER	OE	R	OE	R
EY	E+I	AEH	AEH	AEH
F	F	F	F	F
G	G	G	G	G
HH	H	H	H	H
IH	I	I	I	I
IX	I	EH	EH	EH
IY	IE	IE	IE	IE
JH	D+SCH	SCH	SCH	T
K	K	K	K	K
L	L	L	L	L
M	M	M	M	M
N	N	N	N	N
NG	NG	NG	NG	NG
OW	O+U	OH	OH	OH
OY	EU	E	E	E
P	P	P	P	P
R	R	R	R	R
S	S	S	S	S
SH	SCH	SCH	SCH	SCH
T	T	T	T	T
TH	S	TS	TS	TS
UH	U	UE	UH	UH
UW	UH	UH	UH	UH
V	V	V	V	V
W	U	U	U	U
XL	E2+L	U	OH	U
XM	E2+M	N	UH	N
XN	E2+N	N	N	N
Y	J	J	J	IE
Z	Z	S	S	S
ZH	SCH	SCH	SCH	SCH

Tabelle 4.6: Die 1:1 Abbildungsfunktionen auf einen Blick.

4.2 Erweiterung das akustischen Modells

Durch den Abbildungsansatz lassen sich die englischen Aussprachen nur approximieren, denn die englischen Phoneme können durch die deutschen nicht exakt nachgebildet werden, da sie sich teilweise sehr voneinander unterscheiden. So ist z.B. das deutsche Phonem *SCH* nur ein unzureichender Ersatz für die englischen Phoneme *JH* und *ZH*. Möchte man die korrekten, englischen Aussprachen ohne Verfremdung dem deutschen Spracherkennungssystem hinzufügen, so ist dies nur durch Erweiterung des akustischen Modells möglich. Im Folgenden werden dafür zwei Varianten vorgestellt, zum einen die parallele Verwendung eines englischen und deutschen akustischen Modells, zum anderen die Kombination der beiden zu einem Modell mit Mono- und Polyphonemen.

4.2.1 Parallele Verwendung eines englischen und deutschen akustischen Modells

Da wir bereits über leistungsfähige, monolinguale Spracherkennungssysteme für Deutsch und Englisch verfügen, liegt die Idee nahe, das akustische Modell eines englischen Systems zu nehmen und mit dem unseres deutschen Systems zu kombinieren. Hierbei muss allerdings sichergestellt werden, dass beide Systeme wirklich kompatibel sind. Zur Gewinnung des englischen akustischen Modells trainieren wir deshalb ein neues, englisches System, welches dieselbe Vorverarbeitung und dieselben Transformationsmatrizen wie unser deutsches System benutzt. Die *Linear Discriminant Analysis* (LDA) und *Semi-Tied Covariance* (STC) [Gales 99] Matrizen werden ausschließlich auf deutschen Daten trainiert, da unser Zielsystem ein deutsches ist und unser Testset mehr deutsche als englische Wörter enthält. Wir haben außerdem in früheren, sprachübergreifenden Experimenten festgestellt, dass zumindest die Berechnung der LDA-Matrix nicht besonders abhängig von der Zielsprache ist.

Die so trainierten, englischen Modelle werden den deutschen Modellen hinzugefügt, wobei sie markiert und nach Sprache getrennt gehalten werden. Ebenso wird mit den englischen Phonemen verfahren. Die Trennung nach Sprache kann einfach auf Ebene der Bezeichner erfolgen, indem jedem Bezeichner ein entsprechendes Präfix oder Suffix beigefügt wird, das die Sprachzugehörigkeit ausweist. Abbildung 4.5 stellt dies symbolisch dar. Beim Aussprachewörterbuch gibt es nur zu beachten, dass die in den Aussprachen verwendeten Phonembezeichner entsprechend umbenannt werden, wenn das Wörterbuch wie in Kapitel 3.3 beschrieben neu aufgebaut wird. Das Wörterbuch des englischen Systems findet keine Verwendung, da unser deutsches Wörterbuch bereits alle relevanten, englischen Ausdrücke enthält.

Sowohl unser englisches als auch unser deutsches akustisches Modell ist kontextabhängig und benutzt State-Tying, es gibt also noch zwei phonetische Entscheidungsbäume die zusammengeführt werden müssen. Ein phonetischer Entscheidungsbaum ist ein Binärbaum, der in seinen Knoten Fragen an den phonetischen Kontext stellt. Die linken und rechten Kanten entsprechen jeweils den Antworten „Ja“ bzw. „Nein“. Da der deut-

sche Baum nur nach deutschen Phonemen frägt, kann man von seiner Wurzel aus in jedem Knoten den „Nein“-Pfad nehmen, bis man ein Blatt erreicht. Dort kann dann schließlich mit einer neuen „Nein“-Kante die Wurzel des englischen Baumes angehängt werden, wie in Abbildung 4.5 dargestellt. Auf diese Weise werden englische Kontexte vom deutschen Teilbaum an den englischen Teilbaum weitergereicht. Ein solcher Entscheidungsbaum ist einfach zu konstruieren, hat allerdings den Nachteil, dass er nicht richtig mit gemischten Kontexten, die deutsche und englische Phoneme zugleich enthalten, umgehen kann.

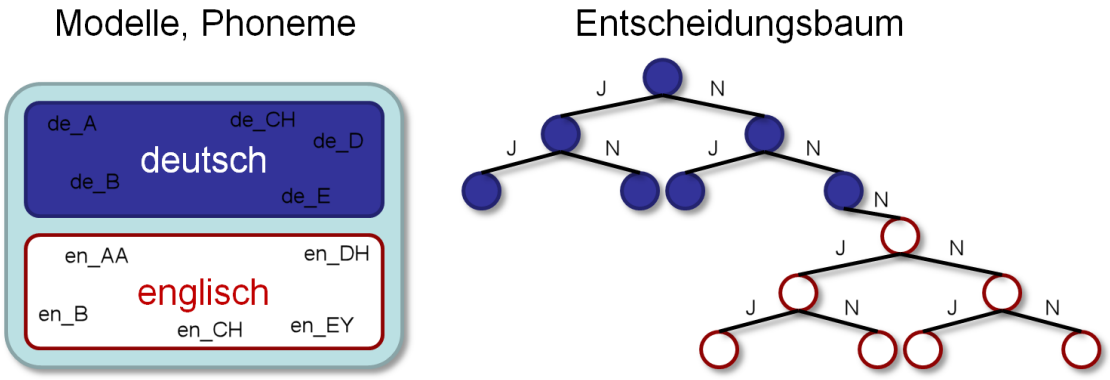

Abbildung 4.5: Symbolische Darstellung des parallelen Systems.

Die Testergebnisse unserer Experimente mit diesem parallelen System werden in Kapitel 5.3 präsentiert.

4.2.2 Mono- und Polyphoneme

In diesem Kapitel wird ein weiterer Ansatz zur Kombination der deutschen und englischen akustischen Modelle vorgestellt. Der Abbildungsansatz stellt ein Extrem dar, denn mit ihm wird nur das deutsche akustische Modell benutzt, um sowohl deutsch als auch englisch abzudecken. Der Parallellansatz ist ein anderes Extrem, denn mit ihm wird die Komplexität des akustischen Modells verdoppelt. Deshalb untersuchen wir nun einen Mittelweg: da sich im Abbildungsansatz gezeigt hat, dass nur ein Teil der englischen Phoneme große Ähnlichkeit zu deutschen Phonemen aufweist, fassen wir nur diese zusammen. Dieses Vorgehen entspricht einer Idee von Dalsgaard und Andersen [Dalsgaard 92], die ihm Rahmen der Sprachidentifizierung die Bezeichnungen *Polyphonem* und *Monophonem* eingeführt haben. Polyphoneme sind hierbei Phoneme, die in zwei oder mehr Sprachen ähnlich genug sind, um als identisch angesehen zu werden, während Monophoneme keine ausreichend ähnlichen Phoneme in anderen Sprachen haben. Die Frage, was dabei als ausreichend ähnlich anzusehen ist, kann nicht pauschal beantwortet werden. In der Praxis wird meist nach verschiedenen Kriterien ein Schwellwert definiert. Liegt der Abstand zweier Phoneme unter diesem Wert, so werden sie zusammengefügt.

Wir folgen diesem Vorgehen und greifen auf die in Kapitel 4.1 vorgestellten Distanzmaße zurück. Als Schwellwert wählen wir den kleinsten Abstand zwischen zwei unterschiedlichen, deutschen Phonemen. Damit werden ein englisches und ein deutsches Phonem nur dann zu einem Polyphonem zusammengefasst, wenn sie sich ähnlicher sind als die deutschen Phoneme untereinander. Tabelle 4.7 zeigt die so ermittelten Polyphoneme. Es fällt auf, dass sich die drei Distanzmaße bei den besten Kandidaten bis auf die vernachlässigbare Reihenfolge einig sind, allerdings nicht bei der Anzahl. Die Kullback-Leibler-Distanz liefert die meisten Polyphoneme und enthält dabei alle Phonempaare, die auch von den anderen beiden Distanzmaßen vorgeschlagen werden. Da wir bereits beim Abbildungsansatz gute Resultate mit diesem Distanzmaß erzielten und alle ermittelten Phonempaare auch bei manueller Überlegung sinnvoll erscheinen, entscheiden wir uns deshalb für die elf Phonempaare der Kullback-Leibler-Distanz.

Wie zuvor markieren wir sämtliche Monophoneme, so dass ihre Sprachzugehörigkeit ersichtlich ist und Kollisionen bei den Bezeichnern vermieden werden. Im Gegensatz zum Parallelsystem aus Kapitel 4.2.1 kombinieren wir hier keine Teile aus bestehenden akustischen Modellen, sondern trainieren ein komplett neues System. Für das Training verwenden wir in gleichen Maßen deutsche und englische Trainingsdaten.

Die Testergebnisse unserer Experimente mit diesem Polyphonem-System werden in Kapitel 5.4 präsentiert.

Euklidisch		Mahalanobis		Kullback-Leibler	
Deutsch	Englisch	Deutsch	Englisch	Deutsch	Englisch
F	F	F	F	F	F
N	N	T	T	T	T
T	T	N	N	N	N
AI	AY	D	D	D	D
D	D	AI	AY	K	K
		K	K	E	EH
		E	EH	AI	AY
		M	M	M	M
		E2	AX	E2	AX
				A	AH
				I	IH

Tabelle 4.7: Ermittelte Polyphonempaare.

Kapitel 5

Experimente

Die in diesem Kapitel beschriebenen Spracherkennerexperimente wurden mit dem *Janus Recognition Toolkit* (JRTk) [Finke 97] und dessen IBIS Decoder [Soltau 01] durchgeführt. Das JRTk wird gemeinsam von den *Interactive Systems Laboratories* an der Universität Karlsruhe (TH), Deutschland und der Carnegie Mellon University in Pittsburgh, Pennsylvania, USA entwickelt und gepflegt.

Alle Testläufe wurden auf einer 45-minütigen, deutschen, technischen Vorlesungen durchgeführt, die an der Universität Karlsruhe (TH) aufgezeichnet wurde. Das Training fand auf einem anderen Set statt, das ebenfalls aus Vorlesungen und Vorträgen bestand. Bei den Tests wurde keine Sprecheradaption verwendet.

Unser akustisches Modell besitzt 2000 Codebücher mit bis zu 64 Gaußglocken. Es ist kontextabhängig und benutzt Triphone. Von dem mit 16 kHz abgetasteten Signal werden alle 10 ms neue Merkmalsvektoren berechnet. Hierfür zerlegen wir das kontinuierliche Sprachsignal durch ein 16 ms Hamming-Fenster und berechnen die skalierte Mel-Frequenz *Minimum Variance Distortionless Response* (MVDR) Einhüllende [Wölfel 09]. Danach werden 20 Cepstralkoeffizienten, die durch cepstrale Mittelwertssubtraktion normalisiert wurden, zusammen mit ihren drei linken und rechten Nachbarn durch eine diskrete Kosinustransformation berechnet. Schließlich wird die Merkmalsdimension durch eine LDA auf 42 reduziert.

Das verwendete 4-Gramm Sprachmodell wurde mit dem SRI Language Modeling Toolkit (SRILM) [Stolcke 02] erstellt. Es berücksichtigt deutsche Fernsehnachrichten, Zeitungstexte sowie manuelle Transkripte um Effekte der gesprochenen Sprache zu modellieren. Es wurde auf Vorlesungen und technische Präsentationen adaptiert.

5.1 Ausgangssystem

Durch unsere in Kapitel 3.2 beschriebene Klassifizierung der Sprachzugehörigkeit kann ein Überblick über den Einfluss englischer Wörter auf die Spracherkennungsleistung ge-

wonnen werden. Wir analysierten eine 6589 Wörter umfassende, 45-minütige Vorlesung an der Universität Karlsruhe (TH), es ergab sich folgende Verteilung:

- 64% oder 4195 Wörter sind eindeutig *deutsch*.
- 21% oder 1397 Wörter sind *ambivalent* und können sowohl deutsch als auch englisch sein.
- 2% oder 110 Wörter sind eindeutig *englisch*.
- 13% oder 887 Wörter konnten weder Deutsch noch Englisch zugeordnet werden und wurden als *unbekannt* eingestuft. Dieser Anteil besteht hauptsächlich aus Häsitationen und Wortfragmenten.

Der Anteil der eindeutig englischen Wörter erscheint recht gering, stimmt aber mit den Ergebnissen einer Analyse von portugiesischen, technischen Vorlesungen an der Universität Lissabon überein; dort wurde ein Anteil von 2,1% ermittelt [Trancoso 06]. Ein Blick auf die individuellen Wortfehlerraten unseres Ausgangssystems in Tabelle 5.1 zeigt, dass diese englischen Wörter sehr viel schlechter erkannt werden als eindeutig deutsche Wörter. Der Einfluss dieser hohen englischen Wortfehlerrate auf die Gesamtfehlerrate von 13,8% ist nicht zu unterschätzen. Gelänge es, die englische Fehlerrate zu halbieren, so würde die Gesamtfehlerrate bereits um 0,5% absolut sinken. Die Tatsache, dass die Wortfehlerrate bei ambivalenten Wörtern dichter am deutschen als am englischen Wert liegt, lässt darauf schließen, dass die meisten, aber nicht alle, dieser ambivalenten Wörter deutsch und nicht englisch ausgesprochen wurden.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	52	1	44	0
Insertions	58	9	37	2
Substitutions				
Deutsch	258	37	91	113
Englisch	7	6	8	7
Ambivalent	68	10	33	56
Unbekannt	5	3	2	4
Gesamtfehler	448	66	215	182
Wortfehlerrate	10,7%	60,0%	15,4%	20,5%

Tabelle 5.1: Fehler sortiert nach Sprachen für das Ausgangssystem (gesamt: 13,8%).

5.2 Abbildungssysteme

Wie Tabelle 5.2 zeigt, lieferte unsere manuelle Abbildungsfunktion in Tests mit 13,2% eine leicht bessere Gesamtfehlerrate als die drei errechneten Funktionen. Dies liegt vermutlich daran, dass die manuelle Abbildung, wie in Kapitel 4.1.1 beschrieben, bestimmte englischen Phoneme durch ein deutsches Phonempaar ersetzte statt nur durch

ein einzelnes deutsches Phonem, was zu einer besseren Approximation führt. Von den errechneten Abbildungsfunktionen lagen die Mahalanobis- und Kullback-Leibler-basierten mit 13,3% Gesamtfehlerrate gleichauf, die euklidische etwas dahinter mit 13,5%. Betrachtet man nur die englische Wortfehlerrate, so führt Kullback-Leibler mit 46,4% sogar vor der manuellen Funktion. Dieser immer noch hohe Wert veranlasste uns zum Erstellen der in Kapitel 4.1.3 vorgestellten gemischten Abbildungsfunktion, die dann tatsächlich die englische Wortfehlerrate auf 34,6% senkte und mit der im Testfeld niedrigsten Gesamtfehlerrate von 12,7% aufwarten konnte. Die Ergebnisse dieses Systems sind in Tabelle 5.3 zusammengefasst.

All diese Testläufe fanden auf dem ursprünglichen akustischen Modell statt. Nach einem Neutraining des akustischen Modells mit einem Wörterbuch, das aus unserer gemischten Abbildungsfunktion resultierte, sank die englische Wortfehlerrate nochmals, auf 27,3%. Allerdings stellten wir fest, dass die Gesamtfehlerrate wieder leicht anstieg, da nun deutsche Wörter schlechter erkannt wurden. Als Grund dafür vermuten wir eine Verringerung der akustischen Eindeutigkeit der deutschen Phoneme, da diese beim Neutraining durch englische Aussprachen verschmutzt werden.

	Wortfehlerrate				
Sprache	Gesamt	Deutsch	Englisch	Ambivalent	Unbekannt
Manuell	13,2%	11,2%	47,3%	14,7%	16,4%
Euklidisch	13,5%	11,3%	51,8%	15,1%	16,7%
Mahalanobis	13,3%	11,2%	48,2%	15,0%	16,6%
Kullback-Leibler	13,3%	11,1%	46,4%	15,0%	17,4%
Gemischt	12,7%	11,1%	34,6%	13,8%	16,1%
Gemischt (neu trainiert)	13,0%	11,7%	27,3%	14,1%	15,3%

Tabelle 5.2: Wortfehlerraten der verschiedenen Abbildungssysteme im Vergleich.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	50	0	38	0
Insertions	54	7	35	0
Substitutions				
Deutsch	260	18	76	84
Englisch	14	5	6	8
Ambivalent	81	8	37	47
Unbekannt	5	0	1	4
Gesamtfehler	464	38	193	143
Wortfehlerrate	11,1%	34,6%	13,8%	16,1%

Tabelle 5.3: Fehler sortiert nach Sprachen für das gemischte Abbildungssystem (gesamt: 12,7%).

5.3 Parallelsystem

Da das akustische Modell des Parallelsystems aus einem deutschen und einem englischen zusammengesetzt wurde, umfasst es im Gegensatz zum Ausgangssystem und den Abbildungssystemen 4000 statt 2000 Codebücher. Es konnte mit 26,4% Wortfehlerrate das insgesamt beste Ergebnis für englische Wörter erzielen. Allerdings produzierte es etwas mehr Fehler auf deutschen Wörtern als die anderen Systeme, wodurch die Gesamtfehlerrate bei 13,4% lag. Die Fehlerraten der anderen Wortklassen wurden im Vergleich zum Ausgangssystem zwar reduziert, jedoch erreichte das gemischte Abbildungssystem hier niedrigere Werte. Die Ergebnisse des Parallelsystems werden in Tabelle 5.4 zusammengefasst.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	54	0	39	0
Insertions	61	10	48	0
Substitutions				
Deutsch	266	11	74	81
Englisch	19	3	5	14
Ambivalent	74	5	37	67
Unbekannt	6	0	2	6
Gesamtfehler	480	29	205	168
Wortfehlerrate	11,4%	26,4%	14,7%	18,9%

Tabelle 5.4: Fehler sortiert nach Sprachen für das Parallelsystem (gesamt: 13,4%).

5.4 Polyphonem-System

Das Polyphonem-System konnte ebenfalls die englische Wortfehlerrate signifikant reduzieren, auf 27,3%. Jedoch enttäuschte es mit vielen Fehlern auf deutschen und ambivalenten Wörtern. Die Gesamtfehlerrate von 15,1% lag dadurch deutlich über der des Ausgangssystems. Die Ergebnisse des Polyphonem-Systems werden in Tabelle 5.5 zusammengefasst. Um eventuelle Vorteile des Parallelsystems auszugleichen, trainierten wir eine weitere Version des Polyphonem-Systems mit einem 4000 Codebücher umfassenden akustischen Modell. Wie in Tabelle 5.6 zu sehen brachte dies keinerlei Verbesserung, sondern produzierte mehr Fehler auf allen Klassen.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	79	0	55	1
Insertions	69	9	37	0
Substitutions				
Deutsch	293	6	98	89
Englisch	18	6	2	7
Ambivalent	114	9	42	55
Unbekannt	4	0	2	2
Gesamtfehler	577	30	236	154
Wortfehlerrate	13,8%	27,3%	16,9%	17,4%

Tabelle 5.5: Fehler sortiert nach Sprachen für das Polyphonem-System mit 2000 Codebüchern (gesamt: 15,1%).

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	79	0	36	1
Insertions	77	11	57	3
Substitutions				
Deutsch	296	6	92	94
Englisch	27	5	4	8
Ambivalent	111	9	46	59
Unbekannt	6	1	2	2
Gesamtfehler	596	32	237	167
Wortfehlerrate	14,2%	29,1%	17,0%	18,8%

Tabelle 5.6: Fehler sortiert nach Sprachen für das Polyphonem-System mit 4000 Codebüchern (gesamt: 15,7%).

Kapitel 6

Schlussbetrachtungen

Im Folgenden werden die Ergebnisse der Arbeit zusammengefasst und es wird ein kurzer Ausblick auf zukünftige Ansatzmöglichkeiten gegeben.

6.1 Ergebnisse der Arbeit

Diese Arbeit stellte mehrere Ansätze vor, um die Erkennungsleistung eines deutschen Spracherkennungssystems auf englischen Wörtern zu verbessern und wertete sie experimentell aus. Der Schwerpunkt lag dabei auf der Anpassung des akustischen Modells und des Aussprachewörterbuchs. Zu diesem Zweck wurden Phonemabstände berechnet und ausgenutzt, wobei drei unterschiedliche Distanzmaße verwendet und verglichen wurden, die euklidische, Mahalanobis- und Kullback-Leibler-Distanz. Diese Maße erwiesen sich dabei als relativ gleichwertig, mit der euklidischen Distanz erwartungsgemäß etwas im Nachteil, da sie keine Varianzen berücksichtigt. Außerdem wurde ein einfaches Verfahren vorgeführt, um die Sprachzugehörigkeit geschriebener Wörter zu bestimmen. Damit konnte gezeigt werden, dass englische Wörter in deutschen, technischen Vorlesungen trotz eines Anteils von nur 2% einen nicht zu unterschätzenden Einfluss auf die Gesamterkenungsleistung haben.

Beim Vergleich der Ergebnisse in Tabelle 6.1 sehen wir, dass alle getesteten Methoden die Erkennungsleistung auf englischen Wörtern stark verbessern. Mit Ausnahme des Polyphonem-Systems wird dabei auch die insgesamte Fehlerrate reduziert. Während der parallele Ansatz die niedrigste Wortfehlerrate für rein englische Wörter hat, ist die gesamte Fehlerrate beim Abbildungsansatz am niedrigsten, sie übertrifft das Ausgangssystem um 1,1% absolut. Interessanterweise verschlechtern alle Ansätze die Erkennungsleistung auf deutschen Wörtern. Aus den Tabellen 5.1, 5.3, 5.4 und 5.5 ist ersichtlich, dass besonders die Ersetzung von englischen Wörtern durch deutsche abnimmt, während gleichzeitig die Ersetzung von deutschen Wörtern durch englische zunimmt.

	Wortfehlerrate				
Sprache	Gesamt	Deutsch	Englisch	Ambivalent	Unbekannt
Ausgangssystem	13,8%	10,7%	60,0%	15,4%	20,5%
Abbildung (gemischt)	12,7%	11,1%	34,6%	13,8%	16,1%
Parallel	13,4%	11,4%	26,4%	14,7%	18,9%
Polyphoneme	15,1%	13,8%	27,3%	16,9%	17,4%

Tabelle 6.1: Wortfehlerrate nach Sprache für die unterschiedlichen Ansätze.

Tabelle 6.2 zeigt, dass alle Ansätze auch eine längere Laufzeit als das Ausgangssystem benötigen. Hauptursache dafür ist das gewachsene Suchvokabular. Beim Abbildungssystem wurden durch die gemischte Abbildungsfunktion sehr viele neue Aussprachevarianten in das Wörterbuch eingeführt. Dadurch wurde zwar die Erkennungsleistung erhöht, aber neben der Auswirkung auf die Laufzeit haben diese Varianten einen weiteren, negativen Effekt. Aus der Literatur [Saraclar 00] ist bekannt, dass die Präsenz von vielen Aussprachevarianten der akustischen Eindeutigkeit eines Wortes abträglich ist und aus Sicht des Spracherkenners die Gefahr der Verwechslung mit anderen Wörtern erhöht. Dies ist vermutlich einer der Gründe, aus denen sich die Erkennung deutscher Wörter verschlechtert hat. Dennoch weisen die Ergebnisse darauf hin, dass die positiven Auswirkungen insgesamt überwiegen können.

System	Ausgang	Abbildung (gemischt)	Parallel/Polyphonem
Echtzeitfaktor	1,50	1,79	1,62
Wörterbuchgröße	63510	97247	70550

Tabelle 6.2: Auswirkung der Wörterbuchgröße auf die Laufzeit.

6.2 Ausblick

Wenn man bedenkt, dass ein vergleichbares englisches Spracherkennungssystem auf rein englischen Daten eine Wortfehlerrate von ca. 10% hat, die beste hier vorgestellte Methode aber nur 26,4% erreichte, dann wird offensichtlich, dass hier noch großes Optimierungspotenzial besteht. Für Abbildungssysteme liefert möglicherweise das Abbilden von englischen, kontextabhängigen Triphonen auf deutsche Triphone bessere Ergebnisse als unsere kontextunabhängigen Funktionen. Das Parallelsystem macht einen viel versprechenden Eindruck, allerdings müsste geklärt werden, was die genauen Ursachen für die verschlechterte Erkennung deutscher Wörter sind und wie man sie bekämpft. Ebenso sollte untersucht werden, warum die Leistung auf englischen Wörtern nicht noch besser ist und ob der auf unsere Weise konstruierte phonetische Entscheidungsbaum negative Auswirkungen hat. Das Polyphonem-System bedarf ebenfalls weiterer Untersuchungen, denn obwohl es insgesamt schlechter als das Ausgangssystem war, lieferte

es doch gute Resultate für englische Wörter. Außerdem wären präzisere und bessere Verfahren zur Identifizierung der Sprachzugehörigkeit nützlich. Durch die Betrachtung des Wortkontexts oder durch akustische Sprachidentifizierung könnten beispielsweise die bisher als ambivalent eingestuften Wörter eindeutig als deutsch oder englisch klassifiziert werden. Schließlich sollte man das Sprachmodell nicht vergessen, denn bisher haben wir uns ausschließlich auf die Akustik konzentriert.

All dies zeigt, dass zukünftige Arbeiten noch viele Möglichkeiten haben, diese Arbeit weiterzuentwickeln und die automatische Transkription von englischen Wörtern in deutschen Vorlesungen zu optimieren.

Anhang A

Deutsche Phoneme

Phonem	Beispiele	Phonem	Beispiele
A	M**a**nn, **A**mt	M	**m**einen, Diplo**m**
AEH	sp**ä**t, prim**ä**r	N	**n**icht, da**nn**
AH	M**ah**l, V**a**ter	NG	Ga**ng**, Werbu**ng**
AI	**Ei**s, Z**ei**t	O	**o**ft, K**o**mma
AU	**au**f, Fr**au**	OH	**oh**ne, K**o**ma
B	**B**ein, glau**b**en	OE	m**ö**chte, B**ö**rse
CH	i**ch**, Mil**ch**	OEH	m**ö**glich, b**ö**se
D	**d**u, schnei**d**en	P	**P**unkt, kna**pp**
E	**E**nde, f**e**st	R	**r**ichtig, **R**est
E2	Ab**e**nd, geh**e**n	S	**s**ehen, Schlu**ss**
EH	**e**ben, F**eh**ler	SCH	**sch**einen, Mat**sch**
ER	ab**er**, flüst**er**n	T	**t**rinken, Bla**tt**
EU	h**eu**te, F**eu**er	TS	**z**ielen, Pla**tz**
F	au**f**, **f**angen	U	**u**nd, Samml**u**ng
G	**g**ehen, fol**g**en	UH	t**u**n, St**uh**l
H	**h**elfen, Frei**h**eit	UE	f**ü**llen, St**ü**ck
I	**i**ch, s**i**tzen	UEH	f**üh**len, fr**üh**
IE	akt**i**v, l**ie**gen	V	**w**arten, **V**ase
J	**J**ahr, **j**ung	X	Bu**ch**, ma**ch**en
K	**k**önnen, ha**ck**en	Z	Analy**s**e, be**s**uchen
L	**l**ang, Ba**ll**		

Tabelle A.1: Verwendete deutsche Phonembezeichner.

Anhang B

Englische Phoneme

Phonem	Beispiele	Phonem	Beispiele
AA	**a**rm, **a**rticle	L	**l**ong, **l**ife
AE	**a**venue, **a**xe	M	**m**an, **m**ouse
AH	b**u**s, **u**p	N	**n**ice, **n**ew
AO	**awe**some, f**o**rce	NG	ba**n**k, playi**ng**
AW	b**ou**nce, d**ow**n	OW	**o**ld, m**o**de
AX	**a**bout, th**e**	OY	b**oy**, p**oi**nt
AXR	captu**re**, lit**er**	P	**p**arty, cla**p**
AY	m**i**ke, ps**y**cho	R	**r**ound, **r**ing
B	**b**rain, a**b**out	S	**s**ell, plu**s**
CH	**ch**ain, tou**ch**	SH	**sh**ip, **sh**ort
D	**d**ay, **d**estiny	T	**t**ime, **t**ip
DH	**th**e, **th**is	TH	**th**ink, mon**th**
EH	**e**rror, **e**xcellent	UH	g**oo**d, w**ou**ld
ER	b**ir**d, t**er**m	UW	l**oo**se, y**ou**
EY	w**ei**ght, t**a**ke	V	**v**ery, o**v**er
F	**f**ire, **f**lag	W	**w**ay, q**u**een
G	**g**old, **g**un	XL	ab**le**, ang**le**
HH	**h**ouse, **h**ome	XM	rhyth**m**, touris**m**
IH	h**i**t, **i**m**a**ge, ab**i**lity	XN	cert**ain**, butt**on**
IX	play**i**ng, abil**i**ty	Y	**y**ear, **y**ou
IY	magaz**i**ne, abilit**y**	Z	i**s**, **z**oo
JH	ma**j**or, mer**ge**	ZH	mea**s**ure, u**s**ual
K	**k**ind, mi**c**ro		

Tabelle B.1: Verwendete englische Phonembezeichner.

Literaturverzeichnis

[Andersen 94] O. Andersen, P. Dalsgaard, W. Barry. On the use of data-driven clustering techniques for language identification of poly- and mono-phonemes for four european languages. Tagungsband: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seiten 121–124, Adelaide, 1994.

[Black 97] A. W. Black, P. A. Taylor. The Festival Speech Synthesis System: System documentation. Human Communciation Research Centre, University of Edinburgh, 1997.

[Cettolo 04] Mauro Cettolo, Fabio Brugnara, Marcello Federico. Advances in the automatic transcription of lectures. Tagungsband: ICASSP, 2004.

[Dalsgaard 92] P. Dalsgaard, O. Andersen. Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organising neural network. Tagungsband: Proceedings of the International Conference on Spoken Language Processing, Seiten 547–550, Banff, Alberta, Canada, 1992.

[Fügen 06] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, A. Waibel. Open domain speech recognition & translation: Lectures and speeches. Tagungsband: Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, May 2006.

[Finke 97] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal. The karlsruhe-verbmobil speech recognition engine. Tagungsband: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seiten 83–86, Munich, Germany, 1997. Janus Recognition Toolkit (JRTk).

[Gales 99] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. IEEE Transactions on Speech and Audio Processing, 7:272–281, 1999.

[Heger 07] Dominic Heger. Speech feature enhancement using particle filters with class-based phoneme models. Studienarbeit, Universität Karlsruhe (TH), 2007.

[Imperl 99] Bojan Imperl. Clustering of context dependent speech units for multilingual speech recognition. Tagungsband: Proceedings of the ESCA-NATO Tutorial Research Workshop on Multi-lingual Interoperability in Speech Technology, Seiten 17–22, Leusden, Netherlands, 1999.

[IPA 05] IPA. The International Phonetic Alphabet, 2005.

[Köhler 96] Joachim Köhler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. Tagungsband: Proceedings of the International Conference on Spoken Language Processing, Seiten 2195–2198, Philadelphia, 1996.

[Le 06] Viet Bac Le, Laurent Besacier, Tanja Schultz. Acoustic-phonetic unit similarities for context dependent acoustic model portability. Tagungsband: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Band 1, 2006.

[Ma 98] Chi Yuen Ma, Pascale Fung. Using english phoneme models for chinese speech recognition. Tagungsband: International Symposium on Chinese Spoken language processing, Seiten 80–82, Hongkong, 1998.

[Németh 03] László Németh. Hunspell: open source spell checking, stemming, morphological analysis and generation., 2003.

[Saraclar 00] M. Saraclar, S. Khudanpur. Pronunciation ambiguity versus pronunciation variability in speech recognition. Tagungsband: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Band 3, Seiten 1679–1682, Istanbul, Turkey, 2000.

[Schultz 06] T. Schultz, K. Kirchhoff. Multilingual Speech Processing. Elsevier, Academic Press, April 2006.

[Soltau 01] H. Soltau, F. Metze, C. Fügen, A. Waibel. A one-pass decoder based on polymorphic linguistic context assignment. Tagungsband: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Seiten 214–217, Madonna di Campiglio, Trento, Italy, 2001.

[Sooful 01] Jayren J Sooful, Elizabeth C Botha. An acoustic distance measure for automatic cross-language phoneme mapping. Tagungsband: PRASA 2001, Seiten 99–102, South Africa, November 2001.

[Stüker 08] Sebastian Stüker. Integrating thai grapheme based acoustic models into the ml-mix framework - for language independent and cross-language asr. Tagungsband: Proceedings of the first International Workshop on Spoken Languages Technologies for Under-Resourced Languages, SLTU, 2008.

[Stolcke 02] A. Stolcke. Srilm - an extensible language modeling toolkit. Tagungsband: Proceedings of the International Conference on Spoken Language Processing, Seiten 901–904, Denver, Colorado, 2002.

[Trancoso 06] Isabel Trancoso, Ricardo Nunes, Luís Neves, Céu Viana, Helena Moniz, Diamantino Caseiro, Ana Isabel Mata. Recognition of classroom lectures in european portuguese. Tagungsband: Proc. of the 9th International Conference on Spoken Language Processing (Interspeech2006-ICSLP), Seiten 281–284, 2006.

[White 08] Christopher M. White, Sanjeev Khudanpur, James K. Baker. An investigation of acoustic models for multilingual code-switching. Tagungsband: Interspeech

2008, Brisbane, Australia, 2008.

[Wölfel 09] M. Wölfel, J.W. McDonough. Distant Speech Recognition. John Wiley & Sons, March 2009.