Institut für Logik,
Komplexität und Deduktionssysteme
Lehrstuhl: Prof. Waibel

Adaption von Kontextentscheidungsbäumen auf neue Sprachen

Studienarbeit

Ausarbeitung : Roald Wolff
Betreuerin : Tanja Schultz

Inhaltsverzeichnis

1 Zusammenfassung

In dieser Studienarbeit wurde ein multilingualer Spracherkenner aufgebaut, der die Sprachen Kroatisch, Koreanisch, Spanisch, Japanisch und Türkisch umfaßt. Die Phonemmenge wurde so gewählt, daß sowohl ähnliche Laute zwischen den Sprachen gemeinsam modelliert wurden als auch einzigartige Laute einer Sprache ein eigenes Modell bekamen. Insgesamt beinhaltet die Phonemmenge 78 Phoneme. Da ein solches System nicht mehr die sprachspezifischen Eigenschaften berücksichtigen kann, nimmt die Erkennungsleistung im Vergleich zu einem monolingualen System ab. Dieser Verlust in der Leistung fällt aber mit 0.4 - 7.5 % gering aus.
Ausgehend von diesem System wurde versucht, eine Adaption auf die Sprachen Portugiesisch und Deutsch vorzunehmen, wobei im Mittelpunkt der Arbeit die Anpassung des Entscheidungsbaums stand. Da Modelle, die nur wenig Trainingsdaten erhalten, eine schlechte Generalisierungsfähigkeit besitzen, wurde zunächst bestimmt, wieviele Polyphone die einzelnen Modelle aus der Adaptionsmenge erhalten. In dem nächsten Schritt wurden alle Modelle, die weniger Polyphone erhalten als eine festlegbare Schranke, aus dem Erkenner entfernt. Somit wird das System robuster, da der verkleinerten Anzahl von Modellen im anschließendem Training mehr Trainingsmaterial zur Verfügung steht. So ließ sich die Erkennungsleistung um absolut 5,5 % bei der Adaption auf Deutsch und um 8,2 % bei der Verwendung von 100 bzw. um 3,6 % bei der Verwendung von 500 portugiesischen Sätzen erhöhen im Vergleich zu den Systemen, bei denen der Entscheidungsbaum nicht angepaßt wurde.

2 Einleitung

Der Stand der Forschung bei LVCSR-Systemen ist inzwischen schon so weit forgeschritten, daß kommerzielle Produkte entstanden sind. Trotzdem sind die bisher vorhandenen Systeme alle extrem von ihrer Sprache abhängig. Dies führt dazu, daß bei jeder neuen Sprache ein hoher Aufwand an Arbeit zu erledigen ist. Ausreichend Sprachdaten müssen gesammelt und segmentiert werden, aus denen initiale Sprachmodelle und akustische Modelle erstellt werden. Eine geeignete Phonemmenge wird erarbeitet, und Tests ergeben, ob die Wahl der Modelle und Phoneme gut war. Insgesamt dauert dieser Prozeß sehr lange. Deshalb entstand die Idee, mit einer einzigen Phonem- und Modellmenge mehrere Sprachen abzudecken. Als Einsatzgebiet eines solchen multilingualen Spracherkenners, der gleichzeitig mehrere Sprachen erkennen könnte, kämen zum Beispiel Auskunftssysteme in Frage, die von Personen aus verschiedenen Nationen benutzt werden. Ein weiterer Vorteil eines solchen Systems sollte die einfache Portierung auf neue Sprachen sein. Die Verwendung von multilingualen Modellen sollte dazu führen, daß diese Modelle auch bei neuen Sprachen eine gute Schätzung darstellen. Damit soll erreicht werden, daß mit wenig Trainingsmaterial ein Spracherkenner für eine fremde Sprache erzeugt werden kann.

Diese Studienarbeit beschäftigt sich mit der Erstellung eines multilingualen Spracherkenners für die Sprachen Kroatisch, Spanisch, Türkisch, Japanisch und Koreanisch. Außerdem wird anhand der Sprachen Deutsch und Portugiesisch untersucht, wie sich das System auf eine fremde Sprache portieren läßt. Diese Ausarbeitung ist wie folgt aufgebaut: zunächst werden im Kapitel 2 einige Arbeiten anderer Institute vorgestellt. Danach schließt sich das Kapitel mit der Beschreibung des multilingualen Spracherkenners an, und im letztem Kapitel wird die Anpassung des Erkenners an neue Sprachen diskutiert. Dabei wird vor allen Dingen untersucht, wie sich der Entscheidungsbaum an eine neue Sprache anpassen läßt.

3 Verwandte Arbeiten

Im folgenden sollen kurz die Arbeiten und Ergebnisse anderer Forschungseinrichtungen auf dem Gebiet der multilingualen Spracherkennung dargestellt werden. Damit soll gezeigt werden, welche Ansätze bisher verfolgt wurden und wie sich der in dieser Studienarbeit entwickelte Spracherkenner davon unterscheidet. Ähnliche Arbeiten zu den hier vorgestellten findet man bei Gokeen [5], Cohen [6] und Bonaventura [7] zu dem Thema der multilingualen Spracherkenung und bei Chollet [8] zu dem Thema der crosslingualen Adaption.
Köhler untersuchte in [4] die Effizienz von multilingualen Phonemmodellen. Dazu verglich er die Erkennungsleistung eines Systems mit 232 monolingualen Mo-

dellen mit einem System mit 95 multilingualen Modellen. In diesen Erkennern wurden die Sprachen Französisch, Deutsch, Italienisch, Portugiesisch, Spanisch und Englisch berücksichtigt, wobei die Aufgabe darin bestand, in diesen Sprachen einzelne Wörter zu erkennen. Der Wortschatz pro Sprache war dabei auf maximal 70 Wörter beschränkt. Jede Sprache wurde mit durchschnittlich 38 kontextunabhängigen Phonemen modelliert, so daß insgesamt ein System mit 232 Phonemen vorlag. Um Gemeinsamkeiten zwischen diesen Phonemen der verschiedenen Sprachen auszunutzen, wurden die sprachspezifischen Modelle in den zugehörigen IPA-Symbolen zusammengefaßt, wodurch ein multilinguales System mit 95 Phonemen entstand. Im Vergleich zum monolingualen System schnitt dieses System nur durchschnittlich absolut 3.2% schlechter ab. Mit diesem Ergebnis konnte gezeigt werden, daß mit einer kleinen Anzahl sprachenunabhängiger Modelle vergleichbare Ergebnisse wie mit einer großen Anzahl von sprachenspezifischen Modellen erzielt werden kann. In der Arbeit wurde weiter untersucht, inwieweit sich das multilinguale System auf eine fremde Sprache anpassen läßt. Als fremde Sprache wurde Deutsch ausgewählt, wozu die deutschen Sätze aus der Trainingsmenge entfernt wurden. Ausgehend von dem trainierten multilingualen Erkenner wurde dann versucht, eine Adaption für Deutsch anzuwenden, indem die Mittelwerte der Dichtefunktionen der HMMs mit einer MAP-Adaption angepaßt wurden. Dadurch ließ sich die Erkennungsrate um ungefähr 10% relativ steigern.

Wheatley versuchte bei einer Arbeit bei Texas Instruments [9], eine Adaption von Englisch auf Japanisch durchzuführen. Die Aufgabe bestand darin, Zahlenkolonnen und Operatorbefehle, wie sie bei Telefonverbindungen vorkommen, zu erkennen. Damit war die Größe des Vokabulars auf 23 japanische Wörter beschränkt, die durch Word-HMMs modelliert wurden. Um einen Vergleich zwischen dem Adaptionsverfahren und herkömmlichen Verfahren machen zu können, wurden auf verschiedene Arten drei Erkenner erstellt. Bei der ersten Variante wurden alle Modelle identisch mit einem Ruhegeräusch initialisiert, wobei sich die Anzahl der Zustände des HMMs an der Anzahl der Phoneme pro Wort orientierte. Diese Art der Initialisierung der Modelle bereitete am wenigsten Aufwand. Wesentlich mehr Aufwand entstand bei der zweiten Variante. Hier wurde versucht, per Hand Phonemlabels der Sprachdaten zu erzeugen, um somit ein initiales Modell der Phoneme zu erhalten. Diese Modelle sind zwar erfahrungsgemäß sehr gut, aber die Vorgehensweise kostet viel Zeit und Expertenwissen und benötigt eine große Datenmenge als Grundlage. Deshalb bietet sich das dritte Verfahren an, bei dem als Grundlage kontextunabhängige, englische Phonemmodelle benutzt werden (Cross-Lingual-System). Die japanischen Wortmodelle wurden in Phonemsequenzen zerlegt, deren einzelne Phoneme dann mit dem passenden englischen Phonem initialisiert wurden. Die englischen Phoneme wurden also benutzt, um eine gute Anfangsschätzung der japanischen Modelle zu bekommen. In der Arbeit wurden dann die Erkennungsleistungen der 3 Systeme in Abhängigkeit von der Anzahl der Trainingsiterationen gemessen. Dabei erreichte

das Cross-Lingual-System vergleichbare Ergebnisse wie das 2. System. Bei einer hohen Anzahl von Trainingsiterationen wurden sogar leicht bessere Ergebnisse erzielt. Dagegen brauchte das System, dessen Modelle alle mit dem Ruhegeräusch initialisiert waren, sehr viel Training, um etwa gleiche Ergebnisse zu erreichen. Es war aber durchweg schlechter als die beiden anderen Erkenner. Mit dieser Arbeit konnte gezeigt werden, daß die Cross-Lingual-Adaption genauso gute Resultate wie ein nach herkömmlichen Verfahren initialisierter Erkenner liefern kann, wobei sich der Aufwand stark reduziert. Zusätzlich wurde noch untersucht, wie sich eine Reduzierung der Trainingsmenge und eine Reduzierung der Sprecheranzahl auswirkt. Dabei wurde festgestellt, daß eine Vergrößerung der Sprecheranzahl mehr Einfluß auf die Leistung des Systems hat als eine Vergrößerung des Trainingsmaterials. Dieses Resultat bedeutet, daß für die Cross-Lingual-Adaption vorzugsweise mehr Sprecher als mehr Daten pro Sprecher benutzt werden sollten. In einem weiteren Versuch wurden unterschiedliche Zuordnungen von den englischen zu den japanischen Phonemen ausprobiert. Da es bei manchen Zuordnungen mehrere Möglichkeiten gab, je nachdem welche Merkmale des Phonems berücksichtigt wurden, sollten die Auswirkungen dieser Zuordnungen untersucht werden. Es war zu beobachten, daß selbst bei nur 2 Trainingsiterationen mit 1000 Sätzen die unterschiedlichen Zuordnungen etwa gleiche Ergebnisse liefern. Somit sind die Zuordnungen zwar mit Bedacht zu wählen, aber alternative Möglichkeiten haben bei ausreichendem Training keinen Einfluß auf die Leistung eines Systems.
Insgesamt läßt sich sagen, daß im Bereich der multilingualen Spracherkennung hauptsächlich kleine Systeme untersucht worden sind. Die Vokabulargröße war auf wenige Wörter beschränkt, und die Gemeinsamkeiten zwischen Phonemen verschiedener Sprachen wurden nur bei kontextunabhängigen Modellen ausgenutzt. Diese Studienarbeit erweitert diese Ansätze, um einen multilingualen Spracherkenner zu erstellen, der mit einem großem Wortschatz und kontextabhängigen Modellen arbeitet. Dabei orientiert sich diese Arbeit an den Resultaten von Schultz [1].

4 Der multilinguale Spracherkenner

In den folgenden Abschnitten wird der Aufbau des multilingualen Erkenners erläutert. Er umfaßt die 5 Sprachen Kroatisch, Spanisch, Türkisch, Japanisch und Koreanisch. Die Trainingsdaten bestehen aus vorgelesenen Sätzen und in dem Wörterbuch sind insgesamt 89.000 Wörter enthalten. In [2] wurden drei Möglichkeiten vorgestellt, die Phonemmenge eines multilingualen Systems zu erstellen. Zum einen kann jedes Phonem, das in mehreren Sprachen vorkommt, ein eigenes Modell erhalten. Dies führt dazu, daß keine Gemeinsamkeiten zwischen den Lauten der Sprachen ausgenutzt werden. Bei der zweiten Variante erhalten die Phoneme eine Markierung für ihre Sprache. Durch das Hinzufügen von Sprachenfragen beim Clusteralgorithmus wird datengetrieben entschieden, ob Ähnlichkeiten zwischen den Lauten der Sprachen modelliert werden. Bei der dritten und hier verwendete Möglichkeit werden Phoneme, die in mehren Sprachen benutzt werden, zu einem gemeinsamen Modell vereinigt. Ein solches System besitzt eine kleinere Phonemmenge und ist damit kompakter.

4.1 Die Datenbasis

Die Sprachdaten für den multilingualen Spracherkenner wurden aus der bestehenden GlobalPhone-Datensammlung [10] genommen. Diese Datensammlung umfaßt die Daten von 1200 Sprechern aus 13 Ländern und wird seit 1996 ständig erweitert. Dafür werden Sprecher ausgewählt, die in ihrem Heimatland leben und die Sprache als Muttersprache beherrschen. Sie haben die Aufgabe, Texte vorzulesen, die aus den Bereichen Politik und Wirtschaft stammen. Der Vorteil von solchen Themen liegt darin, daß hier genug Textdaten zu Verfügung stehen, um ein gutes Sprachmodell zu schätzen. Zu den Transkriptionen der Sprachdaten wurden zusätzlich noch Markierungen für Hintergrund- und artikulatorische Geräusche hinzugefügt.
Für den zu entwickelnden multilingualen Spracherkenner stand eine Datenbasis mit insgesamt 306 Sprechern aus den 5 Ländern zur Verfügung. Deren Verteilung in Sprache und Geschlecht kann aus Tabelle 1 entnommen werden.

Sprache	Geschlecht	# Sprecher	# Sätze
Japanisch	F	9	800
	M	53	4670
Spanisch	F	35	3210
	M	47	2413
Kroatisch	F	39	1788
	M	24	1128
Koreanisch	F	8	711
	M	13	877
Türkisch	F	56	3973
	M	22	1440
Insgesamt		306	21010

Tabelle 1: Datenbasis des multilingualen Spracherkenners

4.2 Phonemset

Die Auswahl der zu modellierenden Phoneme sollte mit großer Sorgfalt geschehen, wobei sich zwei Gütekriterien gegenüberstehen. Auf der einen Seite sollen

möglichst viele charakteristische Eigenschaften einer Sprache übernommen werden. Diese Vorgehensweise wäre optimal erfüllt, falls für jede Sprache getrennt voneinander eine eigene Phonememenge gebildet wird. Als Nachteil würde sich dabei herausstellen, daß eine große Anzahl von Phonemen vorläge, die alle zu modellieren wären, und daß bei der Anwendung einer neuen Sprache keine multilingualen Modelle vorlägen. Stattdessen sollten Gemeinsamkeiten zwischen Phonemen verschiedener Sprachen ausgenutzt werden, indem ein gemeinsames Modell für dieses Phonem gewählt wird. Auf diese Weise wird das System kompakter, und die Modelle sollten auch für neue Sprachen gut generalisieren können.
Bei dem Auswahlprozeß der Phoneme muß also überlegt werden, ob die Ähnlichkeit zwischen zwei Sprachen groß genug ist oder ob ein monolinguales Phonem gewählt wird, um die Einzigartigkeit dieses Lautes modellieren zu können. Eine solche Einteilung der Phoneme liegt bereits in dem IPA-Schema vor, das auf den multilingualen Spracherkenner angewandt werden kann. Die resultierende Phonemmenge ist in Tabelle 2 dargestellt. Sie besteht ingesamt aus 78 Phonemen plus ein Modell für Stille und zwei Modelle für artikulatorische Geräusche.

4.3 Kontextabhängigkeit

Lee konnte zeigen, daß die Erkennungsleistung eines Spracherkenners gesteigert werden kann, falls man den Kontext eines Phonems mitberücksichtigt. In [3] prägte er den Begriff des **generalisierten Triphons**. Es ist zu erwarten, daß auch bei einem multilingualen Spracherkenner durch Berücksichtigung des Kontextes ein Fortschritt zu erzielen ist.

4.3.1 Allgemein

Der Vorteil bei der Verwendung von Phonemen als Spracheinheit liegt in ihrer robusten Trainierbarkeit. Für die Modellierung einer Sprache (z.B. Englisch) reichen 40-50 Phoneme aus, die in den Trainingsdaten (mehrere hunderte Sätze) in ausreichender Anzahl vorkommen. Als Nachteil stellt sich aber ihre Inkonsistenz dar; denn es wird vorausgesetzt, daß ein Phonem in jedem möglichen Kontext gleich ausgesprochen wird. Dies ist aber nicht der Fall, da die Aussprache eines Phonems innerhalb eines Wortes nicht unabhängig von seinen Nachbarphonemen ist. Dies liegt vor allen Dingen daran, daß sich der Vokaltrakt mit seinen Artikulatoren (Lippen, Zunge, Gaumen) nur langsam verändert, so daß Koartikulationseffekte entstehen. So wird z.B. das Phonem /R/ in den beiden Wörtern “Burg“ und “Arbeit“ unterschiedlich ausgesprochen. Falls nur ein Modell für diese beiden Varianten von dem Phonem benutzt werden, so ist dieses zu ungenau. Stattdessen können Modelle benutzt werden, die die direkten Nachbarphoneme berücksichtigen: **Triphone**. Dadurch ständen statt des einen kontextunabhängigen Phonems /R/ zwei kontextabhängige Phoneme R(u,g) und R(a,b) zur Verfügung. Die Genauigkeit der Modelle läßt sich weiter erhöhen, indem nicht nur der Kontext

Phoneme (Worldbetsymbole)	KOR	SPA	KRO	TUR	JAP	$\sum$
n,m,s,l,TS,p,b,t,d,g,k	X	X	X	X	X	
i,e,o	X	X	X	X	X	14
f,j,z		X	X	X	X	
r,u	X	X	X	X		
dZ	X		X	X	X	6
a	X	X	X			
S			X	X	X	
h	X			X	X	
4	X	X			X	4
n,x,L		X	X			
A				X	X	
N	X	X				
V,Z			X	X		
y,7	X			X		
ts			X		X	10
p’,t’,k’,dZ’,s’,oE,oa,4i,uE	X					
E,u,iu,uu,iu,ie,io,ia	X					17
D,G,T,V,r(		X				
ai,au,ei,eu,oi,a+,e+,i+,o+,u+		X				15
palatal c , palatal d			X			2
ix, weichzeichen				X		2
?,Nq, V[,A:,e:,i:,O:,4:					X	8
Monolingual	40	40	30	29	31	
Multilingual						78

Tabelle 2: Phonemmenge des multilingualen Spracherkenners

in direkter Nachbarschaft berücksichtigt wird, sondern auch die Phoneme in einem Abstand von 2 (Quintphone), 3 (Septphone) oder im allgemeinen Fall von variablen Abstand nach links und rechts (Polyphone), wodurch sich Koartikulationseffekte exakt modellieren lassen. Das Problem liegt hierbei aber in der hohen Anzahl der Modelle, die zuviel Speicherplatz beanspruchen würden und aufgrund des begrenzten Trainingsmaterials nicht genau genug geschätzt werden könnten. Wird z.B. ein kontextunabhängiges System mit 40 Phonemen benutzt, so bräuchte bei einer Kontextbreite von 2 das zugehörige kontextabhängige System maximal 40^5 Modelle. Sehr viele von diesen Modellen werden in der Trainingsmenge gar nicht oder nur selten auftreten und stellen deshalb keine robusten Modelle dar. Zwar läßt sich durch Interpolation mit Modellen kleinerer Kontextbreite die Robustheit erhöhen, doch wegen des enormen Speicherverbrauchs werden Tripho-

ne in dieser Form bei der Spracherkennung nicht benutzt.
Die Lösung für das Problem der robusten Kontextmodellierung ohne zu hohen Speicherverbrauch liegt in der Verwendung der **generalisierten Triphone**. Die wichtigste Erkenntnis liegt in der Beobachtung, daß bei einigen Phonemen die gleichen Koartikulationseffekte zu einem Nachbarphonem entstehen. So wird in [3] festgestellt, daß sich z.B. die Phoneme /b/ und /f/ auf einen rechts anschliessenden Vokal gleich auswirken. Diese Ähnlichkeiten werden ausgenutzt, indem geeignete Gruppen gleicher Triphone zu einer gemeinsamen Klasse zusammengefaßt werden. Die resultierenden Spracheinheiten heißen verallgemeinerte oder generalisierte Triphone [3] und führen zu einen System mit wesentlich weniger Modellen, für die deshalb auch mehr Trainingsdaten zur Verfügung stehen. Eine Möglichkeit, ähnliche Kontexte zusammenzufassen, besteht darin, die Daten per Hand zu untersuchen und die Klassen zu bestimmen. Zwar wären auf diese Weise die erzeugten Klassen optimal, aber der Aufwand wäre dafür zu hoch. Stattdessen werden Verfahren angewendet, die automatisch die beste Ballung für die Daten findet. Für diese Aufgabe gibt es zwei Varianten: die agglomerative und die divisive Ballung.

Agglomerative Ballung Das agglomerative Verfahren ist eine bottom-up Prozedur, die eine Schätzung von gewöhnlichen Triphonmodellen voraussetzt. Zu Beginn wird jedes Ballungsgebiet mit einem Triphon initialisiert, wobei auf der Menge aller Modelle ein Abstands- oder Ähnlichkeitsmaß definiert wird, und eine Häufigkeitsanalyse durchgeführt wird. Die beiden Gebiete eines selben Phonems mit dem kleinsten Abstand werden so lange zu einem gemeinsamen Ballungsgebiet zusammengeschlossen bis z.B. die gewünschte Anzahl von Modellen erreicht ist (siehe Abb. 1). Die zu einem Ballungsgebiet gehörenden Triphone werden dann in der Haupttrainingsphase einem gemeinsamen Markovmodell zugeordnet, in dessen Schätzung dann die Vereinigungsmenge aller ihrer Aussprachebeispiele eingeht. Der komplette Algorithmus sieht wie folgt aus:

1. Erzeuge ein HMM für jedes Triphon.
2. Erzeuge Ballungsgebiete und initialisiere jedes mit einem Triphon.
3. Finde die beiden ähnlichsten Ballungsgebiete und vereinige sie.
4. Verschiebe die Elemente von zwei Ballungsgebieten:
   (a) Führe Verschiebung durch, falls die Distanz sich verbessert.
   (b) Wiederhole, bis keine Verschiebungen übrigbleiben
5. Falls Konvergenzkriterium erreicht, dann Ende, sonst zurück zu Schritt 2

Schritt 4 stellt eine heuristische Optimierung dar, die es erlaubt, daß Ele-

mente von einem Ballungsgebiet zu einem anderen Ballungsgebiet wechseln dürfen. Dieser Schritt ist zwar sehr zeitintensiv, führt aber zu besseren Ergebnissen. Um die Ähnlichkeit zwischen zwei HMMs zu bestimmen, wird meistens der Informationsverlust gemessen, der bei der Verschmelzung der beiden Modelle entstehen würde.
$Entropiedistanz = (G_p + G_q)H(p \oplus q) - (G_pH(p) + G_qH(q))$
Die Entropie $H(p)$ eines Modells p wird berechnet als $H(p) = -\sum_x p(x)ld(x)$, wobei x ein Laufindex über die Werte der Gewichtungen der Gaussians eines Modells ist. Die Entropiewerte werden mit den Gewichten G_p und G_q multipliziert, die sich aus der relativen Häufigkeit der Modelle in der Trainingsmenge (berechenbar mit dem Forward-Backward-Alg.) ergeben.

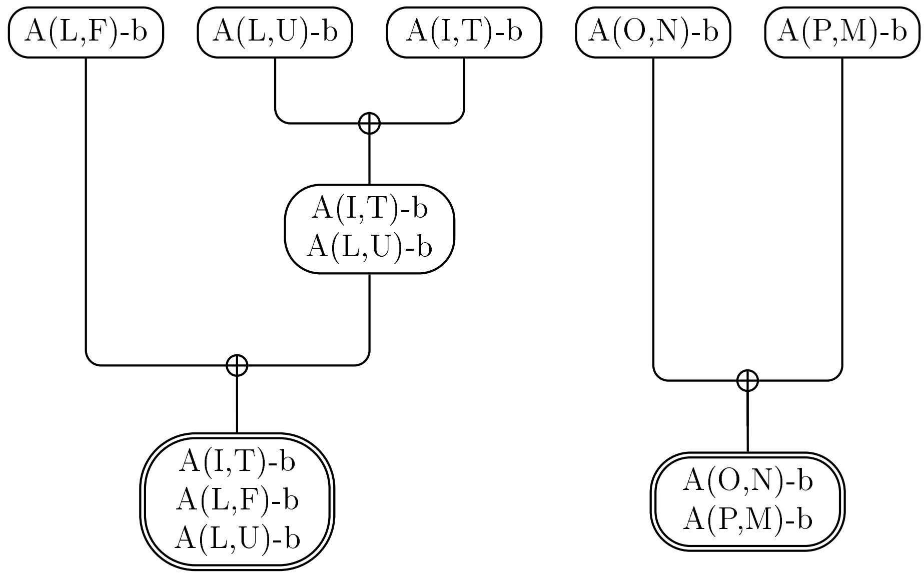

Abbildung 1: Agglomeratives Ballungsverfahren

Divisive Ballung Die zweite Variante der Kontextballung ist ein Top-Down Verfahren, bei dem am Anfang für jedes Phoneme ein Ballungsgebiet erstellt wird, das mit allen möglichen Kontexten initialisiert wird. Anhand eines Fragenkatalogs wird das Ballungsgebiet dann aufgeteilt, wodurch zwei neue Klassen entstehen. Dieser Fragenkatalog benötigt aber Hintergrundwissen, so daß die divisive Ballung nicht rein datengetrieben wie das agglomerative Verfahren laufen kann. Dieses Wissen beschränkt sich auf das Bilden von Phonegruppen, für die angenommen wird, daß sie gleiche Koartikulationseffekte ausüben. Typische Beispiele für solche Gruppen sind die Gruppen der Vokale, Nasale und Frikative. Der Fragenkatalog besteht nun aus Fragen, ob der linke/rechte Kontext ein Element aus einer bestimmten Phonegruppe ist. Für jedes Ballungsgebiet wird die optimale Frage gesucht, indem die Frage mit dem größten Informationsverlust gesucht wird, die dann das Ballungsgebiet in zwei neue Klassen aufteilt. Auf diese Weise ensteht ein sogenannter Entscheidungsbaum, wobei die inneren Knoten Fragen und die Blätter die resultierenden Modelle darstellen (siehe Abb. 2).

1. Initialisiere eine Klasse mit allen Modellen.
2. Berechne den Informationsverlust für alle Fragen.
3. Führe die beste Auftrennung durch.
4. Solange Endkriterium nicht erreicht, gehe zu Schritt 2.

Genauso wie beim agglomerativen Verfahren kann als Endkriterium die gewünschte Anzahl der Modelle dienen. Trotzdem muß zusätzlich beachtet werden, daß bei einer Auftrennung die sich ergebenden, neuen Ballungsgebiete genug Trainingsdaten bekommen. Ansonsten bestände die Möglichkeit, daß die Ballungsgebiete soweit aufgeteilt werden, daß normale Triphone vorlägen. Damit läge das Problem der Robustheit wieder vor.

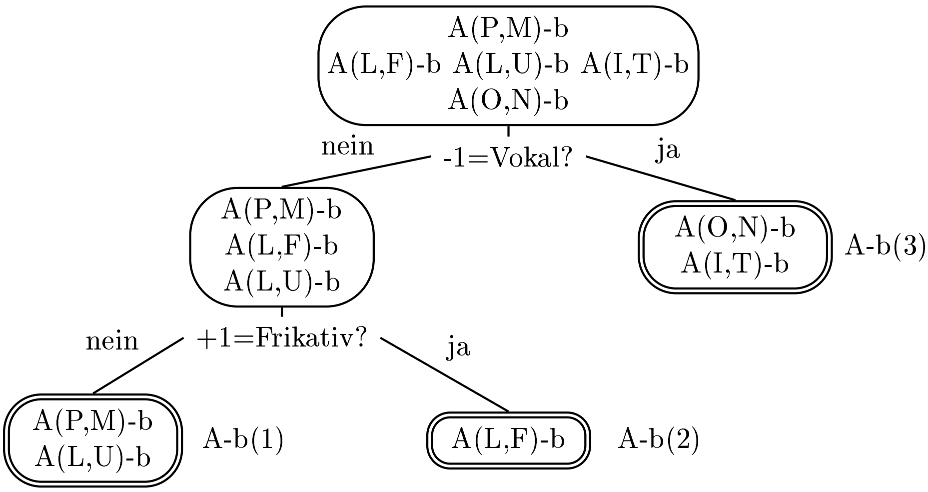

Abbildung 2: Divisives Ballungsverfahren

In [3] wurden die beiden Verfahren miteinander verglichen, wobei beide Methoden gleiche Ergebnisse lieferten. Ein Vorteil des divisiven Verfahren ist es, neue, während des Trainings nicht vorgekommende, Kontexte gut modellieren zu können. Durch Prüfen der Fragen an den Kontext des neuen Polyphons läßt sich der Entscheidungsbaum herabschreiten. Diese Suche endet schließlich in einem Blatt, welches das ähnlichste Modell für dieses neue Polyphon darstellt.
Der JANUS [11] unterstützt ein divisives Ballungsverfahren, welches auch Kontexte über Wortgrenzen hinaus berücksichtigen kann. Dadurch entstehen zwar

eine Vielzahl weiterer Kontexte, aber gerade an Wortübergängen finden oft Verschleifungen statt, die auf diese Weise modelliert werden können. Außerdem werden beim JANUS nicht generalisierte Polyphone sondern sogenannte Senones erzeugt. Dafür wird jedes Phonem des multilingualen Spracherkenners mit einem Links-Rechts-HMM, das aus 3 Zuständen besteht, modelliert. Beim Ballungsverfahren werden nicht die gesamten HMMs zusammengefaßt sondern nur dieselben Subpolyphone (Anfangs-, Mitte-, Endzustand) der HMMs. Die so entstehenden generalisierten Subpolyphone werden Senones genannt.
Der multilinguale Spracherkenner verwendet eine Kontextbreite von 2, wobei die ähnlichen Kontexte mit einem divisiven Ballungsverfahren zusammengefaßt wurden. Der Fragenkatalog wurden noch zusätzlich mit Fragen nach den Sprachen erweitert.

4.3.2 Sprachenfragen

Bisher sorgt die Auswahl der Phoneme dafür, daß ähnliche Laute zwischen den Sprachen gemeinsam modelliert werden. Durch die Verwendung mehrerer Sprachen treten sehr viele, unterschiedliche Kontexte auf, die während des Ballungsvorgangs in die ähnlichsten Klassen vereint werden. Da aber gleiche Kontexte in den Sprachen verschieden ausgesprochen werden können, kann es notwendig sein, für bestimmte Polyphone ein eigenes monolinguales Modell zu erstellen. Deshalb wurden zu dem Fragenkatalog zusätzlich Fragen nach den Sprachen und nach Sprachengruppen hinzugefügt, um zu gewährleisten, daß sowohl Kontexte, die speziell für eine Sprache sind, als auch Kontexte, die typisch für eine Teilmenge der verwendeten Sprachen sind, herausgefiltert werden können. Es entsteht dabei aber eine Inkonsistenz, da z.B. die Frage nach Türkisch zwar alle türkischen Phoneme enthält, aber diese bis auf 2 Phoneme auch von den anderen Sprachen benutzt werden.

In der Tabelle 3 sind die 20 häufigsten Fragen dargestellt, die nach 500, 1000, 1500 und 3000 Auftrennungen vorkommen. Dabei ist eine Frage, die früher im Entstehungsprozeß gestellt wird, bedeutsamer, da ihre Anwendnung an den Entscheidungsbaum zu einem größeren Entropieverlust führt. Es läßt sich zwar aus der Tabelle erkennen, daß die Sprachenfragen verwendet werden, aber eine Bevorzugung einer Sprachengruppe wird nicht sichtbar. Die am Anfang häufig gestellte Frage nach allen fünf Sprachen differenziert die Hintergrundgeräusche von den übrigen Lauten und die Frage nach der Sprachengruppe KO+TU+SP+JA enspricht der Frage nach den beiden einzigen, monolingualen, kroatischen Phoneme zusammen mit dem Ruhegeräusch. Ansonsten sorgt die Frage nach der Sprachengruppe TU+JA für einen hohen Entropieverlust, was ein Indiz dafür sein könnte, daß diese beiden Sprachen ähnliche Kontexte besitzen.

#	500 Fragen	#	1000 Fragen	#	1500 Fragen	#	3000 Fragen
63	wb	152	wb	231	wb	454	wb
33	back-vow	57	back-vow	79	back-vow	141	back-vow
23	front-vow	40	vowel	53	vowel	96	alveodental
20	vowel	35	front-vow	46	consonant	93	consonant
18	unvoiced	34	unvoiced	43	front-vow	93	vowel
18	frik-sibilanten	27	consonant	41	unvoiced	92	JAPANESE
17	open-vow	25	TURKISH	37	JAPANESE	83	front-vow
16	Ko_Tu_Sp_Ja	24	voiced	36	alveodental	82	CROATIAN
15	nasal	23	TU_JA	35	CROATIAN	81	nasal
14	plos-unvoiced	23	frik-sibilanten	35	TURKISH	80	unvoiced
13	consonant	23	JAPANESE	35	plos-unvoiced	67	TU_JA
13	Kr_Ko_Tu_Sp_Ja	23	nasal	34	TU_JA	67	TURKISH
12	TU_JA	22	plos-unvoiced	33	frik-sibil	67	close-vow
12	voiced	22	plosive	33	voiced	66	KOREAN
11	KOREAN	22	open-vow	32	close-vow	62	voiced
11	fric-unvoiced	21	Ko_Tu_Sp_Ja	31	nasal	55	KR_TU
10	m_i	20	CROATIAN	30	open-vow	55	plos-unvoic
10	plosive	18	Kr_Ko_Tu_Sp_Ja	28	plosive	53	KR_KO
10	vow-o	18	alveodental	27	KR_TU	49	plosive
9	KR_TU	18	close-vow	25	Ko_Tu_Sp_Ja	46	KR_JA

Tabelle 3: Häufigkeit der Fragen nach 500, 1000, 1500 und 3000 Auftrennungen des Entscheidungsbaums

4.4 Ergebnisse

Um die Leistungsfähigkeit des multilingualen Spracherkenners zu prüfen, wurde für jede Sprache ein Test durchgeführt. Die Testmengen enthielten 97 türkische, 94 japanische, 39 spanische, 102 koreanische und 52 kroatische Sätze, die nicht in der Trainingsmenge enthalten waren. In jedem der 5 Tests wurde ein sprachspezifisches Wörterbuch verwendet, welches für Türkisch 10.700, für Japanisch 22.100, für Spanisch 18.700, für Koreanisch 5.300 und für Kroatisch 10.900 Wörter umfaßte.
In Tabelle 4 sind die Wortfehlerraten der einzelnen Tests dargestellt. Dabei sind die Erkennungsleistungen für Türkisch, Spanisch, Japanisch und Kroatisch sehr erfreulich, während die Leistung für Koreanisch stark abfällt. Der Grund für die schlechte Erkennungsleistung für Koreanisch liegt vermutlich in einer Inkonsistenz der Datenbasis. Die zugrundeliegenden Sprachdaten wurden während der Studienarbeit verändert und neu organisiert, so daß sie nicht mehr vollständig mit den Transkriptionen übereinstimmten. Außerdem ist die Anzahl der koreanischen Trainingssätze deutlich geringer als die Anzahl der Sätze, die den anderen Sprachen zum Training zur Verfügung stehen. Für alle Sprachen standen auch

monolinguale Systeme zur Verfügung, deren Wortfehlerraten der entsprechenden Tests auch in Tabelle 4 abgebildet sind. Im Vergleich mit den Ergebnissen des multilingualen Systems sind die Werte der monolingualen Erkenner alle besser, da sie schließlich die Eigenschaften der einzelnen Sprachen exakt modellieren können. Aber der Verlust durch den Einsatz der multilingualen Modelle ist nicht hoch. Im besten Fall beträgt der Abstand zwischen mono- und multilingualen System nur 0.4 % (Spanisch) und im schlechtesten Fall nur 7.5 % (Japanisch).
Damit konnte gezeigt werden, daß es möglich ist, ohne großen Verlust in der

System	KRO	SPA	TUR	JAP	KOR
3000	30.5	28.0	22.2	20.5	65.4
7500	30.4	28.8	20.1	17.2	65.5
Mono	26.9	27.6	20.1	13.0	47.3

Tabelle 4: Wortfehlerrate des multilingualen Systems mit 3000 und 7500 Modellen

Erkennungsleistung ein kontextabhängiges, multilinguales System zu erstellen.
In einem weiteren Versuch sollte untersucht werden, ob die vorgegebene Anzahl der Modelle den Clusteralgorithmus zu sehr beschränkt, so daß Modelle entstehen, die zu viele Eigenschaften der verschiedenen Sprachen berücksichtigen müssen. Deshalb wurde neben dem System mit 3000 Modellen ein weiteres System mit 7500 Modellen erstellt. Dies entspricht der Anzahl von Modellen, die bei den monolingualen Systemen benutzt werden (5x1500=7500). Bei so einem großen Erkenner können die Modelle so weit differenziert werden, daß sie monolingual werden, d.h. sie bekommen nur noch Daten von einer einzigen Sprache. Die Wortfehlerraten der Tests für dieses System sind in Tabelle 4 dargestellt. Dabei läßt sich sagen, daß sich der Aufwand, mehr Details der einzelnen Sprachen zu modellieren, nicht lohnt. Die Erkennungsleistung für Spanisch liegt unter den Werten des Systems mit 3000 Modellen, für Koreanisch und Kroatisch sind die Werte fast identisch und für Türkisch steigert sich die Leistung noch um absolut 2.1 % und für Japanisch um 3,3%. Durch die Vielzahl der Modelle ist anzunehmen, daß sich die Akustik zwischen den Modellen zu wenig unterscheidet und deshalb während der Suche sehr leicht ähnliche Modelle miteinander verwechselt werden.

5 Adaption neuer Sprachen

Im folgenden soll getestet werden, wie gut sich der multilinguale Spracherkenner bei neuen Sprachen verhält. Da die Modelle teilweise über mehrere Sprachen trainiert werden und damit sprachenunabhängig sind, sollte der Spracherkenner auch auf Sprachen funktionieren, von denen er bisher keine Daten gesehen hat.

Außerdem soll die Frage untersucht werden, wieviele Trainingssätze notwendig sind, um den Erkenner auf die neue Sprache zu spezialisieren. Die Tests wurden auf den beiden Sprachen Deutsch und Portugiesisch durchgeführt.

5.1 Training auf neuer Sprache

Von den 78 multilingualen Phonemen werden 35 bei Deutsch und 46 bei Portugiesisch benutzt. Für Deutsch stand bereits ein Mapping von den monolingualen auf die multilingualen Phoneme zur Verfügung, während für Portugiesisch das Mapping selber durchgeführt wurde. Die Testmengen bestehen aus 100 Sätzen von 3 Sprechern bei Deutsch und aus 96 Sätzen von 3 Sprechern bei Portugiesisch. Falls das System ohne Training die neuen Sprachen erkennen soll, so sind die Erkennungsleistungen sehr schlecht: 15% bzw. 31.2% (PO/DE).
Die Frage ist nun, wie gut sich der Erkenner mit Hilfe von Training auf die neue Sprache anpassen läßt, wobei als Vergleich ein monolinguales System dient. Dabei sollen natürlich möglichst wenig Daten verwendet werden, und der Aufwand soll auch möglichst gering sein. Deshalb wurde der Erkenner mit 3 Iterationen Viterbitrainings auf 100 und 500 portugiesischen Sätzen (7 bzw. 8 Sprecher) entlang von Labels, die von dem multilingualen System erzeugt wurden, trainiert. Für die Anpassung an Deutsch konnte nur eine Trainingsmenge von 100 Sätzen gewählt werden, da ansonsten das Vokabular hätte erweitert werden müssen. Zusätzlich stimmen die 3 Sprecher aus der Trainingsmenge mit denen aus der Testmenge überein, so daß die Modelle nicht mehr sprecherunabhängig sind. Die Ergebnisse dieser Adaption sind in der Tabelle 5 zusammengefaßt. Die Erkennungsraten von allen Systemen konnte gesteigert werden, wobei sichtbar ist, daß 100 Sätze als Trainingsmaterial nicht ausreichend sind. Zwar sorgt eine Vergrößerung auf 500 Sätze für eine entsprechende Steigerung der Leistung, aber insgesamt sind die Ergebnisse nicht vergleichbar mit den Ergebnissen der monolingualen Systeme. Außerdem ist dieser Ansatz nur dadurch erweiterbar, indem mehr Trainingssätze verwendet werden. Dies widerspricht aber der Forderung, eine Adaption auf eine neue Sprache mit möglichst wenig Trainingsmaterial durchzuführen.

Sprache/Sätze	0	100	500	Monolingual
Portugiesisch	15	34.9	53.2	63.2
Deutsch	31.2	69.0	-	80.0

Tabelle 5: Erkennungsleistung nach Training auf neue Sprachen

5.2 Adaption der Kontextabhängigkeit

Das Problem bei der Portierung des multilingualen Spracherkenners auf eine neue Sprache sind die kontextabhängigen Modelle, da in einer neuen Sprache gänzlich neue Kontexte vorkommen bzw die relative Häufigkeit eines ähnlichen Kontextes verschieden ist. Dagegen kann bei einem kontextunabhängigen System dieses Problem nicht auftreten, da in das akustische Modell eines Phonems sehr viele, unterschiedliche Aussprachevarianten eingehen, so daß die Generalisierungsfähigkeit auch für eine neue Sprache erhalten bleibt. In Tabelle 6 ist dargestellt, wieviele Polyphone aus den Adaptionsmengen in der Trainingsmenge des multilingualen Spracherkenners vorkommen. Da von 765.000 Polyphonen des multilingualen Erkenners weniger als 0.1% abgedeckt werden, wird deutlich, daß die vorhandenen, multilingualen Modelle nur wenig geeignet sind, die neuen Kontexte zu

System	DE100	PO100	PO500
Insgesamt	27.492	41.313	185.565
# Überdeckungen	2.982	3.378	7.431
% Überdeckung	0.38%	0.44%	0.97%

Tabelle 6: Anzahl der Polyphone in den Adaptionsmengen und die Anzahl der Überdeckungen mit der Trainingsmenge des multilingualen Systems

repräsentieren. Da die Menge aller Polyphone die Basis darstellt, um den Entscheidungsbaum zu erstellen, und durch sie festgelegt wird, welche Kontexte ein eigenes, akustisches Modell bekommen, würde nur eine hohe Überdeckungsrate zwischen der Polyphonemenge der neuen Sprache und der Polyphonemenge des multilingualen Erkenners gewährleisten, daß der Entscheidungsbaum auch für die neue Sprache brauchbar ist.
Falls ein kontextabhängiges System bei einem Wechsel des Tasks z.B. eine neue Sprache mit unbekannten Kontexten arbeiten muß, so ist zu erwarten, daß die Leistungsfähigkeit des Systems abnimmt. Denn es ist denkbar, daß ein neuer Kontext durch ein Modell repräsentiert wird, das nicht genug differenziert ist. D.h. dieses Modell muß auf dem neuen Task sehr viele, unterschiedliche Kontexte vertreten, womit es ein schlechtes Modell ist, da es zu allgemein ist. Auf der anderen Seite kann es vorkommen, daß ein Modell nur sehr wenig trainiert wird, da der spezielle Kontext in der neuen Sprache nur noch selten bzw. gar nicht mehr vorkommt. So ein Modell stellt eine schlechte Schätzung dar, so daß es besser aufgelöst wird, und der Kontext durch ein allgemeineres Modell dargestellt wird. Es wird versucht, dieses Problem der schlechten Robustheit mancher Modelle durch folgendes Verfahren zu lösen.

Die relative Häufigkeit der Modelle innerhalb der Adaptionsmenge der neu-

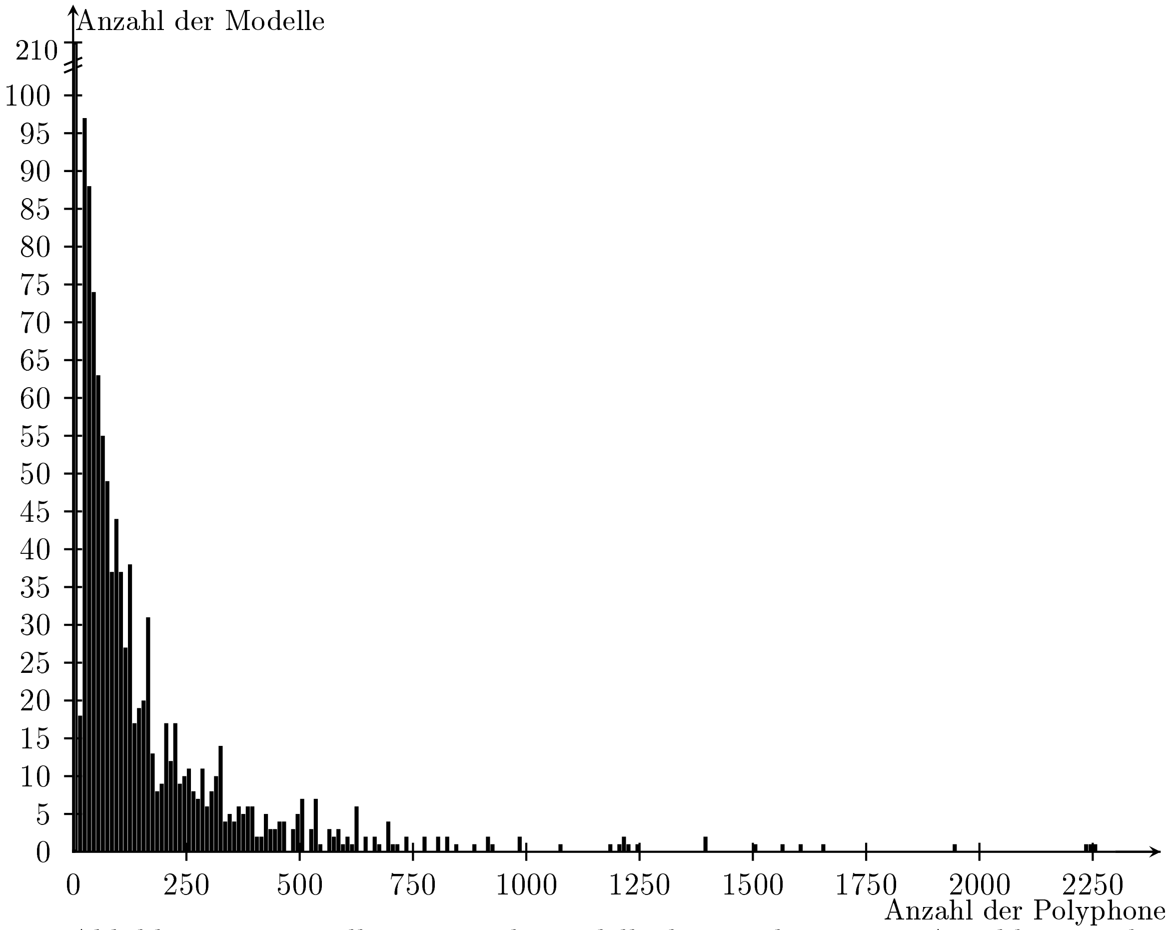

Abbildung 3: Darstellung wieviele Modelle,die eine bestimmte Anzahl von Polyphonen aus der Adaptionsmenge erhalten, vorkommen

en Sprache stellt die Grundlage für die Verkleinerung des Entscheidungsbaumes und das damit verbundene Auffinden von allgemeineren Modellen dar. Dazu wird gezählt, wieviele Trainingsdaten die einzelnen Modelle bekommen. In der Abbildung 3 ist die absolute Häufigkeit der Modelle, die eine bestimmte Anzahl von Polyphonen erhalten, dargestellt. Dabei ist besonders die hohe Anzahl der Modelle auffällig, die nur wenige Polyphone erhalten. So viele Modelle mit dieser Eigenschaft gibt es im Fall des multilingualen Erkenners nicht. Die Bestimmung der Häufigkeiten ist möglich, indem jedem Modell ein sog. PTree angehängt wird, in dem alle zu dem Modell zugehörigen Kontexte gesammelt werden. Über einen Zähler des PTrees ist feststellbar, ob ein Modell in der neuen Sprache viele oder wenige Daten zum Training zur Verfügung gestellt bekommt. Alle diejenigen Modelle, deren Zähler unter einer festgelegten Schranke liegen, werden nun aufgelöst. Die Wahl der Schranke legt fest, inwieweit die Modelle verallgemeinert werden. Wird sie sehr niedrig gewählt, so werden nur die sehr selten benutzten Modelle aufgelöst, wird sie sehr hoch angesetzt, so resultierten daraus kontextunabhängigen Modelle. Die Anzahl der Modelle in Abhängigkeit der Schranke ist in Tab. 4 dargestellt. Der größte Sprung in der Anzahl der ursprünglichen 3000 Modelle ist am Anfang, danach fällt die Anzahl stetig ab, wobei bei hohen Schranken die Abnahme geringer wird. Damit ein kontextunabhängiges System entsteht, wurde die Schranke auf den sehr hohen Wert von 5000 gesetzt.

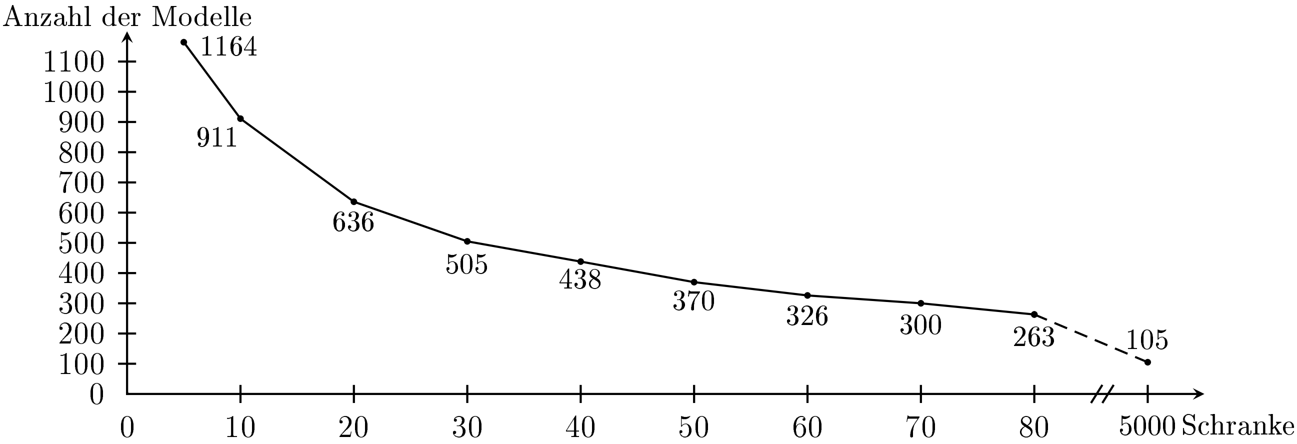

Abbildung 4: Anzahl der Modelle in Abhängigkeit der Schranke für das System DE100

Die Trainingsdaten eines aufgelösten Modells stehen bei einem anschliessenden Training den übrigen Modellen zur Verfügung und sorgen so für eine bessere Robustheit des Systems. Außerdem muß der Entscheidungsbaum angepaßt werden. Die zugehörige Frage, die zu dem Modell verzweigte, kann aufgelöst werden, und die im PTree des Modells gesammelten Kontexte werden auf die übrigen Modelle verteilt (siehe Abb. 5). Da die Werte der Mixturgewichtverteilung der Gausians und die Kovarianzmatizen der Gaussverteilungen eines aufgelösten Modells nicht verlorengehen sollen, wird eine Interpolation durchgeführt.

$$M_1 \cup M_2 = \frac{1}{\lambda_1+\lambda_2}(\lambda_1 M_1 + \lambda_2 M_2)$$

Hierbei bedeutet M_1 das aufzulösende Modell und M_2 ein Modell mit dem Zähler λ_2, das aus dem PTree von M_1 λ_1 viele Kontexte erhält. Durch die Interpolation wird gewährleistet, daß ein aufgelöstes Modell gewichtet in die Schätzung der neuen Modelle eingeht. Die so gewonnenen Modelle dienen als Startwerte für das anschließende Training auf der Adaptionsmenge.

Insbesondere gibt es einige multilinguale Phoneme, die in der neuen Sprache nicht verwendet werden. So werden von den ursprünglichen 78 Phonemen im Deutschen 43 und im Portugiesischen 32 Phoneme nicht benutzt. Da die zugehörigen Modelle kein Trainingsmaterial erhalten, werden auch sie während des Verfahrens aus dem Erkenner herausgenommen. Dieser Vorgang ist sinnvoll, da diese Modelle nicht mehr benötigt werden und das System auf diese Weise kom-

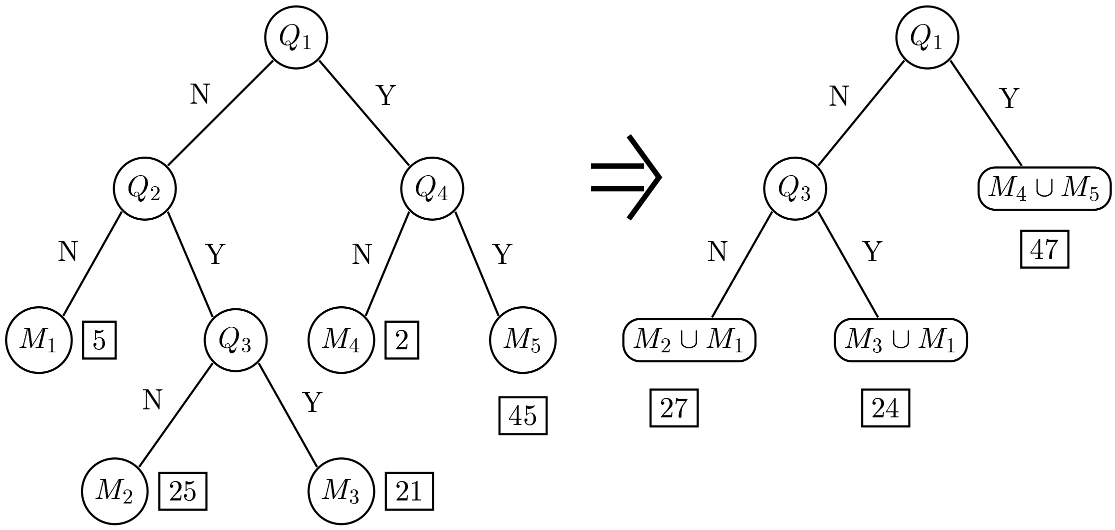

Abbildung 5: Auffinden von allgemeineren Modellen

pakter, schneller bei der Suche und auch zuverlässiger wird.

5.3 Ergebnisse

Das angepaßte System sollte bessere Ergebnisse liefern als ein System, das nur mit den neuen Daten trainiert wurde. Denn die neu entstandenen, allgemeineren Modelle passen besser auf die neuen Kontexte. Es ist zu erwarten, daß die Erkennungsleistung in Abhängigkeit von der Schranke zunächst zunimmt. Dieser Effekt ist darauf zurückzuführen, daß eine größere Schranke für immer allgemeinere Modelle sorgt. Der Anstieg läßt sich aber nicht beliebig steigern, da bei einer zu hohen Schranke ein Modell für zu viele verschiedene Kontexte verantwortlich ist, so daß die Erkennungsrate wieder abnimmt. Wird die Schranke extrem hoch angesetzt, so resultiert daraus letztendlich ein kontextunabhängiges System, welches keine Koartikulationseffekte mehr berücksichtigt und somit von der Leistung schlechter sein sollte. In der Abbildung 6 sind die Ergebnisse des Verfahrens auf Deutsch bei 100 Adaptionssätzen dargestellt. Dabei entspricht der Graph ungefähr dem gewünschten, theoretischen Verlauf. Die Erkennungsleistung von 69.0 % des unangepaßten Systems steigert sich kontinuierlich bis auf einen Wert von 74,5 % bei einer Schranke von 70. Danach fällt sie wieder ab und erreicht bei einem kontextunabhängigen System nur noch einen Wert von 63.2 % . Diese Resultate zeigen, daß mit dem Verfahren eine Steigerung der Erkennungsleistung von 7.9% auf einer neuen Sprache möglich ist. Außerdem zeigen die schlechten Ergebnisse des kontextunabhängigen Systems, daß es durchaus ein Gewinn ist, kontextabhängige Modelle des multilingualen Erkenners zu übernehmen.

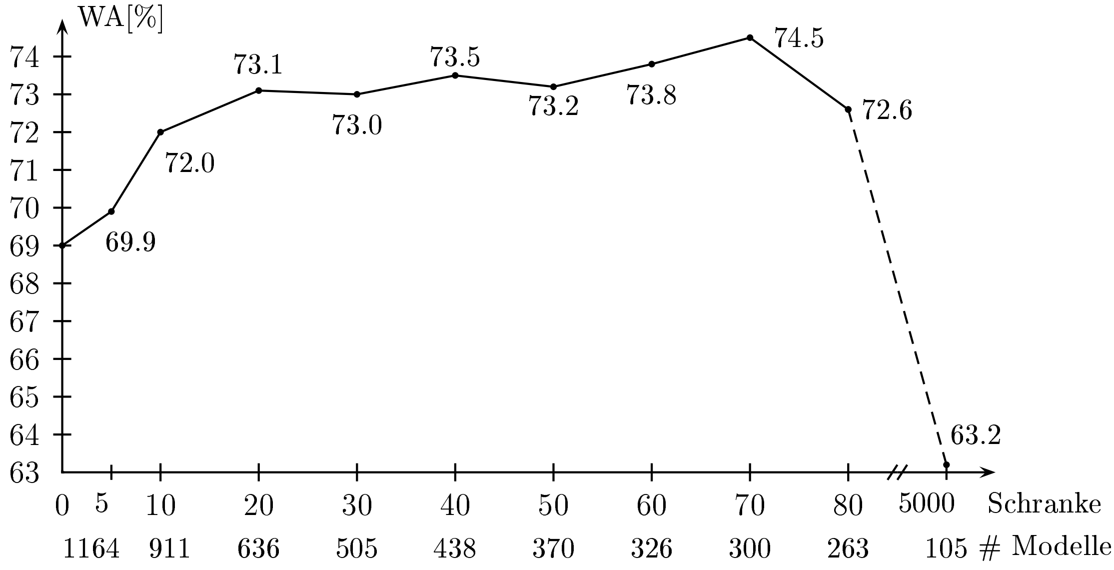

Abbildung 6: Erkennungsleistung nach der Portierung des Entscheidungsbaumes auf Deutsch für verschiedene Schranken

In einem zweiten Versuch wurden die Entscheidungsbäume auf Portugiesisch portiert, bei dem 100 bzw 500 Sätze als Adaptionsmenge dienten. Erfreulich ist, daß immer eine Steigerung der Erkennungsrate erzielt wurde (siehe Abbildungen 8 und 7). Insbesondere konnte durch das Verfahren bei dem System PO100 die Erkennungsleistung um relativ 22.9% gesteigert werden. Dagegen variiert die Erkennungsrate bei der Verwendung von 500 Trainingssätzen in Abhängigkeit von der Schranke nur wenig. Wie zu erwarten war, konnte die höchste Steigerung der Leistung von relativ 6.7% mit einer höheren Anzahl von Modellen erzielt werden als bei der Verwendung bei 100 Trainingssätzen. Da schließlich mehr Trainingsmaterial zur Verfügung steht, können diese Modelle ausreichend geschätzt werden. Eine weitere wichtige Beobachtung ist, daß die kontextunabhängigen Systeme bei einer Schranke von 5000 deutlich schlechtere Ergebnisse erzielen. Somit ist auch hier die Aussage von oben, daß sich die Übernahme von bestimmten kontextabhängigen Modellen lohnt, gültig. Außerdem ist der Zuwachs in der Leistung bei mehr Trainingsmaterial geringer, was dadurch erklärbar ist, daß es leichter ist, von schlechten Modellen eine Verbesserung zu erzielen als von guten Modellen.

Ein Problem bei der Schätzung der Häufigkeit der Polyphone für die neue Sprache stellt die Forderung nach einem geringen Umfang des Trainingsmaterials für die Adaption dar. Es ist davon auszugehen, daß weder bei der Verwendung von 100 noch von 500 Sätzen eine exakte Schätzung erreicht wird. Da aber für die Schätzung keine Sprachdaten sondern einfache Textdaten ausreichen, kann das Problem wie folgt gelöst werden. Auf Basis der Textdaten, die einfacher und billiger zu erhalten sind als Sprachdaten z.B. Zeitungsberichte im Internet, kann die Häufigkeit der Polyphone bestimmt werden. Danach wird der Entscheidungs-

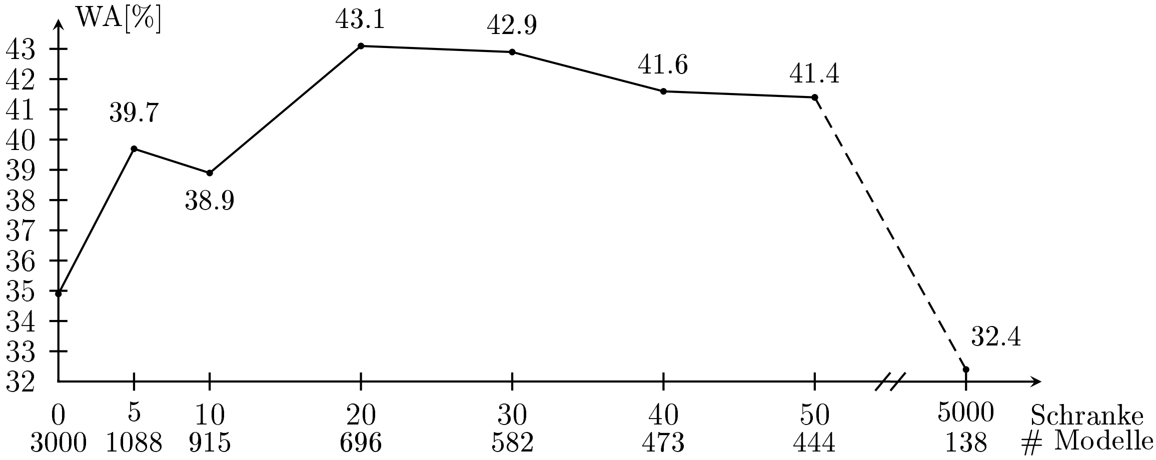

Abbildung 7: Word Accuracy und Anzahl der Modelle nach der Portierung des Entscheidungsbaumes auf Portugiesisch bei 100 Sätzen als Adaptionsmaterial

baum verallgemeinert und auf den Sprachdaten trainiert.
Für diese Testreihe standen 7193 portugiesische Sätze als Textdaten zur Verfügung, die 2.500.000 verschiedene Polyphone beinhalten. Das Sprachmaterial umfaßte 500 Sätze. Somit ist feststellbar, wie sich die Verwendung der Textdaten auswirkt. In der Tabelle 7 sind die Erkennungsleistungen und die Anzahl der verwendeten Modelle dargestellt (PO500+Text). Zusätzlich sind dort die Ergebnisse für die Systeme, die nur auf den Sprachdaten angepaßt wurden, dargestellt (PO500). Falls bei den beiden Verfahren zwei Systeme entstehen, die ungefähr die gleiche Anzahl von Modellen besitzen, so sollte der Erkenner, der die Textdaten zusätzlich benutzen konnte, im Vergleich besser abschneiden. Dies ist aber nur der Fall bei dem Vergleich zwischen dem PO500+Text System bei einer Schranke von 10 und dem PO500 System bei einer Schranke von 100, wo sich die Erkennungsleistung um 0,3 % absolut steigert. Ansonsten schneiden die PO500+Text Systeme schlechter als die PO500 Systeme ab. Besonders die Wortakkuratheit von 52,1 % bei einer Schranke von 50 (PO500+Text) gegenüber den Wert von 56,2 % bei einer Schranke von 5 (PO500) ist enttäuschend. In diesem Fall hat sich sogar eine Verschlechterung durch die Verallgemeinerung des Entscheidungsbaums ergeben.

Insgesamt läßt sich aus den Ergebnissen folgern, daß sich das Benutzen der Textdaten zur Schätzung der Häufigkeit der Polyphone nicht lohnt. Der Grund dafür könnte darin liegen, daß durch die Textdaten Modelle entfernt werden, die von den Sprachdaten gut trainiert werden konnten. D.h., daß die Text- und Sprachdaten nicht gut zusammenpassen. Dieses Argument wiederspricht aber der Tatsache, daß die Verwendung des Textmaterials den Aufbau des Entscheidungsbaums nur wenig verändert. Im Vergleich zwischen dem PO500 System bei ei-

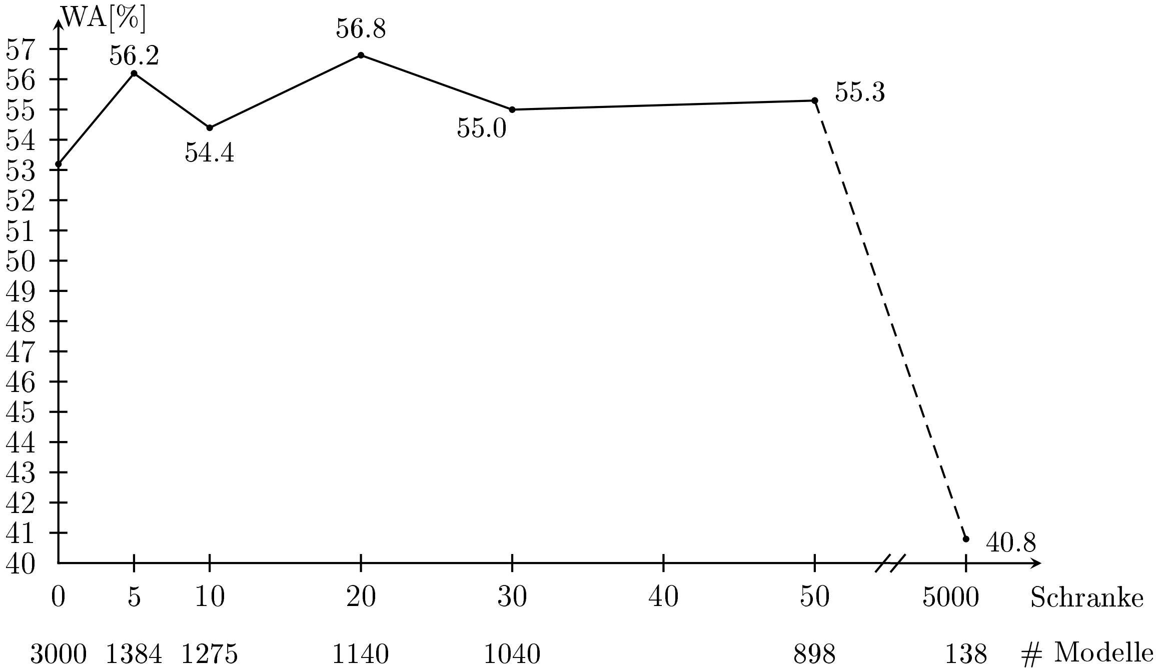

Abbildung 8: Word Accuracy und Anzahl der Modelle nach der Portierung des Entscheidungsbaumes auf Portugiesisch bei 500 Sätzen als Adaptionsmaterial

ner Schranke von 50 und dem PO500+Text System bei einer Schranke von 700 sind z.B. 110 von 138 Entscheidungsbäume der Subphoneme exakt gleich und der höchste Unterschied in der Anzahl der Modelle zwischen den Entscheidungsbäumen der Subphoneme liegt bei 3. Auch die Vergleiche der anderen Systeme zeigen eine ähnlich hohe Übereinstimmung auf. Eine andere interessante Beobachtung ist, daß auch die hohe Anzahl der Polyphone aus den Textdaten nur zu einer geringen Überdeckung von 14.631 Polyphonen des multilingualen Erkenners führen. Dies kann dahin gewertet werden, daß die Sprachen des multilingualen Systems und Portugiesisch sehr unterschiedlich sind und daß eine Adaption auf Portugiesisch schwierig ist.

5.4 Erweitern des Entscheidungsbaums

Bisher wurde versucht, den Entscheidungsbaum zu verkleinern, um allgemeinere Modelle zu erhalten. Diese Modelle wurden durch ihr seltenes Vorkommen in der Trainingsmenge entdeckt. Ein weiteres Problem ist aber, daß es auch multilinguale Modelle gibt, die in der neuen Sprache sehr viele Trainingsdaten bekommen z.B. gibt es bei dem System PO500 17 Modelle die auf mehr als 1000 Kontexten trainiert werden. Bei einer so großen Anzahl von Daten ist anzunehmen, daß das Modell mit sehr vielen unterschiedlichen Polyphonen konfrontiert wird. Die verschiedenen Aussprachevarianten dieser Polyphone führt dazu, daß das Modell zu ungenau wird. Deshalb ist es an dieser Stelle sinnvoll, das Modell weiter zu verfei-

PO500+Text					
Schranke	50	100	200	500	700
# Modelle	1370	1290	1194	976	882
WA[%]	52.1	54.7	54.0	55.9	54.8
PO500					
Schranke	5	10			50
# Modelle	1384	1275			898
WA[%]	56.2	54.4			55.3

Tabelle 7: Word Accuracy und Anzahl der Modelle nach der Portierung des Entscheidungsbaums auf Portugiesisch für verschiedene Schranken

nern. Dies ist möglich, indem an dem Entscheidungsbaum ein Ballungsverfahren, wie im Abschnitt 3.3.1 beschrieben, angewendet wird. Dadurch sollten Modelle entstehen, die auf die speziellen Koartikulationseffekte der neuen Sprache angepaßt sind.
Der Vorteil dieser Vorgehensweise sollte noch dadurch gestärkt werden, wenn zuvor der Entscheidungsbaum verallgemeinert wird, da dann die Trainingsdaten für das Ballungsverfahren auf wenigere Modelle verteilt werden. Somit ist es denkbar, den Entscheidungsbaum zunächst soweit zu verkleinern, daß nur noch kontextunabhängige Modelle vorliegen und danach ein Ballungsverfahren anzuwenden, so daß ein vollständig sprachspezifischer Entscheidungsbaum entsteht. Dies wurde aber in der Arbeit nicht ausprobiert.Von verschiedenen, verallgemeinerten Systemen wurde versucht, den Entscheidungsbaum zu erweitern. Die resultierenden Ergebnisse zeigen, daß keine Verbesserung in der Erkennungsleistung erreicht werden konnte. Der Grund für dieses Verhalten liegt wohl in der kleinen Anzahl von Sätzen in der Trainingsmenge. Da aber das Ballungsverfahren eine Schätzung für die akustischen Modelle aller Polyphone benötigt und die meisten Polyphone sehr selten vorkommen, sind diese Schätzungen sehr schlecht. Dadurch arbeitet das Ballungsverfahren auf keinen zuverlässigen Daten und kann auch keine guten Ergebnisse liefern.
Insgesamt wurde bei 3 Systeme der Versuch unternommen, den Entscheidungsbaum zu erweitern: bei DE100 bei einer Schranke von 20 bzw 50 und bei PO500 bei einer Schranke von 20. Die Erkennungsrate konnte bei den deutschen Systemen von 73.1 auf 73.2 (DE100_b20) bzw von 73.2 auf 73.9 (DE100_b50) gesteigert werden, wobei gleichzeitig die Anzahl der Modelle um 306 bzw 86 Modelle erhöht wurde. Dagegen fiel die Erkennungsrate bei dem portugiesischen System von 56.8 auf 54.7 ab, obwohl 152 Modelle mehr verwendet wurden.

6 Ausblick

Durch das Verallgemeinern der Entscheidungsbäume konnte bei der Adaption des multilingualen Erkenners auf eine neue Sprache die Erkenungsleistung gesteigert werden. Das Verfahren sollte auch erfolgreich bei einem Wechsel des Tasks oder bei einer Veränderung der Hintergrundgeräusche sein. Da die beiden Sprachen Portugiesisch und Deutsch keine große Ähnlichkeiten zu den Sprachen des multilingualen Erkenners aufweisen, sollte das Verfahren auf Sprachen angewendet werden, die z.B. eine höhere Überdeckungsrate der Polyphone liefern. Außerdem muß die Frage untersucht werden, wieviel Adaptionsmaterial notwendig ist, um sowohl die Akustik anzupassen als auch um eine repräsentative Häufigkeitsanalyse der Polyphone durchzuführen. Es besteht zu hoffen, daß der Umfang des Adaptionsmaterials bei ähnlichen Sprachen geringer sein darf als bei völlig verschiedenen Sprachen.
Weitere Untersuchungen sind notwendig, um das Erweitern des Entscheidungsbaums zu bewerten. Insbesondere muß geprüft werden, ob es sich lohnt, zunächst den Entscheidungsbaum auf ein kontextunabhängiges System zu verallgemeinern, um ihn dann speziell für die neue Sprache aufzubauen. Auch hier hängt die Rentabilität dieser Vorgehensweise von dem Umfang des benötigten Materials ab. Falls zuviele Sprachdaten für die Adaption notwendig sind, so wird der eigenständige Aufbau eines sprachspezifischen Systems bessere Leistungen bringen.

7 Literaturverzeichnis

Literatur

[1] T.Schultz und A.Waibel: Das Projekt GlobalPhone: Multilinguale Spracherkennung, Computers, Linguistics, and Phonetics between Language and Speech. Proceedings of the 4th Conference on NLP, Konvens 98, S.179-189, Oktober 1998

[2] T.Schultz und A.Waibel: Multilingual and Crosslingual Speech Recognition, DARPA BN Workshop, Virginia, Febr. 1998

[3] Lee: Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition, IEEE Trans. on Acoustics, Speech, and Signal Processing, April, 1990

[4] J.Köhler: Language Adaption Of Multilingual Phone Models For Vocabulary Independent Speech Regonition Tasks, ICASSP 98, S.417-420, Seattle, Mai 1998

[5] S. Gokeen und J.M. Gokeen:A Multilingual Phoneme and Model Set: Toward a universal base for Automatic Speech Recognition, Automatic Speech Recognition and Understanding, St.Barbara, Dez. 1998

[6] Cohen et al: Towards A Universal Speech Recognizer For Multiple Languages, Watson Research Center, Automatic Speech Recognition and Understanding, St.Barbara, Dez. 1998

[7] P.Bonaventura, F.Gallocchio und G.Micca:Multilingual Speech Recognition For Flexible Vocabularies, Eurospeech 97, S.355-358, Rhodes

[8] A.Constantinescu und G.Chollet: On Cross-Language Experiments and Data-Driven Units for Automatic Language Independent Speech Processing, Automatic Speech Recognition and Understanding, St.Barbara, Dez. 1998

[9] B.Wheatley et al: An Evaluation Of Cross-Language Adaption For Rapid HMM Development In A New Language, ICASSP 94, S.237-240

[10] T.Schultz, M.Westphal und A.Waibel: The GlobalPhone Project: Multilingual LVCSR with JANUS-3, SQEL Plzeň 1997

[11] M.Finke,I.Rogina:Wide Context Acoustic Modeling In Read vs.Spontaneous Speech, IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, 1997