Integration von situationsabhängigen Modalitäten in kontextbasierte Entscheidungsbäume

Diplomarbeit

von

Christian Fügen

Bearbeitungszeitraum:

1. Juni 1999 - 30. November 1999

Institut für Logik, Komplexität und Deduktionssysteme
Fakultät für Informatik
Universität Karlsruhe (TH)
D-76128 Karlsruhe

Betreuer:

Prof. Dr. Alexander Waibel
Dr. Ivica Rogina

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

Karlsruhe, 30. November 1999

Christian Fügen

Abstrakt:

Modalitäten sind sprecherabhängige Einflußfaktoren wie das Geschlecht, der Dialekt oder die Sprechgeschwindigkeit eines Sprechers, aber auch sprecherunabhängige Einflußfaktoren wie zum Beispiel unterschiedliche Aufnahmequalitäten. Ihnen allen gemeinsam ist, daß eine gleichmäßige Berücksichtigung aller Ausprägungen einer Modalität in einem Spracherkennungssystem meist nicht möglich ist. Aufgrund der unterschiedlichen Ausprägungen der Modalitäten, verlieren die Modelle nämlich an Präzision und erschweren damit eine genaue Erkennung.

Normalisierungs- und Filtertechniken, die sich zum Teil adaptiv an eine gegebene Äußerung anpassen können, versuchen die Varianz der einzelnen Modelle zu vermindern. Ist der Wertebereich einer Modalität diskret, so kann es sich lohnen, für jede Ausprägung ein eigenes Spracherkennungssystem zu trainieren.

Nachteilig an der Verwendung von Normalisierungs- und Filtertechniken ist, daß jede Modalität auf eine andere Art und Weise behandelt werden muß. Zum anderen ist es bei der Verwendung von mehreren Spracherkennungssystemen nur bedingt möglich, unterschiedliche Ausprägungen einer akustischen Einheit unter verschiedenen Modalitäten nur durch ein akustisches Modell zu repräsentieren (Parametersharing). Dies wäre aber nötig, um Bereiche des Parameterraums mit geringer Divergenz unter den Modalitäten wesentlich robuster modellieren zu können.

In dieser Arbeit wird ein Verfahren auf Basis von modalitätenabhängigen Kontextentscheidungsbäumen präsentiert, das zum einen die Varianz einzelner Modelle verringert und zum anderen ein Parametersharing für Modelle verschiedener Ausprägungen mehrerer Modalitäten erlaubt. Die Optimierung erfolgt dabei rein datengetrieben, wobei alle Modalitäten auf ein und dieselbe Art und Weise behandelt werden. Mit Hilfe dieses Verfahrens konnte die Gesamtfehlerrate des Referenzsystems um bis zu 8% gesenkt werden. Auch die Modellierung von sprechgeschwindigkeitsanghängigen und signal-rausch-abstands-abhängigen Modellen reduzierte die Fehlerrate um etwa 3% bzw. 4%.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Die sprecherabhängige Erkennung von nicht spontan gesprochener Sprache in wohldefinierten Aufnahmeumgebungen mit begrenztem Wortschatz ist mittlerweile in einem Bereich angekommen, der nur noch relativ wenig Spielraum für Verbesserungen zuläßt. Sobald jedoch eine dieser Einschränkungen nicht mehr eingehalten werden kann, fällt die Erkennungsleistung[1] sichtbar ab. Ferner besitzt ein so begrenztes Spracherkennungssystem nur einen sehr beschränkten Einsatzbereich.

Wird die Einschränkung der Sprecherabhängigkeit aufgehoben, so wirken Modalitäten wie die Sprechgeschwindigkeit, das Geschlecht oder der Dialekt des Sprechers alle in verschiedenem Maße auf den Spracherkennungsprozeß ein. Verändert sich die Aufnahmeumgebung eines Spracherkennungssystems, so spielen Modalitäten wie Hintergrundgeräusche oder der Signal-Rausch-Abstand eine nicht unerhebliche Rolle. Auch die Erkennung von spontan gesprochenen Äußerungen stellt ein Problem dar, weil zum einen Lautverschleifungen, Assimilationen und Elisionen[2] zunehmen und zum anderen die Analyse der Grammatik des Satzes durch Wortwiederholungen, Gedankensprünge und Satzumstellungen erschwert wird. Größere Wortschätze verursachen eine Vergrößerung des Suchraumes und damit bei gleicher Erkennungsgenauigkeit eine Abnahme in der Erkennungsgeschwindigkeit.

Während der Mensch sehr erfolgreich auf solche Situationen adaptieren und selbst dann noch Rückschlüsse auf das Gesagte ziehen kann, wenn er nicht alles ver-

[1] Die Erkennungsleistung wird im folgenden nur durch die Genauigkeit der Erkennung bei gleichbleibender Erkennungsgeschwindigkeit definiert.

[2] Unter Assimilation versteht man die Modifizierung von Lauten in ihren Eigenschaften und unter Elision, die Eliminierung von Lauten.

standen hat, so ist das Spracherkennungssystem in dieser Hinsicht weitestgehend überfordert. Gerade die Fähigkeit des Menschen, mehrdeutige Äußerungen aufgrund des Kontextes im Laufe eines Gesprächs auf eine sinnvolle Deutung zu reduzieren, stellt aktuelle Spracherkennungssysteme noch vor eine nahezu unlösbare Aufgabe. Aus diesem Grund beschränkt sich die heutige Forschungsarbeit fast nur auf Verfahren, die ein Spracherkennungssystem gegenüber den oben erwähnten Modalitäten robuster machen sollen.

Diese Verfahren versuchen durch Normalisierungs- und Filtertechniken die Zunahme an Varianz, die im Parameterraum durch unterschiedliche Ausprägungen einer oder auch mehrere Modalitäten entsteht, zu vermindern. Da die meisten Modalitäten über verschiedene Äußerungen mehrerer Sprecher oder sogar innerhalb einer Äußerung eines Sprechers nicht konstant sind, müssen die eingesetzten Techniken immer an die vorliegende Umgebung adaptiert werden. Ist der Wertebereich einer Modalität diskret, so kann es sich lohnen, für jede Ausprägung einer Modalität ein eigenes Spracherkennungssystem zu trainieren. Dies bietet sich zum Beispiel im Falle unterschiedlicher Dialekte an.

Obige Verfahren besitzen zwei entscheidende Nachteile. Zum einen muß jede Modalität auf eine andere Art und Weise behandelt werden. Zum anderen ist es bei der Verwendung von mehreren Spracherkennungssystemen nur bedingt möglich, unterschiedliche Ausprägungen einer akustischen Einheit[3] unter verschiedenen Modalitäten nur durch ein akustisches Modell zu repräsentieren. Dies wäre aber nötig, um Bereiche des Parameterraums mit geringer Divergenz unter den Modalitäten wesentlich robuster modellieren zu können. Im folgenden wird für die einheitliche Modellierung von unterschiedlichen Ausprägungen einer akustischen Einheit unter verschiedenen Modalitäten der Begriff „Parametersharing" verwendet.

Die Motivation dieser Arbeit war gegeben durch den Wunsch, ein Verfahren zu entwickeln, das ein Parametersharing bei gleichzeitiger Varianzverminderung der Modelle erlaubt. Die oben erwähnten Modalitäten sollten dabei alle auf dieselbe Art und Weise behandelt und Abhängigkeiten unter diesen selbständig berücksichtigt werden. Die Optimierung sollte rein datengetrieben erfolgen.

Vorgehensweise

In der Spracherkennung werden zur Ballung von akustischen Modellen vorwiegend Entscheidungsbäume eingesetzt. Dabei wird vor Beginn des Ballungsvorgangs je-

[3] dies könne zum Beispiel Phoneme oder Subpolyphone sein

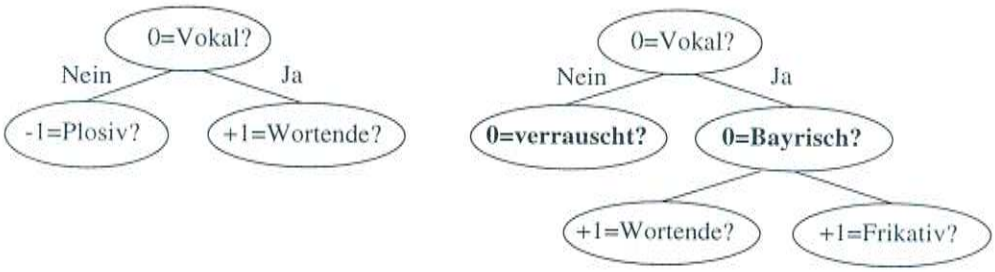


Abbildung 1.1: Modalitätenabhängige Kontextentscheidungsbäume

dem Subpolyphon[4] genau ein akustisches Modell zugeordnet. Die Ballung selbst geschieht mit Hilfe eines fest vorgegebenen Fragenkatalogs. Dieser besteht meist aus Kontextfragen, die jeweils nur binäre Antworten erlauben. Durch Anordnung der Fragen in Struktur eines Binärbaumes entsteht ein Kontextentscheidungsbaum, der es erlaubt, für ein gegebenes Subpolyphon dessen zugehöriges akustisches Modell zu ermitteln. Die Position der Fragen innerhalb des Baumes ist durch ihre Wichtigkeit gegeben, wobei die wichtigste Frage die akustischen Modelle bezüglich eines Distanzmaßes am besten[5] aufspaltet. Die linke Hälfte der Abbildung 1.1 zeigt einen Ausschnitt aus einem solchen Entscheidungsbaum, wobei eine Frage, wie beispielsweise „-1=Plosiv?“ danach fragt, ob das Vorgänger-Phonem ein Plosiv war. Da eine große Menge von akustischen Modellen in immer kleinere Teilmengen aufgespaltet wird, werden solche Ballungsverfahren als divisiv bezeichnet.

Vor der eigentlichen Erkennung werden der vorliegenden Äußerung verschiedene Modalitäten, wie zum Beispiel „weibliche Sprecherin“ und „verrauscht“ zugeordnet. Dabei wurde auch berücksichtigt, daß sich Modalitäten während einer Äußerung ändern können. Eine solche Modalität ist beispielsweise die Sprechgeschwindigkeit.

Jedes akustische Modell, dessen Subpolyphon in dieser Äußerung vorkommt, wird nun mit der gegebenen Modalitätenkombination markiert. Damit ist es möglich akustische Modelle nicht nur nach deren Kontext, sondern auch nach deren Modalitätenkombination aufzuspalten, indem zusätzlich zu den Kontextfragen auch Fragen nach Modalitätenkombinationen in den Fragenkatalog aufgenommen werden. Die rechte Hälfte der Abbildung 1.1 zeigt die modalitätenabhängige Variante eines Kontextentscheidungsbaumes.

[4]Subpolyphone sind kontextabhängige Subphoneme, wobei ein Subphonem ein Teil eines Phonems repräsentiert und der Kontext durch die angrenzenden Phoneme definiert ist.

[5]im Sinne maximaler Distanz

Der Vorteil modalitätenabhängiger Kontextentscheidungsbäume ist, daß rein datengetrieben zum einen die Varianz der akustischen Modelle reduziert und zum anderen ein Parametersharing für Modelle verschiedener Ausprägungen mehrerer Modalitäten erlaubt werden kann. Der Ballungsalgorithmus entscheidet dabei selbständig, an welcher Stelle im Baum eine Modalitätenfrage sinnvoller als eine Kontextfrage ist. Auch Abhängigkeiten zwischen den Modalitäten werden selbständig berücksichtigt.

Die markierten akustischen Modelle lassen sich auch mittels agglomerativer Ballung zu akustisch ähnlichen Gruppen zusammenfassen. Damit ist es möglich sehr spezifische Modelle zu gröberen Modellen zusammenzufassen, um dadurch eine höhere Generalisierungsfähigkeit des Spracherkenners auf ungesehenen Daten zu erreichen. Dies ist vor allem bei einer unüberwachten Ballung sinnvoll, wenn die Klassen selbst nicht im voraus bekannt sind.

Mit Hilfe modalitätenabhängiger Kontextentscheidungsbäume konnte in dieser Arbeit die Fehlerrate des Referenzsystems um bis zu 8% gesenkt werden. Durch Anwendung der agglomerativen Ballung auf akustische Modelle aus 30 initialen Sprachregionen Deutschlands, konnten die Regionen zu 15 größeren zusammengeballt werden. Das war nötig um einen geeigneten Kompromiß zwischen Modellgenauigkeit und Generalisierungsfähigkeit der akustischen Modellierung zu finden. Die Erkennungsleistung eines mit diesen Regionen trainierten Systems erreichte eine Fehlerreduktion von 7%. Auch die Integration von sprechgeschwindigkeitsabhängigen oder signal-rausch-abstands-abhängigen Modalitäten reduzierte die Fehlerrate.

Die Verwendung von Markierungen für akustische Modelle ist schon seit längerem üblich. Es ist in zahlreichen Untersuchungen nachgewiesen worden, daß eine Unterscheidung in Intra- und Inter-Wortmodelle signifikante Verbesserungen der Erkennungsleistung erzielen [Lee88]. Während des Bearbeitungszeitraumes dieser Arbeit wurde eine Untersuchung veröffentlicht, die vorschlägt dieses Konzept auf allgemeine Modalitäten zu erweitern, und Ergebnisse für die Einführung von geschlechtsabhängigen Modellen präsentiert [RC99]. Sie behandelt aber letztendlich nur einen kleinen Teilbereich der in dieser Arbeit untersuchten Problematik.

Inhaltsübersicht

In Kapitel 2 werden die Grundlagen näher gebracht, die zum Verständnis dieser Arbeit beitragen sollen. Die darin beschriebenen Abläufe eines typischen Trainings- und Erkennungsvorgangs sollen dazu dienen, spätere Änderungen, die durch die Integration der Modalitäten notwendig sind, aufzuzeigen.

Um einen Überblick zu bekommen, auf welche Art und Weise bisher versucht wurde, alle Ausprägungen mehrerer Modalitäten gleichermaßen zu berücksichtigen ohne einen großen Verlust in der Erkennungsleistung hinzunehmen, werden in Kapitel 3 einige Arbeiten anderer Autoren vorgestellt.

Kapitel 4 beschäftigt sich mit der Aufgabe, die Modalitäten sinnvoll in den Trainings- und Erkennungsvorgang zu integrieren. Hierbei mußten aufgrund des verwendeten Spracherkenners auch einige Einschränkungen in Kauf genommen werden.

Die Modalitäten, die in dieser Arbeit verwendet wurden, werden in Kapitel 5 erläutert. Hierunter zählen sprecherabhängige Modalitäten, wie das Geschlecht, der Dialekt und die Sprechgeschwindigkeit eines Sprechers, aber auch sprecherunabhängige Modalitäten wie der Signal-Rausch-Abstand einer Äußerung. Es wird beschrieben, wie diese Modalitäten aus vorhandenen Daten gewonnen werden können und wie stark diese mit der Fehlerrate korrelieren.

Die mit den verschiedenen Modalitäten durchgeführten Experimente und die daraus resultierenden Ergebnisse werden in Kapitel 6 aufgeführt.

Kapitel 7 gibt eine kurze Zusammenfassung der Arbeit und Ergebnisse. Offengebliebene Fragen und einen Ausblick auf zukünftige Arbeiten bilden den Abschluß dieser Arbeit.

Kapitel 2

Grundlagen

In diesem Kapitel wird in erster Linie die Ballung akustischer Modelle und die dabei häufig zum Einsatz kommenden Entscheidungsbäume erläutert. Um später auf Veränderungen aufmerksam machen zu können, die sich durch die Integration der Modalitäten in das Spracherkennungssystem ergeben, wird noch je ein typischer Trainings- und Erkennungsvorgang vorgestellt. Dabei wird allerdings ein fundiertes Fachwissen im Bereich der Spracherkennung vorausgesetzt. Deshalb sei an dieser Stelle auf [ST95] und [WL90] verwiesen. Alle in dieser Arbeit erzeugten Spracherkennungssysteme wurden mit Hilfe des JANUS Recognition Toolkits erstellt, welches in gemeinsamer Zusammenarbeit der Universität Karlsruhe und der Carnegie Mellon University Pittsburgh entstanden ist.

2.1 Ballung akustischer Modelle

In Kapitel 1 wurden schon Subpolyphone angesprochen und als kontextabhängige Subphoneme definiert. Dabei repräsentiert ein Subphonem einen Teil eines Phonems. Üblicherweise verwendet man hierbei eine Dreiteilung. Der Kontext wird dabei nicht durch die angrenzenden Subpolyphone, sondern durch die an das zum Subphonem gehörige Phonem angrenzenden Phoneme festgelegt. Jedem Subpolyphon ist ein akustisches Modell zugeordnet. Dabei erfolgt die Modellierung meist durch Mixturen von Gaußverteilungen mit diagonalen Kovarianzmatrizen. Die Mittelwerte der Gaußverteilungen und deren Kovarianzmatrizen bilden das Codebuch des akustischen Modells. Die Mixturgewichte werden von diesen getrennt abgespeichert.

Mit zunehmender Kontextbreite nimmt die Anzahl der in der Trainingsdatenbasis

vorkommenden verschiedenen Subpolyphone und damit die Anzahl der zugehörigen Modelle zu. Infolgedessen nimmt die Anzahl der Trainingsdaten, die auf ein akustisches Modell fallen jedoch ab. Dies führt dazu, daß viele Modelle sehr spezifisch ausfallen und damit meist weniger gut auf ungesehene Daten passen. Umgekehrt paßt ein sehr grobes Modell womöglich besser auf ungesehene oder ein wenig andersartige zum selben Modell gehörende Daten, besitzt jedoch dann eine wesentlich größere Varianz. Eine zu große Varianz kann jedoch auch von Nachteil sein, weil es dadurch vermehrt zu Überschneidungen der verschiedenen Modelle kommen kann. Ziel der Ballung ist es also, die Modellgenauigkeit bei Gewährleistung einer ausreichenden Generalisierungsfähigkeit bis zu einem Optimum zu erhöhen. Die durch Ballung entstehenden Klassen werden, im Falle von Subpolyphonen als Elementareinheiten, als generalisierte Subpolyphone bezeichnet.

Die Ballung akustischer Modelle kann auf mehrere Arten erfolgen. Aufgrund der unterschiedlichen Vorgehensweise wird zwischen agglomerativer und divisiver Ballung unterschieden.

2.1.1 Agglomerative Ballung

Bei der agglomerativen Ballung wird zunächst jeder Ballungsknoten mit einem akustischen Modell initialisiert. Nach Berechnung aller paarweisen Distanzen zwischen den Knoten werden die beiden Knoten mit der kleinsten Distanz zu einem neuen Knoten vereinigt. Um ein besseres Ergebnis zu erzielen, muß für jedes akustische Modell getestet werden, ob durch das Versetzen in einen anderen Knoten die Distanz zwischen den beiden Knoten erhöht werden kann. Ist dies der Fall, so wird das akustische Modell verschoben.

Die agglomerative Ballung wurde in der kontinuierlichen Spracherkennung zum ersten Mal erfolgreich von Kai-Fu Lee eingesetzt [Lee88]. Sie wird jedoch aufgrund zweier wesentlicher Nachteile heute fast nicht mehr zur Kontextballung verwendet. Zum einen nimmt der Rechenaufwand, der bei der Ballung der Modelle entsteht, quadratisch mit der Anzahl der Modelle zu. Der zweite, viel schwerer wiegende Nachteil ist jedoch die suboptimale Modellierung ungesehener Kontexte, da die durch die Ballung resultierenden Zusammenfassungen nur die im Training vorkommenden Kontexte verwendet, weshalb es auch nur für diese möglich ist, eine passende Klasse zuzuordnen. Ungesehene Kontexte müssen dann entweder einer mehr oder weniger willkürlich festgelegten Klasse zugeordnet werden, oder sie werden zusammen mit anderen ungesehenen Kontexten mit einem groben Modell modelliert [Rog98].

2.1.2 Divisive Ballung

Die divisive Ballung geht im Gegensatz zur agglomerativen Ballung den genau umgekehrten Weg und beseitigt damit deren Problem ungesehener Kontexte auf elegante Weise. Die divisive Ballung bedient sich eines Fragenkatalogs, der Kontextfragen zu einzelnen Phonemen und Artikulationsgruppen[1], aber auch Fragen zu artikulatorischen oder nichtartikulatorischen Geräuschen enthalten kann. Die Antworten auf solche Fragen sind meist binärer Natur.

Entscheidungsbäume

Zu Beginn der divisiven Ballung existiert pro Subphonem nur ein Ballungsknoten, der alle akustischen Modelle enthält. Diese Knoten werden jeweils unter Verwendung einer der Fragen aus dem Fragenkatalog in zwei neue Knoten aufgeteilt. Dabei wählt der Ballungsalgorithmus diejenige Frage aus, die bezüglich eines Distanzmaßes die beste Aufteilung[2] der momentan betrachteten akustischen Modelle erreicht. Im aufgeteilten Ballungsknoten wird die verwendete Frage festgehalten, während die beiden Nachfolgerknoten die jeweiligen Teilmengen der akustischen Modelle enthält. Anschließend wird eine Ebene tiefer in dem Baum abgestiegen und für jeden Knoten eine neue Frage zum Aufteilen der akustischen Modelle gesucht. Die Aufteilung wird so lange durchgeführt, bis zum Beispiel kein weiterer Gewinn[3] durch das Teilen einer Menge von akustischen Modellen mehr erzielt werden kann, oder bis ein weiteres Teilen dieser Menge einen zu kleinen Teil bezüglich einer vorher definierten unteren Schranke abspalten würde. Die Menge der akustischen Modelle, die am Schluß in den Blättern übrig bleibt, wird dann zu einer Klasse[4] zusammengefaßt. Aufgrund der Entscheidung, die in jedem Knoten zur Aufspaltung der akustischen Modelle getroffen wird, werden solche Bäume als Entscheidungsbäume bezeichnet. Diese wurden in der Spracherkennung erstmals in [Hon92] und [Ode92] eingesetzt.

Abbildung 2.1 zeigt einen kleinen Ausschnitt aus einem Entscheidungsbaum eines voll kontinuierlichen Spracherkennungssystems. Hierin werden die nichttermina-

[1] Artikulationsgruppen sind zum Beispiel Vokale, Diphthonge, Nasale, Frikative oder Plosive (siehe Anhang B).

[2] im Sinne maximaler Distanz zwischen den aufgespalteten Teilmengen

[3] im Sinne eines zu übersteigenden Schwellwertes oder durch Einsatz einer Kreuzvalidierungsmenge

[4] Die Modellierung dieser Klasse ist dann gegeben durch die Parameter aller in ihr enthaltenen akustischen Modelle. Im allgemeinen Sprachgebrauch wird eine solche Klasse wieder als akustisches Modell bezeichnet.

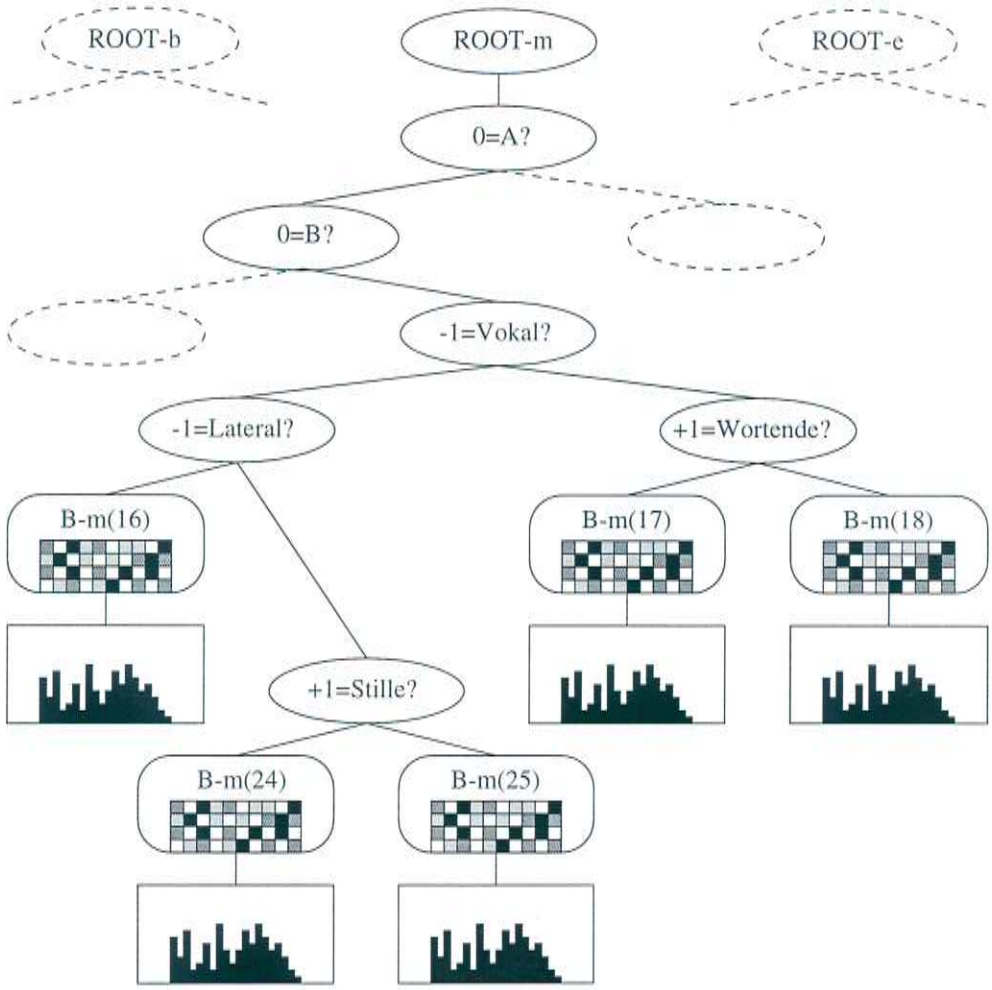


Abbildung 2.1: Ausschnitt aus einem Entscheidungsbaum

len Baumknoten durch Ovale dargestellt, die eine Frage aus dem Fragenkatalog enthalten und dann zwei Nachfolgerknoten, je eine für die positive und negative Antwort auf die Frage, besitzen. Die Blätter enthalten das jeweilige akustische Modell, also die Codebücher, welche in der Abbildung 2.1 durch abgerundete Rechtecke dargestellt sind, und die Mixturgewichte, die durch normale Rechtecke repräsentiert werden.

Der Nachteil der divisiven Ballung ist die Verwendung eines Fragenkatalogs, der die Möglichkeiten einer Aufteilung der akustischen Modelle auf die darin enthaltenen Fragen einschränkt. Demzufolge muß die Zusammenstellung des Fragenkatalogs möglichst genau die akustischen Unterschiede der einzelnen Modelle erfassen.

2.1.3 Distanzmaße

Für die Berechnung der Distanzen zwischen zwei Mengen akustischer Modelle sind in der Spracherkennung zwei verschiedene Distanzmaße, die Entropie-Distanz und die Likelihood-Distanz üblich. Die Entropie-Distanz hat zum Ziel, die Information im Parameterraum eines Spracherkennungssystems zu optimieren, während die Likelihood-Distanz versucht, die Wahrscheinlichkeit der Trainingsdaten zu maximieren [Rog98].

Die Entropie-Distanz wie auch die Likelihood-Distanz sind nicht dazu geeignet, um automatisch festzustellen, wann es sinnvoll ist, einen Ballungsvorgang abzubrechen. Der Gewinn, der durch eine Aufteilung entsteht, wird zwar mit jeder Aufteilung kleiner, jedoch nie negativ und in der Regel auch nie Null. Aus diesem Grund ist der Einsatz einer Kreuzvalidierungsmenge sinnvoll, um den optimalen Abbruchzeitpunkt des Ballungsvorgangs festzustellen. Dieser ist genau dann erreicht, wenn auf der Kreuzvalidierungsmenge eine Aufteilung der akustischen Modelle keinen Gewinn mehr erzielt. Das in dieser Arbeit verwendete Spracherkennungssystem erlaubte es, nur in Verbindung mit dem Likelihood-Distanzmaß eine Kreuzvalidierungsmenge zu verwenden, wobei es prinzipiell auch möglich wäre, diese in Verbindung mit dem Entropie-Distanzmaß einzusetzen.

Die Variante der Ballung von kontextabhängigen Modellen mit Hilfe einer Kreuzvalidierungsmenge unter Verwendung des Likelihood-Distanzmaßes wurde erstmalig in [Rog97] vorgeschlagen. Die Verwendung einer Kreuzvalidierungsmenge ist mit zusätzlichem Rechenaufwand verbunden, da die Optimierung mit Hilfe der Leaving-One-Out-Methode[5] erfolgt, um einem Qualitätsverlust bei der Parameterschätzung, der durch die Aufteilung der zur Verfügung stehenden Trainingsdaten in eine Trainingsuntermenge und eine Kreuzvalidierungsmenge entsteht, entgegenzuwirken. Aus diesem Grund wurde in dieser Arbeit auf eine Kreuzvalidierungsmenge verzichtet. Der optimale Abbruchzeitpunkt für den Ballungsvorgang wurde deshalb mit Hilfe zweier anderer Maße bestimmt. Zum einen ist das die minimale Anzahl an Trainingsdaten, die auf ein akustisches Modell fallen muß, und zum anderen ist das die Gesamtzahl der für das Spracherkennungssystem erlaubten akustischen Modelle. Mit Hilfe dieser Maße konnte bei allen in dieser Arbeit verwendeten Systeme auf dieselbe Art automatisch die Größe des Parameterraumes festgelegt werden.

[5] Dabei wird jeder Teil der zur Verfügung stehenden Trainingsdaten sowohl als Trainingsuntermenge als auch als Kreuzvalidierungsmenge verwendet.

2.2 Ablauf eines Trainingsvorgangs

In diesem Abschnitt wird ein typischer Trainingsablauf, so wie er zum Beispiel in dieser Arbeit für das Referenzsystem verwendet wurde, erläutert.

Soll ein kontextabhängiges Systems trainiert werden, so wird dies meist auf Basis eines fertig trainierten kontextunabhängigen Systems[6] aufgebaut. Der Trainingsablauf eines kontextunabhängigen Systems unterscheidet sich nur geringfügig von demjenigen, der nötig ist, um aus dem kontextunabhängigen System ein kontextabhängiges zu machen. Das einzige Problem ergibt sich bei der Initialisierung der Parameter des kontextunabhängigen Systems. Um eine schnellere und sicherere Konvergenz dieser zu gewährleisten, initialisiert man die Parameter meist mit Hilfe eines anderen, schon trainierten Systems mit ähnlichen akustischen Randbedingungen. Falls ein solches aber nicht zur Verfügung steht, müssen die Parameter entweder mit zufälligen Werten initialisiert werden oder man ordnet manuell Merkmalsvektoren der Trainingsdatenbasis einzelnen Modellen zu, um mit Hilfe dieser eine initiale Schätzung der Parameter zu bekommen.

Im folgenden wird von einem fertig trainierten kontextunabhängigen System ausgegangen. Dabei handelt es sich meist um einen voll kontinuierlichen HMM-Spracherkenner[7], der für jedes Subphonem ein eigenes Codebuch und eine eigene Mixturgewichteverteilung verwendet.

Das Training ist dazu da, um die Parameter des Systems auf den Trainingsdaten zu optimieren, wobei die Parameter durch die Codebücher und die Mixturgewichte aller akustischen Modelle gegeben sind. Das Optimieren der Parameter ist gleichbedeutend mit der Maximierung der Wahrscheinlichkeiten der Viterbi-Pfade[8] der HMMs, wobei es für jede in der Trainingsdatenbasis vorkommende Äußerung ein anderes, durch die Transkription der Äußerung vorgegebenes HMM zu optimieren gilt. Die Maximierung der Wahrscheinlichkeiten erfolgt durch den EM-Algorithmus.

Die Berechnung eines Viterbi-Pfades für ein HMM ist sehr aufwendig. Aus diesem Grund wird für jede in der Trainingsdatenbasis vorkommende Äußerung, die von diesem Algorithmus gefundene zeitliche Zuordnung von Sprachvektoren zu HMM-

[6] Die Modellierungseinheiten von kontextunabhängigen Systemen sind keine Subpolyphone, sondern Subphoneme. Ferner werden keine Kontextfragen im Entscheidungsbaum erlaubt.

[7] Die Abkürzung HMM steht dabei für Hidden Markov Modelle, die dafür verwendet werden, um zustandsgebundene stochastische Prozesse zu modellieren.

[8] Ein Viterbi-Pfad ist definiert als die wahrscheinlichste Zustandsfolge bei einer gegebenen Beobachtung in einem HMM. Die Beobachtung ist in diesem Fall die Folge der Merkmalsvektoren einer Äußerung.

Zuständen (Labels) abgespeichert. So können weitere Trainingsschritte die zuvor gefundene Zuordnung in der Erwartung verwenden, daß sich die Zuordnung im Verlauf des weiteren Trainings nur unwesentlich ändert.

Um die Komplexität und die Modellierungsgenauigkeit des Systems festzulegen, muß noch dessen Architektur definiert werden.

Kontextbreite der Polyphone: Die Kontextbreite bestimmt den Feinheitsgrad in der Modellierung der Subpolyphone und damit der zugeordneten akustischen Modelle. Mit wachsender Kontextbreite steigt die Anzahl der in der Datenbasis vorkommenden Polyphone. Da jedoch die Fehlerrate eines Spracherkennungssystems ab einer Kontextbreite von etwa drei bis vier nur noch insignifikant sinkt, wird für die Kontextbreite meist ein Wert zwischen eins und vier verwendet.

Festlegung der Codebuchgröße: Bei zu großen Codebüchern besteht die Gefahr einer Überanpassung an die Trainingsdaten. Aus diesem Grund sollte die Codebuchgröße explizit an die Trainingsdaten angepaßt werden. Die Entscheidung über die Codebuchgröße kann dabei entweder im voraus [Kem95], das heißt bevor die Codebücher angelegt werden, während des Trainings [WOVY94] oder im nachhinein durch Entfernung unerwünschter Vektoren [Rog98] stattfinden.

Ballung der akustischen Modelle: Die Anzahl der in der Datenbasis vorkommenden Subpolyphone liegt meist bei mehreren hunderttausend. Aus der in Abschnitt 2.1 erwähnten Diskrepanz zwischen Modellgenauigkeit und Generalisierungsfähigkeit ist man an einer Ballung der akustischen Modelle interessiert. Dafür müssen zunächst die Parameter aller Subpolyphone in einer oder mehrerer Epochen trainiert werden. Hierzu werden dem System in jeder Epoche alle Trainingsdaten einmal präsentiert und die Parameter wie oben beschrieben optimiert. Die optimale Anzahl der Epochen sollte auf einer Kreuzvalidierungsmenge bestimmt werden, jedoch wird in der Praxis meist ein auf Erfahrung basierter Wert genommen. Die Ballung der trainierten Modelle zu einzelnen Klassen erfolgt dann divisiv unter Zuhilfenahme eines vorher definierten Fragenkatalogs.

Anzahl der Klassen: Die Anzahl der Klassen ist ausschlaggebend für die Komplexität des Systems. Unter der Bedingung, daß genügend Trainingsdaten auf eine Klasse fallen, kann grob gesagt werden, daß eine Zunahme der Klassenanzahl gleichzeitig mit einer Fehlerratenreduktion verbunden ist. Demzufolge benötigt der Ballungsalgorithmus zum einen eine untere Schranke

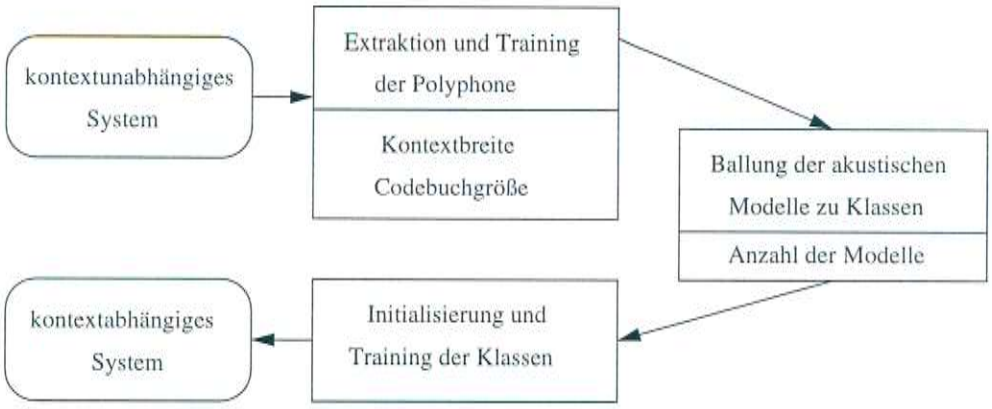


Abbildung 2.2: Schematischer Überblick des Trainingsablaufs

für die Anzahl der Trainingsdaten, die einer Klasse zugeordnet werden und zum anderen die maximale Anzahl der Klassen, die dabei entstehen sollen.

Initialisierung der Klassen: Für jede Klasse wird nun eine eigene Mixturgewichteverteilung und ein eigenes Codebuch verwendet. Die Initialisierung der Codebücher[9] erfolgt mit dem k-Mittelwerte[10] Algorithmus, der die zuvor abgespeicherten zeitlichen Zuordnungen verwendet. Die initialisierten Klassen werden einige Epochen nach obigem Schema trainiert. Das so entstehende System ist ein voll kontinuierlicher kontextabhängiger HMM-Spracherkenner.

Falls innerhalb der Vorverarbeitung eine lineare Diskriminanzanalyse (siehe Abschnitt 2.3.1) verwendet wird, so wird diese vor der Initialisierung der Klassen durch den k-Mittelwerte Algorithmus auf Basis der durch den Ballungsalgorithmus entstandenen Klassen berechnet. Abbildung 2.2 zeigt einen schematischen Überblick des Trainingsablaufs.

2.3 Ablauf eines Erkennungsvorgangs

2.3.1 Vorverarbeitung

Ziel der Vorverarbeitung ist es für alle Teilbereiche einer Äußerung Merkmalsvektoren zu liefern, auf Basis derer die HMMs evaluiert werden können. Die besten

[9] genauer die Mittelwertsvektoren der Gaußmixturverteilungen
[10] auch bekannt unter dem Namen „Basic-ISODATA"

Resultate werden dann erzielt, wenn die Merkmalsvektoren nur die Informationen enthalten, die für eine Unterscheidung einzelner Modelle am wichtigsten sind. Die einzelnen Schritte der Vorverarbeitung werden daher aus den Ergebnissen von Untersuchungen über das menschlichen Hörverhalten abgeleitet. Zunächst werden die Teilbereiche der Äußerung in den Frequenzbereich transformiert. Das so entstehende Kurzzeitspektrum beinhaltet neben den Frequenzüberlagerungen des Aufnahmekanals, auch noch die Frequenzanteile des Sprachsignals, also die Grundfrequenz und die Frequenzanteile des Vokaltraktes. Das menschliche Ohr besitzt für niedrigere Frequenzen ein besseres Auflösungsvermögen als für höhere Frequenzen. Aus diesem Grund werden mit Hilfe des sogenannten Mel-Scalings viele hochfrequente oder wenige niederfrequente Spektralkoeffizienten zu einem Mel-Scale-Koeffizienten gemittelt. Die Grundfrequenz ist für die Spracherkennung von deutscher Sprache nicht wichtig, da sie keine Informationen bezüglich des momentan Gesprochenen enthält. Sie ist über die gesamte Äußerung nahezu identisch. Durch einen Übergang in den Cepstralbereich können die Grundfrequenzanteile mittels sogenannter Tiefpaß-Lifterung weitgehend unterdrückt werden. Die Frequenzüberlagerungen, die durch den Aufnahmekanal oder durch andere Störsignale auftreten können, werden mit Hilfe einer Mittelwertssubtraktion über alle Cepstralkoeffizienten entfernt.

Um die Klassifizierungsaufgabe[11] zu vereinfachen, sollten die einzelnen Klassen möglichst weit voneinander entfernt liegen und die Klassen selber möglichst kompakt sein. Dies wird mit Hilfe der linearen Diskriminanzanalyse (LDA) erreicht, die außerdem noch eine Datendekorrelation bewirkt.

Die im folgenden angegebenen Zahlenwerte wurden für die Vorverarbeitung aller in dieser Arbeit trainierten Spracherkennungssysteme verwendet.

Das vorliegende analoge Sprachsignal wird meist mit 16 kHz abgetastet und mit einer Auflösung von 16 Bit quantisiert. Alle 10 ms wird ein 16 ms breites Hamming-Fenster aus der vorliegenden Äußerung herausgeschnitten. Das darauf basierende Kurzzeitspektrum besitzt 129 Spektralkoeffizienten, die dann auf 13 Mel-Cepstralkoeffizienten reduziert werden. Um den Grad der Veränderlichkeit eines Merkmalsvektors zu einem Zeitpunkt feststellen zu können, werden auf Basis vorhergehender Merkmalsvektoren deren Differenz berücksichtigt. Zusätzlich neben der Annäherung an die erste Ableitung wird auch noch die zweite Ableitung ermittelt, wodurch sich dann insgesamt 39-koeffizientige Merkmalsvektoren ergeben. Bei der anschließenden LDA wird der 39-koeffizientige Merkmalsvektor

[11] Das heißt die Zuordnung eines Merkmalsvektors zu einer Klasse, wobei ein Klasse dabei genau einem Codebuch entspricht.

mit der 39×39 LDA-Matrix[12] multipliziert. Die Koeffizienten des Ergebnisses sind so sortiert, daß der erste Koeffizient die größte und der letzte die kleinste Varianz besitzt. Das hat zur Folge, daß die letzten Koeffizienten bei der Distanzberechnung zweier Vektoren nicht mehr so stark ins Gewicht fallen, und damit eher unwichtig sind. Aus diesem Grund läßt sich ohne größeren Verlust nach der LDA eine Dimensionalitätsreduktion durchführen, bei der hier nur die ersten 32 Koeffizienten übernommen werden.

2.3.2 Dekodierungsprozeß

Die Zuordnung der Merkmalsvektoren aus der Vorverarbeitung zu den einzelnen Klassen ist Aufgabe der akustischen Modellierung. Die Verbindung der akustischen Modelle und die Umsetzung zu einer sinnvollen Hypothese geschieht mit Hilfe des Sprachmodells innerhalb der Suche. Eine vollständige Beschreibung des Dekodierungsprozesses des in dieser Arbeit eingesetzten Spracherkennungssystems ist in [Wos98] zu finden. Hier werden nur die wichtigsten Aspekte angesprochen.

Zunächst muß der Wortschatz und ein dazugehöriges Aussprachewörterbuch definiert werden. Der Wortschatz sollte so groß gewählt werden, daß die Menge der unbekannten Wörter möglichst klein ist. Meist werden hierfür die Trainingsdaten verwendet. Für jedes Wort im Wortschatz muß dessen Aussprache als Phonemsequenz im Aussprachewörterbuch festgelegt werden. Existieren zu einem Wort mehrere Aussprachevarianten, so werden diese ebenfalls in das Aussprachewörterbuch aufgenommen.

Zeitsynchrone Mehrpaß-Suche

Der Suchraum ergibt sich aus allen möglichen Folgen aller Wörter des Wortschatzes. Für jedes Wort ist dessen Phonemsequenz im Aussprachewörterbuch bekannt und für jedes Phonem dessen Aufteilung in Subphoneme. Auch der Kontext der einzelnen Phoneme ist bekannt. Damit sind a priori für jedes Wort aus dem Wortschatz dessen Subpolyphone genau definiert. Bei Verwendung eines Entscheidungsbaumes, der nur aus statischen Fragen[13] besteht, läßt sich damit schon vor der Erkennung für jedes Subpolyphon dessen akustisches Modell ermitteln.

[12] Für die genaue Berechnung der LDA-Matrix sei auf [DH73] verwiesen.

[13] Unter statischen Fragen sind Fragen zu verstehen, deren Antwort unabhängig von der jeweiligen zu erkennenden Äußerung ist. Dies ist im Falle von Kontextfragen gegeben.

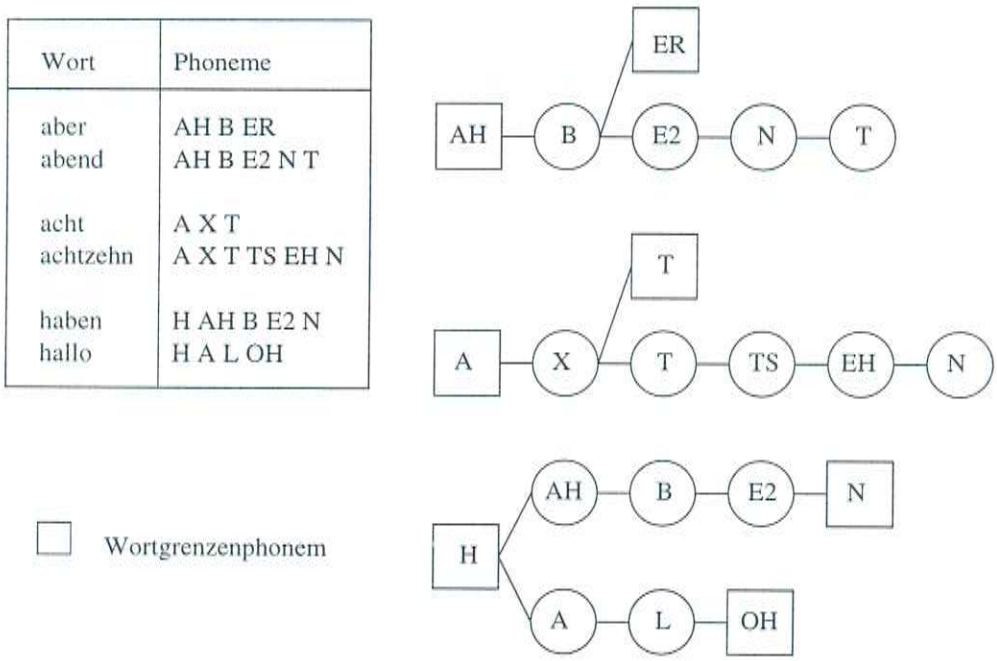


Abbildung 2.3: Transformation der Suchraumes

Mit diesen Vorüberlegungen läßt sich der Suchraum, genauer die dabei verwendeten HMMs, schon vor der eigentlichen Erkennung aufbauen. Verwendet man dabei für jedes Wort einen eigenen HMM-Zustandsgraphen, so wären genau so viele HMMs zu evaluieren, wie Wörter im Wortschatz enthalten sind, wobei nach Beendigung eines Wortes wiederum alle anderen Wörter folgen könnten. Da dies viel zu ineffizient wäre, wird der Suchraum in einen Baum, einen sogenannten „Allophone Tree" transformiert. Dabei werden die HMM-Zustandsfolgen aller Wortanfänge, deren Phonemsequenz identisch sind, zu einer HMM-Zustandsfolge zusammengefaßt. Die Wortenden, die unterschiedlich sind, bilden eigene Zustandsfolgen aus. Ein HMM-Zustand kann danach höchstens einen Vorgänger aber beliebig viele Nachfolger besitzen. Eine solche Transformation ist in Abbildung 2.3 beispielhaft dargestellt. Darin ist auch zu erkennen, daß es nach der Transformation für jedes unterschiedliche Wortanfangsphonem einen solchen Baum geben muß. Damit müssen zu einem Zeitpunkt nicht mehr die HMMs aller Wörter, sondern nur noch die HMMs aller Bäume evaluiert werden. Da es sich bei dem in dieser Arbeit verwendeten Spracherkennungssystem um eine zeitsynchrone Suche handelt, ist zu einem Zeitpunkt die Länge aller Viterbi-Pfade durch die einzelnen HMM-Zustandsgraphen für alle Bäume identisch.

Die Suche selbst läuft in drei Phasen [WCE+93],

- der Treeforward-Phase,
- der Flatforward-Phase und
- der Lattice-Phase

ab. Einen Überblick über die grobe Funktionsweise dieser drei Phasen bietet Abbildung 2.4.

In der ersten Phase werden die Bäume nach möglichen Wörtern, die für die spätere Hypothese in Frage kommen könnten, abgesucht. Dazu müßten theoretisch alle Pfade[14] aller Bäume evaluiert werden, wobei nach der Evaluierung eines Pfades, theoretisch wieder alle Wörter folgen könnten. Das ist natürlich viel zu aufwendig. Darum wird eine Suchraumbeschneidung eingeführt, bei der die Pfade des Suchraumes entfernt werden, deren bis zu einem gegebenen Zeitpunkt berechnete Wahrscheinlichkeit unter eine bestimmte Grenze fällt. Die Wahrscheinlichkeit der einzelnen Wörter wird dabei durch die Evaluierung der HMMs berechnet. Für die Berechnung der Wortübergangswahrscheinlichkeiten kommt das Sprachmodell ins Spiel. Dieses wird meist durch eine stochastische Grammatik definiert, bei der die Wahrscheinlichkeit eines Wortübergangs von dessen vorhergehenden Wörtern abhängt. Eine Einführung in die Theorie der Sprachmodelle ist beispielsweise in [Jel90] zu finden. Durch die Transformation der Wörter in Folgen von Phonemen ergibt sich während der Evaluierung der Pfade das Problem, daß die Wahrscheinlichkeit des Wortübergangs in dieses Wort erst am Ende des Wortes bekannt ist. Dieses Problem wird durch eine Methode namens „Delayed Bigram Approach" [WF96] gelöst, bei der erst, wenn das Wort bekannt ist, die Sprachmodellwahrscheinlichkeiten aufaddiert werden.

Innerhalb der ersten Phase wird der Suchraum quasi nur grob abgesucht, wobei nur ein kleines Bigramm-Sprachmodell und eine relativ starke Suchraumbeschneidung verwendet wird. Hierbei merkt sich der Suchalgorithmus in einer Wortmatrix nur die Eintritts- und Austrittszeitpunkte der Pfade eines bestimmten Wortes.

Die zweite Phase arbeitet auf der Grundlage der erzeugten Wortmatrix und ermittelt dort mit Hilfe des Viterbi-Algorithmus einen Pfad durch die Matrix. Das Ergebnis der Flatforward-Phase wird als Backpointer-Matrix bezeichnet.

In der letzten Phase wird die erzeugte Backpointer-Matrix in einen Worthypothesengraphen umgewandelt, um einfacher damit umgehen zu können. In diesem

[14] Unter einem Pfad in einem Baum, ist ein vollständiger Viterbi-Pfad von der Wurzel des Baumes bis zu einem Blatt durch die daran beteiligten HMM-Zustände zu verstehen.

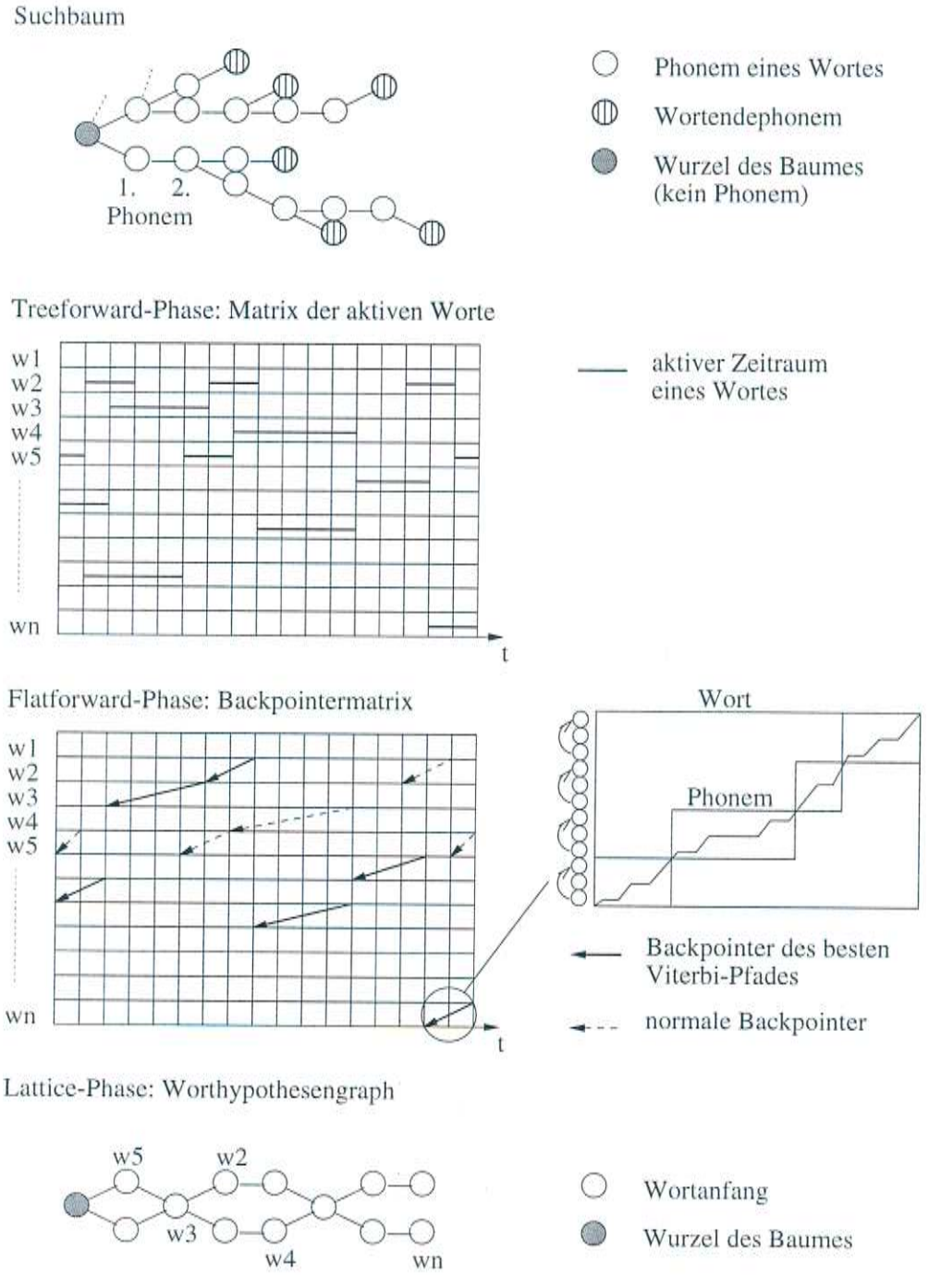


Abbildung 2.4: Ablauf der Suche

werden alle Wörter miteinander verbunden, deren Anfangs- und Endzeitpunkt durch einen Viterbi-Pfad verbunden waren oder die zeitlich gesehen direkt aufeinanderfolgen. Durch Verändern der Sprachmodellparameter kann der Einfluß des Sprachmodells innerhalb des Graphen auf einfache Weise nachträglich verändert werden, wodurch sich eine neue wahrscheinlichste Hypothese ergeben kann. Dieser Vorgang wird auch als Rescoring bezeichnet, weil die Wahrscheinlichkeiten innerhalb des Graphen neu berechnet werden müssen. Die beste Hypothese ergibt sich dann aus dem wahrscheinlichsten Pfad durch den Graphen. Die Gewichtung des Sprachmodells wird dabei auf einer Kreuzvalidierungsmenge einmal optimiert und während der Erkennung immer konstant gehalten. In [Füg98] wurde untersucht, ob sich die beste Gewichtung des Sprachmodells für jede Äußerung mit Hilfe von künstlichen neuronalen Netzen vorhersagen läßt, um mit Hilfe dieser eine bessere Hypothese als mit den fest eingestellten Gewichten zu erhalten.

Kapitel 3

Arbeiten anderer Autoren

In diesem Kapitel werden verschiedene Arbeiten anderer Autoren präsentiert, bei denen allen die Verminderung der Varianz, die im Parameterraum durch unterschiedliche Ausprägungen einer oder mehrerer Modalitäten entsteht, im Vordergrund steht. Zum Einsatz kommen dabei Normalisierungs- und Filtertechniken, aber auch Formen der Spezialisierung, also das Bereitstellen von einzelnen (Teil-) Systemen für unterschiedliche Ausprägungen.

Die Modalitäten lassen sich grob in zwei Gruppen aufteilen: die sprecherabhängigen und die aufnahmeabhängigen Modalitäten. Unter die sprecherabhängigen Modalitäten fallen zum einen die geschlechtsspezifischen Modalitäten, wie zum Beispiel die Vokaltraktlänge und zum anderen die aussprachspezifischen Modalitäten, wie zum Beispiel der Dialekt eines Sprechers. Ein Beispiel für eine aufnahmeabhängige Modalität ist der Signal-Rausch-Abstand einer Äußerung. Auf jede dieser Gruppen wird im folgenden genauer eingegangen. Zuvor werden jedoch noch allgemeine Adaptionsverfahren vorgestellt, die versuchen trainierte Modelle an die jeweils vorliegende Umgebung anzupassen. Sie decken gleichermaßen sprecherabhängige und aufnahmeabhängige Modalitäten ab.

Die Verwendung von markierten Modellen zur Ausweitung der divisiven Ballung auf nicht kontextabhängige Fragen ist schon seit längerem bekannt. Aus diesem Grund wird die Verwendung von Intra- und Inter-Wortmodellen erläutert und auf die Eingangs erwähnte Veröffentlichung von Wolfgang Reichl [RC99] eingegangen, die Ergebnisse für die Integration von geschlechtsabhängigen Modalitäten präsentiert.

Am Ende dieses Kapitels wird auf einen anderen Ansatz eingegangen, um unterschiedliche Modalitäten zu berücksichtigen. Dieser von Michael Finke vorgeschlagene Ansatz beschäftigt sich mit der modalitätenabhängigen Aussprachemo-

dellierung unter Verwendung von Aussprachevarianten, deren Auswahl mit Hilfe einer modalitätenabhängigen Wahrscheinlichkeitsfunktion bestimmt wird [FR97].

3.1 Akustische Adaptionsverfahren

Eine Möglichkeit, mehrere Modalitäten gleichmäßig zu berücksichtigen, ist die Verwendung von Adaptionstechniken. Hierbei werden zunächst allgemeine Modelle trainiert und dann an die jeweilige Umgebung angepaßt. Dies kann zum Beispiel mit Hilfe der Maximum-A-Posteriori Adaption (MAP) [GL94] erfolgen.

Eine andere Möglichkeit ist die Maximum Likelihood lineare Regression [LW94] [LW95], bei der die Systemparameter an die momentan zu erkennende Äußerung angepaßt werden. Dabei werden die Mittelwerte der Gauß-Mixturen so durch eine affine Abbildung transformiert, daß sie auf die vorliegende Äußerung optimal passen.

Das Problem bei der Verwendung solcher allgemeingültigen Modelle in Verbindung mit Adaptionstechniken ist, daß sie modalitätenabhängige Zusammenhänge nicht schon während der Modellierung berücksichtigen können. Außerdem benötigen Adaptionstechniken immer eine gewisse Zeit, bis sie sich an die jeweilige Umgebung optimal angepaßt haben. Aus diesem Grund eignen sie sich nicht dazu, stark veränderliche Modalitäten auszugleichen.

3.2 Sprecherabhängige Modalitäten

3.2.1 Geschlechtsspezifische Modalitäten

Vokaltraktlänge

Die anatomischen Unterschiede des Sprechapparates, genauer die Vokaltraktlänge und die Stimmbänder, zwischen Männer und Frauen führen zu einer unterschiedlichen Stimmlage und unterschiedlichen Resonanzfrequenzen des Vokaltraktes. Die Grundfrequenzanteile werden meist schon in der Vorverarbeitung weitestgehend eliminiert (siehe Abschnitt 2.3.1), so daß nur noch die unterschiedlichen Vokaltraktlängen berücksichtigt werden müssen. Hierzu werden meist adaptive Mechanismen in Verbindung mit Normalisierungstechniken eingesetzt.

Die einfachste Variante, männliche und weibliche Sprecher gleichermaßen zu berücksichtigen, ist die Verwendung von geschlechtsabhängigen Spracherkennungs-

systemen. Für die Erkennung muß auf Basis des Geschlechts des Sprechers das jeweilige Spracherkennungssystem ausgewählt werden. Die Identifikation des Geschlechts kann dabei entweder vor oder nach der Erkennung durchgeführt werden. Erfolgt sie erst nach der Erkennung, so wird auf Basis der Wahrscheinlichkeit der ermittelten Hypothesen die beste ausgewählt. Der Nachteil dabei ist, daß beide Spracherkennungssysteme eingesetzt werden müssen und damit doppelte Rechenzeit bzw. Ressourcen benötigen. Geschieht die Identifikation des Geschlechts schon vor der Erkennung, so reicht zur Bestimmung der Hypothese nur ein Spracherkennungssystem. Die Identifikation kann beispielsweise durch eine Grundfrequenzbestimmung erfolgen (siehe Abschnitt 5.1).

Das Verwenden von zwei getrennten Spracherkennungssystemen hat neben dem zusätzlichen Aufwand, der durch das Training zweier Systeme entsteht noch einen weiteren Nachteil. Durch die getrennte Modellierung stehen jedem Modell nur die Trainingsdaten eines Geschlechts zur Verfügung. Dies kann bei wenigen Trainingsdaten zu einer schlechteren Generalisierungsfähigkeit führen. Aus diesem Grund wäre es nützlich, wenn für Modelle mit wenig Trainingsdaten ein Parametersharing zwischen den Modellen beider Geschlechts stattfinden könnte. Hierzu müssen die Gemeinsamkeiten der beiden Geschlechter bei der akustischen Modellierung berücksichtigt werden, das am einfachsten durch die Verwendung eines gemeinsamen Spracherkennungssystems für beide Geschlechter möglich ist.

Eine andere Möglichkeit ist die Verwendung einer Vokaltraktnormierung [ZW97]. Dabei handelt es sich um eine Transformation des Spektralraumes, die versucht mit Hilfe einer stückweisen linearen Abbildung die Auswirkungen unterschiedlich langer Vokaltrakte zu kompensieren. Der optimale Verzerrungsfaktor wird so eingestellt, daß die Summe der akkumulierten Viterbi-Pfadwahrscheinlichkeiten über alle Viterbipfade des Sprechers maximiert wird. Anschaulich werden durch die sprecherabhängigen Verzerrungsfaktoren die Frequenzen der weiblichen Sprecher abgesenkt und die der männlichen Sprecher angehoben.

Wortwahl, Satzmelodie

In manchen Sprachen, wie zum Beispiel Japanisch können die Unterschiede zwischen den Geschlechtern sprachlich gesehen sogar soweit gehen, daß sich die Aussprache eines Wortes oder der verwendete Wortschatz stark unterscheidet. Außerdem kann sich die von Frauen verwendete Satzmelodie von derjenigen der Männer unterscheiden.

Das erste Problem wird meist durch eine Vergrößerung des Wortschatzes und durch eine Einführung von unterschiedlichen Aussprachevarianten behoben. Die

Wahl der jeweiligen Variante oder sogar des jeweiligen Wortes kann dabei entweder in Abhängigkeit vom Geschlecht des Sprechers [FR97] oder einfach ohne jegliche Beschränkung getroffen werden. Im letzten Fall kann es passieren, daß Worte des anderen Geschlechts in der Hypothese auftauchen.

Für die Berücksichtigung der Satzmelodie werden meist prosodische Merkmale [Kie96] [Kom96] verwendet. Dabei wird besonders das Änderungsverhalten der prosodischen Merkmale innerhalb einer Äußerung analysiert. Es besteht jedoch auch die Möglichkeit, solche Merkmale in der Vorverarbeitung zu berücksichtigen und in den Merkmalsvektor aufzunehmen [Sch99].

3.2.2 Aussprachespezifische Modalitäten

Die sprachlichen oder dialektalen Unterschiede in der Aussprache wurden zum Teil schon im obigen Abschnitt angesprochen. Der größte aussprachebedingte Unterschied zwischen zwei Sprechern ist wohl die Verwendung von unterschiedlichen Sprachen. Verwenden beide Sprecher dieselbe Sprache, so reduzieren sich die Unterschiede auf Dialekte oder Akzente.

Dialekte, Akzente

Innerhalb einer Sprache existieren meist eine Vielzahl an Dialekten oder Akzente. Das Einfügen aller Aussprachevarianten für jede dialektale Ausprägung eines Wortes kommt nicht in Frage, denn dadurch würde sich der Suchraum erweitern und die Ähnlichkeiten zwischen verschiedenen Worten zunehmen, so daß der Dekodierungsprozeß mehr Fehler produzieren würde [SKT98].

In [Ber97] [BSRB98] wurde ein Ansatz beschrieben, der versucht, durch Verwendung eines spezifischen Aussprachewörterbuchs für jeden Dialekt, der Problematik zu vieler Aussprachevarianten entgegenzuwirken. Für jeden Sprecher wurde auf Basis seines Dialekts das entsprechende Aussprachewörterbuch mit den entsprechenden Aussprachevarianten dieses Dialekts selektiert. Die akustische Modellierung blieb jedoch in allen Fällen dieselbe, wodurch sich auch nur insignifikante Verbesserungen der Fehlerrate einstellten.

Eine andere Technik ist die Adaption der akustischen Modelle auf andere Dialekte. Je nach dem Vorhandensein von Trainingsdaten wird zwischen drei Varianten unterschieden.

Bei der ersten Variante, ohne Trainingsdaten, wird ein vorhandenes System auf den neuen Zieldialekt abgebildet. In [FGJ98] wurden dazu Sprecher zu einzelnen

Klassen zusammengeballt, wobei dann für einen neuen Sprecher unbekannten Dialekts mit Hilfe eines Distanzmaßes die Modelle derjenigen Klasse ausgewählt wurden, in die der Sprecher am besten paßt. Aufgrund dessen, daß keinerlei Daten des fremden Dialekts während des Trainings gesehen wurden, kann diese Variante nie an die monodialektalen Ergebnisse heranreichen. Vorteilhaft an dieser Variante ist, daß die Klassenaufteilung dem Ausgangssystem nicht schadet [FGJ98], so daß ein auf diese Weise adaptiertes System auf dem Ausgangs- und dem Zieldialekt eingesetzt werden kann.

Bei der zweiten Variante sind nur wenige Trainingsdaten vorhanden, die dazu genügen müssen, um eine Parameteranpassung durchzuführen. In [FGJ98] wurde dafür die oben schon erwähnte MAP-Adaption verwendet. Bei dem Einsatz eines solchen Systems auf Ausgangs- und Zieldialekt leidet natürlich die Erkennungsleistung auf dem Ausgangsdialekt.

Bei der dritten Variante sind so viele Trainingsdaten vorhanden, daß diese entweder schon während des Trainings des Ausgangssystems verwendet werden können, oder um ein neues Zielsystem zu trainieren. Wird ein neues Zielsystem trainiert, so können die Parameter der Modelle des Ausgangssystems als initiale Parameter dem neuen System dienen, wodurch eine schnellere Konvergenz der Parameter während des Trainings erreicht wird.

Aufgrund der Vielzahl an verschiedenen Dialekten innerhalb einer Sprache ist die Verwendung eines eigenen Spracherkennungssystem für jeden Dialekt jedoch nicht sinnvoll, da zum einen die Voraussetzung genügend vieler Trainingsdaten für jedes System meist nicht erfüllt werden kann und zum anderen der Ressourcenbedarf anwächst. Deshalb faßt man ähnliche Dialekte zu größeren Sprachregionen zusammen und reduziert somit die Anzahl der benötigten Systeme.

Sprachen

Meist wird für jede Sprache ein eigenes System mit eigenem Wortschatz, Aussprachewörterbuch und Sprachmodell verwendet. Will man jedoch aufgrund der bei den Dialekten angeführten Probleme, die durch eine getrennte Modellierung entstehen, alle Sprachen in ein System integrieren, so muß der Phonemsatz einer Sprache auf andere Sprachen abgebildet werden. Manche Phoneme lassen sich jedoch nur sehr schwer anderen Phonemen zuordnen, weshalb in einem solchen System neben den multilingualen akustischen Modellen auch einige monolingualen Modelle verwendet werden. Ferner setzt ein solches multilinguales System auch ein multilinguales Aussprachewörterbuch voraus, wobei jede Sprache ihren

eigenen Wortschatz[1] und ihr eigenes Sprachmodell[2] verwendet.

Eine von Tanja Schultz in [SW98a] vorgeschlagene Lösung zur multilingualen Spracherkennung verwendet sprachenabhängige akustische Modelle, die rein datengetrieben divisiv geballt wurden, wobei zusätzlich zu den Kontextfragen auch Fragen zu den einzelnen Sprachen in den Fragenkatalog aufgenommen wurden. Der multilinguale Spracherkenner umfaßte 6 Sprachen und 78 multilinguale Phoneme. Die besten Resultate wurden dabei mit Hilfe eines multilingualen Aussprachewörterbuches erzielt, wobei die Menge der Wörter, die einem Sprecher zur Verfügung standen durch einen sprachenabhängigen Wortschatz eingegrenzt wurden. Mit diesem Ansatz konnte jedoch keine Verbesserung gegenüber monolingualen Spracherkennungssystemen erreicht werden. Dies lag zum einen an der manuellen Klassifizierung der monolingualen Phoneme zu multilingualen Phonemkategorien aber vor allem auch an den differierenden Kontexten in den verschiedenen Sprachen der Phonemkategorien. Für die letztere Problematik wird in [SW99] eine Lösung präsentiert, die eine Anpassung der Kontextentscheidungsbäume an die jeweilige Sprache vornimmt.

Bei den Dialekten wurde schon die Adaption der akustischen Modelle angesprochen. Diese Technik ist natürlich auch auf verschiedene Sprachen anwendbar. Es ist jedoch in den meisten Fällen leichter, ein Spracherkennungssystem auf einen anderen Dialekt gleicher Sprache zu adaptieren als auf eine andere Sprache, weil unbekannte Dialekte kaum neue, sehr unterschiedliche Kontexte enthalten und damit der Entscheidungsbaum unverändert weiterverwendet werden kann.

Sprechgeschwindigkeit

Die Sprechgeschwindigkeit stellt eine äußerst schwierig zu handhabende Modalität dar, weil sie sich über eine Äußerung nicht so wie die bisher behandelten Modalitäten statisch verhält. In zahlreichen Untersuchungen wurde die Problematik der Sprechgeschwindigkeit behandelt [SS95] [MFM95], wobei vor allem bei Schnellsprechern ein Einbruch in der Erkennungsleistung festzustellen war.

Im Grunde genügt damit eine Abgrenzung der Schnellsprecher von den anderen. Somit könnte man auf Basis einer vorher ermittelten Sprechgeschwindigkeit ein speziell auf Schnellsprechern trainiertes Spracherkennungssystem auswählen. Die Sprechgeschwindigkeit kann entweder auf zuvor bestimmten Labels berechnet oder mit Hilfe eines Maßes, das nur auf dem eigentlichen Sprachsignal arbeitet

[1] für die Selektion der jeweils verfügbaren Wörter der detektierten Sprache
[2] für die getrennte Modellierung der Grammatik

[MFL98], bestimmt werden.

Es wäre jedoch auch denkbar, die Sprechgeschwindigkeit mit in den Merkmalsvektor zu integrieren und somit aufgrund der Dimensionalitätszunahme die Varianzen der akustischen Modelle zu vermindern. Dies ist ohne weiteres möglich, da es sich hierbei um eine wertkontinuierliche Modalität handelt. Bei dieser Art der Integration werden alle Sprechgeschwindigkeiten in der akustischen Modellierung berücksichtigt.

Eine andere Methode ist die Adaption eines vorhandenen Spracherkennungssystems auf Schnellsprecher. Hierzu werden die Zustandsübergangswahrscheinlichkeiten der HMMs so verändert, daß die Modelle besser auf Schnellsprecher passen. Die Anpassung kann zum Beispiel mit Hilfe eines einfachen Skalierungsfaktors für bestimmte phonetische Gruppen [MFM96], oder nach dem Prinzip der Längenmodellierung [ASS95] durch Definition einer expliziten Wahrscheinlichkeitsverteilung für das Verbleiben in einem Zustand erreicht werden.

Spontansprache

Bei spontan gesprochenen Äußerungen treten neben den oben erwähnten Änderungen in der Sprechgeschwindigkeit auch noch andere Phänomene auf.

Das Phänomen der Tempowechsel wurde in [BKK96] anhand der im Rahmen von Verbmobil [VER] gesammelten deutschen Spontandaten untersucht. Es wurde festgestellt, daß vor allem in Passagen, die etwas bereits Gesagtes nochmals wiederholen, etwas schneller gesprochen wird.

Bei spontan gesprochenen Äußerungen bemerkt man eine Zunahme von Lautverschleifungen, Assimilationen und Elisionen [OB99], so daß das Aussprachewörterbuch mit solchen Varianten erweitert werden muß. Außerdem kann es durch Wortwiederholungen, Gedankensprüngen und ganz allgemein Satzumstellungen zu Problemen bei der grammatikalischen Modellierung kommen.

Neben diesen Phänomenen treten auch häufiger artikulatorische Geräusche wie Lachen, Husten und Häsitationen auf. Die Berücksichtigung solcher Geräusche geschieht durch explizite Geräuschmodelle [SR95]. Um diese trainieren zu können, müssen diese entweder in den Transkriptionen ebenfalls aufgeführt sein, oder sie werden als sogenannte optionale Füllwörter in die HMMs integriert (siehe Abschnitt 4.2).

3.3 Aufnahmeabhängige Modalitäten

Aufnahmeabhängige Modalitäten treten meist in Form eines Hintergrundrauschens im Sprachsignal auf. Ist das Rauschen nur schwach vorhanden so beeinflußt es kaum die Ergebnisse des Spracherkenners. Bei stärkerem Hintergrundrauschen ist es für ein Spracherkennungssystem jedoch nahezu unmöglich eine saubere Erkennung durchzuführen, wohingegen der Mensch keine Probleme hat auch stark verrauschte Äußerungen zu verstehen.

Das Beseitigen des Hintergrundrauschens kann mit Hilfe von mehreren Techniken erfolgen. Zum einen kann die oben angesprochene Maximum Likelihood lineare Regression eingesetzt werden. Da das Hintergrundrauschen meist als konstanter additiver Anteil im Spektralraum auftritt, liegt es nahe, diesen Anteil vom Spektrum wieder zu subtrahieren. Ein solcher Ansatz wird in [Bol79] beschrieben. Hierzu wird die spektrale Information, die dazu benötigt wird, um die spektralen Anteile des Hintergrundrauschens zu beschreiben, aus den Stille-Bereichen des Sprachsignals ermittelt.

Auch die Cepstrale Mittelwertsubtraktion (siehe Abschnitt 2.3.1) beseitigt zum Teil solche Frequenzüberlagerungen, da die additiven Anteile, die im Spektrum auftreten auch im Cepstrum vorhanden sind. Bei Verwendung des Signal-Rausch-Abstands als Maß für das Hintergrundrauschen, könnte man diesen, wie schon bei der Sprechgeschwindigkeit in den Merkmalsvektor integrieren, um dadurch die Varianz der akustischen Modelle zu vermindern.

In die Sparte der aufnahmeabhängigen Modalitäten fallen auch Hintergrundgeräusche, die vom Mikrophon zusätzlich erfaßt wurden. Hier kann es sich um alle Arten von Geräuschen handeln, wie zum Beispiel andere Gespräche, das Klingeln von Telefonen oder Türgeräusche. Nichtartikulatorische Geräusche lassen sich genauso wie die artikulatorischen Geräusche durch explizite akustische Modelle berücksichtigen [SR95]. Die größten Probleme bereiten den heutigen Spracherkennern jedoch andere Gespräche oder Sprecherüberlappungen, die in einer Äußerung vorhanden sind, da es sich hierbei um in der Lautstärke und Art veränderliche Geräusche von längerer Dauer handelt.

3.4 Verwendung markierter Modelle

Markierte Modelle werden schon seit längerem, vor allem zur Wortgrenzenmodellierung eingesetzt. Hierbei handelt es sich um eine statische Modalität, die als Markierung mit in das Aussprachewörterbuch übernommen wird. Auch das

Geschlecht ist eine statische Modalität. Jedoch macht es hier keinen Sinn eine solche Modalität ebenfalls als Markierung in das Aussprachewörterbuch zu übernehmen, da sich der Sprecher und damit auch das Geschlecht von Äußerung zu Äußerung ändern kann.

3.4.1 Intra- und Inter-Wortmodelle

Intra- und Inter-Wortmodelle werden seit [Lee88] in den meisten Spracherkennungssystemen eingesetzt. Mit Hilfe dieser lassen sich Koartikulationseffekte innerhalb und am Rande von Worten getrennt modellieren. Dies ist vor allem dann sinnvoll wenn ein Wortendephonem nicht mehr so deutlich ausgesprochen wird, wie das gleiche Phonem innerhalb eines Wortes. Da es sich bei den Wortgrenzenmarkierungen um äußerungsunabhängige Modalitäten handelt, können diese schon in das Aussprachewörterbuch integriert werden. Die Ballung von wortgrenzenabhängigen und nicht wortgrenzenabhängigen akustischen Modellen kann völlig datengetrieben mit einem der in Abschnitt 2.1 erwähnten Verfahren erfolgen. Bei der divisiven Ballung ist es außerdem üblich explizit Fragen zu Wortgrenzen zu erlauben. Es zeigte sich, daß eine solche Modellierung zu deutliche Verbesserungen in der Erkennungsleistung führt. Die zusätzliche Unterscheidung zwischen Wortanfängen und Wortenden bringt jedoch kaum mehr einen zusätzlichen Gewinn.

3.4.2 Geschlechtsspezifische Modalitäten

In [RC99] wurde ein Ansatz präsentiert, der markierte Modelle verwendet, um diese divisiv zu Ballen. Dabei wurden zusätzlich zu den Kontextfragen auch Fragen zu den einzelnen Markierungen in den Fragenkatalog aufgenommen. Das Training erfolgte über einen zweistufigen Ballungsalgorithmus, der sich von dem in Abschnitt 2.2 erläuterten Training dadurch unterscheidet, daß die divisive Ballung der akustischen Modelle, die initiale Codebuchbestimmung durch den k-Mittelwerte Algorithmus und das anschließende EM-Training miteinander vereinigt wurden [RC98]. Die Bestimmung der Codebuchgröße erfolgte durch einen iterativen Wachstumsprozeß, bei dem zu Anfang alle akustischen Modelle durch eine Gaußverteilung modelliert und in den folgenden Schritten immer weiter aufgeteilt werden [WOVY94]. Für die Ballung der akustischen Modelle wurde das Likelihood-Distanzmaß verwendet.

Neben den oben erwähnten äußerungsunabhängigen Wortgrenzenmarkierungen wurden auch Ergebnisse für äußerungsabhängige Geschlechts-Markierungen prä-

sentiert. Das Geschlecht eines Sprechers wurde dabei als bekannt vorausgesetzt. Die Fehlerratenreduktion lag dabei, je nach verwendeter Evaluierungsmenge zwischen 3.1% und 8.6% bei Verwendung eines 20000 Wörter umfassenden englischen Wörterbuchs.

Die dort aufgeführte Analyse der Entscheidungsbäume zeigte, daß Vokale und Diphthonge am meisten von einer getrennten Modellierung profitierten, während bei Stopps und Frikativen bis zu 34% der Modelle für beide Geschlechter identisch waren. Dies ist dadurch zu begründen, daß die geschlechtsspezifischen Merkmale in stimmhaften Lauten viel stärker hervortreten als in stimmlosen Lauten.

Dieser Ansatz beschränkte sich nur auf die Verwendung von statischen Modalitäten. Die Problematik, die bei der Berücksichtigung von zeitveränderlichen Modalitäten, also Modalitäten, deren Wert sich während einer Äußerung ändern kann, entsteht (siehe Kapitel 4), wurde nicht angesprochen.

3.5 Modalitätenabhängige Aussprachemodellierung

Die Problematik für das Hinzufügen von vielen Aussprachevarianten zum Aussprachewörterbuch wurde oben angesprochen. Ein anderes Problem ergibt sich durch die unterschiedliche Verwendung von Aussprachevarianten in den Transkriptionen. Diese beiden Probleme werden in [FR97] durch eine modalitätenabhängige Aussprachemodellierung angegangen. Die Berücksichtigung von wortübergangsabhängigen Aussprachephänomenen kann nur über die Modellierung von wortübergangsabhängigen Polyphonen erreicht werden und damit nur die daran beteiligten Phoneme und nicht die ganzen Wörter verwenden. Aus diesem Grund wurden wortübergangsabhängige Aussprachephänomene durch sogenannte Multiwörter in das Aussprachewörterbuch mit aufgenommen. Für andere Aussprachevarianten wurde ein Regelwerk aufgestellt, dessen Anwendung auf eine Basisform, die häufigsten Aussprachevarianten produziert.

Damit war es möglich, fehlerhafte Transkriptionen durch ein sogenanntes „Flexible Transcription Alignment" zu korrigieren. Hierbei handelt es sich um die Berechnung eines Viterbi-Pfades durch ein mit Multiwörtern, Aussprachevarianten und optionalen Geräuschen angereicherten HMMs. Die Aussprachevarianten und die Multiworte wurden dabei auf Basis der Transkription der momentan vorliegenden Äußerung und den zuvor definierten Regeln ermittelt. Damit konnte eine neue Transkription ermittelt werden, die zusätzlich auch noch die Ausspracheva-

rianten optimal berücksichtigt.

Für die Erkennung einer Äußerung sollten die Aussprachevarianten und Multiwörter ebenfalls mit einbezogen werden. Hierzu wird unter Verwendung der oben ermittelten Transkriptionen für jede Regel deren Wahrscheinlichkeit in Abhängigkeit des momentan vorliegenden phonetischen Kontextes und der in der dazugehörigen Äußerung auftretenden Modalitätenkombination geschätzt. Dazu werden zunächst alle in den Transkriptionen vorkommenden Varianten, das heißt die zu einer Regel gehörigen Kontexte und Modalitätenkombinationen, gesammelt. Um die Wahrscheinlichkeit einer Regel in Abhängigkeit des momentan vorliegen Kontextes und der Modalitätenkombination zu bestimmen, werden Entscheidungsbäume verwendet, die mit Hilfe der oben extrahierten Varianten aufgebaut werden.

Auf Basis des vorliegenden Kontextes und der extrahierten Modalitätenkombination kann nun durch Absteigen in den erzeugten Entscheidungsbäumen die Wahrscheinlichkeit einzelner Aussprachevarianten vorhergesagt werden. Damit werden einige Aussprachevarianten bevorzugt ausgewählt, während einige andere benachteiligt werden. Die Ergebnisse dieses Ansatzes zeigten, daß es durchaus lohnenswert ist, Aussprachevarianten in Abhängigkeit von einzelnen Modalitäten explizit zu gewichten, um deren Auswahl zu bevorzugen. Die Fehlerrate des Referenzsystems konnte dabei um 13% reduziert werden, wobei nur sprechgeschwindigkeitsabhängige Modalitäten verwendet wurden.

Kapitel 4

Integration der Modalitäten

Das vorherige Kapitel zeigte schon einige Probleme auf, die bei einer gerechten Berücksichtigung von allen Ausprägungen mehrerer Modalitäten entstehen können. Bei der Modellierung durch getrennte Systeme für einzelne Teile des Wertebereichs der Modalitäten kann kein Parametersharing stattfinden. Dies wäre jedoch für einzelne Bereiche des Parameterraums, in die nur wenige Trainingsdaten fallen, sehr sinnvoll, um robustere Modelle zu erhalten. Ein anderes Problem entsteht zum Beispiel bei multidialektalen Spracherkennungssystemen durch die zusätzliche Aufnahme von Aussprachevarianten für die einzelnen Dialekte. Diese sind nötig, um die Erkennungsleistung zu verbessern. Werden jedoch zu viele bzw. zu viele ähnliche Aussprachevarianten aufgenommen, so wird zum einen der Suchraum aufgebläht und zum anderen erhöht sich die Verwechslungsgefahr zwischen den einzelnen Einträgen des Aussprachewörterbuchs. Dies führt oftmals zu einem Abfall in der Erkennungsleistung. Der Einsatz von adaptiven Verfahren kommt bei zeitveränderlichen Modalitäten nicht in Frage, weil zu spät Veränderungen der Modalitäten erfaßt werden. Die Verwendung von Normalisierungsverfahren erlaubt keine Berücksichtigung von modalitätenabhängigen Eigenschaften. Diese gehen fast alle durch die Normalisierung verloren. Es wäre jedoch angebracht, wenn die Ballung auch solche Eigenschaften berücksichtigen könnte. Ein weiterer Nachteil ist, daß die Normalisierung meist über die ganze Äußerung auf dieselbe Art und Weise erfolgt, so daß einige Bereiche der Äußerung besser und andere wiederum schlechter normalisiert werden. Dadurch geht die Zielsetzung, nämlich eine Varianzverminderung der Modelle zu erreichen, teilweise wieder verloren. Eine Varianzverminderung ist deshalb nötig, weil Überlappungen im Parameterraum, die durch zu große Varianzen der akustischen Modelle enstehen, die Erkennungsleistung negativ beeinflussen.

Das Ziel dieser Arbeit war die Entwicklung eines Verfahrens, das ein Parametersharing bei gleichzeitiger Varianzverminderung erlaubt. Modalitäten wie das Geschlecht, die Sprechgeschwindigkeit oder der Signal-Rausch-Abstand sollten alle auf dieselbe Art und Weise behandelt werden, wobei Abhängigkeiten der Modalitäten untereinander automatisch berücksichtigt werden sollten. Zusätzlich dazu sollte es möglich sein, auch zeitveränderliche Modalitäten sinnvoll zu berücksichtigen.

In diesem Kapitel wird gezeigt, wie die Integration der Modalitäten in das verwendete Spracherkennungssystem erreicht wurde. Die Zielsetzung erforderte zunächst einige Vorüberlegungen über zu stellende Forderungen, über das zur Realisierung verwendete Konzept und über die durch das verwendete Spracherkennungssystem und dessen Rechnerumgebung vorgegebenen Einschränkungen.

4.1 Vorüberlegungen

4.1.1 Forderungen

An die Integration der Modalitäten in den Spracherkennungsprozeß wurden die folgenden Forderungen gestellt:

Modularität: Die Integration der Modalitäten sollte möglichst modular erfolgen, das heißt ohne deren Benutzung sollte das Spracherkennungssystem wie gewohnt verwendet werden können.

Erkennungsgeschwindigkeit: Der Dekodierungsprozeß sollte durch die Integration der Modalitäten nicht oder nur wenig an Geschwindigkeit verlieren. Wollte man ein langsameres modalitätenabhängiges Spracherkennungssystem mit einem anderen vergleichen, so müßte vorausgesetzt werden, daß beide in etwa gleich viel Zeit für die Erkennung einer Äußerung benötigen. Dazu müßte der Suchraum des modalitätenabhängigen Systems wesentlich mehr beschnitten werden, wodurch der Gewinn, der durch die Integration der Modalitäten erreicht wurde, wieder verloren ginge.

Zeitveränderliche Modalitäten: Zeitveränderliche Modalitäten sind solche, deren Wert sich während einer Äußerung ändern kann. Hierzu zählt zum Beispiel die Sprechgeschwindigkeit oder ein Sprecherwechsel während einer Äußerung. Da es Ziel dieser Arbeit ist, die Varianzen der einzelnen akustischen Modelle zu verringern, sollten auch zeitveränderliche Modalitäten

berücksichtigt werden. Dies bedeutet aber auch, daß die Extraktion der Modalitäten über die gesamte Aufnahme erfolgen muß, weshalb diese möglichst schnell erfolgen sollte.

Schritthaltende Erkennung: Wird ein Spracherkennungssystem in der Praxis eingesetzt, so ist es oftmals wünschenswert, die Erkennung noch während der Sprecher spricht zu beginnen (schritthaltende Erkennung). Dadurch wird die Antwortzeit des Spracherkennungssystems wesentlich verkürzt. Für die Modalitäten sollte also ein Mechanismus vorgesehen sein, der eine schritthaltende Aktualisierung dieser erlaubt.

4.1.2 Konzept

Extraktion der Modalitäten

Die Berücksichtigung von zeitveränderlichen Modalitäten verlangt deren Extraktion über die gesamte Äußerung. Eine Äußerung wird von der Vorverarbeitung durch ein über die Aufnahme gleitendes Fenster in mehrere sich überschneidende Intervalle[1] unterteilt, wodurch es als sinnvoll erscheint, die Extraktion der Modalitäten ebenfalls auf diese Intervalle als Grundeinheit zu beschränken. Somit wird jedem Intervall ein Wert zugeordnet, der sich im Falle zeitveränderlicher Modalitäten von Intervall zu Intervall unterscheiden kann. Ändert sich eine Modalität innerhalb der vorliegenden Äußerung nicht, so wird jedem Intervall derselbe Wert zugeordnet.

Da die Extraktion der Modalitäten möglichst schnell erfolgen sollte, sollte es gerade bei über der Zeit konstanten Modalitäten möglich sein, diese nur auf einem Teil der Äußerung zu extrahieren. Die Zuordnung der Werte zu den restlichen Intervallen sollte dann automatisch erweitert werden. Außerdem sollten zur Extraktion der Modalitäten, von Modalität zu Modalität unterschiedliche Verfahren eingesetzt werden können.

Aus diesen Gründen erfolgt die Extraktion der Ausprägungen einer Modalität über eine sogenannte Aktualisierungsfunktion, die die jeweiligen Eigenheiten, die es bei der Extraktion einer bestimmten Modalität zu beachten gilt, berücksichtigt. Diese erlaubt es auch auf sehr einfache Weise, neue Modalitäten in das Spracherkennungssystem zu integrieren oder die Extraktionsverfahren bereits bestehender Modalitäten abzuändern.

[1] Im folgenden wird dafür auch der Begriff des „Fensterausschnittes" verwendet. Im englischen ist die Bezeichnung „Frame" üblich.

Modalitätenabhängige Modelle

Die Berücksichtigung der Modalitäten in der akustischen Modellierung erfolgt durch Markierung der akustischen Modelle. Durch die Markierung der akustischen Modelle nimmt die Flexibilität der akustischen Modellierung zu. Die akustischen Modelle müssen nicht mehr nur nach deren Kontext geballt werden, sondern es können auch andere Merkmale bzw. Modalitäten berücksichtigt werden, die ein akustisches Modell in seiner Varianz beeinflussen.

Für das Training wird dabei jedes in einer Äußerung vorkommende Subpolyphon mit dessen Modalitätenkombination[2] markiert, wobei die Subpolyphone durch die Transkription der Äußerung gegeben sind. Die Ballung der akustischen Modelle geschieht divisiv unter Verwendung eines vorher definierten Fragenkatalogs. Durch das zusätzliche Aufnehmen von Modalitätenfragen in den Fragenkatalog ist es dann möglich, akustische Modelle nicht nur nach deren Kontext, sondern auch nach deren Modalitätenkombination aufzuspalten. Somit ergeben sich modalitätenabhängige Kontextentscheidungsbäume. Modalitätenfragen beziehen sich immer auf das aktuelle Phonem bzw. Subpolyphon, da zum einen die Modalitäten über einen längeren Zeitraum konstant sind und zum anderen die Aussprache eines Phonems nicht davon abhängt, mit welcher Modalitätenkombination was in der Vergangenheit gesprochen wurde oder in der Zukunft noch gesprochen wird. Beispielsweise ist die Aussprache eines Phonems nicht davon abhängig, ob das Vorgängerphonem von einem männlichen oder weiblichen Sprecher gesprochen wurde, sondern nur davon abhängig welcher Sprecher das aktuelle Phonem wirklich ausgesprochen hat. Abbildung 4.1 zeigt einen Ausschnitt eines modalitätenabhängigen Entscheidungsbaumes, der als Grundlage für die folgenden Beschreibungen dienen soll.

Bei der Erkennung können dann auf Basis der extrahierten Modalitätenkombinationen einer vorliegenden Äußerung im Entscheidungsbaum die Modelle ermittelt werden, die mit Hilfe dieser Modalitätenkombinationen trainiert wurden. Für die Erkennung der Äußerung werden also nur die akustischen Modelle der momentan vorliegenden Modalitätenkombinationen berücksichtigt.

Die jeweiligen Modalitätenkombinationen müssen allerdings schon vor der Berechnung der Wahrscheinlichkeiten der akustischen Modelle bekannt sein. Dies gilt gleichermaßen für die Erkennung und das Training.

[2]Der Einfachheit halber wird die Kombination der Ausprägungen aller verwendeter Modalitäten für eine Äußerung mit Modalitätenkombination bezeichnet.

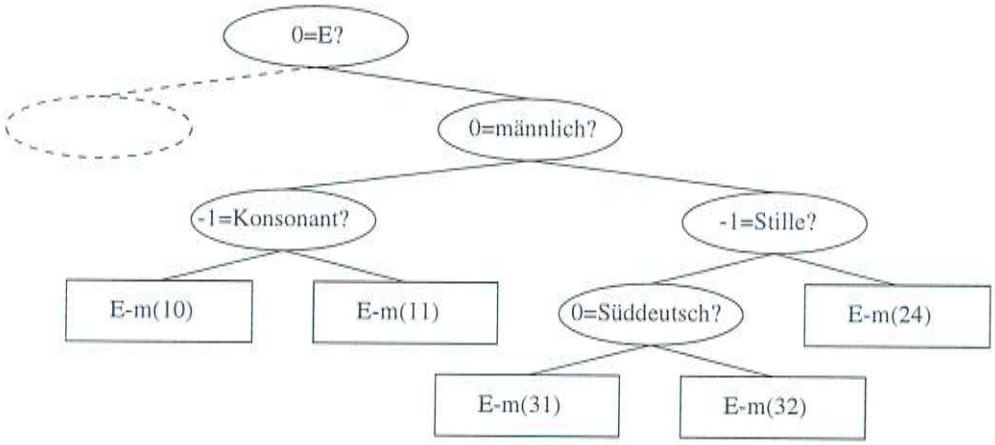


Abbildung 4.1: Modalitätenabhängiger Entscheidungsbaum

Modularer Aufbau

Für die Integration der Modalitäten werden drei Objekte eingeführt.

Modalität: Das Modalitäten-Objekt ist für das Aufrufen der entsprechenden Aktualisierungsfunktion und für die Verwaltung der extrahierten Werte verantwortlich. Auch wird hier der jeweilige Modus, nach dem die Aktualisierung der Werte erfolgen sollte, festgelegt. Als Modus kommt zum Beispiel eine schritthaltende Aktualisierung oder eine globale Aktualisierung in Frage.

Modalitätenmenge: Die Modalitätenmenge faßt alle Modalitäten zusammen und bietet Funktionen für die gemeinsame Behandlung aller in ihr enthaltenen Modalitäten an.

Lookup-Tabelle: Um die Erkennungsgeschwindigkeit nicht zu erhöhen, ist es nötig, eine sogenannte Lookup-Tabelle einzuführen, in der für jede mögliche Modalitätenkombination und jedes mögliche Subpolyphon, dessen zugehöriges akustisches Modell nachgeschlagen werden kann.

4.1.3 Einschränkungen

Aufgrund des in dieser Arbeit verwendeten Spracherkennungssystems und der dazugehörigen Rechnerumgebung müssen einige Einschränkungen in Kauf genommen werden.

Wertkontinuierliche Modalitäten: Der Entscheidungsbaum erlaubt nur Fragen mit binären Antworten, wobei die Anzahl der Fragen durch die Anzahl der zur Verfügung stehenden Markierungen begrenzt ist. Somit ist es bei wertkontinuierlichen Modalitäten nicht möglich, jeden Wert durch eine Frage explizit abzudecken. Durch eine Quantisierung des Wertebereichs kann diese Einschränkung jedoch weitestgehend überwunden werden. Das Problem dabei ist jedoch, daß sich die Auswirkungen des dabei entstehenden Quantisierungsfehlers auf die Erkennungsleistung nicht vorhersagen lassen.

Kontextbreite, Dimensionalität der Modelle: Durch die Markierung der akustischen Modelle steigt die Anzahl der in der Trainingsdatenbasis vorkommenden verschiedenen Modelle stark an. Dadurch ergeben sich zwei Probleme. Zum einen ist die große Anzahl an Modellen irgendwann nicht mehr geeignet handhabbar und zum anderen steigt der Speicherplatzaufwand der benötigten Parameter auf unerträgliche Ausmaße. Aus diesem Grund wurden für alle in dieser Arbeit trainierten Spracherkennungssystemen nur Triphone als Grundeinheit verwendet. Ferner bestanden die Codebücher der akustischen Modelle vor der Ballung nur aus 16 statt 32 Referenzvektoren je Codebuch. Erst nach der Ballung wurde die Anzahl der Referenzvektoren auf 32 erhöht.

Anzahl der Modalitäten: Die Anzahl der verschiedenen Modalitäten ist auf eine kleine Anzahl beschränkt. Diese ist, wie später noch ersichtlich wird, gegeben durch den zur Verfügung stehenden Speicherplatz.

4.2 Praktische Realisierung

4.2.1 Modalitätenabhängiges Training

Das Training erfordert für jede Äußerung das Aufbauen eines HMM-Zustandsgraphen, wobei mit Hilfe des Entscheidungsbaumes jedem Zustand ein Modell zugewiesen wird. Der Aufbau des HMM-Zustandsgraphen wird durch die jeweilige Transkription einer Äußerung fest vorgegeben. Hierzu wird jedes Wort aus der Transkription unter Verwendung des Aussprachewörterbuchs in eine Phonemsequenz umgewandelt, welche dann wiederum in Subpolyphone aufgespalten wird.

Das Anfertigen von guten Transkriptionen ist sehr zeitaufwendig. Trotz aller dabei aufgebrachten Sorgfalt unterlaufen den Transkribierern immer noch ei-

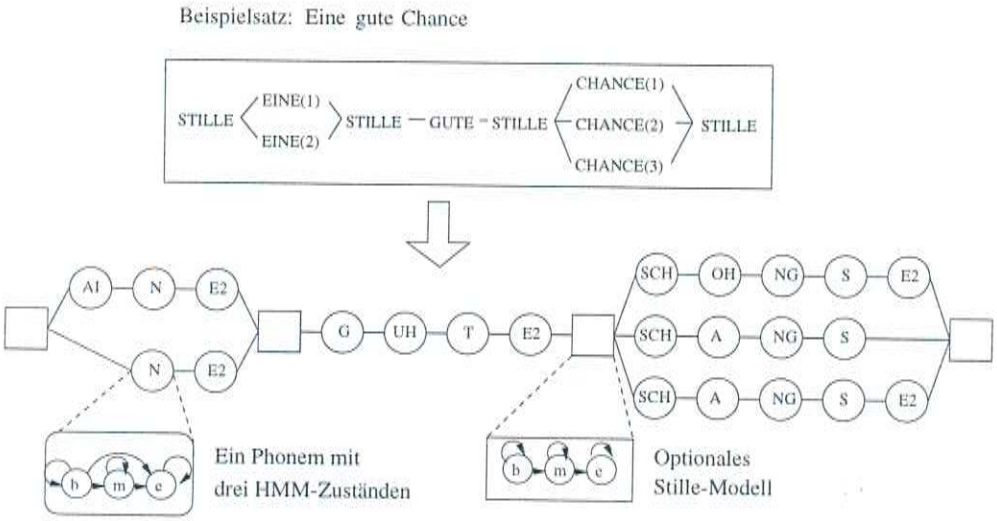


Abbildung 4.2: Vollständiges HMM eines Satzanfangs

nige Fehler. In manchen Fällen ist die Aussprache eines Wortes einfach nicht genau zu identifizieren. Aus diesem Grund werden beim Aufbau des HMM-Zustandsgraphen alle möglichen Aussprachevarianten eines Wortes in der Transkription mitberücksichtigt. Ferner sind kurze Pausen oder Atemgeräusche kaum hörbar, weshalb sie in der Transkription nur selten auftauchen. Deshalb werden beim Aufbau der HMM-Zustandsgraphen sogenannte Füllwörter berücksichtigt. Dabei handelt es sich meist um Sprachpausen oder Atemgeräusche, die immer zwischen zwei Wörter der Transkription eingebaut werden. Ein solches HMM mit Aussprachevarianten und der Stille als Füllwort ist für einen Teil eines Satzes in Abbildung 4.2 dargestellt. Bei der Evaluierung des HMMs durch den Viterbi-Algorithmus wird dann jedem Fensterausschnitt der Aufnahme genau ein HMM-Zustand zugeordnet.

Die Zuordnung der akustischen Modelle zu einzelnen HMM-Zuständen erfolgt also ohne einen Zusammenhang zum jeweils vorliegenden Sprachsignal. Bei modalitätenabhängigen akustischen Modellen ergibt sich dann das Problem, daß die Zuordnung zwischen einzelnen HMM-Zuständen und extrahierten Merkmalskombinationen der vorliegenden Äußerung nicht gegeben ist. Diese Information wird jedoch gerade bei zeitveränderlichen Modalitäten benötigt.

Für die Lösung dieses Problems werden die zuvor abgespeicherten Labels eines kontextabhängigen Systems ohne Modalitäten verwendet. Da im allgemeinen eine Veränderung einer Modalität innerhalb eines Wortes ignoriert werden kann, genügt es Start- und Endzeitpunkte aller Wörter der Transkription zu bestimmen.

Mit Hilfe dieser Zuordnung ist es dann möglich, für jedes Wort dessen Modalitätenkombination vor dem Aufbau eines HMM-Zustandsgraphen zu bestimmen. Während des Aufbaus des HMM-Zustandsgraphen wird dann zusätzlich zur Kontextinformation auch die jeweilige Modalitätenkombination durch Markierungen in das Subpolyphon integriert. Durch Abstieg im Entscheidungsbaum kann dann das richtige Modell für dieses modalitätenabhängige Subpolyphon gefunden werden.

4.2.2 Modalitätenabhängige Erkennung

In Abschnitt 2.3.2 wurde bereits der Dekodierungsprozeß beschrieben. Hierbei fiel auf, daß der Aufbau der Suche nur dann möglich ist, wenn der Entscheidungsbaum nur statische Fragen enthält. Genau das ist aber bei Verwendung von Modalitätenfragen nicht immer der Fall. Um einen Aufbau der Suche trotz Modalitätenfragen zu ermöglichen, wird der Entscheidungsbaum geeignet transformiert.

Transformation des Entscheidungsbaumes

Die Transformation des Entscheidungsbaumes wird im folgenden anhand des in Abbildung 4.1 dargestellten Beispielbaums erläutert werden. Hierzu werden die Kontextfragen so weit nach unten geschoben, bis keine Kontextfrage mehr zwischen zwei Modalitätenfragen vorhanden ist. Dadurch entstehen an den Baumenden Teilbäume, die nur Modalitätenfragen und die dazugehörigen Modelle enthalten. Jeder dieser Teilbäume wird jetzt durch ein imaginäres akustisches Modell ersetzt. Die Teilbäume werden, um keine Geschwindigkeitseinbußen zu erhalten, in eine Tabelle umgesetzt. Die Spalten der Tabelle sind durch alle Modalitätenkombinationen und die Zeilen der Tabelle durch alle imaginären Modelle gegeben.

Bei dem Hinunterschieben der Modalitätenfragen ist natürlich auf einen konsistenten Baum zu achten. Die Einschränkung, daß nur Fragen mit binären Antworten erlaubt sind, vereinfacht die Transformation. Zusätzlich dazu sind auch noch UND- und ODER-Fragen erlaubt, die im Falle von gemischten Fragen, das heißt Kontext- und Modalitätenfragen noch aufgeteilt werden müssen.

Eine Aufteilung von UND-Fragen ist relativ einfach. Dies wird durch Aneinanderhängen der Einzelfragen erreicht, wobei der Nein-Teil für alle Fragen identisch ist. Dabei ist es nicht nötig alle Teilfragen der UND-Frage extra auszubilden, sondern es genügt, wenn die Kontextfragen von den Modalitätenfragen getrennt

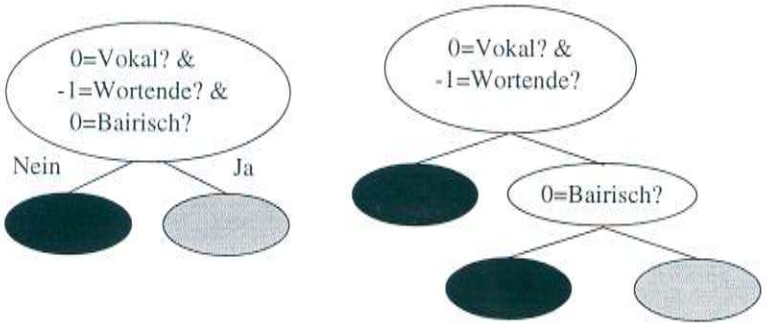


Abbildung 4.3: Aufteilung einer UND-Frage

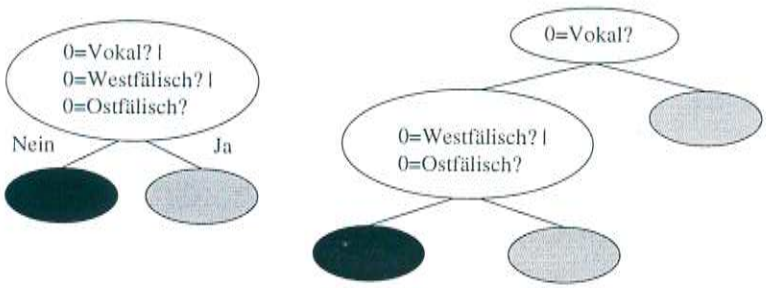


Abbildung 4.4: Aufteilung einer ODER-Frage

werden. UND-Fragen, die nur Kontext- oder nur Modalitätenfragen enthalten, sind unkritisch. Abbildung 4.3 zeigt das Ergebnis einer solchen Aufteilung.

Die Aufteilung einer ODER-Frage geschieht nach demselben Schema, nur ist jetzt der Ja-Teil aller Fragen identisch und die einzelnen Teilfragen werden immer an den Nein-Teil gehängt. In Abbildung 4.4 ist ein Beispiel einer solchen Aufteilung aufgeführt.

Die Transformation selbst, also das Verschieben einer Modalitätenfrage kann sich der Leser am besten anhand der Abbildung 4.5 klarmachen. Die Transformation geschieht dabei gemäß den Booleschen Gesetzen, wobei immer der linke Nachfolger den Elter, also die Modalitätenfrage, ersetzt. Die Fragen „0=männlich?“ und Frage „0=Süddeutsch?“ aus Abbildung 4.1 wurden soweit nach unten geschoben, bis keine anderen Fragen außer Modalitätenfragen zwischen zwei Modalitätenfragen mehr vorhanden waren.

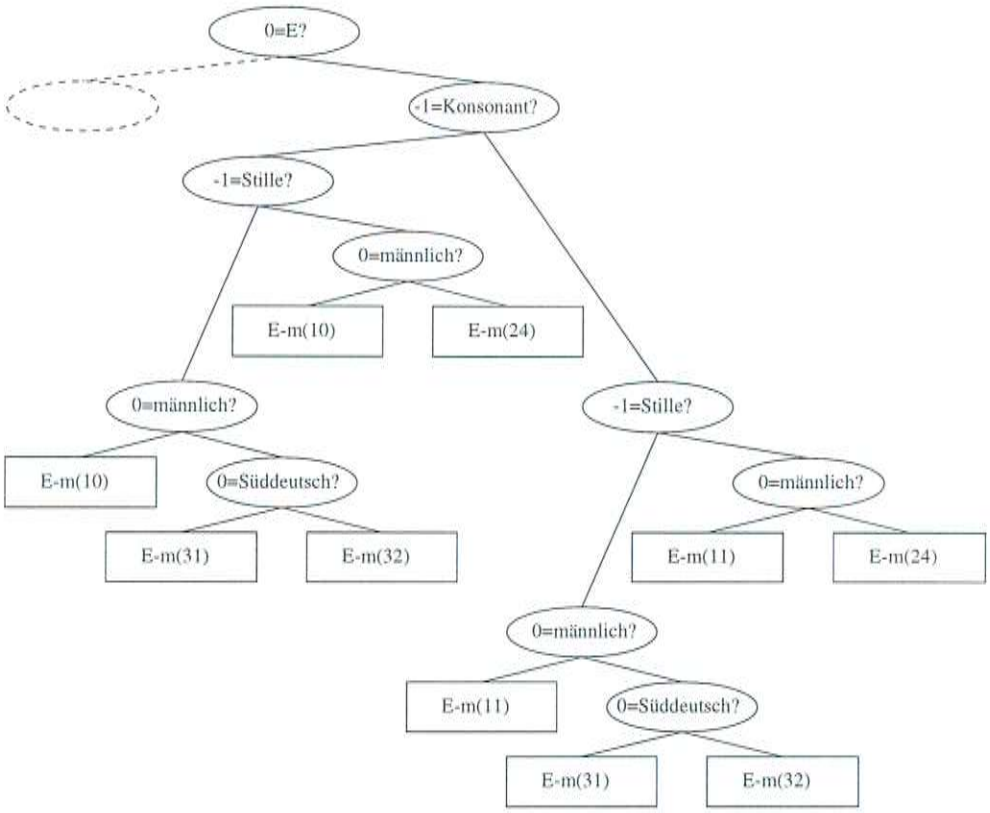


Abbildung 4.5: Verschieben der Modalitätenfrage aus Abbildung 4.1

Aufbau der Lookup-Tabelle

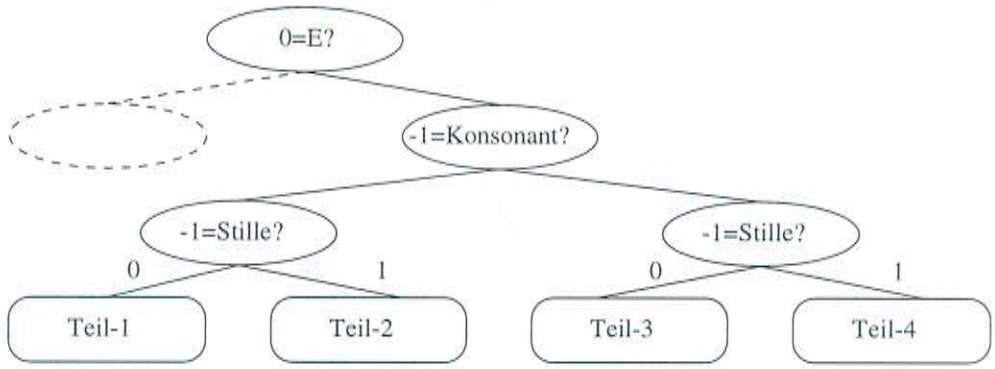
Sind alle Modalitätenfragen nach unten geschoben worden, so lassen sich die Teile des Entscheidungsbaumes, die nur Modalitätenfragen enthalten, abspalten und durch imaginäre akustische Modelle ersetzen.

In Kapitel 2.3.2 wurde schon erwähnt, daß es aus Effizienzgründen nötig ist, den Suchbaum statisch und vor der eigentlichen Erkennung aufzubauen. Die einzelnen Verweise auf die akustischen Modelle werden dabei mitintegriert. Jedesmal, wenn für ein Fensterausschnitt die Suche auf ein imaginäres akustisches Modell trifft, wird vor Berechnung der Wahrscheinlichkeit für diesen Fensterausschnitt das imaginäre Modell durch seinen realen Gegenpart ersetzt. Dieser hängt immer von der jeweiligen Modalitätenkombination ab, die für diesen Fensterausschnitt bestimmt wurde.

Für die Effizienz der Suche wäre es nicht vorteilhaft, wenn das entsprechende reale Modell immer erst durch einen Abstieg in einem der Teilbäume gefunden werden müßte. Aus diesem Grund wird im voraus eine Tabelle aufgebaut, die für jeden Teilbaum und für jede Modalitätenkombination das reale akustische Modell enthält. Damit reduziert sich das Durchsuchen eines Teilbaumes auf das einfache Identifizieren eines Tabelleneintrags. Dieser letzte Schritt, das heißt die Umsetzung der Teilbäume in eine Tabelle, ist in Abbildung 4.6 für den Baum aus Abbildung 4.5 dargestellt.

Die Tabelle ist auch verantwortlich für die Beschränkung der Anzahl der Modalitäten. Der Speicherverbrauch einer solchen Tabelle ist immens. Bei beispielsweise 20 binären Modalitäten, also ungefähr 10^6 Kombinationen würde eine Zeile der Tabelle bereits auf einen Speicherumfang von ungefähr 4 MB, bei Verwendung von 4 Byte für einen Eintrag, kommen. Würden jetzt bei der Transformation des Entscheidungsbaumes 100 Teilbäume entstehen, so würde die Tabelle ungefähr 400 MB verbrauchen. Da die Tabelle aufgrund dessen, daß sie relativ oft von der Suche frequentiert wird, in den Hauptspeicher passen sollte, ist es nötig, die Anzahl der verschiedenen Modalitäten zumindest bei der Erkennung[3] auf unter 20 zu begrenzen. Eine andere Möglichkeit ist es, die Tabelle zu komprimieren, um so den Speicheraufwand zu minimieren. Die Kompression bzw. die für das Lesen der Tabelleneinträge benötigte Dekompression sollte die Erkennungsgeschwindigkeit des Sprachsystems jedoch nicht negativ beeinflussen.

[3] denn nur hier wird die Tabelle benötigt



männlich	0		1	
Süddeutsch	0	1	0	1
Teil-1	E-m(10)	E-m(10)	E-m(31)	E-m(32)
Teil-2	E-m(10)	E-m(10)	E-m(24)	E-m(24)
Teil-3	E-m(11)	E-m(11)	E-m(31)	E-m(32)
Teil-4	E-m(11)	E-m(11)	E-m(24)	E-m(24)

Abbildung 4.6: Umsetzung der Teilbäume aus Abbildung 4.5 in eine Tabelle

Kapitel 5

Untersuchte Modalitäten

In diesem Kapitel werden einige der in Kapitel 3 vorgestellten Modalitäten exemplarisch untersucht. Zu betonen ist, daß der in dieser Arbeit verwendete Ansatz auch auf eine Vielzahl anderer Modalitäten erweiterbar ist, so daß es im Grunde kein Problem darstellt, die aufgeführten Experimente mit anderen Modalitäten durchzuführen. Das Schwierigste dabei ist meist die Extraktion der Modalitäten aus einer Äußerung.

Für die Untersuchungen wurden die Daten aus dem Verbmobil-Projekt [VER] verwendet. Hierbei handelt es sich um spontane Dialoge von Terminabsprachen, die an mehreren deutschen Universitäten gesammelt wurden. Die zur Verfügung stehenden Daten wurden in drei disjunkte Mengen, die Trainingsmenge, die Kreuzvalidierungsmenge und die Evaluierungsmenge aufgeteilt. Die Daten dieser drei Teilmengen sind Tabelle 5.1 zu entnehmen. Darin wurde die Sprechgeschwindigkeit in mrate [MFL98] und der Signal-Rausch-Abstand in dezibel gemessen.

	Training	Kreuzvalidierung	Evaluierung
Länge	ca. 57 Std.	ca. 35 Min.	ca. 40 Min.
Äußerungen	21246	327	372
Sprecher (männl./weibl.)	784 (455/329)	8 (4/4)	22 (10/12)
Sprecher bekannt. Dialekt	448	7	16
versch. Dialekte	30	6	8
Sprechgeschw. $[min, max]$	$[1.21, 8.54]$	$[1.54, 6.45]$	$[1.87, 6.98]$
SNR $[min, max]$	$[-11.61, 64.15]$	$[-0.19, 40.90]$	$[0.00, 55.94]$

Tabelle 5.1: Daten der Trainings-, Kreuzvalidierungs- und Evaluierungsmenge

Die in diesem Kapitel aufgeführten Korrelationen der Modalitäten zur Fehlerrate wurden unter Verwendung des in Abschnitt 6.2 erläuterten Referenzsystems berechnet.

5.1 Geschlecht

Wie in Abschnitt 3.2.1 erwähnt, ist die Konstruktion eines Spracherkennungssystem, das gleich gute Ergebnisse auf beiden Geschlechtern liefert, sehr schwierig. Aus diesem Grund ist eine Aufteilung der akustischen Modelle durch Einführung von Modalitätenfragen zum Geschlecht eines Sprechers sinnvoll.

In der vorliegenden Arbeit wurde das Geschlecht als binäres Merkmal, also entweder männlich oder weiblich, modelliert. Es besteht jedoch auch die Möglichkeit, anstatt des diskreten Wertebereichs des Geschlechts einen kontinuierlichen Wertebereich zu verwenden, wobei hierfür dann die Grundfrequenz oder die Faktoren zur Normalisierung der Vokaltraktlänge genommen werden können. Eine solche Modellierung wäre dann sinnvoll, wenn innerhalb eines Geschlechts anatomische Unterschiede des Sprechapparates, die zu einer unterschiedlichen Stimmlage oder Aussprache führen, einen Einfluß auf die Erkennungsleistung haben.

Extraktion des Geschlechts

Für die Extraktion des Geschlechts sind mehrere Verfahren bekannt. Am üblichsten ist die Bestimmung der Grundfrequenz des Sprechers und dann darauf basierend das Geschlecht. Die Bestimmung der Grundfrequenz eines Sprachsignals geschieht meist mit Hilfe des Cepstrums [Nol67], der Autokorrelation [Rab77] oder der Kreuzkorrelation [Tal95]. Einen Überblick über die Möglichkeiten zur Grundfrequenzbestimmung bietet [Hes83]. Es ist jedoch auch möglich, mit Hilfe der oben erwähnten Faktoren zur Vokaltraktnormierung eine Geschlechtsbestimmung durchzuführen.

5.2 Dialekte

Deutschland ist ein Land, das aus vielen kleinen Teilgebieten entstanden ist. Dies spiegelt sich auch in den Sprachregionen wider. Die Vorlage für die Aufteilung Deutschlands in verschiedene Sprachregionen war die in Abbildung 5.1 gezeigte Karte aus [BO99] von Susanne Burger. Darin wird Deutschland in 20 Sprachre-

Sprachregion	Anzahl	Sprachregion	Anzahl
unbekannt	343	Nordbayern	6
Ausland	9	Norddeutschland	8
Baden	3	Nordniedersachsen	12
Baden-Württemberg	7	Oberbayern	28
Bayern	35	Ostfalen	10
Berlin	3	Ostfranken	15
Brandenburg	1	Pfalz	9
Franken	3	Rhein	32
Friesland	1	Rheinland	80
Hessen	6	Rheinland-Pfalz	1
Mittelbayern	11	Sachsen	1
Mittelfranken	4	Schleswig-Holstein	30
München	64	Schwaben	23
Niederalemanien	3	Südfranken	10
Niederrhein	11	Westfalen	28
Niedersachsen	17		

Tabelle 5.2: Anzahl der Sprecher pro Sprachregion in der Datenbasis

gionen aufgeteilt. Jede dieser Regionen wurden nun die Sprecher der Datenbasis zugeordnet, wobei durch inkonsistente Bezeichnungen in den Sprecherdaten noch 11 weitere Regionen hinzugenommen wurden. Beispielsweise war anstatt des Dialekts des Sprechers das Bundesland angegeben, in dem er gelebt hat, wobei sich das Bundesland aus mehreren Sprachregionen zusammensetzt und somit keine eindeutige Zuordnung möglich war. Insgesamt konnten 448 der 784 verschiedenen Sprecher klassifiziert werden. Tabelle 5.2 zeigt die Anzahl der Sprecher die auf eine Sprachregion fallen.

Extraktion der Dialekte

Im allgemeinen lassen sich alle Varianten, die für eine Sprachenidentifikation verwendet werden, auch auf Dialekte übertragen. Zu erwarten ist jedoch, daß die Resultate aufgrund der größeren Ähnlichkeit der Dialekte etwas schlechter ausfallen. Einen guten Überblick über das Gebiet der Sprachenidentifikation bietet [ZB99]. Die besten Erfolge werden dabei durch den Einsatz von mehreren monolingualen Spracherkennungssystemen erzielt, wobei die Klassifikation mittels der von diesen Systemen erzeugten Wahrscheinlichkeiten der Hypothesen erfolgt.

Eine andere, in dieser Arbeit nicht evaluierte Möglichkeit wäre die Identifikation der Dialekte mit Hilfe der dialektabhängigen akustischen Modelle. Dazu müßte eine Hypothese aus einem dialektunabhängigem Spracherkennungssystem vorlie-

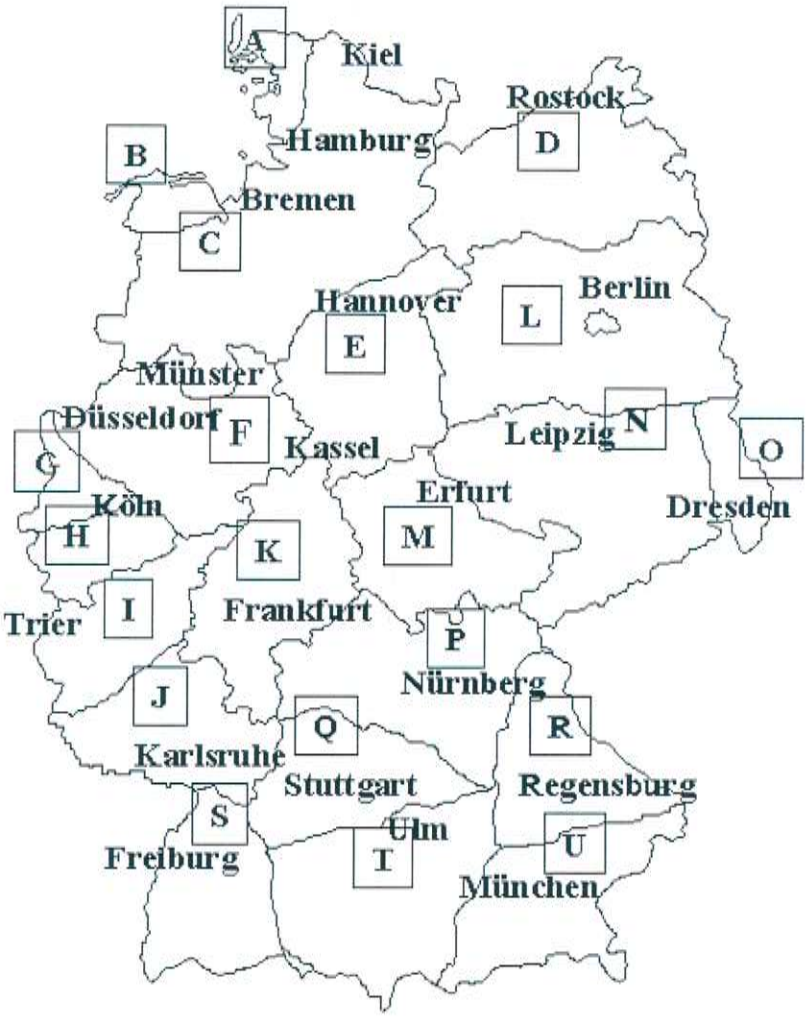


Abbildung 5.1: Sprachregionen Deutschlands: A Nordfriesland, B Ostfriesland, C Nordniedersachsen, D Mecklenburg, E Ostfalen, F Westfalen, G Niederrhein, H Mittelfranken, I Moselfranken, J Pfalz, K Hessen, L Brandenburg, M Thüringen, N Obersachsen, O Sorbien, P Ostfranken, Q Südfranken, R Nordbayern, S Niederalemanien, T Schwaben, U Mittelbayern

gen, für welche dann im modalitätenabhängigen System dessen Wahrscheinlichkeit für jeden Dialekt berechnet werden würde. Die Berechnung könnte durch Evaluation des mit Hilfe der Hypothese aufgebauten HMMs unter Verwendung des Viterbi-Algorithmus erfolgen. Der momentan vorliegenden Äußerung würde dann derjenige Dialekt zugeordnet werden, dessen Viterbi-Pfad die höchste Wahrscheinlichkeit besäße.

Da einzelne Dialekte sich oftmals sehr ähnlich sind, könnte es passieren, daß die Wahrscheinlichkeiten sehr nahe beieinander liegen. Die Klassifikation würde damit immer unsicherer werden, wobei es durchaus auch zu Fehlklassifikationen kommen könnte. Durch Definition eines Konfidenzmaßes für die Sicherheit bzw. Unsicherheit der Klassifikation könnte zum Beispiel im Falle von großer Unsicherheit die dialektunabhängige Erkennung einfach fortgesetzt werden.

Diese Art der Klassifikation von Äußerungen zu verschiedenen Ausprägungen der Modalitäten ließe sich auch auf andere Modalitäten, wie zum Beispiel das Geschlecht anwenden.

5.3 Sprechgeschwindigkeit

Die Sprechgeschwindigkeit stellt im Gegensatz zum Geschlecht und Dialekt des Sprechers eine zeitveränderliche wertkontinuierliche Modalität dar. Da zum einen nur eine kleine Anzahl an Markierungen für die akustischen Modelle zur Verfügung stehen und zum anderen der Entscheidungsbaum nur binäre Fragen erlaubt, muß der Wertebereich quantisiert werden. Hierzu wurde der gesamte Wertebereich in 12 Intervalle unterteilt, wobei der jeweilige Repräsentant eines Intervalls dessen Mittelwert ist. In Abbildung 5.2 ist die Korrelation zwischen der Sprechgeschwindigkeit und der Fehlerrate graphisch veranschaulicht. Die Zahlen an der x-Achse repräsentieren die 12 Intervalle. Es ist deutlich zu erkennen, daß eine Abweichung vom Mittelwert gleichzeitig auch eine Zunahme der Fehlerrate bedeutet.

Extraktion der Sprechgeschwindigkeit

Die Berechnung der Sprechgeschwindigkeit erfolgte mit einem rein äußerungsbasierten Verfahren, dem sogenannten MRATE [MFL98]. Dieses kombiniert verschiedene Methoden, um die Sprechgeschwindigkeit zu schätzen und korreliert sehr stark mit der Berechnung der Sprechgeschwindigkeit durch die Anzahl der Silben pro Sekunde.

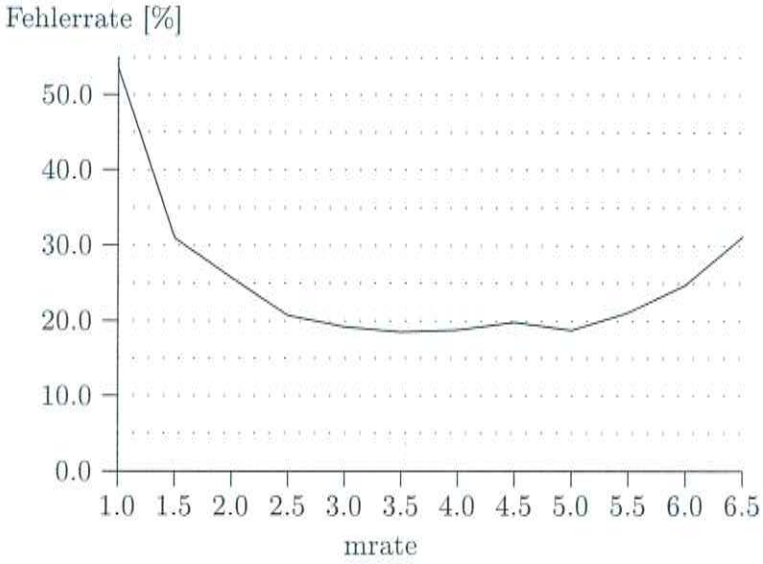


Abbildung 5.2: Korrelation zwischen der Sprechgeschwindigkeit und der Fehlerrate. Größere Werte bedeuten einen höhere Sprechgeschwindigkeit.

Genauso gut könnte man die Sprechgeschwindigkeit mit Hilfe vom System erzeugten Labels auf Phonembasis berechnen, wobei die mittlere erwartete Länge eines Phonems vorher auf den Trainingsdaten ermittelt werden müßte. Der Nachteil dieser Variante ist der wesentlich höhere Zeitaufwand, da für die Ermittlung der Sprechgeschwindigkeit eine fertige Hypothese vorliegen müßte. Aus diesem Grund kommt diese Variante der Sprechgeschwindigkeitsbestimmung in dieser Arbeit nicht in Frage.

5.4 Signal-Rausch-Abstand

Der Signal-Rausch-Abstand (SNR) einer Äußerung ist meist konstant, so daß es sich hierbei um eine wertkontinuierliche Modalität handelt. Genauso wie bei der Sprechgeschwindigkeit muß eine Quantisierung des Wertebereichs stattfinden. In diesem Fall entstanden 15 gleichgroße Intervalle. In Abbildung 5.3 ist ein deutliches Abfallen der Fehlerrate für größere SNRs festzustellen. Die Zahlen an der x-Achse repräsentieren hier die 15 Intervalle.

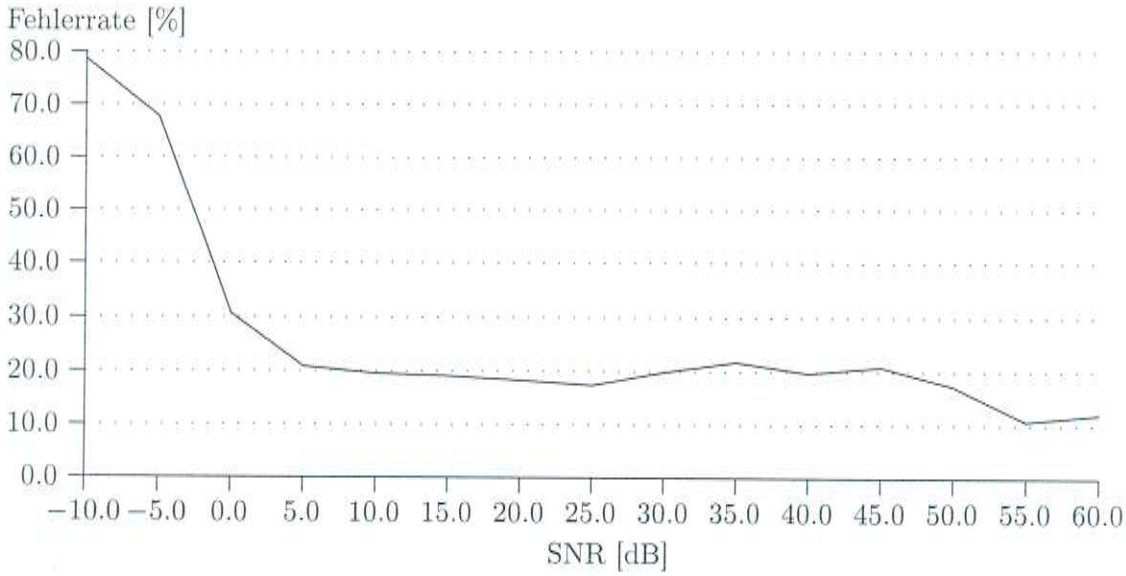


Abbildung 5.3: Korrelation zwischen dem Signal-Rausch-Abstand und der Fehlerrate

Extraktion des Signal-Rausch-Abstands

Zur Bestimmung des SNRs einer Äußerung wurde in das vorhandene Spracherkennungssystem ein Sprache-Stille-Detektor auf Basis der logarithmierten mittleren Betragsamplituden und der Anzahl der Nulldurchgänge integriert. Für eine genaue Erläuterung dieses Verfahrens wird auf den Anhang A verwiesen. Die Anzahl der falsch klassifizierten Fensterausschnitte betrug dabei 10%. Auf Basis dieser Detektion wurde der Signal-Rausch-Abstand einer Äußerung bestimmt.

Kapitel 6

Experimente und Ergebnisse

In diesem Kapitel werden die Experimente und deren Ergebnisse präsentiert. Um die Anzahl der Trainingsdaten pro Sprachregionen zu erhöhen und um sinnvolle Fragen für die divisive Ballung der Modalitäten zu erhalten, wurden die einzelnen Sprachregionen agglomerativ zu größeren Regionen zusammengefaßt.

Das Training der Spracherkennungssysteme erfolgte immer auf dieselbe Weise und mit denselben Parametern. Das verwendete Referenzsystem und dessen Fehlerraten werden in Abschnitt 6.2 erläutert. Zusätzlich sind in diesem Abschnitt noch einige Hinweise zur Wahl der Parameter und der Anzahl der Modelle für die modalitätenabhängigen Spracherkennungssysteme zu finden.

Für jede der im vorherigen Kapitel aufgeführten Modalitäten wurde ein Experiment durchgeführt, in dem nur jeweils eine Modalität in das Spracherkennungssystem integriert wurde. Auf diese Art konnten die Ergebnisse besser analysiert werden.

6.1 Agglomerative Ballung der Sprachregionen

Ein modalitätenabhängiges Spracherkennungssystem, das für jede aus Tabelle 5.2 aufgeführten Sprachregion eigene akustische Modelle bereit hält, wird vermutlich eine sehr schlechte Erkennungsleistung erzielen. Der Grund dafür liegt darin, daß die Sprecher sehr unterschiedlich auf die einzelnen Sprachregionen verteilt sind, so daß es Regionen mit mehr oder weniger Sprecher und damit mit mehr oder weniger Trainingsdaten gibt. Die naheliegendste Lösung für dieses Problem ist eine Zusammenfassung von mehreren kleineren Sprachregionen zu größeren Gebieten. Hier stellt sich aber die Frage, wie diese Gebiete aus den Sprachregionen

gebildet werden sollen. Zum Beispiel könnte die Ballung aufgrund der geographischen Lage oder durch Analyse der Sprachstämme der Sprachregionen erfolgen. Das Problem dabei ist jedoch, daß die Granularität der Aufteilung nicht genau definiert werden kann. Außerdem wird dem Spracherkennungssystem eine feste Aufteilung vorgegeben, bei der nicht sichergestellt werden kann, ob nicht eine andere Aufteilung eine bessere Erkennungsleistung erzielt hätte. Aus diesem Grund wird in dieser Arbeit eine Ballung der Sprachregionen rein auf ihren akustischen Gegebenheiten vorgenommen. Diese sind durch die darin enthaltenen Sprecher vorgegeben.

Zunächst werden für jede der Sprachregionen eigene, mit der jeweiligen Sprachregion markierte, akustische Modelle trainiert. Dafür genügt ein kontextunabhängiges System, wodurch dann die Anzahl der verschieden Modelle durch die Anzahl der Sprachregionen und die Anzahl der verschiedenen Phoneme nach oben begrenzt ist. Um eine Ballung der Sprachregionen an sich und nicht in Abhängigkeit der vorhandenen Phoneme zu ermöglichen, werden alle akustischen Modelle einer Sprachregion zu einem akustischen Modell zusammengefaßt. Die Zusammenfassung erfolgte nur durch aneinanderhängen der Mixturgewichte und der Codebücher, um einen Qualitätsverlust der Modelle durch eine Dimensionalitätsreduktion zu vermeiden. Damit ist für jede Sprachregion genau ein akustisches Modell vorhanden, welche nun mittels agglomerativer Ballung zusammengefaßt werden können.

In Abbildung 6.1 ist das Resultat der agglomerativen Ballung von 31 akustischen Modellen der Sprachregionen zu sehen. Es entstand eine Dreiteilung Deutschlands in Nord-, Mittel- und Süddeutschland. Verfolgt man die Schritte des Ballungsalgorithmus zurück, so lassen sich neben den drei großen Regionen auch noch mehrerer kleinere Teilregionen ausmachen. Diese sind neben den drei großen Regionen in den Tabellen 6.1, 6.2 und 6.3 aufgeführt.

Jede der so entstandenen Teilregionen wurde als Modalität in das Spracherkennungssystem integriert. Die Ergebnisse eines so trainierten Systems werden in Abschnitt 6.4 vorgestellt.

Bei diesem Verfahren zur Ballung von Sprachregionen werden nur die akustischen Eigenschaften der Modelle genutzt. Dies ist zwar auf der einen Seite erwünscht, kann jedoch auch Schwierigkeiten bei einer unausgewogenen Trainingsmenge verursachen. Es könnte zum Beispiel passieren, daß einzelnen Regionen nicht aufgrund ihres ähnlichen Dialekts, sondern aufgrund dessen, daß in diesen Regionen nur Sprecher eines Geschlechts vorhanden sind, zusammengeballt werden. Deshalb sollten für eine optimale Ballung in jeder Region genügend Trainingsdaten

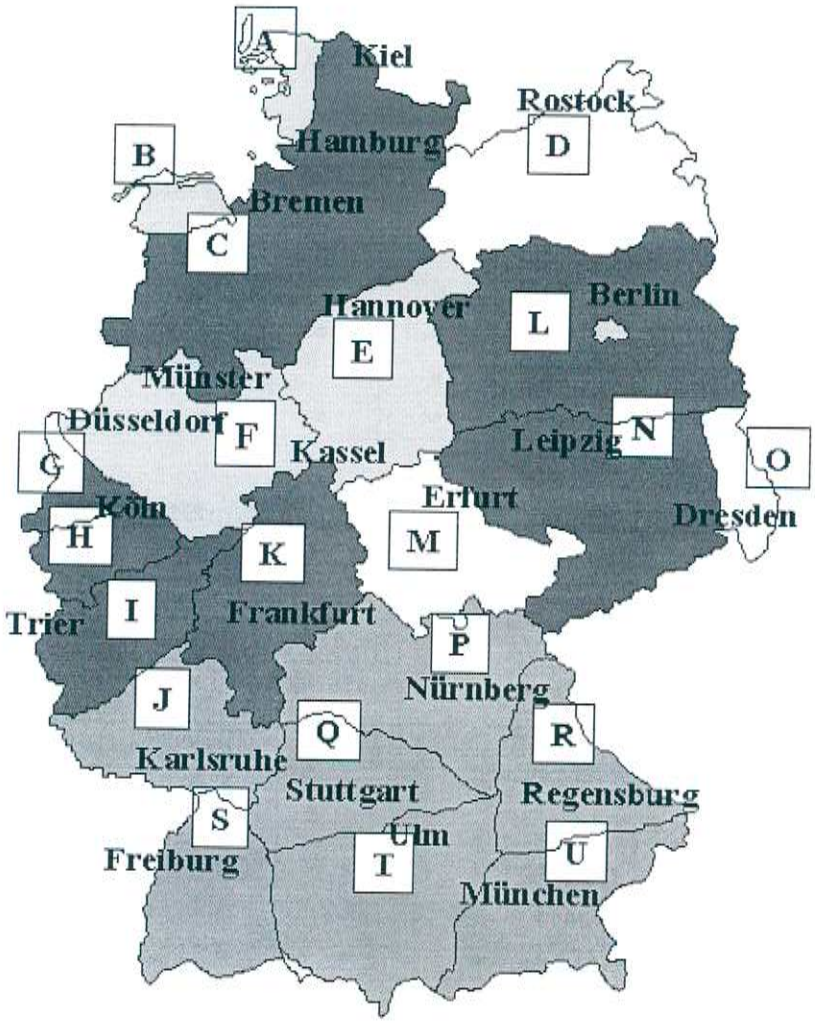


Abbildung 6.1: Agglomerativ geballte Sprachregionen

	Nord	Nord5	Nord4	Nord3	Nord1	Nord2
Baden	•	•	•	•	•	•
Brandenburg	•	•				
Franken	•	•	•	•	•	
Hessen	•	•	•			•
Mittelfranken	•	•	•	•	•	•
Niederrhein	•	•	•	•		
Niedersachsen	•	•	•	•	•	•
Norddeutschland	•					
Nordniedersachsen	•	•	•	•	•	•
Rheinland	•	•	•	•	•	•
Rheinland-Pfalz	•	•	•	•	•	•
Sachsen	•	•				
Schleswig-Holstein	•	•	•	•	•	

Tabelle 6.1: Geballte Sprachregionen (Norddeutschland)

	SMitte	SMitte1	Mitte	Mitte1	Mitte2
Baden-Württemberg	•	•	•	•	
Berlin	•	•	•	•	•
Bayern	•	•	•		
Franken	•				
Friesland	•	•	•	•	•
Niederrhein					•
Norddeutschland				•	•
Ostfalen	•	•	•	•	
Rhein	•	•	•	•	•
Westfalen	•	•	•	•	•
Süd1 (s. Tabelle 6.3)	•	•			

Tabelle 6.2: Geballte Sprachregionen (Mitteldeutschland)

	Süd	Süd3	Süd1	Süd2
Ausland	•	•		
Hessen	•			
Mittelbayern	•	•	•	
München	•	•		•
Niederalemanien	•	•	•	
Nordbayern	•	•		•
Oberbayern	•	•		•
Ostfranken	•	•		•
Pfalz	•	•	•	
Schwaben	•	•	•	
Südfranken	•	•	•	

Tabelle 6.3: Geballte Sprachregionen (Süddeutschland)

vorhanden sein. Die in dieser Arbeit verwendeten Verbmobil-Daten waren dafür nur bedingt geeignet, da in einigen Regionen nur sehr wenige Sprecher vorhanden waren.

6.2 Referenzsystem

Für alle in dieser Arbeit aufgeführten Ergebnisse wurden die Spracherkennungssysteme auf dieselbe Art und Weise, nach dem in Abschnitt 2.2 beschriebenen Verfahren, trainiert. Um vor allem den Einfluß von geschlechtsabhängigen Modalitäten genau zu analysieren, wurde von einer Vokaltraktnormalisierung abgesehen. Die Vorverarbeitung der Systems wurde bereits in Abschnitt 2.3.1 beschrieben.

Für das Referenzsystem wurde die Anzahl der akustischen Modelle auf 2000 begrenzt und die Dimensionalität dieser auf 32 festgelegt. Das Referenzsystem erzielte auf den Evaluationsdaten eine Fehlerrate von 14.5%.

Vergleicht man dieses Ergebnis mit Ergebnissen anderer Spracherkenner auf der Verbmobil-Datenbasis, so erscheint die Fehlerrate gegenüber den anderen sehr gering zu sein. Dies liegt daran, weil die in dieser Arbeit verwendeten Evaluationsdaten hauptsächlich aus der ersten Verbmobil-Phase stammten und somit weniger spontan waren. Das Referenzsystem erzielte jedoch auf der Kreuzvalidierungsmenge, die mehrheitlich aus Daten der zweiten Verbmobil-Phase bestand, ähnliche Ergebnisse, wie das aktuelle Spracherkennungssystem der Universität

Karlsruhe, das bei der Evaluation 1998 verwendet wurde. Die Kreuzvalidierungsmenge konnte jedoch aufgrund einer kleinen Sprecherüberlagerung mit der Trainingsmenge nicht zur Evaluation verwendet werden.

Bei den modalitätenabhängigen Systemen wurde die Anzahl der akustischen Modelle so festgelegt, daß einem Sprecher a priori genauso viele Modelle zur Verfügung standen wie im modalitätenunabhängigen Referenzsystem. Das bedeutet, daß einer Äußerung eines Sprecher unabhängig von der dazugehörigen Modalitätenkombination immer etwa 2000 Modelle zur Verfügung stehen. Damit hängt die Gesamtzahl der Modelle eines modalitätenabhängigen Spracherkennungssystems davon ab, wieviele Modelle modalitätenabhängig waren und bei wievielen Modellen eine Parametersharing erfolgte.

Die unten aufgeführten Tabellen für den Prozentsatz an modalitätenabhängigen Modellen pro Phonem, wurden alle mit derselben unteren Schranke für die Anzahl der Trainingsdaten, die auf ein akustisches Modell fallen müssen, ermittelt. Je nach verlangter Gesamtzahl an akustischen Modellen wurde der Baum früher oder später unten abgeschnitten.

6.3 Geschlecht

Die erste, genauer untersuchte Modalität war das Geschlecht. Bei Berücksichtigung von etwa 2000 simultan verwendbaren akustischen Modellen für jeden Sprecher, kam dieses System auf eine Gesamtzahl von 3000 Modellen. In Abbildung 6.2 ist deutlich zu erkennen, daß sich durch Verwendung geschlechtsabhängiger Modalitäten die Fehlerrate für beide Geschlechter gegenüber der des geschlechtsunabhängigen Systems deutlich senken ließ. Im Falle der männlichen Sprecher wurde eine Fehlerratenreduktion um 5.4% und im Falle der weiblichen Sprecherinnen sogar um 9.0% erreicht. Damit reduzierte sich die Gesamtfehlerrate um 7.6%. Im Vergleich mit zwei geschlechtsabhängig trainierten Systemen konnte nur die Fehlerrate der männlichen Sprecher um 3.7% gesenkt werden, wodurch sich die Gesamtfehlerrate des modalitätenabhängigen Systems gegenüber der Gesamtfehlerrate beider geschlechtsabhängiger Systeme insignifikant erhöhte. Die Integration der Modalitäten führte damit zu einem etwas ausgewogenerem Verhältnis in der Fehlerrate zwischen Männer und Frauen, ohne die Gesamtfehlerrate gegenüber rein geschlechtsabhängig trainierten Systemen signifikant zu erhöhen.

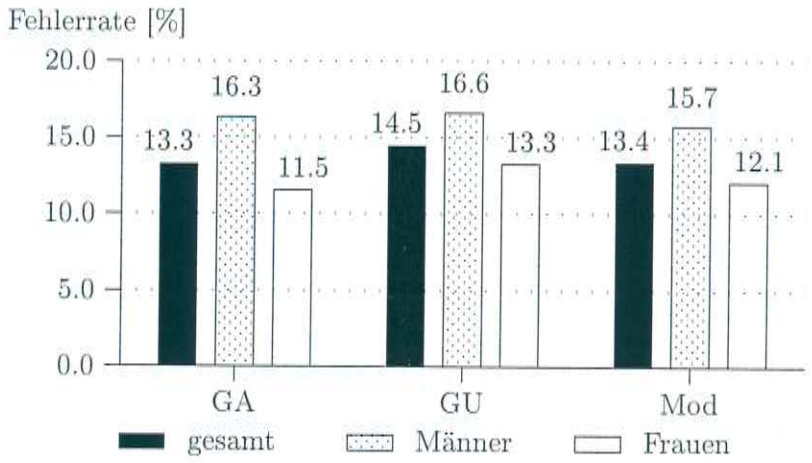


Abbildung 6.2: Vergleich des Modalitäten-Systems (Mod) mit den beiden geschlechtsabhängigen (GA) und dem geschlechtsunabhängigen (GU) System

Diskussion der Ergebnisse

Analysiert man die Position der geschlechtsabhängigen Fragen im Entscheidungsbaum, so läßt sich feststellen daß mehrheitlich die Vokale und Diphthonge[1] von einer geschlechtsabhängigen Modellierung profitieren, während der Prozentsatz an geschlechtsabhängigen Modellen für Plosive weitaus geringer ist (siehe Tabelle 6.4). Selbst wenn die Gesamtzahl der Modelle auf 1000 reduziert wird, ist der Gesamtanteil an geschlechtsabhängigen Modellen im System noch relativ groß. Dies zeigt, daß das Geschlecht für den Ballungsalgorithmus ein äußerst wichtiges Unterscheidungskriterium ist und somit geschlechtsabhängige Modalitätenfragen sehr weit oben im Entscheidungsbaum auftauchen.

Dieses Ergebnis wird auch von Untersuchungen zu geschlechtsspezifischen Unterschieden zur Aussprache unterstützt. Stimmhafte Laute, das heißt Vokale und Diphthonge sind durch die vorhandenen Grundfrequenzanteile und Vokaltraktresonanzfrequenzen in ihrer Aussprache wesentlich mehr vom Geschlecht eines Sprechers abhängig als stimmlose Laute (Frikative, Plosive, usw.). Frikative werden aufgrund ihrer längeren durchschnittlichen Dauer noch eher geschlechtsabhängig modelliert als Plosive[2]. Diese Abhängigkeiten in der Modellierung zu bestimmten Artikulationsgruppen stimmt auch mit den Schlußfolgerungen überein, die in [RC99] gezogen wurden.

[1] Eine Tabelle der Artikulationsgruppen mit den dazugehörigen Phonemen ist im Anhang B zu finden.

[2] Im Anhang C ist hierzu eine Tabelle der durchschnittlichen Dauer aller Phoneme angegeben.

#Modelle	A	AEH	AH	AI	AU	B	CH	D	E	E2	EH	ER2
1000	33	57	67	76	72		31		55		21	25
2000	61	75	84	91	91		55	5	76	18	60	56
3000	76	77	90	97	93	16	75	28	83	48	73	69

#Modelle	EU	F	G	H	I	IE	J	K	L	M	N	O	OE	OH
1000	40			10		31	15	7	31		15	59		
2000	86			31	19	59	52	4	60	27	45	72	50	33
3000	89	41	15	44	55	71	68	27	63	51	73	79	67	41

#Modelle	P	R	S	SCH	T	TS	U	UE	UEH	UH	V	X	Z
1000		18	8	15						10		24	
2000	7	37	37	67			15			45	13	61	16
3000	16	54	80	74	32	47	33	40		48	35	74	45

Tabelle 6.4: Prozentsatz der geschlechtsabhängigen Modelle pro Phonem

Allgemein kann gesagt werden, daß es durchaus profitabel ist, dem Ballungsalgorithmus zu erlauben geschlechtsabhängige Modelle zu modellieren. In einem kurzen Experiment das in beiden Fällen, das heißt im modalitätenabhängigen System und im geschlechtsunabhängigen System eine Vokaltraktnormalisierung der Sprecher verwendet, konnte ebenfalls festgestellt werden, daß sich die Fehlerrate des modalitätenabhängigen Systems gegenüber der des geschlechtsunabhängigen Systems senken ließ. Die Fehlerreduktion fiel jedoch wesentlich geringer aus als im oben erwähnten Fall ohne Vokaltraktnormalisierung. Dies ist dadurch zu erklären, daß die Vokaltraktnormalisierung einige die geschlechtsspezifischen Unterschiede aus den Lauten eliminiert. Außerdem könnte man vermuten, daß der erzielte Gewinn durch geschlechtsabhängige Modelle in einem System mit Vokaltraktnormalisierung deshalb geringer ausfällt, weil durch die binäre Aufteilung der Modelle der positive Einfluß der Normalisierung verloren geht. Der Ballungsalgorithmus wird eine Teilung der Modelle nach Geschlecht aber auch mehr bevorzugen als erwünscht, weil in manchen Fällen der Entropiegewinn durch eine Aufteilung der Modelle in nahezu zwei gleich große Teile[3], gegenüber einer anderen, für die Erkennungsleistung des Systems bessere Aufteilung, überwiegt. Aus diesem Grund wäre es vielleicht vorteilhaft, das Geschlechts nicht über diskrete Modalitäten auszudrücken sondern auf einen kontinuierlichen Wertebereich zu erweitern. Damit hätte das System auch die Möglichkeit Aufteilungen innerhalb eines Geschlechts vorzunehmen. Dies könnte zum Beispiel mit einer der in

[3]bezüglich der zugehörigen Trainingsdaten

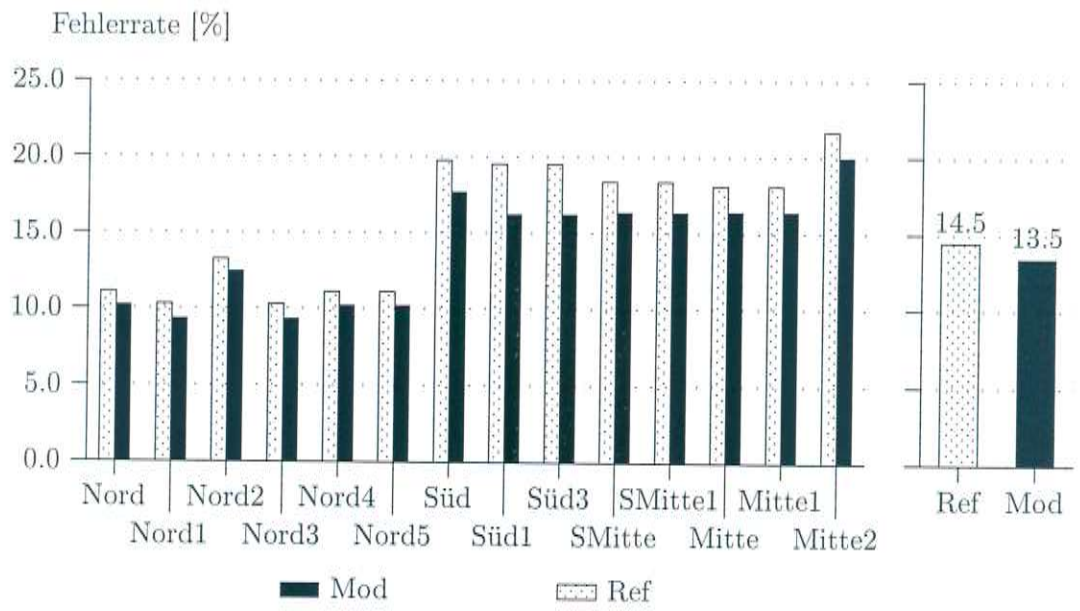


Abbildung 6.3: Fehlerratenreduktion für verschiedene Sprachregionen

Abschnitt 5.1 beschriebenen Methoden erfolgen.

6.4 Dialekte

Eine Aufteilung der Modelle je nach Dialekt des Sprechers liefert nahezu dieselben Ergebnisse, wie die Integration von geschlechtsabhängigen Modalitäten in das Spracherkennungssystem. Die Gesamtfehlerrate konnte hierbei um 6.9% gesenkt werden. In Abbildung 6.3 läßt sich auch deutlich erkennen, daß die Fehlerrate für alle Dialekte gesenkt werden konnte; im Falle von süddeutschen Sprechern sogar um bis zu 17%. Das dialektabhängige Modalitäten-System besaß 2200 akustische Modelle.

Diskussion der Ergebnisse

Allgemein liegt die Fehlerrate der Süd- und Mitteldeutschen Sprecher um einiges höher als die der norddeutschen Sprecher, was wohl daran liegt, daß die Aussprache der norddeutschen Sprecher wesentlich deutlicher ist als die der süddeutschen Sprecher.

Diese Ergebnisse werden auch von Untersuchungen zu dialektspezifischen Ausspracheabweichungen vom Standarddeutsch unterstützt. In [BO99] wurde für

	AEH	AH	AI	CH	D	E2	EH	ER2	IE	J
Nord		3.5	1.5		2.4	1.1			1.5	7.4
Mitte		5.8						1.7		7.4
SMitte	8.3	4.7	7.7		2.4		1.8	3.5	1.5	
Süd	8.3	5.8	13.9	3.9	3.7	3.4	1.8	3.5	1.5	

	K	N	OH	P	R	S	T	TS	UH	V	Z
Nord	6.6	1.8		3.0		1.4	3.4	6.0		1.7	20.4
Mitte		0.6				4.1					2.0
SMitte		1.8	2.2		1.1				5.6		
Süd		2.4	2.2		1.1				5.6		

Tabelle 6.5: Prozentsatz der dialektabhängigen Modelle nach Sprachregionen für das 2200 Modelle umfassende Modalitäten-System

süddeutsche Sprecher eine wesentlich höhere Anzahl an Enkliktikons[4] und Assimilationen[5] nachgewiesen als für norddeutsche Sprecher.

Teilweise lassen sich Enkliktikons und Assimilationen durch Aussprachevarianten erfassen. Eine alleinige Erweiterung des Aussprachewörterbuchs durch dialektspezifische Aussprachevarianten bringt jedoch keinen signifikanten Gewinn [BSRB98]. Das mag daran liegen, daß durch teilweise fehlerhafte Transkriptionen keine geeignete Modellierung der Modelle erfolgte und somit erst bei einer Aufteilung dieser Modelle in dialektabhängige Varianten ein Erfolg festzustellen ist. Die hier durchgeführten Experimente wurden sogar ohne eine Erweiterung des Aussprachewörterbuchs durch zusätzliche dialektspezifische Varianten vorgenommen. Im verwendeten 14600 umfassenden Aussprachewörterbuch besaßen nur 5% aller Wörter Aussprachevarianten, wobei unter diesen die mittlere Anzahl von Varianten pro Wort bei 1.2 lag.

In [BSRB98] werden einige Regeln für Phonemwechsel zu einzelnen Dialekten aufgeführt. Gerade bei den süddeutschen Dialekten ändert sich vor allem die Aussprache von Vokalen und Diphthongen, jedoch treten durchaus auch Veränderungen von Nasalen und gerade im bairischen beim R auf.

Diese Beobachtungen decken sich in etwa mit den vom Spracherkennungssystem getroffenen Entscheidungen zur dialektabhängigen Modellierung einzelner Phoneme. In Tabelle 6.5 ist zu erkennen, daß eine explizite Modellierung von süddeut-

[4]reine Verschmelzung zweier Wörter

[5]Phonemwechsel bedingt durch den Einfluß benachbarter Phoneme am Verschmelzungspunkt zweier Wörter

#Modelle	A	AEH	AH	AI	AU	B	CH	D	E	E2	EH	ER2
1500			6	6				6		3		
2200		25	33	29			13	10		15	4	10
3000	30	36	51	55	28	8	37	29	7	51	35	22

#Modelle	EU	F	G	H	I	IE	J	K	L	M	N	O	OE	OH
1500								10						
2200						6	41	16			8			4
3000	44	34	7	12	25	29	62	31	14	18	45	17	55	4

#Modelle	P	R	S	SCH	T	TS	U	UE	UEH	UH	V	X	Z
1500	10					6							24
2200	9	2	13		8	20				11	5		55
3000	16	20	51	16	40	42	24	29	17	28	17	17	57

Tabelle 6.6: Prozentsatz der dialektabhängigen Modelle pro Phonem

schen akustischen Modellen hauptsächlich bei Vokalen und Diphthongen erfolgt, während bei Plosive und Frikative eher norddeutsche akustische Modelle vom Rest abgespalten werden. Vor allem die T- und S-Laute, und hier vor allem das stimmhafte S (=Z), die im Norddeutschen wesentlich ausgeprägter ausgesprochen werden, werden explizit mit norddeutschen Modellen modelliert. Die meisten Varianten sind jedoch im Süddeutschen zu finden, weshalb der Prozentanteil an süddeutschen Modellen auch größer ausfällt, als der der norddeutschen.

Vergleicht man die Prozentsätze der dialektabhängigen Modellen aus Tabelle 6.6 mit denen aus Tabelle 6.4 der geschlechtsabhängigen Modellen, so fällt auf, daß ein Parametersharing für wesentlich mehr dialektabhängige Modelle stattfindet, als für geschlechtsabhängige, obwohl beide Modalitäten-Systeme annähernd dieselbe Fehlerrate besitzen. Wird die Anzahl der Gesamtmodelle im dialektabhängigen Modalitäten-System erhöht, so nimmt der Anteil der dialektabhängigen Modelle für alle Phoneme rapide zu. Daraus läßt sich schließen, das dialektabhängige Fragen erst relativ spät im Entscheidungsbaum gegenüber den Kontextfragen bevorzugt ausgewählt werden und trotzdem den Parameterraum in geeigneter Weise aufsplitten.

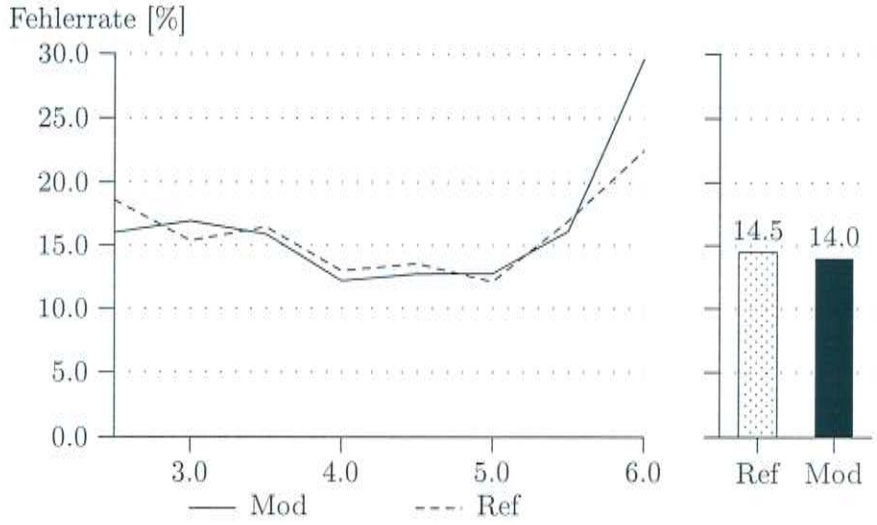


Abbildung 6.4: Fehlerratenreduktion für verschiedene Sprechgeschwindigkeiten

6.5 Sprechgeschwindigkeit

Die Ergebnisse für die sprechgeschwindigkeitsabhängige Modellierung der akustischen Modelle waren nicht so erfolgreich, wie diejenigen der anderen Modalitäten-Systeme. Die Gesamtfehlerrate reduzierte sich um 3.4%, wobei der Verlauf der Fehlerrate in Abhängigkeit zur Sprechgeschwindigkeit in Abbildung 6.4 auch deutliche Unebenheiten zeigte.

Diskussion der Ergebnisse

So profitierten in erster Linie die Normalsprecher[6] von einer solchen Modellierung, während bei den Schnellsprechern die Fehlerrate deutlich zunahm. Dies lag vor allem daran, daß für diesen Bereich zwar genügend Trainingsdaten zur Verfügung standen, diese jedoch nur von sehr wenig verschiedenen Sprechern waren. Damit war eine optimale Schätzung der Modelle nicht gewährleistet. Auch dieses System besaß eine Gesamtzahl von 2200 akustischen Modellen.

Da es sich bei der Sprechgeschwindigkeit um eine wertkontinuierliche Modalität handelte, stellt sich die Frage, wie der vorhandene Wertebereich durch sprechgeschwindigkeitsabhängige Fragen aufgeteilt wurde. In den Fragenkatalog wurden neben den Fragen zu nur einem der 12 Intervalle (siehe Abschnitt 5.3) auch noch explizit Fragen zu mehreren miteinander verbundenen Intervallen aufgenommen.

[6] die Sprecher mit durchschnittlicher Sprechgeschwindigkeit

Häufigkeit	Intervall
6	[1.75, 3.75]
5	[1.75, 4.25]
4	[4.25, 7.25]
3	[2.25, 6.75]
3	[2.25, 3.25]
3	[2.25, 3.75]
2	[3.75, 6.75]
2	[3.75, 5.25]
2	[3.75, 4.25]

Tabelle 6.7: Häufigkeit der 9 meist verwendeten Fragen zur Sprechgeschwindigkeit für das 2200 Modelle umfassende Modalitäten-System

Tabelle 6.7 gibt Aufschluß darüber, wie häufig welche Fragen zu einem Intervall während des Ballungsvorgangs verwendet wurden.

Darin ist deutlich zu erkennen, daß vor allem der mittlere Wertebereich am häufigsten durch Fragen abgedeckt wurde. Fragen, die den unteren oder oberen Wertebereich betreffen, überschneiden sich oftmals auch mit dem mittleren Wertebereich. Somit ist es nicht verwunderlich, daß die Fehlerrate in diesen Bereichen mit niedriger und hoher Sprechgeschwindigkeit kaum von einer sprechgeschwindigkeitsabhängigen Modellierung profitiert.

In Tabelle 6.8 zeigt sich, daß vor allem Vokale und Diphthonge sprechgeschwindigkeitsabhängig modelliert werden. Dies liegt daran, daß die Aussprachedauer dieser Phoneme sehr stark von der Sprechgeschwindigkeit abhängt (siehe Anhang C). Eher selten werden Nasale, Plosive und Frikative modalitätenabhängig modelliert. Dies hängt mit der mittleren Dauer dieser Phoneme zusammen. Länger andauernde Phoneme variieren in ihrer Dauer viel mehr als kürzer andauernde Phoneme. Bei diesen kann es jedoch passieren, daß sie der schnellen Aussprache ganz zum Opfer fallen.

Wird die Gesamtanzahl der Modelle des sprechgeschwindigkeitsabhängigen Modalitäten-System erhöht, so ergibt sich dasselbe Bild wie im Falle der Dialekte, bei dem die Anzahl der modalitätenabhängigen Modelle rapide zunimmt. Somit tauchen auch hier die Fragen zur Sprechgeschwindigkeit mehrheitlich sehr weit unten im Entscheidungsbaum auf.

#Modelle	A	AEH	AH	AI	AU	B	CH	D	E	E2	EH	ER2
1500			4									
2200	3		20	6						2	8	
3000	30	15	36	31	11	13	31	31	9	38	29	13

#Modelle	EU	F	G	H	I	IE	J	K	L	M	N	O	OE	OH
1500														
2200						6	25			2	10	6		
3000	40	14	17	7	21	20	61	22	3	20	37	12	67	13

#Modelle	P	R	S	SCH	T	TS	U	UE	UEH	UH	V	X	Z
1500													
2200					2		5			12		6	
3000	9	21	23	6	29	28	23	36		37	16	36	10

Tabelle 6.8: Prozentsatz der sprechgeschwindigkeitsabhängigen Modelle pro Phonem

6.6 Signal-Rausch-Abstand

Die Berücksichtigung des Signal-Rausch-Abstands einer Äußerung bei der Ballung der akustischen Modelle reduzierte die Gesamtfehlerrate des Modalitäten-Systems um 4.1%. In Abbildung 6.5 ist der Zusammenhang zwischen dem Signal-Rausch-Abstand und der Fehlerrate aufgezeichnet.

Diskussion der Ergebnisse

Die Fehlerrate des Modalitäten-Systems reduzierte sich vor allem in den Bereichen mit sehr hohem Rauschanteil in der Aufnahme. In den Bereichen mit einem sehr geringem Rauschanteil stieg die Fehlerrate etwas an.

Dies mag zum Teil daran liegen, daß die Bestimmung des Signal-Rausch-Abstands mit Hilfe eines Sprach-Stille-Klassifikators durchaus auch Fehler macht, und somit zur Berechnung des Signal-Rausch-Abstands auch falsch klassifizierte Bereiche verwendet werden können. Damit kann es passieren, daß einige Äußerungen mit starkem Hintergrundrauschen Bereichen mit wenig Hintergrundrauschen zugeordnet werden. Da die meisten Äußerungen in diesem Bereich jedoch nur wenig Hintergrundrauschen besitzen, kann es bei einer unausgewogenen Evaluierungsmenge durchaus zu solchen Diskrepanzen in der Fehlerrate kommen. Eine Evaluierungsmenge, die den ganzen Wertebereich gleichmäßig abdeckt, steht je-

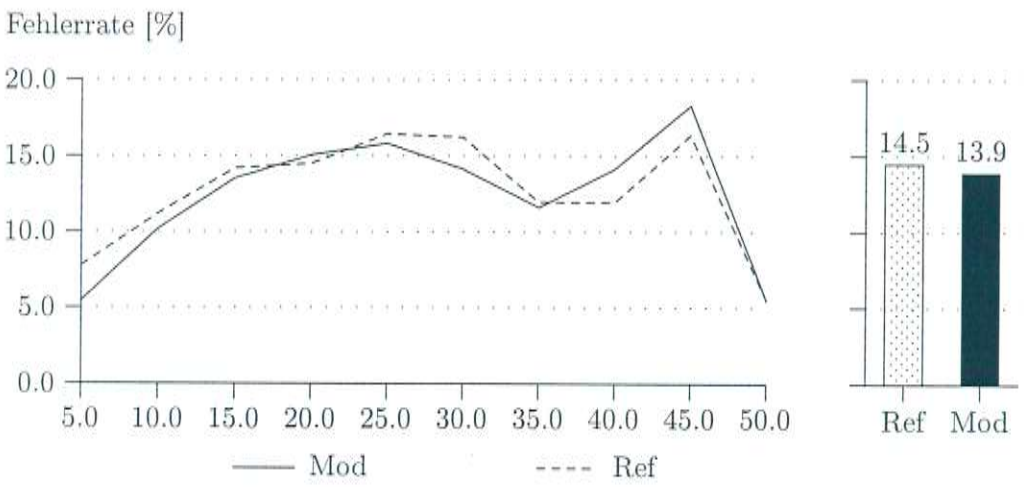


Abbildung 6.5: Fehlerratenreduktion für verschiedenen SNRs

Häufigkeit	Intervall
16	[−12.5, 27.5]
16	[−12.5, 22.5]
5	[−12.5, 17.5]
2	[22.5, 62.5]
2	[−12.5, 32.5]
1	[−12.5, 37.5]

Tabelle 6.9: Häufigkeit der 6 meist verwendeten Fragen zum Signal-Rausch-Abstand für das 2200 Modelle umfassende Modalitäten-System

doch meist nicht zur Verfügung, da diese sehr umfangreich sein müßte und somit potentielle Trainingsdaten verloren gingen.

Auch hier stellt sich die Frage, mit welchen Modalitätenfragen der Wertebereich des Signal-Rausch-Abstand abgedeckt wurde. In Tabelle 6.9 sind hierzu die 6 meist verwendeten Modalitätenfragen aufgelistet. Die meisten Fragen werden darin zu Bereichen mit sehr hohem Rauschanteil gestellt, wobei der Ballungsalgorithmus eine Zweiteilung des Wertebereichs vorgenommen hat. Die Grenze lag in etwa bei einem Signal-Rausch-Abstand von 22.5 dB.

In Tabelle 6.10 ist der Prozentsatz der SNR-abhängigen Modelle pro Phonem aufgelistet. Auch hier zeigt sich, daß die Modalitätenfragen erst sehr weit unten im Entscheidungsbaum gestellt werden. Der Anteil der SNR-abhängigen Modelle ist bei stimmlosen Konsonanten und Frikativen am größten. Genau mit diesen Lauten hat ein modalitätenunabhängiges Spracherkennungssystem die meisten

Probleme, wodurch es oftmals zu Verwechslungen mit Geräuschmodellen kommen kann. Wird die Gesamtanzahl der Modelle im System erhöht, so werden auch immer mehr Vokale und Diphthonge SNR-abhängig modelliert. Da diese meist eine sehr lange Aussprachedauer besitzen, kommen die Rauschanteile dort besonders zum tragen und setzen sich in der akustischen Modellierung fest. Dies ist oftmals an einer wesentlich höhere Varianz der akustischen Modelle zu erkennen[7]. Durch eine SNR-abhängige Modellierung werden die akustischen Modelle wieder aufgeteilt und können so sehr viel genauer modelliert werden.

#Modelle	A	AEH	AH	AI	AU	B	CH	D	E	E2	EH	ER2
1500			4					4				
2200	3		15			4		14		2		3
3000	32		34	23	11	4	38	29	9	36	21	10

#Modelle	EU	F	G	H	I	IE	J	K	L	M	N	O	OE	OH
1500											2			
2200							31	7		2	10			
3000	22	19	18	7	16	13	61	14	3	21	38	11	67	14

#Modelle	P	R	S	SCH	T	TS	U	UE	UEH	UH	V	X	Z
1500												8	
2200			11		8	9				12		6	
3000		18	38	12	34	23	17	37		35	14	15	11

Tabelle 6.10: Prozentsatz der SNR-abhängigen Modelle pro Phonem

[7]das akustische Modell wird quasi auseinandergezogen

Kapitel 7

Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Verfahren präsentiert, daß es ermöglicht sprecherabhängige und sprecherunabhängige zeitveränderliche Modalitäten in den Spracherkennungsprozeß zu integrieren. Dadurch können die Varianzen der akustischen Modelle verringert werden. Außerdem entscheidet das Verfahren automatisch für welche akustischen Modelle ein Parametersharing erfolgen sollte. Die Optimierung der Modellgenauigkeit unter Gewährleistung einer ausreichenden Generalisierungsfähigkeit findet dabei rein datengetrieben durch den Ballungsalgorithmus statt. Dieser berücksichtigt auch automatisch Abhängigkeiten zwischen verschiedenen Ausprägungen einer oder mehrerer Modalitäten.

Für die Problematik zu grober Modelle aufgrund der gleichmäßigen Berücksichtigung aller Ausprägungen einer oder mehrerer Modalitäten, wurden in Kapitel 3 andere varianzvermindernde Verfahren vorgestellt. Erwähnt wurden zunächst akustische Adaptionsverfahren, die allgemeingültig trainierte akustische Modelle an die momentan vorliegende Umgebung anpassen. Sprecherabhängige Modalitäten, wie zum Beispiel das Geschlecht können durch Normalisierungsverfahren berücksichtigt werden. Für aussprachespezifische Modalitäten, wie der Dialekt des Sprechers bietet sich eine Adaption der akustischen Modelle auf andere Dialekte, verbunden mit dem Einfügen von zusätzlichen Aussprachevarianten für den neuen Dialekt, an. Wertkontinuierliche Modalitäten, wie zum Beispiel die Sprechgeschwindigkeit und der Signal-Rausch-Abstand lassen sich auch in den Merkmalsvektor aufnehmen, wodurch dieser in seiner Dimensionalität erweitert wird. Für alle hier aufgezählten Modalitäten lassen sich auch unterschiedliche Spracherkennungssysteme trainieren, die jeweils nur einen Teil des Wertebereichs erfassen.

Diese aufgeführten Verfahren besitzen jedoch mehrere Nachteile. Adaptionstech-

niken benötigen meist eine gewisse Zeit, um sich an die jeweilige Umgebung anzupassen. Werden solche Techniken auf zeitveränderliche Modalitäten angewendet, so kann es passieren, daß sich der Wert der Modalität schon bevor die Anpassung erfolgt ist, wieder verändert hat. Da Normalisierungsverfahren meist nur stückweise lineare Abbildungen über den gesamten Wertebereich verwenden, gibt es Bereiche, die etwas besser und Bereiche die etwas schlechter normalisiert werden. Eine gleich gute Normalisierung aller Ausprägung einer Modalität kann dadurch nicht statt finden. Die Verwendung von unterschiedlichen Spracherkennungssystemen für Teile des Wertebereichs der Modalitäten verursacht während des Trainings einen großen Zeitaufwand und während der Erkennung einen hohen Bedarf an Ressourcen. Ferner müssen für alle Bereiche genügend Trainingsdaten vorhanden sein. Das Training unterschiedlicher Systeme erlaubt es nur unter erschwerten Bedingungen, für einzelne Bereiche des Parameterraums dieselben Trainingsdaten für alle Systeme zu verwenden, um damit ein Parametersharing der akustischen Modelle für dünn besetzte Bereiche im Parameterraum zu erlauben. Ein modalitätenabhängiges System könnte nun diese Probleme beseitigen, in dem es einige Modelle modalitätenabhängig modelliert und andere wiederum nicht.

Um dies zu erreichen, wurde in Kapitel 4 beschrieben, wie akustische Modelle mit den Modalitäten der vorliegenden Äußerung markiert und wie Modalitätenfragen in die Entscheidungsbäume integriert wurden. Somit konnten die akustischen Modelle nicht nur nach deren Kontext, sondern auch nach deren Modalitätenkombination divisiv geballt werden. Durch die Integration der Modalitätenfragen in den Entscheidungsbaum verliert dieser seine statische Eigenschaft, das heißt, die Antwort auf eine Frage kann sich von Äußerung zu Äußerung ändern. Diese ist jedoch nötig, um den Suchraum aufzubauen und geeignet zu komprimieren, so daß keine HMMs doppelt evaluiert werden müssen. Aus diesem Grund wurden die Modalitätenfragen durch eine Baumtransformation soweit nach unten geschoben, bis keine andere Frage mehr unter einer Modalitätenfrage vorhanden war. Dies machte es möglich, die so entstandenen Teilbäume, die nur Modalitätenfragen und die dazugehörigen akustische Modelle enthalten, durch imaginäre akustische Modelle zu ersetzen. Die abgespaltenen Teilbäume wurden aus Effizienzgründen in eine Lookup-Tabelle umgesetzt, deren Zeilen durch die Anzahl der imaginären Modelle und deren Spalten durch alle Modalitätenkombinationen gegeben waren. Jedes Mal, wenn während des Dekodierungsprozesses die Wahrscheinlichkeit eines imaginären Modells berechnet werden soll, kann nun durch Angabe der Modalitätenkombination der vorliegenden Äußerung das imaginäre Modell durch das entsprechende reale Modell ersetzt werden.

In Kapitel 5 wurden zur Evaluierung dieses Verfahrens exemplarisch die folgenden Modalitäten untersucht:

- Geschlecht
- Dialekte
- Sprechgeschwindigkeit
- Signal-Rausch-Abstand

Es wurde beschrieben, wie diese Modalitäten aus vorhandenen Äußerungen zu extrahieren und in das Spracherkennungssystem zu integrieren sind.

Zu betonen ist jedoch, daß dieses Verfahren auch auf weitere sprecherabhängige oder auch sprecherunabhängige Modalitäten anwendbar ist. In Frage kommen da beispielsweise verschiedene prosodische Merkmale, um akustische Eigenschaften, wie hyperartikulatorische Teile [SW98b] einer Äußerung explizit zu modellieren, oder Konfidenzmaße die Auskunft über die Spontaneität der momentan vorliegenden Äußerung geben. Es konnte gezeigt werden, daß gerade wertkontinuierliche Modalitäten wie die Sprechgeschwindigkeit und der Signal-Rausch-Abstand eine außerordentliche Korrelation mit der Fehlerrate besitzen.

Für jede der oben aufgeführten Modalitäten wurde in Kapitel 6 ein System trainiert, das für jeden Sprecher a priori dieselbe Anzahl an Modellen zur Verfügung hatte, wie das Referenzsystem. Damit standen für jeden Sprecher unabhängig von seiner Modalität etwa 2000 Modelle zur Verfügung. Die Integration von geschlechtsabhängigen Modalitäten in das Spracherkennungssystem erbrachte den größten Gewinn. Die Fehlerrate konnte um 7.6% reduziert werden. Die Verwendung von dialektabhängigen Modellen erbrachte eine Reduktion der Fehlerrate um 6.9%. Auffallend war dabei, daß die Fehlerrate für alle in den Evaluationsdaten vorkommenden Dialekte gesenkt werden konnte. Die Integration der Sprechgeschwindigkeit brachte nicht die erwarteten Erfolge. Hier konnte die Fehlerrate um 3.4% gesenkt werden, wobei die Fehlerrate gerade für Äußerungen mit einer hohen Sprechgeschwindigkeit zunahm. Die Integration des Signal-Rausch-Abstands reduzierte wie erhofft die Fehlerrate für Bereiche mit erhöhtem Hintergrundrauschen. Jedoch nahm die Fehlerrate für besonders klare Äußerungen zu. Insgesamt ergab sich jedoch eine Reduktion der Fehlerrate um 4.1%. In Abbildung 7.1 sind die ganzen Ergebnisse zusammengefaßt dargestellt.

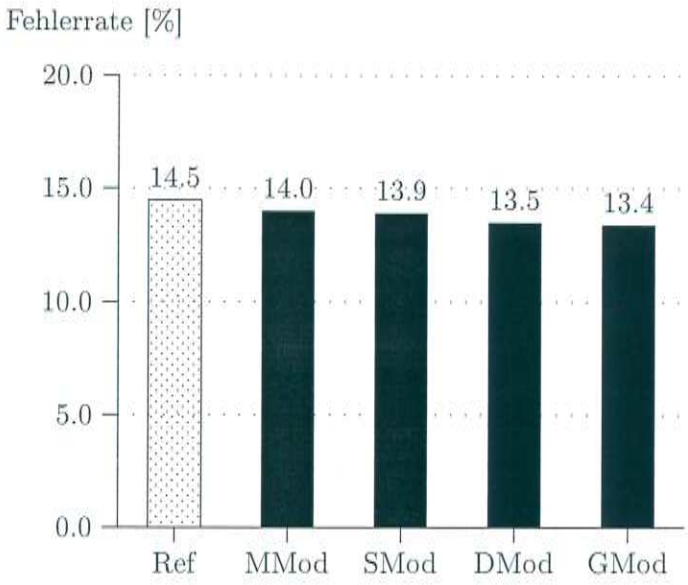


Abbildung 7.1: Fehlerratenreduktion für verschiedene modalitätenabhängige Systeme gegenüber dem Referenzsystem (Ref). MMod (Sprechgeschw.), SMod (SNR), DMod (Dialekte), GMod (Geschlecht)

Diskussion der Ergebnisse

Für die in dieser Arbeit untersuchten Modalitäten konnte in allen Fällen die Fehlerrate gesenkt werden, wobei durch die Integration der Modalitäten die Fehlerraten über den Wertebereich der Modalitäten etwas mehr ausgeglichen werden. Damit ist anzunehmen, daß es noch eine Reihe weiterer Modalitäten gibt, bei denen ein ähnliches Resultat zu erwarten ist. Die Schwierigkeit, die sich gerade bei wertkontinuierlichen Modalitäten ergibt, ist jedoch, eine geeignete Aufteilung des Wertebereichs dieser zu finden. Hierzu mußte der Wertebereich in Intervalle aufgeteilt werden, und aufeinanderfolgende Intervalle als Fragen in den Fragenkatalog aufgenommen werden. Die hier getroffene Annahme, daß sich nur aufeinanderfolgende Intervallbereiche akustisch ähnlich sind, könnte jedoch das Spracherkennungssystem in seiner Modellierungsfreiheit beschränken. Für Modalitäten wie der Signal-Rausch-Abstand oder die Sprechgeschwindigkeit ist das zwar nicht zu erwarten, weil es intuitiv keinen Sinn machen würde, zum Beispiel besonders schnell und besonders langsam ausgesprochene Phoneme zusammenzufassen. Es könnte jedoch auch Modalitäten geben, bei denen eine solche Modellierung zu einer besseren Erkennungsleistung führen könnte.

Eine weitere Schwierigkeit ist es, die Granularität der Aufteilung des Wertebereichs und damit die Größe des Quantisierungsfehlers festzulegen. Hierzu wurden

in dieser Arbeit keine Untersuchungen gemacht. Der Wertebereich wurde einfach in möglichst kleine, gleich große Intervalle unterteilt. Es könnte jedoch auch von Vorteil sein, die Intervallgrenzen nicht äquidistant auf dem Wertebereich zu verteilen, sondern in Abhängigkeit der in einem Intervall vorhandenen Trainingsdaten festzulegen. Dies würde dazu führen, daß Bereiche mit einer hohen Anzahl von Trainingsdaten feiner aufgeteilt werden würden als andere. Der Ausreißer in der Fehlerrate, der bei dem sprechgeschwindigkeitsabhängigen Modalitäten-System bei Schnellsprechern zu sehen war, könnte eventuell genau auf diese Weise beseitigt werden.

Für diskretwertige Modalitäten mit besonders vielen Ausprägungen, bei denen die Trainingsdaten ungleichmäßig auf die Ausprägungen verteilt sind, muß ebenfalls eine Ballung von Regionen mit wenig Trainingsdaten zu Regionen mit vielen Trainingsdaten erfolgen. Dies wurde in dieser Arbeit für die Ballung der Sprachregionen zu größeren Gebieten durchgeführt. Die hierbei verwendete agglomerative Ballung der modalitätenabhängigen akustischen Modelle, faßte diese nur aufgrund ihrer akustischen Ähnlichkeit zusammen. Trotzdem konnte gezeigt werden, daß eine solche Ballung angewandt auf die Menge der Sprachregionen Deutschlands durchaus zu plausiblen Ergebnissen führt. Die vom Ballungsalgorithmus erzeugte Dreiteilung besaß wohldefinierte Grenzen, die in den meisten Regionen auch einer linguistischen Analyse standhalten könnten. Gleichzeitig konnten durch Rückverfolgung des Ballungsvorgangs auch kleinere Teilgebiete ausgemacht werden, die zusätzlich in den Fragenkatalog mit aufgenommen wurden. So war es auch möglich, kleine Überschneidungen zwischen den drei großen Gebieten zuzulassen.

Ausblick

Der ideale Einsatzbereich eines solchen modalitätenabhängigen Systems sind Umgebungen mit hoher Variabilität in der Aussprache und in der Akustik der Äußerungen, wie sie zum Beispiel bei Auskunftssystemen zu finden sind. Dort rufen Menschen aus den unterschiedlichsten Sprachregionen Deutschlands an, wobei die jeweilige Leitungsqualität von Anruf zu Anruf variiert. Durch eine während des Gesprächs stattfindende automatische Analyse des Hintergrundrauschens und des Dialekts des Sprechers könnten die entsprechenden akustischen Modelle ausgewählt werden.

Das hier vorgestellte Verfahren könnte auch dazu benutzt werden Äußerungen, dessen Ausprägungen einer Modalität nicht bekannt sind, mit Hilfe eines bereits bestehenden modalitätenabhängigen Systems zu klassifizieren. Hierzu müßte für

jede Ausprägung der Modalität ein Viterbi-Pfad mit den dazugehörigen Modellen ermittelt werden, wobei die Äußerung dann derjenigen Modalität zugeordnet werden würde, dessen Viterbi-Pfad die größte Wahrscheinlichkeit besäße. Der Viterbi-Pfad wird nur entlang einer von einem modalitätenunabhängigen Spracherkennungssystem ermittelten Hypothese für diese Äußerung bestimmt. Somit basiert die Klassifikation rein auf Basis akustischer Ähnlichkeiten der modalitätenabhängigen Modelle mit der Äußerung.

Die Ballung akustische Ähnlichkeiten könnte auch für die multilinguale Spracherkennung von Vorteil sein. Darin muß immer eine Zuordnung von Phonemen unterschiedlicher Sprachen getroffen werden. Diese wird zur Zeit fast nur manuell festgelegt, und schränkt damit das Spracherkennungssystem schon im voraus unnötig ein. Durch Anwendung der agglomerativen Ballung auf die fertig trainierten akustischen Modelle der einzelnen Phoneme verschiedener Sprachen, könnte die Zuordnung von Phoneme durch das Spracherkennungssystem selbst gefunden werden.

In allen durchgeführten Experimenten wurde keine Anpassung der Basis-Sprachmodellparameter vorgenommen. Bei der Treeforward-Phase wurde immer dieselbe Gewichtung des Sprachmodells zur akustischen Modellierung verwendet. Falls eine Korrelation der Sprachmodellparameter mit den extrahierten Modalitäten besteht, so könnte man sich gerade beim Signal-Rausch-Abstand vorstellen, daß eine höhere Gewichtung des Sprachmodells für Äußerungen mit hohem Rauschanteil zu einer besseren Erkennungsleistung führt, wie bei einer unveränderten Einstellung der Sprachmodellparameter. Deshalb könnte eine Mitoptimierung der Sprachmodellparameter in Abhängigkeit der extrahierten Modalitäten von Vorteil sein.

Die Zuordnung der Sprachmodellparameter zu einzelnen Modalitätenkombinationen könnte sogar mit Hilfe des in dieser Arbeit vorgestellten Verfahrens passieren. Hierzu würde für jede Äußerung neben der dazugehörigen Modalitätenkombination dessen beste Einstellung der Sprachmodellparameter ermittelt. Durch Definition eines Distanzmaßes könnten die Sprachmodellparameter nun mit Hilfe von Fragen zu einzelnen Modalitäten zusammengeballt werden. Während der Erkennung ließe sich dann für jede Äußerung durch Abstieg im Baum anhand der extrahierten Modalitätenkombination die gewünschten, bestmöglichen Sprachmodellparameter ermitteln. Bei richtiger Wahl des Distanzmaßes werden diejenigen Modalitäten, welche am meisten mit den Sprachmodellparametern korreliert sind, auch als erstes in Form einer Frage im Entscheidungsbaum auftauchen.

Das Problem der Quantisierung des Wertebereichs von wertkontinuierlichen Mo-

dalitäten könnte auch auf andere Weise beseitigt werden. So könnte man entweder wertkontinuierliche Antworten auf Fragen im Entscheidungsbaum erlauben oder durch Interpolation mehrerer akustischer Modelle von verschiedenen Stützstellen im Wertebereich die dazwischenliegenden Modelle erzeugen.

Die unterschiedliche akustische Modellierung der Modelle je nach Dialekt bietet dem System auch die Möglichkeit, fehlende Aussprachevarianten der jeweiligen Dialekte durch eine andere akustische Modellierung vorhandener Aussprachevarianten zu ersetzen. Trotzdem wäre es sicherlich gerade im Falle von unterschiedlichen Dialekten sinnvoll dialektabhängige Aussprachvarianten in das Aussprachewörterbuch aufzunehmen. Diese würden dann über einen Selektionsmechanismus nur für den jeweiligen Dialekt verwendet werden.

Literaturverzeichnis

[ASS95] **A. Anastasakos, R. Schwartz und H. Shu**. Duration Modeling in Large Vocabulary Speech Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 628–631. IEEE, 1995.

[Ber97] **N. Beringer**. Die dialektale Färbung des Deutschen – ein Problem für die automatische Spracherkennung? Diplomarbeit, Universität Stuttgart, Dezember 1997.

[BKK96] **A. Batliner, A. Kießling und R. Kompe**. Tempo und Tempowechsel in Verbmobil-Dialogen. *Verbmobil Memo 110*, 1996.

[BO99] **S. Burger und D. Oppermann**. Regional Variants of German: Categories of Pronunciation Deviation from Standard German. In *Proceedings of the ICPhS*, 1999, San Francisco.

[Bol79] **S. F. Boll**. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 27, S. 113–119. IEEE, April 1979.

[BSRB98] **N. Beringer, F. Schiel und P. Regel-Brietzmann**. German Regional Variants – A Problem for Automatic Speech Recognition? In *Proceedings of the International Conference on Speech and Language Processing*. IEEE, November 1998, Sydney.

[DH73] **R. O. Duda und P. E. Hart**. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[Füg98] **C. Fügen**. Optimieren der Sprachmodellparameter durch künstliche neuronale Netze. Studienarbeit, Oktober 1998.

[FGJ98] **V. Fischer, Y. Gao und E. Janke**. Speaker-Independent Upfront Dialect Adaption in a Large Vocabulary Continuous Speech Recognizer. In *Proceedings of the International Conference on Speech and Language Processing*. IEEE, November 1998, Sydney.

[FR97] **M. Finke und I. Rogina**. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 3, S. 1743–1746. IEEE, April 1997, München.

[GL94] **J.-L. Gauvain und C.-H. Lee**. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. In *IEEE Transactions on Speech and Audio Processing*, Band 2, S. 291–298. IEEE, 1994.

[Hes83] **W. Hess**. *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.

[Hon92] **H.-W. Hon**. Vocabulary-Independent Speech Recognition: The VOCIND System. Technical Report CMU-CS-92-108, Carnegie Mellon University, Pittsburgh, PA, März 1992.

[Jel90] **F. Jelinek**. Self-Organized Language Modeling for Speech Recognition. In A. Waibel und K.-F. Lee, Herausgeber, *Readings in Speech Recognition*. Morgan Kaufman, 1990.

[Kem95] **T. Kemp**. Data-Driven Codebook Adaptation in Phonetically Tied SCHMMs. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 377–479. IEEE, Mai 1995, Detroit, USA.

[Kie96] **A. Kießling**. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Dissertation, Universität Erlangen-Nürnberg, 1996.

[Kom96] **R. Kompe**. *Prosody in Speech Understanding Systems*. Dissertation, University of Erlangen-Nürnberg, 1996.

[Lee88] **K.-F. Lee**. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. CMU-CS-88-148, Carnegie Mellon University, Pittsburgh, PA, April 1988.

[LW94] **C. Leggetter und P. Woodland**. Speaker Adaptation of Continuous Density HMMs Using Linear Regression. In *Proceedings of*

the International Conference on Speech and Language Processing, Band 2, S. 451–454, 1994, Yokohama.

[LW95] **C. Leggetter und P. Woodland**. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185, 1995.

[MFL98] **N. Morgan und E. Fosler-Lussier**. Combining Multiple Estimators of Speaking Rate. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Mai 1998, Seattle.

[MFM95] **N. Mirghafori, E. Fosler und N. Morgan**. Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes. In *Proceedings of EUROSPEECH95*, S. 491–494, September 1995, Madrid.

[MFM96] **N. Mirghafori, E. Fosler und N. Morgan**. Towards Robustness to Fast Speech in ASR. In *International Conference on Acoustics, Speech, and Signal Processing*, S. 335–338, Mai 1996, Atlanta.

[Nol67] **A. M. Noll**. Cepstrum Pitch Determination. *Journal of the Acoustical Society of America*, 14:293–309, 1967.

[OB99] **D. Oppermann und S. Burger**. What Makes Speech Data Spontaneous? In *Proceedings of the ICPhS*, 1999, San Francisco.

[Ode92] **J. J. Odell**. The Use of Decision Trees with Context Sensitive Phoneme Modeling. Diplomarbeit, Department of Engineering, Cambridge University, Cambridge, UK, August 1992.

[Rab77] **L. R. Rabiner**. On the Use of Autocorrelation Analysis for Pitch Detection. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band ASSP-25, S. 24–33, Februar 1977.

[RC98] **W. Reichl und W. Chou**. Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, May 1998, Seattle.

[RC99] **W. Reichl und W. Chou**. A Unified Approach of Incorporating general Features in Decision Tree Based Acoustic Modeling. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1999.

[Rog97] **I. Rogina**. Automatic Architecture Design by Likelihood-Based Context Clustering with Crossvalidation. In *Proceedings of EUROSPEECH97*, September 1997, Rhodos.

[Rog98] **I. Rogina**. *Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular*. Dissertation, University of Karlsruhe, Germany, 1998.

[Sch99] **K. Schubert**. Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung. Diplomarbeit, University of Karlsruhe, Germany, November 1999.

[SKT98] **F. Schiel, A. Kipp und H. G. Tillmann**. It's not the Model, it's the data. In *Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Mai 1998, Kerkrade, Netherlands.

[SR95] **T. Schultz und I. Rogina**. Acoustic and Language Modeling of Human and Nonhuman Noises for Human-To-Human Spontaneous Speech Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 293–296. IEEE, Mai 1995, Detroit, USA.

[SS95] **M. A. Siegler und R. M. Stern**. On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 1, S. 612–615. IEEE, Mai 1995, Detroit.

[ST95] **E. G. Schukat-Talamazzini**. *Automatische Spracherkennung*. Vieweg, 1995.

[SW98a] **T. Schultz und A. Waibel**. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In *Proceedings of the International Conference on Speech and Language Processing*. IEEE, November 1998, Sydney.

[SW98b] **H. Soltau und A. Waibel**. On the Influence of Hyperarticulated Speech on Recognition Performance. In *Proceedings of the International Conference on Speech and Language Processing*. IEEE, November 1998, Sydney, Australia.

[SW99] **T. Schultz und A. Waibel**. Language adaptive LVCSR through Polyphone Decision Tree Specialization. In *MIST Workshop Proceedings*, S. 85–90. MIST, 1999, Leusden, Niederlande.

[Tal95] **D. Talkin**. A Robust Algorithm for Pitch Tracking (RAPT). *Speech Coding and Synthesis*, S. 495–518, 1995.

[VER] **VERBMOBIL**. Verbmobil Homepage. http://www.dfki.uni-sb.de/verbmobil.

[WCE+93] **M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel und W. Ward**. Recent Advances in JANUS: A Speech Translation System. In *ARPA Workshop on Human Language Technology Systems*. Morgan Kaufmann, März 1993.

[WF96] **M. Woszczyna und M. Finke**. Minimizing Search Errors due to Delayed Bigrams in Real-Time Speech Recognition Systems. In *International Conference on Acoustics, Speech, and Signal Processing*, S. 137–140. IEEE, Mai 1996.

[WL90] **A. Waibel und K.-F. Lee**. *Readings in Speech Recognition*. Morgan Kaufmann, 1990.

[Wos98] **M. Woszczyna**. *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition*. Dissertation, University of Karlsruhe, Germany, 1998.

[WOVY94] **P. Woodland, J. Odell, V. Valtchev und S. Young**. The HTK Large Vocabulary Recognition System: An Overview. In *ARPA Spoken Language Technology Workshop*, März 1994, Princeton, New Jersey.

[ZB99] **M. A. Zissman und K. M. Berkling**. Automatic Language Identification. In *Multi-lingual Interoperabilty Speech Technology Workshop Proceedings*, 1999, Leusden, Niederlande.

[ZW97] **P. Zhan und M. Westphal**. Speaker Normalization Based on Frequency Warping. In *International Conference on Acoustics, Speech, and Signal Processing*, S. 1039–1042. IEEE, April 1997.

Anhang A

Sprache-Stille-Klassifikation

Der in das verwendete Spracherkennungssystem integrierte Sprache-Stille-Klassifikator basiert auf den logarithmierten mittleren Betragsamplituden und der Anzahl der Nulldurchgänge.

Hierzu wird für jeden Fensterausschnitt des Sprachsignals die logarithmierte mittlere Betragsamplitude und die Anzahl der Nulldurchgänge ermittelt. Danach werden die Betragsamplituden aufsteigend sortiert, wobei die Nulldurchgänge parallel mitsortiert werden. Durch Anwendung des 2-Mittelwerte-Algorithmus auf die sortierten Betragsamplituden erhält man je einen Mittelwert für die Sprachanteile und einen Mittelwert für die Stilleanteile des Sprachsignals. Die Intervallgrenze der beiden Mittelwerte definiert die Schwelle für eine Unterscheidung zwischen stimmhaften Lauten und Stille oder stimmlosen Lauten (Schwelle 1). Durch die parallele Sortierung der Nulldurchgänge sind die Nulldurchgänge im Stilleintervall der Betragsamplituden vorgegeben. Auf diesem Intervall wird nun ebenfalls der 2-Mittelwerte-Algorithmus angewendet und somit in zwei Teile zerlegt. Die Grenze zwischen den beiden Teilen bildet nun die Schwelle für die Unterscheidung zwischen Stille und stimmlosen Lauten (Schwelle 2). Die durch diese beiden Schwellen vorgegebene Sprache-Stille-Klassifikation besitzt nun noch einige Unebenheiten durch relativ kurze Ausreißer. Diese werden durch eine anschließende Glättung entfernt. Abbildung A.1 bietet nochmals einen Überblick über den Ablauf dieses Verfahrens.

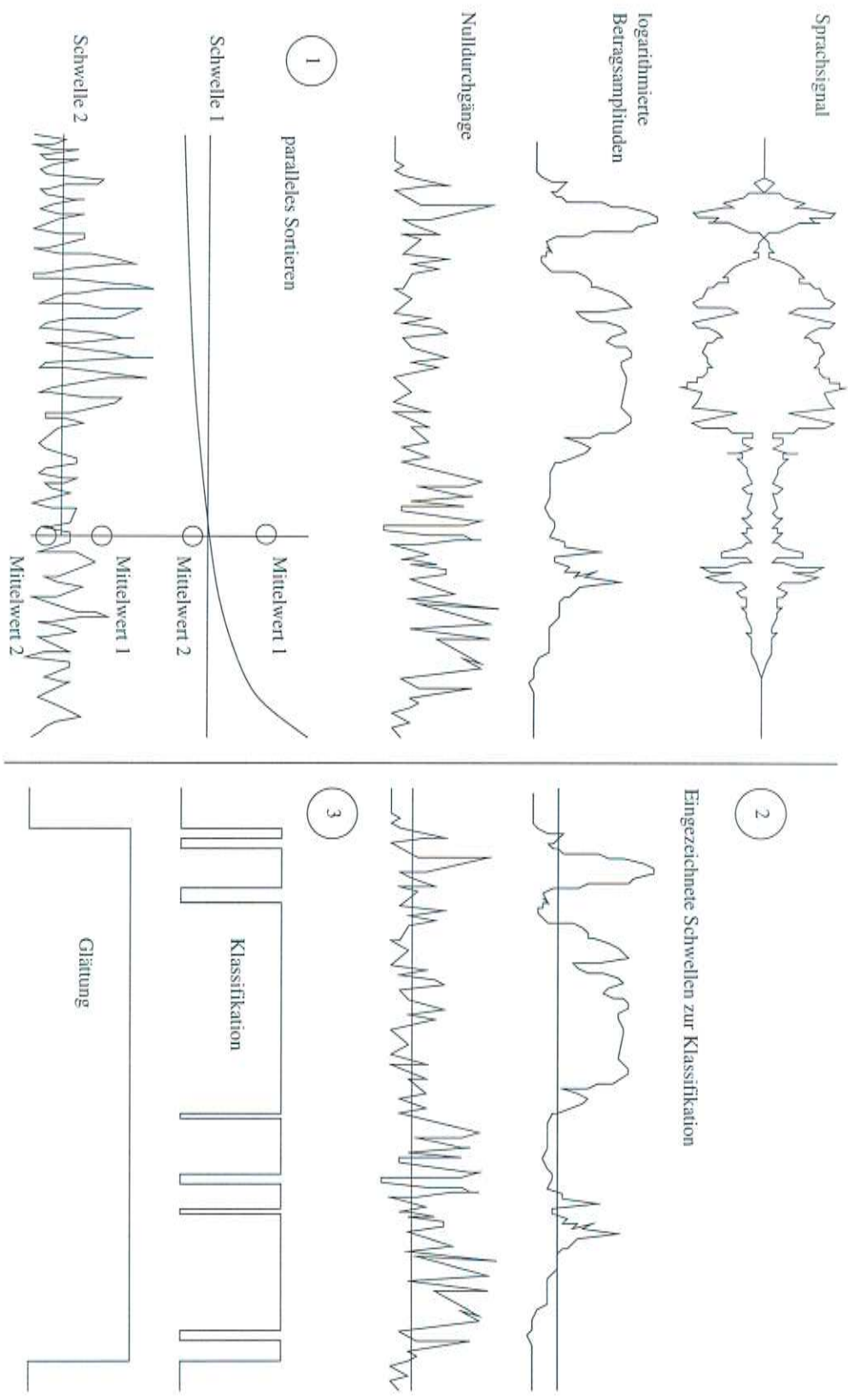


Abbildung A.1: Sprache-Stille-Klassifikation

Anhang B

Einige Artikulationsgruppen

In diesem Abschnitt werden einige der in dieser Arbeit verwendeten Artikulationsgruppen mit deren dazugehörigen Phonemen aufgelistet.

Artikulationsgruppe	dazugehörige Phoneme
Affrikate	TS TSCH
Diphthonge	AEHR AHR AI AR AU EHR ER EU IHR IR OHR OR TS TSCH UEHR UHR UR
Behauchte	K P T
Dental Alveolare	D L N R S T TS Z
Frikative	CH F H J R S SCH TS TSCH V X Z
Glottale	H
Konsonanten	B CH D F G H J K L M N NG P R S SCH T TS TSCH V X Z
stimmhafte Kons.	L M N NG R
stimmlose Kons.	CH F H K P S SCH T TS TSCH X
Nasale	ANG M N NG
Labiale	B M P
Laterale	L
Palatale	CH J
Plosive	B D G K P T TS TSCH
Schwas	AEHR AHR AR E2 EHR ER ER2 IHR IR OHR OR UHR UR
Vibranten	R
Vokale	A AEH AEHR AH AHR AI ANG AR AU E E2 EH EHR ER ER2 EU I IE IHR IR O OE OEH OH OHR OR U UE UEH U EHR UH UHR UR
runde Vokale	AU EU O OE OEH OH OHR OR U UE UEH UEHR UH UHR UR
unrunde Vokale	A AEH AEHR AH AHR AI ANG AR AU E E2 EH EHR ER ER2 I IE IHR IR

Tabelle B.1: Artikulationsgruppen und dazugehörige Phoneme

Anhang C

Untersuchungen zur Phonemdauer

In C.1 ist für jedes Phonem dessen durchschnittliche Dauer angegeben. Diese wurden auf Basis der vom Spracherkennungssystem erzeugten Zuordnungen der HMM-Zustände zu den einzelnen Fensterausschnitten der Äußerung auf den Trainingsdaten ermittelt.

	A	AEH	AH	AI	AU	B	CH	D	E	E2
Dauer	6.9	8.1	9.9	11.2	12.5	6.6	8.0	6.0	6.8	6.2
Varianz	11.1	24.0	47.6	40.3	45.1	39.6	33.8	77.3	10.9	35.0

	EH	ER2	EU	F	G	H	I	IE	J	K
Dauer	7.9	7.7	11.3	9.3	6.8	6.8	6.1	7.2	12.0	9.4
Varianz	32.6	49.2	13.9	19.8	24.5	18.5	11.6	29.3	54.7	42.5

	L	M	N	O	OE	OH	P	R	S	SCH
Dauer	6.9	8.16	9.8	6.8	8.0	9.1	8.6	7.1	9.6	9.2
Varianz	19.1	36.8	55.1	8.9	9.4	45.1	36.8	22.7	45.7	17.2

	T	TS	U	UE	UEH	UH	V	X	Z
Dauer	8.5	10.1	7.1	5.6	6.5	8.8	6.3	8.9	7.7
Varianz	39.1	25.9	31.2	6.4	30.5	32.3	88.1	92.4	13.4

Tabelle C.1: Durchschnittliche Dauer und Varianz der Phoneme in Fensterabschnitten (Frames) von 10 ms Länge