Automatisierte Erstellung von Korpora für die Prosodieforschung

Dissertationsschrift
zur Erlangung der Würde
eines Doktors der Philosophie (Dr. phil.),
eingereicht bei der Fakultät für Philosophie
der Universität Stuttgart

vorgelegt von
Stefan Rapp

Stuttgart
September 1998

Inhaltsverzeichnis

1 Einleitung **1**

2 Eine phonologische Intonationsbeschreibung für das Deutsche **14**

2.1 Inventar an Akzenttönen 16

2.2 Inventar an Grenztönen 18

2.3 Oberflächen- und Tiefenstuktur, Linking 19

2.4 Downstep und Register 20

2.5 Zusammenfassung 22

3 Ein System zur automatischen Segmentierung gesprochener Sprache in Wörter, Silben und Phoneme **24**

3.1 Segmentierungsverfahren 24

3.1.1 Implizite Verfahren: Signalnahe Segmentierung 25

3.1.2 Explizite Verfahren: Segmentieren durch Erkennen 25

3.2 Systemarchitektur 27

3.2.1 Aufbereitung des orthographischen Texts 27

3.2.2 Aufbereitung des Sprachsignals 30

3.3 Training 30

3.3.1 Topologie der Phonemmodelle 31

3.3.2 Modellierung von Sprechpausen 32
3.3.3 Evaluation des Trainings 33
3.4 Segmentierung 33
3.5 Diskussion 35

4 Eine Parametrisierung der Sprachgrundfrequenz 37
4.1 Anforderungen des Intonationsmodells 37
4.2 Wahl der Parametrisierungsfunktion 39
4.3 Phonetische Interpretation der Parameter 42

5 Eine Beschreibung der Intensität und Stimmqualität 43
5.1 Einführung 43
5.2 Untersuchungsmaterial 49
5.3 Schiefe und Schließrate als Korrelate der Betonung 49
5.4 MFCC als Korrelat der Betonung 51
5.5 Training und Test an einem größeren Korpus 56
5.6 Zusammenfassung 59

6 Automatische Bestimmung der phonologischen Intonationsbeschreibung 62
6.1 Vorüberlegungen 64
6.1.1 Überblick über Lernverfahren 64
6.1.2 Automatisch extrahierbare Merkmale für die Vorhersage der Intonationsbeschreibung 66
6.1.3 Weitere Aspekte 70
6.1.4 Verwandte Arbeiten 73
6.2 Voruntersuchungen zur Auswahl eines geeigneten Lernansatzes . . 74

6.2.1 Statistische Modellierung mit HMM 74

6.2.2 Silbensynchrone statistische Modellierung mit HMM . . . 76

6.2.3 Prädikatenlogisches Lernverfahren 83

6.2.4 Attribut-Wert-Lernverfahren 88

6.2.5 Bewertung 94

6.3 Auswahl von Merkmalen 95

6.3.1 Berücksichtigung von Kontext 95

6.3.2 Berücksichtigung von Tags 96

6.3.3 Intensitätsmaße 100

6.3.4 Silbendauer vs. Nucleusdauer 102

6.4 Von sprecherabhängiger Erkennung zu sprecherunabhängiger Erkennung 103

6.4.1 Sprechereinflüsse auf die Dauermerkmale 104

6.4.2 Sprechereinflüsse auf die Grundfrequenzmerkmale 104

6.4.3 Experimentelle Untersuchungen zur Kompensation von Sprechereinflüssen 107

6.4.4 Vorteile durch die Normierung 109

6.4.5 Experimente zur Bewertung des Bedarfs an Trainingsmaterial 111

6.5 Zusammenfassung 117

7 Korpuserstellungs- und Korpuszugriffshilfen 120

7.1 Selektion wiederkehrender Nachrichten 120

7.1.1 Einleitung 121

7.1.2 Aufteilen der Nachrichtensendung in Einzelnachrichten . 125

7.1.3 Naive Methode 126

7.1.4 Realignment-Methode 128

7.1.5 Ausnutzung der Unterschiede zu anderen Nachrichten . . 130

7.1.6 Experimentelle Auswertung 132

7.1.7 Bewertung 145

7.2 Erstellung einer Übersicht über die Intonationsvariation auf phonologischer Ebene 148

8 Schlußbemerkungen 150

A Übersicht über aufbereitete Korpora 154

A.1 Über das digitale Satelliten-Radio aufgenommene und aufbereitete Sprachdaten 154

A.1.1 Manuell aufbereitete Radionachrichten (zentrales Teilkorpus) 155

A.1.2 Automatisch aufbereitetes Nachrichtenkorpus 156

A.1.3 Sternzeit 157

A.2 Aufbereitete Fremdaufnahmen 157

A.2.1 Die Leiden des jungen Werther 158

A.2.2 Die Buttergeschichte 158

A.2.3 Saarbrückener Map Task 158

A.2.4 Verbmobil–Dialoge 158

A.3 Aufbereitete Eigenaufnahmen 159

A.3.1 Das Dicke Kind 159

A.3.2 Anaphernresolution 160

A.3.3 „Erst“-Daten 160

A.3.4 Glottale Parameter von Pitchakzenten 161

A.3.5 Zeitungsartikel 161

Dank

Mein Dank gebührt an erster Stelle Herrn Prof. Dr. habil. Grzegorz Dogil. Ich habe ihm aus mehrerlei Hinsicht sehr viel zu verdanken. Professor Dogil hat mich mit einem Seminar, das die Grundlagen der akustischen Phonetik und das Lesen von Spektrogrammen vermittelte, erstmals für die stark interdisziplinär geprägten Gebiete der experimentellen Phonetik und der Sprachtechnologie begeistert. Professor Dogil gab mir nach meinem Studium die Gelegenheit für eine interessante Forschungstätigkeit im Sonderforschungsbereich 340, „Sprachtheoretische Grundlagen der Computerlinguistik“, Teilprojekt C4, „Ein Prosodiemodul für die Diskursrepräsentationstheorie“, während der ein Großteil der in dieser Arbeit vorgestellten Werkzeuge entwickelt, erprobt, eingesetzt und verbessert wurden. Ganz besonders bedanken möchte ich mich bei ihm für seine hervorragende Betreuung, die ich mir in keinem Punkt hätte besser vorstellen können. Bedanken möchte ich mich bei ihm auch für die großzügige Förderung, die ich als Student genießen konnte. Bei Herrn Prof. Dr. Mats Rooth möchte ich mich für die Begutachtung der Arbeit bedanken.

Als nächstes möchte ich mich bei Dr. Jörg Mayer für die kollegiale und produktive Zusammenarbeit im Projekt bedanken. Er hat auch, zusammen mit Karin Müller und Antje Schweizer, die prosodische Analyse von weit über einer Stunde Sprachmaterial bewältigt, ohne die der in dieser Arbeit verfolgte Korpus-basierte Ansatz nicht möglich gewesen wäre. Außerdem bedanken möchte ich mich bei allen Kolleginnen und Kollegen des Lehrstuhls und des Instituts, besonders bei Dr. Wolfgang Wokurek, Dr. Krzysztof Marasek, Dr. Michael Jessen, Dr. Gregor Möhler sowie bei Dr. Peter Regel-Brietzman. Sie alle haben mir in einer sehr freundlichen Umgebung durch ihre Expertise und Anregungen in zahllosen Diskussionen die Arbeit wesentlich erleichtert.

Der Abteilung Intelligente Systeme des Instituts für Informatik, Prof. Dr. Egbert Lehmann, Dipl.-Inform. Irene Weber und Martin Bauer, danke ich für die Bereitstellung von Rechenzeit und die Durchführung der Foil-Experimente.

Dipl.-Ing. Peter Buchner und Dipl.-Ing. Peter Haubrich von der Sony International (Europe) GmbH danke ich für die Einräumung einer Teilzeitregelung, die mir eine zügige Fertigstellung der Arbeit ermöglichte.

Meiner Frau Irene Weber danke ich für die geduldige Unterstützung und die Bereitschaft, die Last von drei bis vier schönen Berufen gemeinsam zu tragen. Neben ihren wissenschaftlichen Beiträgen hat sie durch ihre fachkundigen Kommentare die Lesbarkeit der vorliegenden Arbeit wesentlich verbessert.

Kapitel 1

Einleitung

Thema dieser Arbeit ist die automatisierte Erstellung von Korpora für die Prosodieforschung. Die Prosodie ist ein Teilbereich der Sprachwissenschaft, der in letzter Zeit viel Beachtung fand, da er Beziehungen zu fast allen anderen Bereichen der Linguistik aufweist ([Lev89] bzw. S. 124) und man sich von ihm Erklärungen bisher nur unzureichend verstandener Phänomene erhofft. Die Beeinflussung der Prosodie durch die Vielzahl an Faktoren erschwert jedoch eine klare Analyse. Sie wird im wesentlichen aus zwei unterschiedlichen Richtungen betrieben. Die erste Herangehensweise versucht, in kontrollierten Laborexperimenten einen beeinflussenden Faktor zu variieren, während alle anderen möglichst konstant bleiben. Aufgrund der Vielzahl der Beeinflussungen ist es nicht leicht, geeignete Aufgaben zu konstruieren, so daß in der Folge ein solches Vorgehen oftmals in einer für die Versuchspersonen künstlichen Kommunikationssituation resultiert, die die Ergebnisse einer solchen Untersuchung angreifbar macht. Bei der zweiten Herangehensweise wird versucht, in möglichst großen Korpora, die für die Prosodieforschung natürlich nicht nur aus geschriebenem Text, sondern vor allem aus aufgenommener Sprache bestehen müssen, nach Belegen zur Unterstützung oder Widerlegung von Hypothesen zu suchen. Die methodologische Begründung für dieses Herangehen ist, daß sich alle vom untersuchten, variierten Faktor verschiedenen Einflüsse neutralisieren oder zumindest stabilisieren, wenn die Datenmenge nur ausreichend groß ist. Da die Äußerungen eines solchen Korpus idealerweise einer für die untersuchten Sprecher natürlichen Kommunikationssituation entstammen, die zudem für die Sprecher möglichst unbemerkt erhoben worden sein sollten, entfällt ein wichtiger Einwand gegen die erste Herangehensweise. Auch dienen Korpora als Belegquelle für die Illustration von Hypothesen. Durch den Verweis auf gänzlich unabhängig produzierte Realisierungen, die eine bestimmte Hypothese stützen oder falsifizieren, entzieht man sich dem Vorwurf, konstruierte

Beispiele zu verwenden, die in tatsächlichen Kommunikationssituationen nie auftreten. Man muß dabei jedoch bedenken, daß nicht alle beeinflussenden Faktoren der Analyse einer bestimmten Äußerung restlos zugänglich sind. Die vermutlich beste Untersuchungsmethode liegt in der Kombination der beiden Herangehensweisen, also im Finden von Belegexemplaren in Korpora für eine in Laborexperimenten gefundene Hypothese oder im Überprüfen von bei Korpusanalysen gefundenen Hypothesen durch kontrollierte Laborexperimente.

Die prosodische Analyse kontrollierter Laborexperimente erfolgt traditionell mit den Mitteln der experimentellen Phonetik. Das heißt, für jede Versuchsäußerung werden Daten durch Messen akustisch-physikalischer Korrelate erhoben und anschließend einer im allgemeinen statistischen Analyse zugeführt. Obwohl die Datenerhebung schon zu einem Großteil durch den Einsatz mächtiger automatischer Analysehilfen (digitale phonetische Meßgeräte) unterstützt wird, ist sie in der Regel ein sehr aufwendiger Prozeß, der den zeitlichen Umfang einer Untersuchung häufig dominiert. Um die bei der Sprachproduktion zwangsläufig auftretende zufällige Variation von Äußerung zu Äußerung durch die statistische Analyse hinreichend zu erfassen, werden zudem häufig Wiederholungen gefordert, die die Analysearbeit zusätzlich vergrößern.

Ist für die Auswertung von Laborexperimenten die Anwendung automatischer Analyseverfahren nützlich, so wird diese bei der Aufbereitung großer Sprachdatensammlungen schließlich unausweichlich. Ziel einer automatisierten Analyse ist es einerseits, Kapazitäten auf wissenschaftlich interessantere Fragestellungen zu konzentrieren, und andererseits objektivierte Entscheidungskriterien zu besitzen, die eine möglicherweise unbewußt einseitige Analyse unwahrscheinlicher machen.

Die Verfügbarkeit großer Korpora hat einen nicht zu unterschätzenden Einfluß sowohl auf die Theoriebildung als auch auf die praktische Implementierung von Systemen gehabt und wird diesen auch in der Zukunft besitzen. Beispiele sind die für die Sprachwissenschaft verhältnismäßig jungen Felder der statistischen Verarbeitung geschriebener Sprache und der Spracherkennung. Für die statistische Verarbeitung geschriebener Sprache ist die Verfügbarkeit großer Textkorpora eine naheliegende Voraussetzung, da die Bestimmung der Parameter einer statistischen Modellierung umso genauer wird, je größer die Datenbasis ausfällt. Obwohl die Qualität der Systeme, die in diesem neuen Zweig der angewandten Sprachwissenschaft entwickelt wurden, in Einzelfällen in Frage gestellt werden kann, so sind die erzielten Ergebnisse beachtlich, vor allem vor dem Hintergrund der kurzen Entwicklungsgeschichte des Feldes. Bei der Spracherkennung ist die Situation möglicherweise noch drastischer. Nachdem lange Zeit mehr oder minder erfolglos versucht wurde, das Expertenwissen von Phonetikern und Phonologen, Aku-

stikern und Elektrotechnikern in regelbasierte Systeme zu codieren, gelang erst mit der statistischen Modellierung durch Hidden-Markov-Modelle (HMM) und mit den konnektionistischen Verfahren, der Modellierung durch künstliche Neuronale Netzwerke (ANN), ein wesentlicher Fortschritt auf dem Weg zu einem auf gesprochene Sprache reagierenden System. Der zentrale Punkt für den Erfolg dieser Verfahren ist ihre Fähigkeit, Systemparameter selbständig aus großen Sprachdatensammlungen zu lernen. Die Spracherkennung beginnt nun, eine in der industriellen Praxis einsetzbare Technologie zu werden.

Auch auf dem Gebiet der Sprachsynthese werden in jüngerer Zeit korpusbasierte Techniken angewandt. Neuere Arbeiten beschreiben Verfahren, mit denen aus einem möglichst großen Korpus von Sprachaufnahmen eines Sprechers durch Auswahl kleiner, gut zusammenpassender Einheiten und anschließendem Zusammensetzen sehr natürlich klingende Sprache synthetisiert werden kann. Damit diese Systeme zufriedenstellend arbeiten, müssen die zugrundeliegenden Korpora jedoch geeignet aufbereitet sein, sowohl segmental als auch prosodisch. Der Prosodie scheint gerade auf dem Gebiet der Synthese von Sprache eine große Bedeutung zuzukommen. Verfügt ein System über eine unbefriedigende Prosodiekomponente, wird es von Benutzern schlecht akzeptiert werden.

Ausgangsbasis für den Aufbau eines Korpus für die Prosodieforschung ist im allgemeinen gänzlich unanalysiertes Sprachmaterial, das auf verschiedenen Wegen erhoben wird. Im einfachsten Fall kann direkt Sprache mit einem Mikrophon aufgenommen werden. Für die Verarbeitung und den problemlosen Zugriff wird das Sprachsignal digitalisiert und auf dem Massenspeicher eines Rechnersystems gespeichert. Da Sprachaufnahmen mitunter an Stellen gewonnen werden, die vom Rechnersystem örtlich getrennt sind, wird die Sprache häufig zunächst analog auf Tonbändern oder besser digital, z.B. auf DAT-Bändern (Digital Audio Tape), gespeichert und erst anschließend auf das Rechnersystem übertragen. Um die Analysierbarkeit und Verwendbarkeit des Korpus für weitere Zwecke wie die Sprachsynthese nicht unnötig einzuschränken, sind gewisse Mindestanforderungen an die akustische Qualität einer Sprachaufnahme zu stellen. Dies betrifft die Raumakustik, die Übertragungsqualität des Mikrophons und eventuell verwendeter Aufnahmegeräte, sowie die Digitalisierung und Speicherung auf dem Rechnersystem. Idealerweise sollten Sprachaufnahmen in Räumen ohne Nachhall („Schalltoter Raum“) oder wenig Hall (Studio, möblierter Wohnraum mit Teppichboden) und ohne Nebengeräusche (Straßenlärm, Klimaanlage, Lüftungsgeräusche technischer Geräte, Sprachgeräusche anderer Sprecher) gemacht werden. Das Mikrophon sollte eine möglichst lineare Aufnahmecharakteristik aufweisen, also alle für Sprache relevanten Frequenzbereiche gleich stark in elektrische Signale umsetzen. Auch eine eventuelle Zwischenspeicherung soll die Frequenzcharakteristik und die zeitliche Struktur des Sprachsignals möglichst wenig beeinflussen.

Eine weitere Möglichkeit für die Akquirierung von Sprachmaterial sind Radio und Fernsehen. Ein frühes Beispiel für ein prosodisch aufbereitetes Korpus aus dieser Quelle (das aus heutiger Sicht mit erstaunlich wenig technischer Hilfe erstellt wurde) stellt die Arbeit von Royé dar [Roy83]. Vorteil bei dieser Art der Datenerhebung ist neben dem geringen Aufwand die meist gute technische Aufnahmequalität (Raumakustik, Mikrophonqualität, Nebengeräusche), mitunter auch eine qualitativ hochwertige Übertragungstechnik etwa bei Ausnutzung des Digitalen Satellitenradio (DSR). Auch kann durch die Einseitigkeit des Übertragungskanals eine Beeinflussung des Sprechers weitgehend ausgeschlossen werden. Nachteilig sind die möglicherweise unnatürliche Kommunikationssituation und daß sich ein Teil der Signalerhebung der Einflußnahme entzieht. Beispielsweise läßt der Einsatz von Kompressoren und Klangfilterung in den Sendeanstalten Untersuchungen über Energie und frequenzabhängige Energieverteilungen zumindest fragwürdig erscheinen.

Weitere Quellen sind Literaturlesungen, die auf Audiocasetten, Compact Disk und CD-ROM erhältlich sind. Auch diese Aufnahmen sind meist von guter technischer Qualität und weisen eine für die Prosodieforschung interessante Vielfalt auf. Die Darbietung erfolgt meist durch Schauspieler. Mitunter ist die Sprache aufgrund der Entstehungszeit und Form des Werkes nur eingeschränkt mit Alltagssprache vergleichbar.

Bereits existierende Korpora für das Training von Spracherkennern können nur eingeschränkt für die Prosodieforschung und die korpusbasierte Sprachsynthese verwendet werden. Sie bestehen zum Teil aus isoliert gesprochenen Einzelwörtern, beispielsweise Ziffern, oder kurzen Sätzen. Für die Untersuchung der Prosodie sind jedoch größere Einheiten interessant, etwa vorgelesene Geschichten oder ganze Dialoge. Manche Spracherkennungskorpora sind für ganz spezifische Anwendungsgebiete erstellt worden und sind deshalb aus technischen Gründen (z. B. Sprache aufgenommen über schlechte Telefonverbindungen) oder aufgrund der unnatürlichen Kommunikationssituation (z. B. Einzelwortkommandos) für die Prosodieforschung ungeeignet.

Besonders interessant für die Untersuchung von Prosodie sind mehrfache Produktionen, entweder durch den gleichen Sprecher oder von verschiedenen Sprechern. An solchen wiederholten Äußerungen kann die Variabilität von Prosodie untersucht werden, also, welche Ereignisse zufällig, sprecherspezifisch oder linguistisch determiniert sind.

Neben den äußeren Kriterien wie technische Qualität, kommunikative Situation der Sprecher und Auswahl und Umfang des Sprachmaterials gibt es noch weitere Anforderungen, die an ein Korpus für die Prosodieforschung zu stellen sind:

- Sämtliche Annotationen sollen möglichst konsistent vorgenommen sein.
- Das Auffinden von Wortrealisierungen, auch in spezifischen linguistischen Kontexten, soll möglich sein.
- Alle automatisch meßbaren Daten sollen einer z. B. statistischen Analyse zugänglich sein.
- Alle automatischen und manuellen Annotationen sollen für die Analyse zur Einschränkung der Daten verwendet werden können.
- Das Auffinden von Realisierungen (lokaler) prosodischer Ereignisse soll möglich sein (zur Analyse von Parametern, Lernen für datengetriebene Synthese).
- Das Korpus soll es erlauben, globale prosodische Eigenschaften zu untersuchen (zur Theoriebildung durch Analyse von Parametern, Lernen für datengetriebene Synthese).
- Die Strukturen sollen erweiterbar sein (z.B. um Daten von neuen phonetischen Meßgeräten zu integrieren).

Im folgenden werden die Punkte nacheinander eingehender diskutiert. Die Forderung nach einer konsistenten Annotation konkurriert mit dem Wunsch, bereits geleistete Analysearbeit auszunutzen. Beispielsweise wird es nicht immer möglich sein, Etikettierungen von anderen Korpora zu übernehmen, wenn unterschiedliche Etikettierungsrichtlinien oder ein abweichendes Symbolinventar verwendet wurden. Dies hat zur Folge, daß gegebenenfalls Analysearbeit wiederholt werden muß oder nur Teile einer Analyse (z.B. Segmentierung in Wörter) benutzt werden können.

Um eine effiziente Aufbereitung nicht nur der (teil-)analysierten Daten sondern auch gänzlich unaufbereiteten Sprachmaterials zu gewährleisten, ist es wichtig, möglichst große Anteile der Analysearbeit zu automatisieren. Für manche Teilaufgaben existieren dazu bereits erprobte Verfahren, für andere ist die Forschung noch nicht so weit entwickelt, daß sie für die Zwecke der Korpuserschließung zufriedenstellende Lösungen bieten kann. Das angestrebte Vorgehen bei der Korpuserschließung muß sein, möglichst alle technologisch beherrschbaren Schritte automatisch, das heißt ohne menschliche Eingriffe, auszuführen. Diejenigen Teilaufgaben, für die die Forschung noch keine zufriedenstellenden Techniken hervorgebracht hat, müssen gezwungenermaßen weiterhin durch aufwendige menschliche Analysearbeit bewältigt werden. Diese menschliche Analysearbeit, die zum größten Teil Expertenwissen erfordert, ist der begrenzende Faktor, der die bewältigbare Korpusgröße bestimmt. Jede Aufgabe, die automatisch

bewältigt werden kann, verringert die Arbeitslast der Experten und hat damit eine Vergrößerung der analysierbaren Sprachdatenmenge zur Folge. Andererseits ist für Untersuchungen, die an einem Korpus durchgeführt werden, nicht nur der Umfang von Interesse, sondern auch die Qualität der Analyse. Häufig ist die Analyse, die von Experten geleistet wird, einer automatischen Analyse überlegen. *Eine automatische Analyse ist dann einzusetzen, wenn sie an die Leistungen von Experten heranreicht oder diese sogar übertrifft.* Auch bei Teilaufgaben, bei denen die Güte einer automatischen Analyse deutlich unter der eines Expertenurteils liegt, kann manchmal der Einsatz eines automatischen Verfahrens den Arbeitsaufwand des Experten verringern, ohne die Qualität der Analyse zu verschlechtern, indem nämlich automatisch eine Analyse erstellt wird, die der Experte dann kontrolliert und gegebenenfalls in Teilen ausbessert. Abhängig vom Umfang der erforderlichen Korrekturen, kann ein solches Vorgehen deutlich die Arbeitsbelastung des Experten vermindern.

Eine Anforderung an ein Prosodiekorpus ist, daß es den gezielten Zugriff auf einzelne Realisierungen eines Wortes erlauben soll. Die Erfüllung dieser Forderung ermöglicht es beispielsweise zu untersuchen, welche prosodische Realisierung Satzadverbien in bestimmten Kontexten aufweisen. Dazu müssen etwa alle Vorkommen des Wortes „erst“ in einer umfangreichen Sprachdatenbank aufgefunden werden. Ein Abhören der Aufnahmen ist ab einer gewissen Korpusgröße nicht mehr praktikabel. Es muß also einen Zugriffsmechanismus geben, der ein irgendwo in der großen Sprachsignalmenge geäußertes Wort findet. Dazu gibt es mindestens drei verschiedene Vorgehen:

1. Suche im Sprachsignal zur Zugriffszeit
2. Indexierung des Sprachsignals durch Wörter
3. Indexierung des Sprachsignals durch Phonemnetzwerke

Erstens kann man zur Zugriffszeit im Sprachsignal Stellen suchen, die zu einem gesuchten Wort passen. Die dazu notwendige Technik, das sogenannte keyword-spotting oder Schlüsselworterkennen, ist relativ gut entwickelt und wird zum Beispiel beim Überwachen von Telefonverbindungen eingesetzt. Der Nachteil dieses Verfahrens ist, daß das relativ aufwendige Suchen im Sprachsignal für jedes neue gesuchte Wort wiederholt werden muß, was ein für ein effizientes Arbeiten mit dem Korpus inakzeptables Antwortverhalten zur Folge hat. Auch werden vom keyword spotting nicht notwendigerweise alle Vorkommen eines Wortes gefunden (misses), oder es werden Sprachsignalabschnitte gefunden, die das gesuchte Wort gar nicht enthalten (false alarms). Vor der eigentlichen Analyse können

false alarms bei verhältnismäßig kleiner Trefferzahl auch bei größeren Korpora noch durch Abhören manuell ausgeschlossen werden. Nicht gefundene Vorkommen bleiben einer weiteren Analyse jedoch unzugänglich.

Zweitens ermöglicht die Indizierung des Sprachsignals durch Orthographie einen verhältnismäßig schnellen Zugriff, da wesentlich weniger Daten durchsucht werden müssen. Für sehr große Datenmengen können auch die effizienten Verfahren angepaßt werden, die beim Zugriff auf reine Textkorpora verwendet werden (z.B. Indizierung von Wörtern über Hashtabellen und Tries). Für sehr kurze Sprachaufnahmen, z. B. einzelne kurze Sätze, ist die reine Verschriftung der Sprachdaten unter Umständen ausreichend. Für längere Sprachaufnahmen müssen die Anfangs- und Endzeiten der gesprochenen Wörter bekannt sein, damit man die interessierenden Sprachsignalabschnitte leicht auffindet und sie auch einer weiteren automatischen Verarbeitung zuführen kann. Das Finden einer orthographischen Repräsentation zu einem Sprachsignal leistet die Spracherkennung.

Für das Datenmaterial, das für ein Prosodiekorpus wünschenswert ist, ist die Technologie der automatischen Spracherkennung noch nicht gut genug entwickelt. Es sollten keine Beschränkungen des Wortschatzes oder der Grammatik bestehen. Funktionierende Spracherkennungssysteme beschränken die Vielfalt der gesprochenen Sprache auf unterschiedliche Weise. Nahezu alle Spracherkennungssysteme nutzen die phonotaktische und lexikalische Beschränkung einer Sprache aus, indem sie Wortmodelle als grundlegende Einheiten verwenden, die ihrerseits jedoch, in einem Lexikon definiert, aus Phonemmodellen zusammengesetzt sind. Dieses Vorgehen hat gegenüber der Erkennung von beliebigen Phonemfolgen den Vorteil, daß phonotaktisch oder lexikalisch unmögliche Phonemfolgen bei der Erkennung ausgeschlossen sind. Gegenüber der Verwendung von Wortmodellen hat dieses Vorgehen den Vorteil, daß nicht im Trainingsmaterial vorhandene Wörter erkannt werden können und daß für einzelne Phoneme mehr Trainingsmaterial verfügbar ist als für einzelne Wörter. Desweiteren beschränken manche Systeme das Erkennungsproblem auf die Einzelworterkennung. Das bedeutet, daß der Sprecher zwischen zwei Wörtern jeweils eine kurze Pause lassen muß. Mit solchen Systemen kann unter Berücksichtigung von Wortabfolgewahrscheinlichkeiten (statistische Sprachmodelle) eine Worterkennungsrate von etwa 95% bei einem Wortschatz von mehreren 10 000 Wörtern erzielt werden. Die Forderung nach einer Pause zwischen den Wörtern verhindert jedoch den Einsatz dieses Verfahrens für die Aufbereitung von beliebigem Sprachmaterial. Andere Spracherkennungssysteme modellieren nur ein beschränktes Anwendungsgebiet, beispielsweise das Diktieren von radiologischen Befunden. Auch diese Systeme sind nicht direkt für die Aufbereitung eines von diesem Anwendungsgebiet abweichenden Prosodiekorpus geeignet. Beim Aufbereiten von unbeschränktem Audiomaterial muß immer wieder mit dem Auftreten von Wörtern gerechnet werden, die

nicht im Vokabular des Spracherkennungssystems enthalten sind, sogenannte Out-of-vocabulary-Wörter (OOV-Wörter). Für diese problematische Wortklasse muß Phonemerkennung verwendet werden, um eine Beschreibung des unbekannten Wortes zu erlangen. Das Zusammenspiel zwischen Wort- und Phonemerkennung erfordert eine sorgfältige Abstimmung, Verfahren zur Behandlung unbekannter Wörter und die Integration in Spracherkennungssysteme sind Gegenstand der aktuellen Forschung.

Wenn man ein weiteres schnelles Vorankommen auf dem Gebiet der Spracherkennungstechnologie bei großem Wortschatz unterstellt, bietet die Erkennung neben dem Zugriff auf Wortrealisierungen weitere Vorteile: Ein Zugriff auf einzelne Phrasen, Füße, Silben, Mora, Phoneme oder beliebige andere linguistische Einheiten kann bei Einsatz einer phonembasierten Modellierung bei der Spracherkennung leicht erreicht werden. Solange die Spracherkennung noch unbefriedigend ist, kann man das Korpus entweder von Erkennungsfehlern manuell bereinigen (bei guter Erkennung) oder gleich komplett die automatische Spracherkennung durch die menschliche Spracherkennung ersetzen. Diese sogenannte Transliteration muß nicht von phonetisch geschulten Experten durchgeführt werden, sondern kann, dazu vermutlich noch wesentlich effizienter, von Stenotypisten erledigt werden. Für die Ziele dieser Arbeit, die automatische Erstellung von Sprachkorpora für die Prosodieforschung, ist dieser Zustand natürlich unbefriedigend und nur als Übergangsstrategie akzeptabel, bis die Spracherkennungstechnologie auf einem ausreichenden Niveau angelangt ist. Bei Äußerungen, deren orthographische Beschreibung bekannt ist, kann die Spracherkennungstechnologie jedoch schon jetzt dazu verwendet werden, die zu den einzelnen Wörtern gehörenden Signalabschnitte aufzufinden. Das Vorgehen dazu wird in Kapitel 3 beschrieben. Da dieses Verfahren ohne menschliche Eingriffe funktioniert, erschließen sich gelesene Äußerungen einer Analyse auch ohne Transliteration, da ja der Text schon bekannt ist und geeignet aufbereitet werden kann. Besonders Literaturlesungen sind hier eine ergiebige Materialquelle, wenn der Text in maschinell verarbeitbarer Form vorliegt. Eine Transliteration wird jedoch bei spontansprachlichem Material sowie bei gelesenem Material, dessen Vorlage unzugänglich ist (z.B. Radionachrichten), weiterhin notwendig sein. Für mehrfach auftretende Äußerungen wie die Meldungen in Nachrichtensendungen beschreibt Abschnitt 7.1 ein Verfahren, das den Transliterationsaufwand minimiert.

Die dritte Möglichkeit zur Indizierung der Datenbasis, nämlich durch Verwendung von Phonemhypothesennetzwerken, ist eine recht junge Technik, die eine zur Anfragezeit schnellere Schlüsselworterkennung ermöglicht. Dazu wird während der Datenaufbereitung mit Hilfe eines Phonemerkenners für die Sprachdaten ein Hypothesennetzwerk aus Phonemen erstellt. In diesen Phonemnetzwerken wird dann zur Anfragezeit nach Schlüsselwörtern gesucht. Es handelt sich also gewisserma-

ßen um eine Zwischenstrategie zwischen vollständiger Erkennung zur Aufbereitungszeit und vollständiger Schlüsselworterkennung zur Anfragezeit. Durch die grundsätzliche Beschränkung auf einen Phonemerkenner zur Analysezeit vermeidet man das Out-of-Vocabulary-Problem, da man den lexikalischen Zugriff bis zur Suchanfrage aufschiebt. Durch Verwendung von Hypothesennetzen möchte man verhindern, daß Fehlerkennungen des Phonemerkenners zu einem Fehlschlagen der Schlüsselwortsuche führt. Die Indizierung durch Phonemnetze kann man also einerseits als Erkennung zur Aufbereitungszeit betrachten, bei der der Lexikonzugriff erst zur Anfrage erfolgt, andererseits als Schlüsselwortsuche, bei der ein Teil bereits zur Aufbereitungszeit durchgeführt worden ist.

Wenn man nur an einem Auffinden von Wortrealisierungen interessiert ist und das Nichtfinden einzelner Realisierungen toleriert, so ist die Schlüsselworterkennung unter Zuhilfenahme von zur Aufbereitungszeit berechneten Phonemhypothesennetzen die Methode der Wahl, da sie ohne Transliteration auskommt und dennoch kurze Anfragezeiten ermöglicht. Die Methode versagt jedoch, wenn man gleichzeitig Informationen über unterhalb der Wortebene liegende Einheiten gewinnen will, beispielsweise Informationen über alle unbetonten Silben, alle Phonemrealisierungen in einem bestimmten Lautkontext etc., da durch die Verwendung von Phonemhypothesennetzwerken noch nicht über die lexikalische Phonem- oder Silbenstruktur entschieden ist. Für einen uneingeschränkten Zugriff auf Wortrealisierungen und Realisierungen kleinerer Einheiten ist eine Erkennung zur Aufbereitungszeit die geeignetste Methode, auch wenn sie für manches Material zum derzeitigen Forschungsstand der Spracherkennung menschliche Hilfestellung verlangt. Durch Auswahl des Sprachmaterials (gelesenes Material oder spontansprachliches Material, das schon transliteriert ist) kann diesem Nachteil entgegengewirkt werden.

Eine weitere Anforderung an ein Prosodiekorpus ist die Möglichkeit zur Auffindung lokaler prosodischer Ereignisse. Auch hier können unterschiedliche Suchstrategien betrieben werden. Wieder wurde in dieser Arbeit zugunsten einer Analyse zur Aufbereitungszeit entschieden, indem prosodische Ereignisse in der Aufbereitungsphase im Korpus annotiert werden. Die Annotation sollte nicht einer proprietären, sondern einer unabhängig entwickelten allgemein einsetzbaren Systematik folgen, um sicherzustellen, daß der Austausch von Ergebnissen mit anderen Forschergruppen möglich und die Beschreibung für die untersuchte Sprache auch auf längere Sicht adäquat ist. Die vorliegende Arbeit verwendet als Beschreibung lokaler prosodischer Ereignisse die phonologische Intonationsbeschreibung nach dem Stuttgarter System [May95], das eine Umsetzung der von Féry in [Fér93] vorgelegten phonologischen Analyse der Intonation des Deutschen nach dem Tonsequenzmodell [Pie80] darstellt. Kapitel 2 stellt die grundlegenden Konzepte des Intonationsmodells kurz vor. Es gibt zwei Arten lokaler

prosodischer Ereignisse: Prominenz, also die akustische Hervorhebung von Teilen einer Äußerung, und Phrasierung, also die Gliederung einer Äußerung. Beide müssen in einem Korpus für die Prosodieforschung repräsentiert sein. Um die Korpuserstellung zu automatisieren, müssen auch die Beschreibungen für die lokalen prosodischen Ereignisse automatisch bestimmt werden. Da das manuelle Erstellen einer solchen prosodischen Beschreibung ein sehr aufwendiger und Expertenwissen erfordernder Prozeß ist, bildet seine Automatisierung die wichtigste Teilaufgabe und eigentliche Herausforderung bei der automatisierten Erstellung von Prosodiekorpora. Sie wird hier dadurch gelöst, indem zunächst aus den Sprachdaten automatisch Merkmale extrahiert werden, die mit den Prosodiekorrelaten zeitliche Strukturierung, Grundfrequenz(bewegungen), Intensität und Stimmqualität zusammenhängen. Die entwickelten Extraktionsverfahren sind den Korrelaten entsprechend in den Kapiteln 3 bis 5 beschrieben. Die Merkmalsextraktionsverfahren stellen eine phonetische Prosodieanalyse dar, als deren Ergebnis Parameter resultieren, die zumeist phonetisch interpretierbare, akustisch-physikalische Meßwerte beschreiben. Sie bilden die Eingabe für die eigentliche Prosodieerkennung, die in Kapitel 6 beschrieben ist. Diese phonologische Komponente des Analysesystems leitet aus den phonetischen Parametern mit Silben assoziierte Symbole ab, welche die Intonation auf einer abstrakteren Ebene beschreiben.

Die Extraktion von phonetischen Parametern ist heute — nicht nur bei der Analyse der Prosodie, wie sie in dieser Arbeit beschrieben wird — zumeist technologisch beherrscht, also mit ausreichender Genauigkeit automatisch möglich. Generell gilt, daß die Ableitung rein akustisch-physikalischer Korrelate aus dem Sprachsignal algorithmisch recht gut gelingt. Die Parameter, die üblicherweise in der Phonetik behandelt werden, sind allerdings häufig bereits auf einem abstrakteren Niveau als dem rein akustisch-physikalischen und algorithmisch mitunter schwerer zu erfassen. Beispielsweise ist die Bestimmung von Formantenkandidaten technisch machbar, das Problem ist hinreichend klar spezifiziert und läßt sich algorithmisch lösen, indem zu beliebigen Zeitpunkten die n Pole einer LPC-Analyse $n$ter Ordnung aus einem kurzen, den Zeitpunkt umgebenden Signalabschnitt berechnet werden. Zum Beispiel werden alle 10 ms acht Pole und somit acht Formantenkandidaten ausgegeben. Die Auswahl unter den Kandidaten, also die Bestimmung der perzeptiv zur Unterscheidung von Phonemen wichtigen ersten, zweiten und dritten Formanten, ist jedoch fehlerträchtig und eine automatische Bestimmung ist schwieriger, da die Aufgabe nicht mehr so leicht objektiv zu spezifizieren ist. Die Auswahl für die ersten 3 Formanten der ersten Messung könnte beispielsweise auf den ersten, dritten und vierten Pol der LPC-Analyse fallen. Bei der 10 ms später auftretenden zweiten Messung könnte die Auswahl dagegen den zweiten, dritten und fünften Pol als Meßwerte für die ersten drei For-

manten ergeben. Diese Wechsel sind zum einen durch die Variabilität der Sprache begründet, liegen zum anderen aber auch daran, daß manche Konzepte der Phonetik an die Artikulation angelehnt sind und somit nicht immer unmittelbar aus dem von Mikrophonen aufgefangenen Sprachschall algorithmisch „bottom up“ abgeleitet werden können. Vielmehr schränkt die Kenntnis der artikulatorischen Vorgänge „top down“ die Plausibilität von Meßwerten ein. Das stellenweise Verengen und Erweitern des Mund-Rachenraumes geschieht nicht beliebig schnell und somit können die durch die Mund-Rachenraumkonstellation hervorgerufenen Formanten sich nicht abrupt ändern. Für die Auswahl der Formantenkandidaten ergeben sich dadurch kontextuelle Abhängigkeiten, die eine exakte algorithmische Beschreibung der Auswahl unter den Kandidaten erschweren. Obgleich schon die phonetischen Parameter, im Beispiel die Formantenlage bzw. Formantenbewegungen, eine gewisse Abstraktion vom Sprachsignal erfahren haben, kann auf eine noch abstraktere Repräsentationsebene geschlossen werden, und zwar die der Phonologie. Aus der Information über die Formantenlage und -bewegung (sowie weitere phonetische Parameter) läßt sich schließen, welchem Phonem das betrachtete Sprachsignal zuzuordnen ist.

Ähnlich ist die Situation bei der Sprachgrundfrequenzanalyse wie sie in Kapitel 4 beschrieben wird. Die Analyse der Sprachgrundfrequenz durch algorithmische Grundfrequenzschätzer (Pitch tracker) ist relativ gut beherrscht [Hes83]. Sie liefert zum Beispiel alle 10 ms einen Grundfrequenzwert in Hz. Dieser Meßwert spiegelt nicht notwendigerweise wieder, mit welcher Frequenz die Stimmlippen periodisch auf- und zugeschlagen sind (artikulatorisches Korrelat), geschweige denn mit welcher Höhe der Sprecher das betrachtete Stückchen Sprache produzieren wollte (Funktion der Sprache), vielmehr ist der „gemessene“ Wert nicht mehr als eine von dem verwendeten Grundfrequenzbestimmungsalgorithmus ausschließlich aus dem betrachteten Sprachsignal berechnete Schätzung. Die in der Prosodieforschung verwendete phonetische Beschreibung der Grundfrequenz befaßt sich jedoch nicht mit einzelnen Grundfrequenzschätzungen sondern bereits mit einer Abstraktion davon. Beispielsweise interessiert ein starker Anstieg der Grundfrequenz oder ein gegenüber der Umgebung erhöhtes Niveau. Ausgehend von den Grundfrequenzniveaus und Grundfrequenzbewegungen wird auf der nächsten Abstraktionsebene auf die Zugehörigkeit zu einer phonologischen Klasse entschieden, beispielsweise ein steigender Satzakzent oder eine hohe Phrasengrenze. Im hier entwickelten System bewerkstelligt diesen Schritt von der „Phonetik der Prosodie“ zum phonologischen Intonationsmodell ein Verfahren des maschinellen Lernens. Das Lernverfahren lernt aus einer von menschlichen Experten aufbereiteten Stichprobe, die Beziehungen zwischen phonetischen Parametern und phonologischen Klassen. Diese gelernten Beziehungen können dann auf prosodisch unaufbereitetes Material angewandt werden und realisieren so eine auto-

matische prosodische Analyse, die sehr viel umfangreichere Datenmengen verarbeiten kann als in vertretbarer Zeit von Menschenhand zu bewältigen wären.

Das in dieser Arbeit vorgestellte System verwendet weder bei der phonetischen Analyse noch bei der phonologischen Intonationsbeschreibung proprietäre Datenstrukturen, so daß es einfach möglich ist, die Ergebnisse weiterer phonetischer Meßinstrumente hinzuzufügen. Die Messungen lassen sich kombinieren und entweder in ein Datenbankmanagementsystem (DBMS) importieren oder direkt mit UNIX-Tools analysieren. So können prosodische Realisierungen von Wortarten und Wörtern, prosodische Ereignisse, ausgesuchte Phonemkonstellationen etc. in großen Datenmengen gesucht und geeignet eingeschränkt werden z.B. bezüglich der Ausprägung eines bestimmten phonetischen Parameters. Das Ergebnis der Suche kann dann entweder zur manuellen Inspektion visualisiert oder aber statistisch analysiert werden.

Die vorliegende Arbeit beschreibt eine Implementation der vorgestellten Verfahren zur Aufbereitung und prosodischen Analyse von umfangreichen Sprachdatenmengen. Obwohl die vorgestellten Verfahren prinzipiell nicht nur für eine Einzelsprache wirksam sind, wurden doch manche Parameter und Komponenten ans Deutsche angepaßt. Das in Kapitel 3 vorgestellte Segmentierungssystem Alphons ist nur für deutsche Äußerungen sinnvoll einsetzbar, da das angeschlossene Lexikon nur Einträge dieser Sprache umfaßt, auch die enthaltene regelgesteuerte Graphemphonemkonvertierung ist speziell für das Deutsche entwickelt worden. Schließlich umfassen die HMM nur den Phonemsatz des Deutschen, da sie an ausschließlich deutschsprachigem Material trainiert wurden. Bei der Grundfrequenzparametrisierung wurde die Größe des jeweils betrachteten Fensters auf zwei Silben festgelegt, eine Entscheidung, die nicht für alle Sprachen sinnvoll sein muß. Schließlich ist das eingesetzte phonologische Intonationssystem ausschließlich für das Deutsche entwickelt. Da die Beziehungen zwischen phonetischen Parametern und phonologischen Klassen automatisch von bereitzustellenden Trainingsdaten gelernt werden, ist jedoch auch eine Übertragung von Teilen oder des vollen Systems auf andere Sprachen und/oder andere Intonationssysteme möglich. Es ist jedoch offensichtlich, daß etliche der in dieser Arbeit angestellten Untersuchungen wiederholt werden müssen, um für die neue Konstellation eine adäquate Parameterauswahl zu etablieren.

Neben der eigentlichen Aufbereitung der Sprachdaten, welche die Segmentierung sowie die phonetische und phonologische Prosodiebeschreibung umfaßt, sind für eine effiziente Erstellung und einen effizienten Umgang mit dem aufbereiteten Material weitere Komponenten nützlich, die ebenfalls im Zuge dieser Arbeit entstanden und in Kapitel 7 beschrieben sind. Abschnitt 7.1 beschreibt ein Verfahren, das es erlaubt, wiederholt verlesene Radionachrichtenmeldungen bereits vor

einer Transliteration automatisch zu detektieren. Das Verfahren wird eingesetzt, um automatisch aufgezeichnete Nachrichtensendungen eines Rundfunksenders automatisch dahingehend zu analysieren, daß wiederholt verlesene Nachrichtenmeldungen zu Clustern zusammengefaßt werden. Die in einem Cluster enthaltenen Nachrichtenmeldungen unterscheiden sich nur minimal, idealerweise sind sie mehrfach identisch verlesene Meldungen. Diese Cluster sind aus Forschungssicht interressant, da sie die prosodische Variation unter Beibehaltung des semantischen und syntaktischen Settings offenbaren. Außerdem reduziert das Zusammenfassen erheblich den Aufwand bei der Transliteration, da pro transliterierter Meldung gleich mehrere (typisch 3-5) Sprachsignale automatisch weiteranalysiert werden können.

In Abschnitt 7.2 wird ein Werkzeug vorgestellt, das eine Übersicht über die phonologische Intonationsbeschreibung mehrerer Sprachsignale erzeugt. Ausgehend von Wortsegmentierung und phonologischer Intonationsbeschreibung, die beide als Labeldateien vorliegen, erzeugt das Werkzeug eine textuelle Zusammenfassung der prosodischen Analyse, indem es zeilenweise die Orthographie mit darunterstehender Intonationsbeschreibung der einzelnen Realisierungen kombiniert.

Nach einer Zusammenfassung der Ergebnisse dieser Arbeit in Kapitel 8 bietet der Anhang eine Übersicht über das mit den implementierten Verfahren analysierte, insgesamt etwa 12 Stunden umfassende Sprachmaterial.

Kapitel 2

Eine phonologische Intonationsbeschreibung für das Deutsche

Es gibt zahlreiche Möglichkeiten, die Intonation gesprochener Sprache zu beschreiben. Eine Dimension, in der sich Beschreibungsformen für Intonation unterscheiden, betrifft die Frage, was unter Intonation zu verstehen ist, das heißt, welche akustisch-physikalischen Größen von der Intonation kontrolliert werden. Intonation kann einerseits lediglich die Grundfrequenz bzw. die wahrgenommene Tonhöhe (Pitch) beschreiben oder andererseits zusätzliche Faktoren wie Lautstärke oder zeitlich-rhythmische Aspekte umfassen. In dieser Arbeit wird unter Intonation die Hervorhebung von Teilen einer Äußerung (Prominenz) sowie die Gliederung einer Äußerung (Phrasierung) verstanden, die dem Hörer durch verschiedene akustische Hinweise, den akustischen Korrelaten der Intonation, übermittelt werden. Als wichtigste Korrelate der Intonation sind zu nennen: die Tonhöhe bzw. als ihr meßbares Korrelat die Sprachgrundfrequenz, die Intensität, bzw. als ihr meßbares Korrelat die Lautstärke, die Längung oder Kürzung von Lauten, Silben, Äußerungsteilen und Sprechpausen, sowie Änderungen der Stimmqualität. Im allgemeinen wird der Tonhöhe die wichtigste Rolle bei der Markierung von Prominenz und Phrasierung zuerkannt, jedoch kann die Wichtigkeit der Korrelate von Sprache zu Sprache unterschiedlich sein, und möglicherweise sogar von Sprecher zu Sprecher wechseln.

Die andere Dimension, in der sich die Beschreibungsformen unterscheiden, ist der Abstraktionsgrad. Eine sehr zuverlässige Beschreibung der akustischen Korrelate der Intonation kann erreicht werden, wenn die Korrelate einzeln nahe am

aufgenommenen Mikrophonsignal beschrieben werden, beispielsweise durch Angabe der Sprachgrundfrequenz in Hz, oder durch Angabe der Lautstärke in dB. Für eine kontinuierliche Bereitstellung dieser Meßwerte gibt es recht genau und zuverlässig arbeitende Algorithmen. Am anderen Pol der Abstraktionsdimension steht eine Beschreibungsform, die die Information darstellt, die durch die Mittel der Intonation mit all ihren Korrelaten vom Sprecher an den Hörer übermittelt werden sollte.

Damit eine so abstrakte Beschreibung der Intonation nützlich ist, ist Wissen darüber nötig, wie sich einerseits die Beschreibung aus einer gesprochenen Äußerung ableiten läßt, und andererseits, wie aus der abstrakten Formulierung eine entsprechende Äußerung zu generieren ist. Zum derzeitigen Stand der Forschung ist jedoch noch nicht ausreichend verstanden, welche Kategorien eine so abstrakte Beschreibung besitzen sollte und wie die Umsetzung von und zur gesprochenen Sprache vonstatten gehen soll. Eine solche Beschreibung könnte als das Ziel der Intonationsforschung angesehen werden, zusammen mit einem kompletten Verständnis der Prozesse, die die Beschreibung mit dem Sprachsignal verbinden. Das unzureichende Verständnis dieser Prozesse ist der eigentliche Grund für die Notwendigkeit von Intonationsbeschreibungssystemen. Um einen Erkenntnisgewinn für die fehlenden Teile der beiden Prozesse zu erhalten, muß klar sein, wie die Daten beschrieben werden können.

Das Problem bei einer sehr signalnahen Beschreibung ist, daß der Erkenntnisgewinn durch die Vielzahl an Daten erschwert oder gar verhindert wird. Eine für den derzeitigen Forschungsstand sinnvolle Beschreibung der Intonation muß eine weit abstraktere Ebene annehmen, damit sie zur Überbrückung der noch unverstandenen Teile der Prozesse beitragen kann. Geeignet dafür ist die Ebene der Phonetik/Phonologie, auf der die im folgenden beschriebenen Systemen anzusiedeln sind.

Die Beschreibungsform, die in letzter Zeit die breiteste Unterstützung erzielte, ist sicherlich das ToBI System (Tones and Break Indizes) von Pierrehumbert und Beckman [Pie80, BP86, PB88, SBP+92]. Das ToBI-System wurde zur Beschreibung der Intonation des Englischen entwickelt, kann aber mit Einschränkungen auch zur Beschreibung anderer Sprachen eingesetzt werden. Das sprachenunabhängige Grundkonzept von ToBI ist das Tonsequenzmodell (TSM) [Pie80], das die Intonation als diskrete Abfolge atomarer Töne, meist H (für engl. high) und L (für engl. low), darstellt. Eingebettet ist das TSM in die autosegmentale Phonologie [Gol90]. Ebenfalls auf dem TSM beruht die Analyse von Féry für das Deutsche [Fér93]. Férys System verwendet ein anderes Symbolinventar als das englische ToBI. Ein weiterer wesentlicher Unterschied ist, daß Féry zwei Ebenen der Intonationsbeschreibung einführt, eine Oberflächen- und eine Tiefenstruktur,

die eine kompositionelle Analyse vereinfachen[May97]. Betrachtet man das System nur bezüglich der Oberflächenstruktur, so ist es mit dem bekannten ToBI-System in etwa vergleichbar. Für das Verständnis dieser Arbeit genügt es darum, eine der TSM-basierten Beschreibungsarten zu kennen. Da das implementierte System mit deutschen Sprachdaten operiert und die deutsche Intonationsbeschreibung vorhersagt, werden im folgenden das Intonationsmodell von Féry [Fér93] und die in Stuttgart entwickelten Erweiterungen [May95] vorgestellt. Zunächst wird das Inventar an Pitchakzenten eingeführt, die der Prominenzbeschreibung dienen. Dann folgt das Inventar an Grenztönen zur Beschreibung der Phrasierung. Als nächstes werden dann die Tiefen- und Oberflächenstruktur behandelt sowie die Linkingregeln, die die beiden Ebenen miteinander verbinden. Schließlich werden das Diakritikum für „Downstep“ und erlaubte Register beschrieben, die einige quantitative Aspekte der Intonation kategorisieren. Eine vollständigere Beschreibung mit Beispielen für die einzelnen Ettikettierungssymbole des prosodischen Transkriptionssystems gibt [May95]. Zu den phonologischen Grundlagen und zur Abgrenzung vom englischen System sei nochmals auf [Fér93] verwiesen.

2.1 Inventar an Akzenttönen

Im Féryschen Modell gibt es in der Tiefenstruktur fünf grundlegende Akzente, denen an der Oberflächenstruktur sieben Akzente und zwei abgespaltene Trailtöne entsprechen. Von den fünf Grundakzenten sind drei in ihrem Gebrauch stark eingeschränkt, so daß sich das Inventar im wesentlichen auf zwei Grundakzente beschränkt. Sie werden geschrieben:

L*H für einen steigenden Akzent,

H*L für einen fallenden Akzent.

Die Notation spiegelt gleichzeitig wider, wie die Akzente phonetisch realisiert werden. In dieser Notation entspricht der Stern * einer Markierung der betonten Silbe, er wird hinter der tonalen Spezifikation angegeben. Ein steigender Akzent L*H ist somit eine tiefe betonte Silbe gefolgt von einer hohen Silbe, ein fallender Akzent H*L eine hohe betonte Silbe gefolgt von einer tiefen Silbe. Die tatsächliche phonetische Realisierung kann dabei durchaus variieren. So kann etwa bei einem steigenden Akzent der Anstieg vollständig in der betonten Silbe liegen oder aber erst in der auf die betonte Silbe folgenden Silbe auftreten. Für die phonologische Beschreibung als steigenden Akzent hat dieser Unterschied keine Bedeutung. Das phonologische Symbol L*H ist dabei als ganzes mit einer Silbe (der betonten

Silbe) assoziiert. Das TSM beruht auf der autosegmentalen Phonologie. Das von dort bekannte Prinzip des Spreading (Spreizung) kann auch hier angewandt werden: Folgen beispielsweise auf eine mit einem fallenden Pitchakzent H*L assoziierte Silbe noch weitere unbetonte Silben, also Silben, die bezüglich ihrer tonalen Ausprägung nicht spezifiziert sind, so breitet sich der Trailton, das ist der letzte in einem Akzent spezifizierte Ton, aus und belegt die tonale Spezifikation der folgenden Silben (in diesem Fall mit einem L-Ton). Der gesternte Ton kann (mit einer Ausnahme, s.u.) nicht gespreizt werden, er dominiert stets nur eine Silbe. In einer Konfiguration, bei der mehrere mit H*L assoziierte Silben durch dazwischenliegende unbetonte Silben verbunden sind, ergibt sich durch das Spreading des Trailtones des jeweils vorhergehenden Akzentes, daß die akzentuierten Silben als H, ihre Umgebung als L spezifiziert sind. In solchen Konfigurationen ist in der Grundfrequenzmessung ein Gipfel über der betonten Silbe zu sehen, der fallende Akzent H*L kann also manchmal auch (in Abhängigkeit vom Vorkontext) als Gipfelakzent bezeichnet werden. Die drei seltener auftretenden Grundakzente sind:

L*HL für den steigend-fallenden Akzent,

HH*L für den früh fallenden Akzent,

H*M für den stilisierten Akzent.

Unter den drei selteneren Akzenten ist der L*HL am häufigsten. Eine tiefe betonte Silbe wird gefolgt von einer hohen Silbe, die ihrerseits wieder von einer tiefen Silbe gefolgt wird. Der Trailton L dominiert aufgrund des Spreading-Prinzips auch weitere folgende unbetonte Silben. Der steigend-fallende Akzent kann auch als Später-Gipfel-Akzent bezeichnet werden, denn sein Gipfel wird später erreicht als bei einem H*L mit vorausgehendem L-Niveau. Auch wenn die Bezeichnung eine phonetische Realisierung in drei Silben nahelegt, so ist die Realisierung meist schon in der zweiten Silbe vollständig abgeschlossen.

Der HH*L-Akzent setzt eine metrisch schwache vorangehende Silbe voraus. Charakterisiert wird der HH*L durch einen Tonabfall, der im allgemeinen bereits vor der betonten Silbe beginnt und sich dann bis in die Silbe fortsetzt, die auf die betonte Silbe folgt. Meist ist die metrisch schwache Silbe etwas höher als die betonte, die jedoch trotzdem noch mit einem hohen Ton assoziiert ist. Der Abfall findet im allgemeinen gegenüber einem H*L-Akzent früher statt, so daß dieser Akzent auch als Früher-Gipfel-Akzent bezeichnet werden kann.

Der H*M-Akzent tritt im Vergleich zu den anderen beiden weniger häufigen Akzenten noch seltener auf, da sein Gebrauch pragmatisch sehr stark eingeschränkt

ist. Einzig als „Vokativ-“ oder „Rufakzent“, wenn man jemanden beim Namen ruft, ist er häufig zu beobachten. Das sonst nicht auftretende M (für engl. mid) in seiner Bezeichnung ist nicht die einzige Anomalie des Akzents. Die weitere Besonderheit betrifft das Spreading. Bei diesem Akzent spreizt sich nicht wie sonst der Trailton, sondern der gesternte Ton. Für einen H*M-Akzent, der mit der (penulten) Wortakzentsilbe von „Irene!“ assoziiert ist, ergeben sich L H M als Assoziationen mit der ersten, zweiten bzw. dritten Silbe. Für „Antonia!“ (viersilbig gesprochen, mit antepenulter Betonung) ergibt das Spreading L H H M als Assoziationen zu den Silben und nicht, wie sonst zu erwarten, L H M M. Aufgrund seines seltenen Auftretens und den klar definierten Anwendungsfällen ist der H*M-Akzent mit seinen Besonderheiten für diese Arbeit unbedeutend.

Aus Gründen der Vollständigkeit werden noch die beiden nur in der Oberflächenstruktur möglichen Akzente und die beiden abgespaltenen Trailtöne aufgezählt:

L* für den tiefen Akzent,

H* für den hohen Akzent,

..H für den abgespaltenen hohen Trailton,

..L für den abgespaltenen tiefen Trailton.

Sie werden im Zusammenhang mit den Linkingregeln in Abschnitt 2.3 erklärt.

2.2 Inventar an Grenztönen

Grenztöne markieren das Ende oder den Anfang eines prosodisch zusammenhängenden Äußerungsteiles, die sogenannte Intonationsphrase (IP). Äußerungen sind aus unmittelbar aufeinanderfolgenden Intonationsphrasen aufgebaut; Intonationsphrasen können nicht ineinander enthalten sein; es gibt keine Struktur, die sich über die Intonationsphrasen hinweg aufbaut. Zur Kennzeichnung von Intonationsphrasen dienen im von Féry vorgeschlagenen System die drei Symbole %H, % und H%. Für das im Anhang A näher beschriebene, in Stuttgart manuell annotierte Radionachrichtenmaterial wurden zwei weitere Symbole hinzugefügt, – und L%, die einen leichteren Austausch mit anderen Forschergruppen ermöglichen sollten. Durch das Hinzufügen des Symbols – rückt das System näher an das englische ToBI-System heran, da dieses Symbol auch dort die Abgrenzung zweier intermediärer Intonationsphrasen beschreibt, die im Féryschen System implizit angenommen und nicht explizit gekennzeichnet werden. Analog zum Aufbau

von Äußerungen aus nicht hierarchisch strukturierten Intonationsphrasen, sind im englischen ToBI-System und in der Stuttgarter Ettikettierung Intonationsphrasen ihrerseits aus nicht hierarchisch strukturierten intermediären Intonationsphrasen (ip) aufgebaut. Anders betrachtet kennzeichnen die ein % enthaltenden Symbole eine starke Grenze, das – kennzeichnet eine schwache Grenze. Jede ip (und damit auch jede IP) enthält mindestens einen Akzent. Als Markierungen für Intonationsphrasen stehen damit zur Verfügung:

%H für eine hoch beginnende IP,

% für ein tonal nicht spezifiziertes Ende einer IP,

H% für ein (weiter) ansteigendes Ende einer IP,

L% für ein (weiter) abfallendes Ende einer IP,

– für das Ende einer ip.

Es gibt kein Symbol für eine tief beginnende IP. Diese ist der Regelfall und braucht deshalb nicht ausdrücklich gekennzeichnet zu werden. Das nicht tonal spezifizierte Ende einer IP, %, erhält seinen phonetischen Wert durch Spreading des vorangehenden Trailtones. Eine Konstellation H*L % ergibt nach dem Abfall des Akzentes aufgrund des gespreizten L-Tones ein gleichbleibend niedriges Grundfrequenzniveau. Geht dem % hingegen ein L*H-Akzent voraus, so ist auch das Ende der IP hoch, ebenso wie die zwischen dem Akzent und dem Grenzton liegenden Silben. Ein mit H% gekennzeichnetes Ende einer IP besitzt einen deutlichen Anstieg. Ist das vorangehende Niveau aufgrund eines gespreizten H-Tones bereits hoch, so muß ein weiterer Anstieg erfolgen. Ein mit L% gekennzeichnetes Ende einer IP ist durch einen deutlichen Abfall charakterisiert. In der Regel geht dem L% ein (zu spreizender) L-Trailton voraus. Das vorangehende Niveau muß weiter deutlich abfallen. Das – kennzeichnet eine schwache Grenze, die intermediäre Phrasen (ip), aber keine Intonationsphrasen (IP) trennt. Häufig erkennt man die intermediäre Phrasengrenze an einem abrupten Wechsel der Tonhöhe.

2.3 Oberflächen- und Tiefenstuktur, Linking

Wie bereits erwähnt, ist eine von Féry vorgeschlagene Neuerung die Einführung zweier Ebenen in die Intonationsbeschreibung. Die Oberflächenstruktur beschreibt die sichtbare, phonetisch realisierte Struktur, die sich aufgrund sogenannter Linkingregeln aus der Tiefenstruktur ableitet. Aus einer Tiefenstruktur können,

beeinflußt beispielsweise vom Sprechtempo, verschiedene Oberflächenstrukturen erzeugt werden. Diese verschiedenen Strukturen besitzen dennoch die gleiche Interpretation.

Es gibt zwei Linkingregeln, die nur auf die Trailtöne der beiden Akzente L*H und H*L wirken können: partielles Linking und komplettes Linking. Als weitere Einschränkung kann Linking nicht auf den letzten Akzent einer ip (nuklearer Akzent) angewandt werden. Damit kann sich Linking nicht über eine Phrasengrenze hinweg auswirken. Die beiden Regeln werden in der [May97] entnommenen Abbildung 2.1 an einem Beispiel gezeigt, das in der Tiefenstruktur zwei steigende Akzente enthält. Durch Anwenden keiner oder einer der beiden Regeln können drei Oberflächenstrukturen erzeugt werden, die alle dieselbe Interpretation besitzen, jedoch phonetisch sehr unterschiedlich realisiert sind. Die Regeln lassen sich unter Berücksichtigung der oben genannten Einschränkungen analog auch bei anderen Konstellationen anwenden. Insbesondere muß der zweite Akzent nicht notwendigerweise ein L*H- oder H*L-Akzent sein.

Beim partiellen Linking (Abb. 2.1 (b)) wird der Trailton vom Akzent abgespaltet und wandert bis an die unmittelbar vor dem nächsten Akzent liegende tonal unspezifizierte Silbe. Zurück bleibt der gesternte Ton. Da dieser nicht gespreizt werden darf, bleiben die dazwischenliegenden Silben tonal unterspezifiziert und werden phonetisch als allmählicher Übergang vom Niveau des gesternten Tones zum Niveau des Trailtones realisiert.

Beim kompletten Linking (Abb. 2.1 (c)) wird der Trailton ebenfalls abgespaltet. Der Unterschied zum partiellen Linking besteht darin, daß der Trailton mit dem ersten Ton des folgenden Akzentes verschmilzt (bei gleicher tonaler Kategorie) oder ganz wegfällt (bei unterschiedlicher tonaler Kategorie wie im Beispiel in Abbildung 2.1). Die dazwischenliegenden Silben bleiben unterspezifiziert und werden wie im letzten Fall ebenfalls als allmählicher Übergang realisiert.

Zur Annotation der Oberflächenstruktur werden die oben bereits aufgezählten zusätzlichen Etiketten L*, ..H, H* und ..L benötigt.

2.4 Downstep und Register

Die bis hierher vorgestellte Intonationsbeschreibung ist im wesentlichen qualitativ und zunächst nicht quantitativ. Es gibt im Stuttgarter Intonationsbeschreibungssystem zwei Arten, quantitative Aspekte zu kodieren, die dennoch kategoriell aufzufassen sind: Downstep und Register.

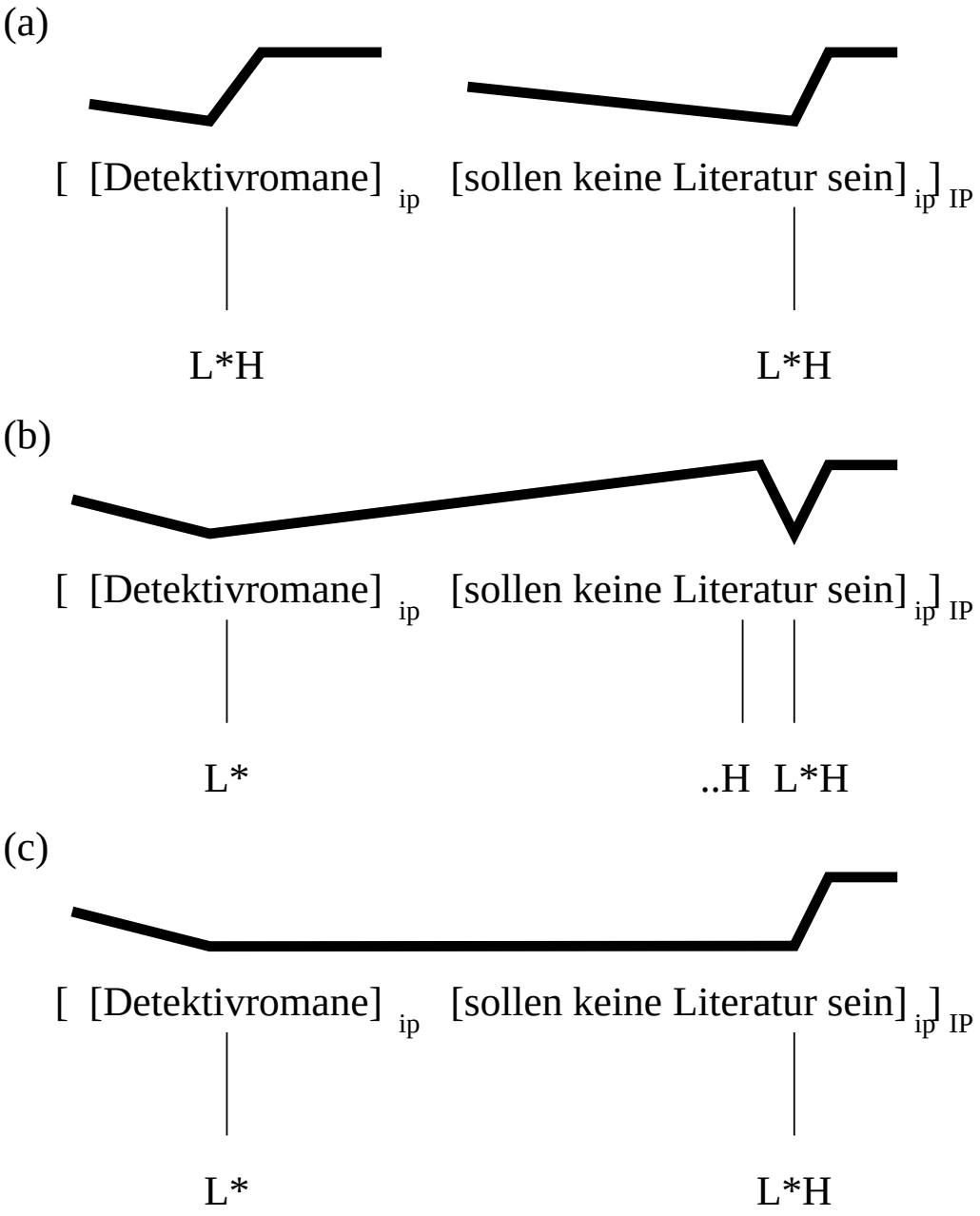


Abbildung 2.1: Die drei möglichen Oberflächenstrukturen zweier L*H-Akzente in der Tiefenstruktur: (a) ohne Anwendung von Linking, (b) nach Anwendung von partiellem Linking und (c) nach Anwendung von komplettem Linking.

Downstep ist ein Diakritikum, das auf alle H-Töne mit Ausnahme des ersten H-Tones einer ip angewandt werden darf und das Niveau der H-Töne bis zum Ende der aktuellen IP verringert. Es wird mit einem Ausrufezeichen vor der tonalen Spezifikation notiert. Auch eine Anwendung des Diakritikums auf nachfolgende Akzente der gleichen ip ist möglich, in diesem Fall wird das Niveau der folgenden H-Töne weiter reduziert. Ein Downstep von L-Tönen ist prinzipiell ausgeschlossen.

Register ist ein Merkmal der intermediären Phrase und wirkt auf alle tonalen Spezifikationen der darin enthaltenen Akzente und Grenztöne. Das Register ändert lokal die Stimmlage, also das Niveau von H- und L-Tönen. Es kann auch andere Parameter wie Sprechgeschwindigkeit oder Stimmqualität beeinflussen. Diese Aspekte sind jedoch noch nicht komplett ausgearbeitet. Es gibt verschiedene Register, die ein Sprecher verwenden kann:

h für eine normale Stimmlage (high),

l für eine tiefere Stimmlage (low),

l(h) für eine erweiterte Stimmlage (expanded),

l(l) für eine eingeschränkte tiefe Stimmlage (low compressed),

h(h) für eine nach oben verschobene Stimmlage (extra high),

h(l) für eine erweiterte Stimmlage (expanded),

h(hl) für eine stark erweiterte Stimmlage (high expanded).

Der Gebrauch der Registeretiketten, bestehend aus nicht geklammerten zugrundeliegenden Merkmalen sowie geklammerten modifizierenden Merkmalen, unterliegt gewissen Restriktionen. Für eine ausführlichere Diskussion sei auf [May97] verwiesen.

2.5 Zusammenfassung

In diesem Kapitel wurde ein phonologisches Intonationsbeschreibungssystem für das Deutsche vorgestellt, daß von Féry entwickelt [Fér93] und von Mayer [May95, May97] erweitert wurde. Für das erweiterte System stehen mit [May95] Ettikettierungsrichtlinien für die manuelle prosodische Etikettierung zur Verfügung,

anhand derer das Stuttgarter Prosodiekorpus, bestehend aus Radionachrichten und anderem, größtenteils gelesenem Material, erstellt wurde. Das Intonationsbeschreibungssystem kann als erprobtes System angesehen werden, für das einerseits ein automatischer Generierungsprozess existiert [Möh98], und andererseits ein automatischer Erkennungsprozess vorliegt, der in der vorliegenden Arbeit beschrieben ist. Das Intonationssystem scheint bei oberflächlicher Betrachtung den Schwerpunkt auf die Beschreibung der Tonhöhe zu legen. Bei genauerer Betrachtung wird jedoch schnell klar, daß auch die anderen Korrelate der Intonation erfaßt werden, also zeitliche Strukturierung und Intensität/Stimmqualität. Die Lage der Akzente und die Markierung von Grenzen sind zwar zu einem Großteil durch die Grundfrequenz(bewegungen) bestimmt, aber auch wesentlich durch zeitliche Strukturierung (Längungen, Sprechpausen) beeinflußt.

Kapitel 3

Ein System zur automatischen Segmentierung gesprochener Sprache in Wörter, Silben und Phoneme

In diesem und den nächsten beiden Kapiteln werden Verfahren zur Extraktion von Merkmalen vorgestellt, die die Prosodiekorrelate zeitliche Strukturierung, Grundfrequenz und Intensität/Stimmqualität beschreiben. Zentral ist das automatische Segmentierungssystem Alphons, das neben der Bereitstellung von Zeitmerkmalen auch der Etablierung eines Zugriffsmechanismus auf das Sprachsignal dient. Zur Berechnung der Grundfrequenz- und Intensitätsmerkmale wird der Zugriffsmechanismus ausgenutzt, indem die von Alphons erzeugte Silbensegmentierung verwendet wird. Das Segmentierungssystem wird in diesem Kapitel beschrieben.

3.1 Segmentierungsverfahren

Im wesentlichen werden bei der Segmentierung zwei Ansätze unterschieden, nämlich explizite und implizite Verfahren [Eng95]. Das für diese Arbeit implementierte System ist als ein explizites Verfahren einzuordnen, denn es handelt sich um eine Segmentierung durch Erkennung. Implizite Segmentierungsverfahren werden in dieser Arbeit nicht behandelt und deshalb nur sehr kurz zur Abgrenzung vom implementierten Vorgehen im nächsten Abschnitt angerissen. Eine ausführlichere Diskussion der in der Literatur berichteten Verfahren gibt [Eng95, Kap. 4].

3.1.1 Implizite Verfahren: Signalnahe Segmentierung

Implizite Verfahren für das Auffinden von Segmentgrenzen im Sprachsignal sind dadurch gekennzeichnet, daß sie keine externen Informationsquellen wie etwa Information über die in einem Sprachsignal geäußerten Wörter oder Phoneme benötigen. Sie versuchen, lediglich aus dem Sprachsignal Segmentgrenzen zu ermitteln. Die Verfahren bilden das Vorgehen von Phonetikern nach, die beim Interpretieren eines Spektrogramms zunächst das Spektrogramm in Abschnitte einteilen und erst anschließend auf eine Zuordnung der Abschnitte zu Phonemen entscheiden. Die Verfahren, die diesen Ansatz verfolgen, beruhen zumeist auf der Annahme, daß sich die spektrale Charakteristik innerhalb eines Segments weniger stark ändert als an der Grenze von einem Segment zum nächsten. Operationalisiert werden kann die Ähnlichkeit spektraler Charakteristiken durch verschiedene Verfahren, wie zum Beispiel künstliche neuronale Netze, informationstheoretische Maße wie die Entropierate [Pap84] oder *mutual information* [BK94].

3.1.2 Explizite Verfahren: Segmentieren durch Erkennen

Im Gegensatz zum impliziten Ansatz werden bei den expliziten Verfahren externe Wissensquellen herangezogen. Eine weit verbreitete Vorgehensweise bei diesen Verfahren ist die Segmentierung durch Erkennung. Dabei wird angenommen, daß die geäußerten Wörter oder Phoneme bereits bekannt sind. Die Aufgabe, bekannte Sprache in Wörter oder Phoneme zu segmentieren, ist eng verwandt mit dem Problem der Spracherkennung [Rab89, HTK]. Grob gesagt handelt es sich bei der Segmentierung durch Erkennung um Spracherkennung mit stark eingeschränktem Suchraum. Bei der Spracherkennung interessiert, welche Wörter in einem aufgenommenen Sprachsignal vorkommen, während die Anfangs- und Endzeiten der enthaltenen Segmente belanglos sind. Bei der Segmentierung ist es gerade umgekehrt, die Wortfolge wird als bereits bekannt angenommen, und es interessieren lediglich die Anfangs- und Endzeiten jedes Segmentes. Dies kann ein Phonem oder eine beliebige andere linguistische Kategorie sein, die mit dem Sprachsignal abgeglichen werden soll, wie z.B. Silbe, Fuß, Morphem oder Phrase. Durch das Beschränken aller möglichen Pfade auf eine einzige mögliche Folge von Kategorien reduziert man das Suchproblem bei der Spracherkennung erheblich, da man nur noch die Zeitdimension als einzigen Freiheitsgrad berücksichtigen muß.

Da man die Aufgabe des Segmentierens in Wörter oder Phoneme als vereinfachte Spracherkennung betrachten kann, ist es naheliegend, erfolgreiche Spracherkennnungsparadigmen für die Segmentierung zu verwenden. Dieser Ansatz

wurde für das entwickelte System Alphons gewählt, das die in der Spracherkennung erfolgreiche Methode der statistischen Modellierung mit Hidden-Markov-Modellen (HMM) [Rab89, HTK] für die Segmentierung verwendet. Das Verfahren ist in zwei Phasen unterteilt, die Trainingsphase die nur einmal durchzuführen ist, und die Segmentierungsphase. In der Trainingsphase wird für jedes Phonem ein Markov-Prozeß bestimmt, dessen Emission zusätzlich durch eine Zufallsvariable beschrieben ist. Das heißt, daß neben dem von einer diskreten Verteilungsfunktion bestimmten Übergang in einen Folgezustand des Prozesses auch die Ausgabe jedes Zustandes durch eine (hier mehrdimensionale und stetige) Verteilungsfunktion beschrieben ist. Die beiden Zufallsvariablen bewirken, daß man nur aus der Beobachtung der Emissionen nicht auf die Abfolge der aufgesuchten Zustände schließen kann, die Zustandsfolge ist somit verdeckt (hidden). Die Parameter der Modellierung, also die Übergangswahrscheinlichkeiten und die Parameter der Wahrscheinlichkeitsdichtefunktion für die Ausgabe in den Zuständen, werden in einer Trainingsphase an einer größeren Sprachstichprobe abgeschätzt, für die die Segmentierung in Phoneme bekannt ist. Damit weicht bei dem entwickelten System die Trainingsphase der HMM kaum vom Trainingsverfahren bei der Entwicklung von Spracherkennern ab.

Sowohl für das Training als auch für das tatsächliche Segmentieren sind einige Vorverarbeitungsschritte nötig. Da die Modellierung auf Phonemen beruht und von der zu analysierenden Äußerung zunächst nur eine orthographische Beschreibung bekannt ist, muß diese erst in eine phonemische Repräsentation überführt werden, das heißt, eine Graphem-Phonem-Konvertierung ist erforderlich. Auch das Sprachsignal erfährt eine Vorverarbeitung, bei der das digitalisierte Sprachsignal einer Kurzzeit-Spektralanalyse unterzogen wird, deren Ergebnis dann in Merkmalsvektoren zusammengefaßt wird. Diese in der Spracherkennung übliche Repräsentation des Sprachsignals versucht, die Ähnlichkeit innerhalb der Klassen zu maximieren (allophonische Variation) und die Ähnlichkeit zwischen den Klassen zu minimieren (Distinktivität der Phoneme). Für die Segmentierung werden die trainierten Modelle entsprechend der phonemischen Repräsentation aneinandergehängt und mittels Viterbi-Decodierung auf die aus dem Sprachsignal extrahierten Merkmalsvektoren angewandt [HTK]. Durch Rückverfolgung des wahrscheinlichsten Pfades läßt sich die gemäß der Modellierung wahrscheinlichste Segmentierung des Sprachsignals in Phoneme bestimmen und daraus auch die Segmentierung in größere Einheiten ableiten.

Alphons setzt die Kenntnis der gesprochenen Wörter bzw. der Phonemfolge voraus, kann aber auch so abgewandelt werden, daß es ohne diese Wissensquelle funktioniert. Da nämlich die Segmentierung durch Erkennung letztlich die gleiche Technologie wie die Spracherkennung verwendet, kann auch ein beliebiger Spracherkenner benutzt werden, um eine Segmentierung in Wörter oder Phoneme

vorzunehmen. Da die Erkennungsgenauigkeit sowohl bei Wort- als auch bei Phonemerkennern heutiger Technologie ohne Domäneneinschränkung nicht ausreichend ist,[1] kann die Ausprägung der Phoneme nicht als zuverlässig angenommen werden. Jedoch können durchaus auch die Grenzen von falsch erkannten Phonemen mit einer tatsächlichen Grenze übereinstimmen. Ist man also nur an einer Segmentierung interessiert, kann möglicherweise auch dieses Verfahren zufriedenstellende Ergebnisse liefern.

3.2 Systemarchitektur

Das entwickelte Segmentierungssystem Alphons beruht auf dem Hidden Markov Model Toolkit V1.5 (HTK) von Entropic Inc. [HTK] und dem 359 611 Wortformen umfassenden deutschen Teil des CELEX-Lexikons [BPvR93], das für die Graphem-Phonem-Konvertierung verwendet wird. Abbildung 3.1 zeigt schematisch die Systemarchitektur des Segmentierungssystems. Die beiden Eingabeströme des Systems, nämlich der orthographische Text und das digitalisierte Sprachsignal, sind in Abb. 3.1 am oberen Rand dargestellt. Diese werden vorverarbeitet und dienen dann als Eingabe für den HTK Viterbi-Decoder, der anschließend die eigentliche Segmentierung ausführt. Die nächsten beiden Abschnitte beschreiben die Vorverarbeitungsstufen für den orthographischen Text bzw. das Sprachsignal.

3.2.1 Aufbereitung des orthographischen Texts

Der orthographische Text muß zunächst in eine Folge von Phonemen konvertiert werden. Aus der Phonemfolge wird dann eine reguläre Grammatik konstruiert, die beschreibt, wie die den Phonemen entsprechenden HMM für die Segmentierung zu verketten sind. Die Graphem-Phonem-Konvertierung erfolgt in mehreren Stufen. Die meisten Grapheme werden konvertiert durch Nachschlagen in einem Vollformenlexikon mittels binärer Suche. Aufgrund seiner großen Abdeckung wird dazu das recht umfangreiche CELEX–Lexikon verwendet, das $359\,611$ Wortformeneinträge enthält. Vor dem Zugriff auf das CELEX–Lexikon wird in einem Ausnahmelexikon nachgeschlagen, in dem die Transkription zusätzlicher Wörter festgehalten werden kann. Diese Zugriffsanordnung ermöglicht es, fehlerhafte Einträge im CELEX–Lexikon zu maskieren. Wörter, die nicht in den Lexika

[1] Die ohne die lexikalische Einschränkung der Sprache arbeitenden Phonemerkenner erreichen etwa bis zu 75% Phonemakkuratheit.

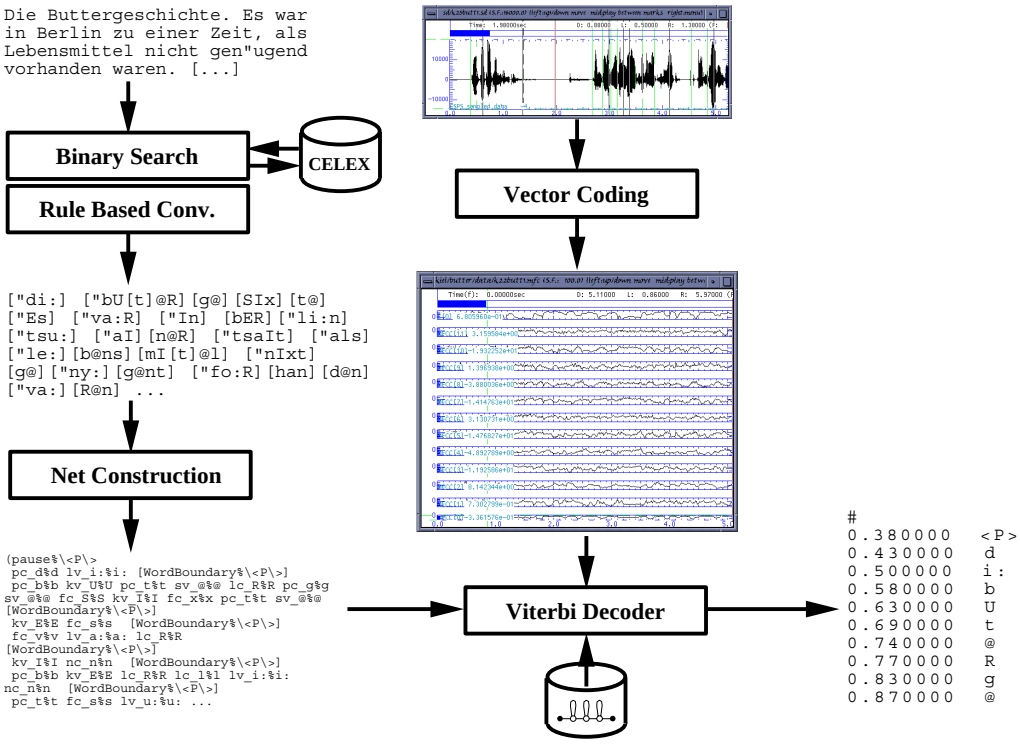


Abbildung 3.1: Überblick über die Systemarchitektur.

aufgefunden werden (meist sind dies Eigennamen oder Komposita), werden in der nachfolgenden Stufe regelgesteuert umgesetzt. Die regelgesteuerte Graphem-Phonem-Konvertierung ist eine lex-Implementation derjenigen Regeln aus dem Ausspracheduden [Man90], die ohne aufwendige morphologische, etymologische oder semantische Analyse auskommen. Lex, ein sogenannter Scannergenerator, ist ein UNIX-Standardwerkzeug, das für die Compilerentwicklung verwendet wird. Lex vereinfacht die Implementation der lexikalischen Analysephase eines Compilers, indem es erweiterte reguläre Ausdrücke, die die lexikalischen Einheiten einer Programmiersprache beschreiben (z. B. Bezeichner, Schlüsselwörter, Operatoren, Zahlen), automatisch in ein C-Programm umwandelt, das einen endlichen Automaten zur Erkennung dieser lexikalischen Einheiten simuliert. Bei der Graphem-Phonem-Konvertierung dienen die erweiterten regulären Ausdrücke zur Beschreibung der Grapheme oder Graphemfolgen, die auf ein Phonem abzubilden sind. Dabei kann der graphemische Rechtskontext durch den Lookahead der erweiterten regulären Ausdrücke und der Linkskontext durch Speicherung in C-Variablen für die Auswahl des einzusetzenden Phonems verwendet werden. Ein weiteres, ebenfalls mit lex implementiertes Programm fügt in die regelbasiert umgewandelten Phoneme Silbengrenzen ein, indem es die einfache, aber wirkungsvolle Heuristik der Maximierung des Silbenonsets verwendet; zusätzlich muß dabei der Silbenonset am Anfang eines in CELEX verzeichneten Wortes auftreten. Die verwendeten Regeln und Einzelheiten der Implementation sind in [Tab95] ausführlicher beschrieben.

Alle Konvertierungsstufen, also das Nachschlagen in den Lexika, die regelgesteuerte Umwandlung und das Einfügen von Silbengrenzen, sind in einer Pipeline angeordnet. Token, die von einer Stufe nicht übersetzt werden können, verbleiben im Strom für die Bearbeitung durch die folgenden Stufen. Dadurch ist es einfach, Komponenten auszuschalten, auszutauschen oder weitere Komponenten hinzuzufügen, z.B. die korrekte Aussprache von gebräuchlichen Abkürzungen, Datumsangaben, Zahlen etc. Weiterhin erlaubt diese Architektur, Wörter, die nicht im Lexikon stehen und von der regelgesteuerten Umsetzung schlecht erfaßt werden, direkt im orthographischen Text zu transkribieren, beispielsweise fremdsprachliches Material, das weder im Lexikon aufgefunden noch von der deutschen Graphem-Phonem-Konvertierung angemessen transkribiert wird. Neben dieser phonemischen Annotationsmöglichkeit im orthographischen Text gibt es noch die Möglichkeit, Kontrollstrukturen zu vermerken: optionale Textstücke, Wiederholungen, Disjunktionen und Klammerung. Dies hat sich für die Analyse von stark repetitiven Laboraufnahmen und die Erschließung externer Korpora als sehr nützlich erwiesen [Rap96].

Ausgabe des gesamten Konvertierungsverfahrens ist eine phonemische Umschrift, die Silbenmarkierungen mit ambisyllabischen Konsonanten enthält. Es werden

keinerlei Aussprachevarianten erzeugt, es sei denn, diese sind durch Kontrollstrukturen und Transkription im orthographischen Text ausdrücklich gewollt.

Der Viterbi-Decoder des HTK-Systems erwartet eine reguläre Grammatik welche die möglichen Wort- oder Phonemabfolgen beschreibt und damit den Suchraum der Erkennung festlegt. Für den Segmentierungsfall ist dies im wesentlichen die einfache Aneinanderreihung der Phonemmodelle entsprechend der aus dem Text erzeugten phonemischen Umschrift. Die Grammatik wird also für jedes zu segmentierende Sprachsignal neu aus der jeweiligen phonemischen Umschrift erzeugt. Die generierte Grammatik erlaubt zudem ein optionales Einfügen eines Sprechpausen modellierenden HMM zwischen zwei Wörtern. Die Annotationsmöglichkeiten für Kontrollstrukturen werden 1:1 in die reguläre Grammatik übernommen.

3.2.2 Aufbereitung des Sprachsignals

Als zweite Eingabe benötigt die Viterbi-Decodierung Merkmalsvektoren, die das Sprachsignal beschreiben. Das implementierte System verwendet eine in der Spracherkennung verbreitete Kodierung spektraler Information [ICA97]. Das Sprachsignal wird konvertiert in zwölf Mel-Frequenz-skalierte Kepstralkoeffizienten sowie die Gesamtenergie. Aus diesen Parametern werden delta- und delta-delta-Werte berechnet, die mit der ersten beziehungsweise zweiten Ableitung korrespondieren. Insgesamt fallen also pro Zeiteinheit 39 Parameter an. Der zeitliche Abstand zwischen zwei Vektoren beträgt 10 ms, die Werte werden aus einem Hamming-gewichteten Ausschnitt aus dem Sprachsignal berechnet, das die Länge 25.6 ms aufweist. Somit überlagern sich zwei aufeinanderfolgende Fenster um mehr als die Hälfte. Um dem artikulatorisch-akustisch bedingten Energieabfall höherfrequenter Signalanteile entgegenzuwirken, wird das Zeitsignal vor der Analyse mit einer Präemphase von 0.97 verzerrt. Die Analyseordnung bei der Berechnung der kepstralen Koeffizienten wurde entsprechend der gängigen Praxis (z. B. [HTK]) auf die Anzahl der verwendeten Mel-Frequenz-skalierten Kepstralkoeffizienten mal 2, also auf 24, eingestellt.

3.3 Training

Als Trainingsmaterial dienten die sogenannten Berlin- und Marburg-Sätze. Dabei handelt es sich um je 100 phonetisch balancierte Sätze, die auf der CD-ROM „The

Kiel Corpus of Read Speech Vol. 1“ [Koh94, Koh95] als digitalisiertes Sprachsignal mehrerer Sprecher mit enger phonetischer Transkription vorliegen. Es wurden nur 1 200 der 2 400 auf der CD-ROM enthaltenen Äußerungen verwendet, und zwar die der 6 männlichen Sprecher und der 6 weiblichen Sprecherinnen der jeweils jüngeren Altersgruppe (also 100 Sätze pro Sprecher(in)). Das Trainingsmaterial enthält insgesamt 6144 gesprochene Wörter. Die HMM wurden initialisiert mit der elaborierten phonetischen Etikettierung, die vom Institut für Phonetik und digitale Sprachverarbeitung in Kiel erstellt wurde. Die weiteren Trainingsschritte jedoch wurden mit der stärker phonologisch orientierten Transkription durchgeführt, wie sie durch die Graphem-Phonem-Konvertierung erzeugt wird. Da es für diese Phonemfolge keine Handsegmentierung wie für die phonetische Kieler Etikettierung gibt, wurde die „reestimation“-Phase übersprungen und direkt eine „embedded reestimation“ durchgeführt [HTK]. Durch diese Vorgehensweise wird eine stark implizite Modellierung erzielt, da das Lexikon keinerlei Aussprachevarianten enthält, die häufig im Trainingsmaterial anzutreffen sind. Das explizite Training von phonetischen Varianten wäre mit dem Verfahren natürlich auch möglich, entweder durch mehrfache Lexikoneinträge oder durch regelgesteuerte Erzeugung von Varianten aus der kanonischen Lexikon-Transkription (vgl. [Wes95]). Für unsere Belange hat sich dies jedoch als nicht notwendig erwiesen. Zudem wird in der Literatur auch die Meinung vertreten, phonetische Variation könne aus einer phonologisch segmentierten Transkription systematisch abgeleitet werden [Cam95]. Es sind jedoch weitere Untersuchungen notwendig, um diese Frage zu klären.

3.3.1 Topologie der Phonemmodelle

Phoneme werden wie in der Spracherkennung gebräuchlich durch links-rechts-Modelle mit drei emittierenden Zuständen modelliert. Eine Ausnahme bilden die Diphtonge, die im allgemeinen deutlich länger als andere Phoneme sind. Sie werden durch HMM mit 4 emittierenden Zuständen modelliert. Der Phonemkontext wird in der Modellierung nicht berücksichtigt. Als Wahrscheinlichkeitsdichtefunktion werden (einfache) Normalverteilungsfunktionen mit diagonaler Kovarianzmatrix verwendet. Drei verschiedene Modelltopologien wurden getestet, die in Abb. 3.2, (1)–(3) wiedergegeben sind. Die Modelle sind (1) ohne skip transitions, (2) mit drei skip transitions, und (3) sogenannte t-Modelle, also Modelle, die komplett übersprungen werden können. Überraschenderweise wurden die besten Ergebnisse mit dem einfachsten Modell erzielt (Modell (1) in Abb. 3.2). Sobald das automatische Verfahren mehr Freiheit bekam, wurden die Segmentierungen schlechter.

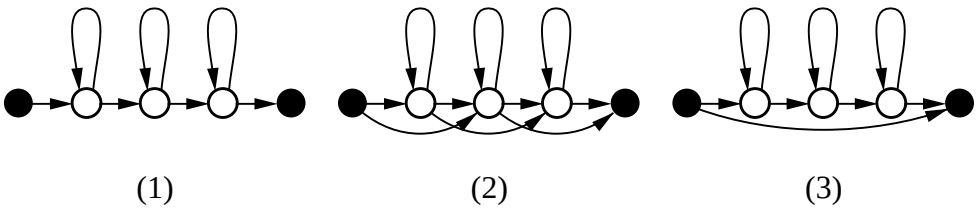


Abbildung 3.2: Modelltopologien für Phoneme.

3.3.2 Modellierung von Sprechpausen

Im verwendeten Trainingsmaterial sind Atemgeräusche nicht etikettiert. Dies führte bei der automatischen Segmentierung zunächst zu Problemen. Mit einer Pausenmodellierung mit einfacher Topologie wie in Abb. 3.3 (1) ergab sich bei der automatischen Segmentierung von Testdaten, daß häufig bei Sequenzen Stille–Einatmen–Stille lediglich die Stille vor oder nach dem Atmen als Sprechpause etikettiert wurde. Das Atemgeräusch und das andere Stillesegment wurde vorhergehenden oder nachfolgenden Segmenten zugeschlagen, insbesondere, wenn es sich bei diesen um Plosive oder Frikative handelte.

Diesem Problem wurde wie folgt begegnet. Zunächst wurde die kompliziertere Modelltopologie gewählt, die in Abb. 3.3 (2) gezeigt ist. Für die Initialisierung der Ausgabewahrscheinlichkeiten des ersten und dritten Zustandes (diese sind in Abb. 3.3 (2) mit einem „S“ markiert) wurden die Wahrscheinlichkeitsdichtefunktionen aus dem einfacheren Modell (Abb. 3.3 (1)) eingesetzt. Zusätzlich wurden diese gebunden, das heißt, daß sie beim nachfolgenden Training nur zusammen verändert werden können.[2] Die Wahrscheinlichkeitsdichtefunktion für den mittleren Zustand, in Abb. 3.3 (2) mit einem „B“ markiert, wurde vom mittleren Zustand des Modells für den Frikativ `/h/` kopiert, da die Akustik bei diesem Phonem vermutlich am ehesten der Akustik beim Einatmen entspricht. Zudem ist die allophonische Variation bei diesem Laut groß, da sich der Vokaltrakt bei diesem Phonem stark an die Umgebung anpaßt. Somit ist eine große Varianz bei den Wahrscheinlichkeitsdichtefunktionen zu erwarten, die deshalb das Atemgeräusch gut erfassen sollten. Durch die nachfolgenden Trainingsstufen hat sich das auf diese Art synthetisierte Modell sehr zufriedenstellend an die Modellierung von Stille–Einatmen–Stille angepaßt, wie durch Untersuchungen an Testmaterial beobachtet wurde. Durch den erlaubten Übergang vom ersten in den dritten Zustand war

[2] Die Implementierung stellt sicher, daß die Parameter nicht unterschiedlich werden können, indem sie bei den beiden Zuständen jeweils einen Zeiger auf eine einzige Datenstruktur einrichtet, die die Parameter der Wahrscheinlichkeitsdichtefunktion enthält.

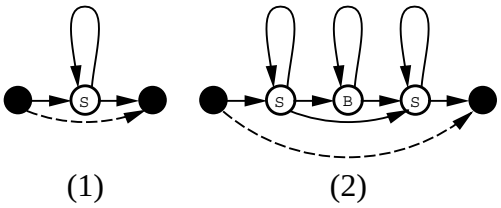


Abbildung 3.3: Modelltopologien für Pausen.

das Modell auch in der Lage, Sprechpausen, bei denen kein erkennbares Atemgeräusch enthalten war, korrekt zu segmentieren.

3.3.3 Evaluation des Trainings

Beim Abschätzen der Modellparameter ist es notwendig zu entscheiden, wann der iterative Trainingsprozeß abgebrochen werden soll. Abbildung 3.4 zeigt Lernkurven für unterschiedliche Genauigkeitsschwellwerte. Die unterste Kurve entspricht einem Genauigkeitsschwellwert von 10 ms um die manuelle Segmentierung. Darüber sind die Kurven für Genauigkeiten von 15 ms, 16 ms, 20 ms, 32 ms, 48 ms und 64 ms gezeichnet. Sie wurden berechnet durch automatische Segmentierung von neuem, nicht in der Trainingsmenge enthaltenem Sprachmaterial („Die Buttergeschichte“, [Koh95]) und anschließendem Vergleich mit der manuellen Segmentierung. Auch die Sprecher waren nicht in der Trainingsdatenmenge vertreten. Die Graphik verdeutlicht, daß der beste Zeitpunkt, das Training abzubrechen, unmittelbar nach der ersten Iteration liegt. Für eine Genauigkeit von 20 ms erreicht das Verfahren 76.82%. Das bedeutet, daß etwa ein Viertel der automatisch gefundenen Segmentgrenzen 20 ms oder weiter von der manuellen Segmentierung entfernt liegen.

3.4 Segmentierung

Es existieren zwei Versionen des Systems, die am Lehrstuhl Experimentelle Phonetik regelmäßig eingesetzt werden. Eine Version, `alignphones`, bestimmt die Anfangs- und Endzeiten von Phonemen aus Sprachsignal und orthographischem Text. Wenn der orthographische Text nicht vorliegt, wird der Benutzer gebeten, diesen einzugeben, während das Sprachsignal abgespielt wird. Die andere Version, `alignwords`, bestimmt Anfangs- und Endzeiten ganzer Wörter. Dies wird dadurch bewerkstelligt, daß eine andere reguläre Grammatik für den Viterbi-Decoder erzeugt wird, die diesen anweist, anstatt für jedes erkannte Phonem ein

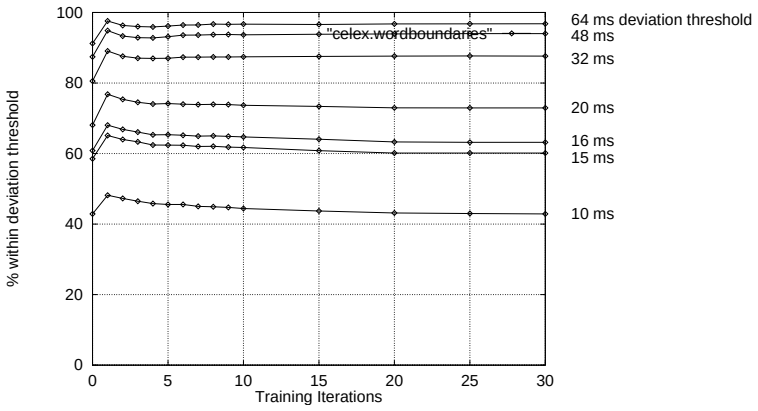


Abbildung 3.4: Anteil der Wortgrenzen, die nicht weiter als 10, 15, 16, 20, 32, 48, 64 ms von der manuellen Segmentierung entfernt sind, ermittelt für verschiedene Anzahlen von Trainingsiterationen.

Phonemetikett, lediglich ein Wortetikett nach dem letzten erkannten Phonem eines Wortes auszugeben. Als Resultat erhält man eine Wortsegmentierung, die mit der Phonemsegmentierung kompatibel ist, das heißt, die Wortgrenzen koinzidieren mit Phonemgrenzen. Silbenetiketten werden nicht direkt vom Segmentierungssystem produziert, da das HTK-System überlappende Segmente nicht behandeln kann. Zahlreiche Phonologen vertreten die Ansicht (u.a. [Ven90, Hal92]), daß das Deutsche ambisyllabische Konsonanten hat, also Konsonanten, die sowohl zur Koda der vorhergehenden Silbe als auch zum Anfangsrand der nachfolgenden Silbe gerechnet werden müssen, was ein Überlappen der Silben zur Folge hat, das „Gelenk“ in der Terminologie Vennemanns [Ven90]. Da die Graphem-Phonem-Konvertierung Silbengrenzen unter Berücksichtigung ambisyllabischer Konsonanten enthält, wird eine explizite Silbenetikettierung aus der Phonemsegmentierung berechnet.

Ein verbleibendes Problem ist die nicht sehr feine zeitliche Auflösung des Systems, die durch den Abstand der Merkmalsvektoren vorgegeben ist. Einige Experimente mit kleineren Abständen (5 bzw. 1 ms anstatt 10 ms) haben gezeigt, daß die Robustheit des Segmentierungssystems abnimmt, wenn man die Genauigkeit auf diese Weise zu verbessern versucht. Abbildung 3.5 zeigt, daß ein Abstand von 10 ms für alle Schwellwerte die besten Ergebnisse liefert. Lediglich ein Merkmalsvektorabstand von 5 ms scheint ebenfalls noch zufriedenstellend zu sein. Es ist bei den Kurven für 1 und 5 ms Abstand zu beachten, daß die „embedded reestimation“ die Systemleistung nicht verbessert, die Werte für 0 Iterationen sind jeweils die höchsten. Für einen Abstand von 10 ms wird das beste Resultat nach einer Trainingsiteration erzielt. Die hier dargestellten Werte unterscheiden

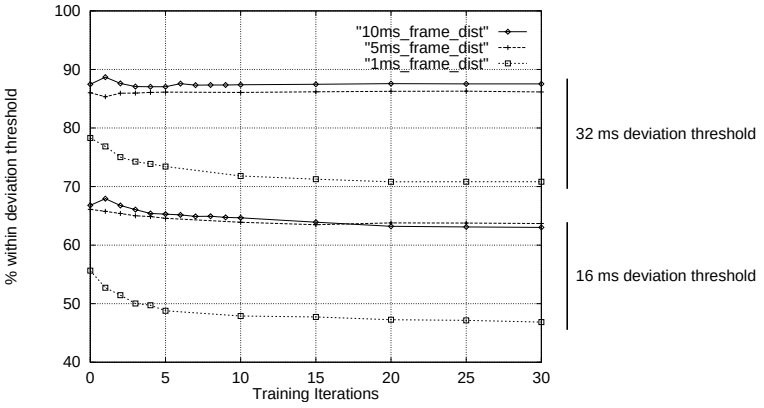


Abbildung 3.5: Anteil der Grenzen innerhalb einer tolerierten Genauigkeit für verschiedene Fensterabstände. Die oberen Kurven entsprechen einem Schwellwert von 32 ms, die unteren einem Schwellwert von 16 ms.

sich leicht von denen in Abbildung 3.4, da vor dem „embedded“ Training ein Trainingsschritt zur Abschätzung der Transitionswahrscheinlichkeiten eingefügt werden mußte.

3.5 Diskussion

In diesem Kapitel wurde Alphons vorgestellt, ein robustes Phonem- und Wortsegmentierungssystem, das aus leicht verfügbaren Ressourcen [Koh95, BPvR93] und Werkzeugen [HTK] aufgebaut wurde. Wie Tabelle 3.1 zeigt, kann das Segmentierungssystem mit kommerziellen Produkten und mit Forschungssystemen anderer Institute konkurrieren, die für andere Sprachen entwickelt wurden. Die Verwendung eines großen Lexikons erlaubt eine gute Abdeckung auftretender Wortformen. Die Wahl einer spezifischen Lexikonimplementierung (CELEX) erlaubt es, Sprache nicht nur segmental, sondern auch syllabifiziert, morphologisch und morphosyntaktisch zu annotieren. Obwohl die phonetische Repräsentation des CELEX-Lexikons streng phonologisch orientiert ist, hat die Implementation gezeigt, daß eine solche Repräsentation ein recht robust arbeitendes System ergibt, auch wenn Sprache oder Text fehlerhaft oder die Äußerungen reduziert sind. Alphons ist flexibel und nicht komplett vom verwendeten Lexikon abhängig, da das Lexikon entweder durch ein anderes oder sogar komplett durch eine regelgesteuerte Graphem-Phonem-Umsetzung wie etwa [Tab95] substituiert werden kann. Beim Verzicht auf ein Lexikon ist zu erwarten, daß die Qualität speziell der Phonemtranskription schlechter ausfällt, da sich hier Fehler der regelbasier-

		10 ms	15 ms	20 ms	16 ms	32 ms	48 ms	64 ms
Dalsgaard, Andersen, Barry (1991)	Dänisch	48.7	58.9	65.5				
	Englisch	62.6	72.5	77.5				
	Italienisch	34.2	45.8	52.0				
Kvale (1993)	Dänisch	70.3	81.0	86.1				
	Englisch	66.2	76.4	82.3				
	Italienisch	64.2	77.7	84.5				
	Norwegisch	68.1	80.4	86.4				
Entropic's Aligner	Englisch				71	90		97
Alphons a	Deutsch	53.9	70.6	80.9	73.7	91.3	96.5	98.4
Alphons b		59.1	75.9	84.4	78.4	91.1	94.5	96.0
Alphons c		60.8	77.9	86.3	80.4	92.4	95.4	96.7

Tabelle 3.1: Gegenüberstellung der Genauigkeit von verschiedenen Segmentierungssystemen. Die Tabelle enthält die prozentualen Anteile der automatisch gefundenen Grenzen, die nicht weiter als der in der ersten Reihe angegebene Schwellwert von der manuellen Segmentierung entfernt sind. Die Werte für die oberen beiden Systeme sind [Kva93], die des nächsten Systems [Ent94] entnommen. Die Ergebnisse für das in dieser Arbeit beschriebene System Alphons wurden ermittelt, nachdem alle Parameter festgelegt wurden (10 ms Merkmalsvektorabstand, eine embedded reestimation) mit vorher nicht betrachtetem Material („Nordwind und Sonne“, [Koh95]). Reihe a: Lexikonzugriff mit CELEX wie oben beschrieben, für die Auswertung wurden nur Wortgrenzen berücksichtigt. Reihe b: forced alignment mit Phonemen der manuellen Segmentierung, alle Grenzen berücksichtigt. Reihe c: wie b, jedoch Sprechpausen nicht berücksichtigt.

ten Umsetzung stärker niederschlagen als bei der Wortsegmentierung. Das System erzielt eine recht gute Abschätzung der Silben- und Wortlängen, da für die größeren Einheiten das 10 ms-Raster der Merkmalsvektoren geringere Auswirkungen besitzt. Für die gestellte Aufgabe, die Untersuchung silben- und wortbasierter prosodischer Phänomene, ist die Genauigkeit des Segmentierungssystems ausreichend. Für ein genaues phonetisches Etikettieren mit Genauigkeitsanforderungen unter 10-20 ms sollten andere Werkzeuge verwendet werden. Neben dem Gebrauch der Segmentierung für Zugriff und prosodische Annotation von Korpora für die Prosodieforschung wird das Segmentierungssystem auch für die datengetriebene konkatenative Sprachsynthese verwendet [Kam97, Möh98].

Kapitel 4

Eine Parametrisierung der Sprachgrundfrequenz

In diesem Kapitel wird ein automatisches Verfahren zur Ermittlung einer phonetischen Beschreibung des Grundfrequenzverlaufs beschrieben, das Voraussetzung für die im nächsten Kapitel beschriebene Prosodieerkennung ist, aber auch zur genaueren, phonetischen Beschreibung der spezifischen Ausprägung von prosodischen Ereignissen dienen kann. Mit der F_0-Parametrisierung wird eine phonetische Zwischenstufe zwischen dem tatsächlichen Sprachsignal und der daraus geschätzten Sprachgrundfrequenz einerseits und andererseits den phonologischen Ereignissen, also den ToBI-Etiketten, eingeführt. Mit der F_0-Parametrisierung soll eine phonetische Beschreibung der Intonationskontur entwickelt werden, die die beiden folgenden Anforderungen erfüllt:

- Sie soll hinreichend explizit definiert sein, so daß eine algorithmische Bestimmung für eine große Datenmenge möglich ist.
- Sie soll kompatibel mit dem verwendeten phonologischen Intonationsmodell sein.

4.1 Anforderungen des Intonationsmodells

Wegen der phonologischen Orientierung des Intonationsmodells ist die grundlegende Zeitstrukturierung für die phonetische Parametrisierung die Silbensegmentierung, die durch die automatische Segmentierung mit Alphons gefunden wur-

de. Als weitere Eingabe verwendet die F_0-Parametrisierung die Ausgabewerte eines Grundfrequenzbestimmungsalgorithmus (engl. *pitch trackers*), der die Grundfrequenz des Sprachsignals für kurze, als stationär angenommene Ausschnitte abschätzt. Ausgabe der F_0-Parametrisierung sind sieben phonetisch interpretierbare Werte, die den Grundfrequenzverlauf für eine Silbe beschreiben.

Idealerweise sollte die Parametrisierung für Silben, die mit dem gleichen Pitchakzent- oder Grenztontyp assoziiert sind, ähnliche Werte liefern, und Silben, die unterschiedlich prosodisch kategorisiert sind, sollten unterschiedliche Parametrisierungen aufweisen.

Zum Verständnis des gewählten Parametrisierungsansatzes ist es hilfreich, sich die grundlegenden Akzent- und Grenztontypen des Stuttgarter Intonationssystems aus Kapitel 2 noch einmal in Erinnerung zu rufen. Bei den Pitchakzenten unterscheidet das System fünf verschiedene Typen: L*H (steigender Akzent), H*L (fallender Akzent), L*HL (steigend-fallender Akzent / später Gipfel), HH*L (früher Gipfel) und H*M („stylized contour"). Die ersten beiden Akzente, L*H und H*L, sind dabei die mit Abstand häufigsten. L*HL und HH*L sind deutlich seltener anzutreffen, und der letzte, H*M, tritt noch seltener auf, da seine Verwendung pragmatisch stark restringiert ist. H*M wird nur bei Vokativen eingesetzt, also z.B. wenn jemand beim Vornamen gerufen wird. In den Sprachdaten, die bis heute am IMS prosodisch analysiert wurden, traten die obigen Akzente in einer Verteilung von 59 zu 36 zu 3 zu 2 zu 0 Prozent auf. Da die H- (für high) und L- (für low) Töne jeweils mit (mindestens) einer Silbe assoziiert sind und die betonte Silbe mit einem nachfolgenden Stern markiert ist, ist schon aus der Benennung der Pitchakzente ersichtig, daß die Grundfrequenz der betonten und der auf sie folgenden Silbe die Unterscheidung des Pitchakzenttyps stark beeinflußt.[1] Für zwei Akzente, HH*L und L*HL, scheint weiter entfernt liegende F_0-Information relevant zu sein. Bei HH*L trägt Information aus vor der akzentuierten liegenden Silbe zur Unterscheidung von einem H*L-Akzent bei. Ähnliches kann aus der Benennung des L*HL abgeleitet werden, bei dem die Grundfrequenz der übernächsten Silbe hinter der betonten Silbe die Unterscheidung von L*H ermöglicht. Die Erfahrung mit dem manuellen prosodischen Etikettieren zeigt jedoch, daß der Späte-Gipfel-Akzent L*HL fast immer vollständig auf der akzentuierten und der ihr nachfolgenden Silbe realisiert wird und nicht auf der akzentuierten und den zwei auf sie folgenden Silben. Es ist in der Praxis also sehr selten, daß F_0-Information, die weiter als eine Silbe von der akzentuierten Silbe entfernt ist, zur Unterscheidung zwischen L*H und L*HL beiträgt. Bei HH*L jedoch scheint sehr wohl

[1] Ohne Beachtung des „complete linking" (s. Abschnitt 2.3) und eines erlaubten spreading des gesternten H-Tons des sehr seltenen H*M, wenn die betonte Silbe antepenult liegt.

F_0-Information der vor der akzentuierten liegenden Silbe für die Unterscheidung von einem H*L eine Rolle zu spielen, wie die manuelle Etikettierungs-Erfahrung zeigt.

Für die Parametrisierung ergibt sich aus diesen Überlegungen, ein zwei Silben umfassendes Fenster zu betrachten, das mit der ersten, potentiell akzentuierten, Silbe assoziiert ist. Falls die Trennfähigkeit der Parametrisierung zwischen H*L und HH*L nicht ausreichen sollte, kann für die Grenzfälle immer noch ein Schritt nachgeschaltet werden, der die Parametrisierung der vor der akzentuierten Silbe liegenden Silbe berücksichtigt.

4.2 Wahl der Parametrisierungsfunktion

Als erster Schritt zum Auffinden der Parametrisierung wird eine Zeittransformation durchgeführt, die diejenigen Grundfrequenzmessungen des Pitch trackers, die in der untersuchten Silbe liegen, linear auf das Intervall (-1,0) abbildet. Ebenso werden für die folgende Silbe die Grundfrequenzmessungen in das Intervall (0,1) transformiert. Ziel der Transformationen ist, bei der F_0-Parametrisierung von der Silbendauer zu abstrahieren und außerdem leicht interpretierbare Parameter zu erhalten. Bei den Transformationen wird eine mögliche Ambisyllabizität berücksichtigt.

Für den Bereich der untersuchten und der ihr folgenden Silbe wird eine parametrisierte Funktion bestimmt, die die Grundfrequenzmessungen des Pitch trackers möglichst gut approximiert. Die Wahl der Parametrisierungsfunktion geschieht unter Berücksichtigung des Intonationsmodells und physiologischer Beschränkungen. Anforderungen des Intonationsmodells an die Parametrisierungsfunktion sind die Möglichkeit zur Modellierung von Tonhöhenänderungen und die Möglichkeit zur Modellierung von Gipfeln. Aufgrund physiologischer Beschränkungen sind mehr als 2 Grundfrequenzbewegungen pro Silbe nicht notwendig. Die Parametrisierungsfunktion setzt sich aus drei Einzelfunktionen zusammen, die im folgenden motiviert und beschrieben sind.

Es ist klar, daß die Parametrisierung besonders gut zwischen den beiden verbreitetsten Pitchakzenten, H*L und L*H, unterscheiden sollte. Funktionen, die einen Anstieg oder Abfall modellieren können und dabei stetig und differenzierbar sind, sind beispielsweise die Arcustangensfunktion oder die hyperbolische Tangensfunktion. Zur Beschreibung eines Anstiegs oder Abfalls wurde die hyperbolische Tangensfunktion gewählt, die durch drei Parameter α, β und γ festgelegt wird. Diese Parameter bestimmen die Amplitude, also die Höhe eines Anstiegs oder

Abfalls, die Lage des Wendepunkts, also den Zeitpunkt eines Anstiegs oder Abfalls, und die Skalierung des Zeitmaßstabs, also die Steilheit eines Anstiegs oder Abfalls.

Diese drei Parameter genügen bereits, um einen Großteil der L*H- und H*L-Akzente voneinander zu unterscheiden. Wenn jedoch in dem Zweisilbenfenster ein Gipfel oder ein Tal auftritt, hervorgerufen durch zwei bitonale Pitchakzente bei unmittelbar aufeinanderfolgenden Silben, oder durch einen auf zwei Silben realisierten L*HL-Akzent (Später-Gipfel-Akzent), ist es nicht ersichtlich, ob der Tangens Hyperbolicus den Anstieg oder den Abfall modellieren sollte. Eine Funktion, die einen Gipfel oder auch ein Tal modellieren kann, ist die Funktion e^{-x^2}, die auch von der Wahrscheinlichkeitsdichtefunktion einer Normalverteilung bekannt ist. Diese Funktion ist die zweite Komponente der Parametrisierungsfunktion. Auch dieser Teil der Parametrisierungsfunktion ist durch drei Parameter kontrolliert, δ, ε und ζ. Sie beschreiben ebenfalls Amplitude, also Höhe des Gipfels bzw. Tales, Lage des Maximums, also den Zeitpunkt eines Gipfels oder Tales, und die Zeitskalierung, also die Steilheit bzw. Breite eines Gipfels oder Tales.

Diese beiden Parametrisierungsfunktionen können die Änderungen der Grundfrequenz beschreiben, aber keine Aussagen über das Niveau der Grundfrequenz machen. Die Beschreibung des Grundfrequenzniveaus übernimmt eine dritte Parameterisierungsfunktion. Eine naheliegende Wahl hierfür ist eine Gerade, $ax+\eta$. Wenn man eine Deklination, also ein allmähliches Abfallen der Grundfrequenzgesamtkontur annimmt, könnte man festlegen, daß die Steigung der Geraden, a, negativ und für alle Silben einer Phrase konstant sein muß. Auch könnte man festlegen, daß die Steigung a global konstant sein soll. A priori ist keine Phraseninformation bekannt. Somit gibt es die Optionen, die Steigung a global festzulegen, oder a für jedes zu parametrisierende Silbenpaar frei wählbar zu lassen. Im Falle einer frei wählbaren Steigung konkurriert eine beliebig schief liegende Gerade mit der Tangens-Hyperbolicus-Funktion, die die Grundfrequenzänderung beschreiben soll. Da im Zweifelsfall einfache Lösungen bevorzugt werden sollten, wird hier ein global festgelegtes $a = 0$ angenommen und dadurch die Gerade zur konstanten Funktion degeneriert. Die verbleibende Konstante, η, korreliert mit dem F_0-Niveau in dem Zweisilbenfenster.

Die drei Parametrisierungsfunktionen werden additiv zu einer einzigen Parametrisierungsfunktion zusammengefaßt. So ergibt sich insgesamt die folgende Parametrisierungsfunktion:

$$f(t) = \alpha \tanh(\beta(t-\gamma)) + \delta e^{-(\varepsilon(t-\zeta))^2} + \eta$$

Die mathematische Beschreibung dieser Parametrisierungsfunktion erlaubt es, algorithmisch für jede Silbe optimale Werte für die sieben Parameter α, β, γ, δ, ε,

ζ und η zu suchen, so daß die Parametrisierungsfunktion den gemessenen Grundfrequenzverlauf möglichst gut approximiert. Die Güte der Approximation wird anhand einer zu minimierenden Abstandsfunktion bewertet. Als Abstandsfunktion wird die Summe der Fehlerquadrate zu den gemessenen Grundfrequenzwerten gewählt.

Die Optimierung der Parameter geschieht mit dem iterativen Nelder-Meade-Algorithmus [NM, DW87]. Dies ist ein lokal arbeitender Optimierungsalgorithmus, der jedoch nicht das Finden des globalen Optimums garantiert, also nicht notwendigerweise die beste phonetische Parmetrisierung erbringt. Daher wird durch heuristische initiale Parameterschätzung versucht, mit der Parametersuche in einem Gebiet zu beginnen, das nahe dem globalen Optimum liegt. Zudem wird nach jedem Iterationsschritt erzwungen, daß die Parameter in einem realistischen Bereich liegen. Beispielsweise wird sichergestellt, daß η innerhalb des Grundfrequenzbereichs des Sprechers liegt und außerdem auch innerhalb des lokalen Grundfrequenzbereichs des betrachteten Zweisilbenfensters. Auf diese Art wird die Empfindlichkeit gegenüber groben Ausreißern bei der Grundfrequenzschätzung stark reduziert. Abbildung 4.1 zeigt die Parametrisierungsfunktion als Ganzes und aufgeteilt in die drei Einzelfunktionen für eine mit einem steigenden Akzent assoziierte Silbe.

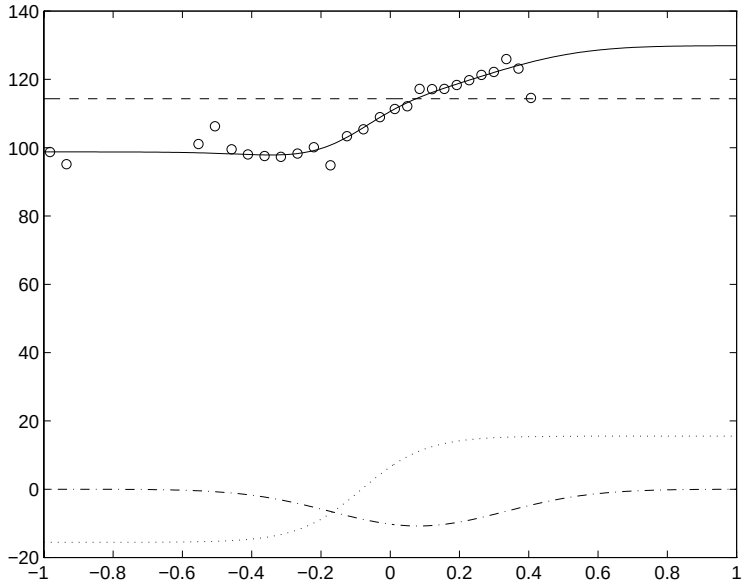


Abbildung 4.1: Ausgabe des Pitch trackers (Kreise) und Parametrisierungsfunktion (durchgezogene Linie), bestehend aus: $\tanh$ (gepunktelte Linie), e^{-x^2} (gepunktelt und gestrichelte Linie), Konstante (gestrichelte Linie) für eine mit L*H assoziierte Silbe.

4.3 Phonetische Interpretation der Parameter

Die phonetisch motivierten Parameter sollen die Erkennung der phonologischen Intonationsbeschreibung erleichtern, aber auch die spezifische phonetische Realisierung eines bestimmten phonologischen Typs beschreiben können. Auf diese Weise kann zum Beispiel das Phänomen „Downstep“ sowie der Gebrauch von Registern quantitativ untersucht werden. Die phonetische Interpretation der Parameter ist wie folgt:

- α korreliert mit dem tonalen Unterschied zwischen der akzentuierten und postakzentuierten Silbe. $alpha$ beschreibt die Abweichung beider Silben vom gemeinsamen Grundniveau η. 2α ist somit die Höhe eines Anstiegs oder Abfalls.
- β korreliert mit der Steilheit eines Anstiegs oder Abfalls.
- γ korreliert mit der zeitlichen Alignierung eines Anstiegs oder Abfalls.
- δ korreliert mit der Höhe eines Gipfels / Tals.
- ε korreliert mit der Steilheit eines Gipfels / Tals.
- ζ korreliert mit der zeitlichen Alignierung eines Gipfels / Tals.
- η korreliert mit dem generellen F_0-Niveau.

Ein psycholinguistisch motivierter Ansatz für die heuristische Schätzung der initialen Parameter des Nelder-Meade-Algorithmus wäre, die Suche nach den Parametern mit einem für den als nächstes erwarteten Pitchakzenttyp prototypischen Parametersatz zu starten. Dabei sollte die Erwartung nur vom Vorkontext, also von den linguistischen Funktionen der vorangehenden Wörter abhängen. Da jedoch der Zusammenhang zwischen linguistischer Funktion und prosodischer Ausprägung noch nicht klar ist, kann dieser Ansatz noch nicht verfolgt werden.

Mit der in diesem Kapitel entworfenen F_0-Parametrisierung steht eine algorithmisch bestimmbare Beschreibung des Grundfrequenzverlaufs zur Verfügung, die auch auf große Datenmengen anwendbar ist. Die Parametrisierung abstrahiert von der zeitlichen Strukturierung, indem sie Zeitbezüge in abstrakten Silbeneinheiten mißt. Die pro Silbe resultierenden sieben Parameter sind auch menschlichen Experten zugänglich, da sie eine phonetische Interpretation besitzen.

Kapitel 5

Eine Beschreibung der Intensität und Stimmqualität

5.1 Einführung

Jüngere Arbeiten, z. B. [Slu95, Cla96, CDJ+], haben gezeigt, daß glottale Merkmale mit Wortbetonung und Satzbetonung korrelieren und die Korrelation sogar stärker ist als die Korrelation zwischen Betonung und gemessener Lautstärke [Slu95]. Eine automatisch ableitbare Beschreibung glottaler Parameter zur Beschreibung der Intensität ist deshalb für die Erkennung prosodischer Ereignisse interessant. Die Korrelation zwischen Betonung und glottalen Merkmalen entsteht dadurch, daß ein Sprecher beim Produzieren einer betonten Silbe einen höheren subglottalen Druck aufbaut. Die dadurch veränderte Aerodynamik im Bereich der Glottis bewirkt ein anderes Schwingungsverhalten der Stimmlippen. Die Stimmlippen schließen abrupter, das Anregungssignal für den Vokaltraktresonator wird dadurch verändert. Akustisch wirkt sich die Änderung des Anregungssignals in einer Anhebung der höheren Frequenzen aus, das heißt, der spektrale Abfall zu den höheren Frequenzen hin, der bei Sprachsignalen stets auftritt, ist bei betonten Silben geringer als bei unbetonten. Naturgemäß kann das geänderte Schwingungsverhalten an der Glottis die Akustik nur bei stimmhaften Phonemen beeinflussen. Am deutlichsten sollte sich dies an Vokalen beobachten lassen, da hier der Mundrachenraum nicht stark verengt ist. Abbildung 5.1 illustriert die unterschiedlich ausgeprägten Spektra bei betonten (Abb. 5.1 oben) und unbetonten (Abb. 5.1 unten) Vokalen.

Leider sind glottale Parameter aufgrund der Filterung im Vokaltrakt nur relativ aufwendig aus dem Sprachsignal zu extrahieren [Slu95]. Sluijter [Slu95] postu-

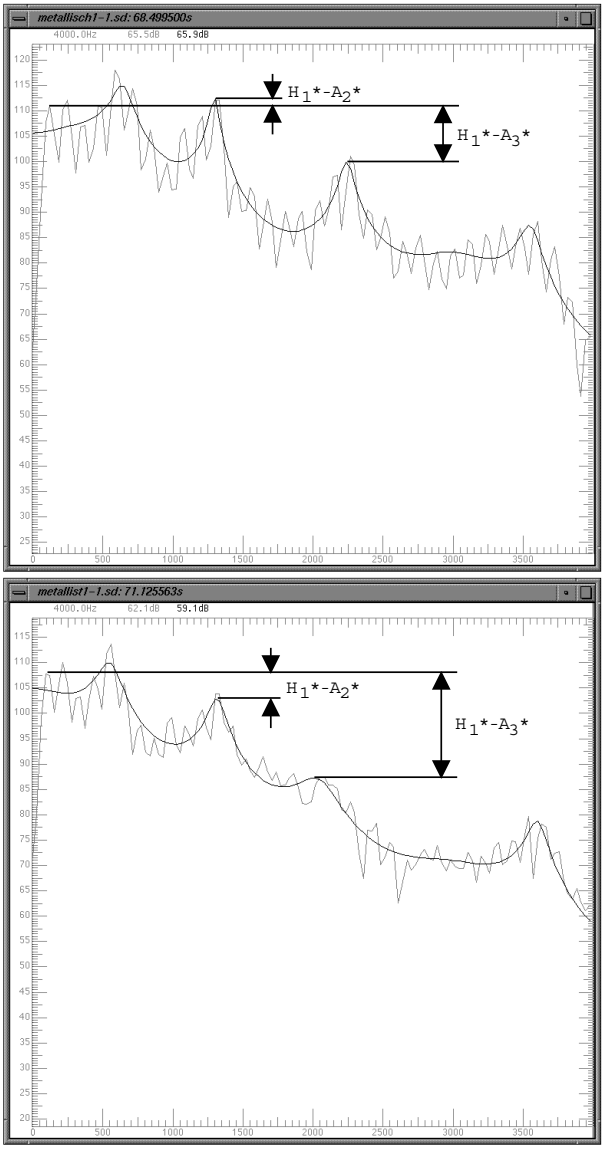


Abbildung 5.1: Die Abbildung oben zeigt den Abfall des Spektrums eines betonten Vokals. Unten ist der stärker ausgeprägte spektrale Abfall des gleichen Vokals in unbetonter Stellung zu sehen. Die Intensität fällt zu den höheren Frequenzen (rechts in den Darstellungen) hin ab. Der Abfall ist bei betonten Vokalen jedoch nicht so stark wie bei unbetonten.

OQ	Öffnungsquotient (*open quotient*)	H1* - H2*
SK	Schiefe des glottalen Impuls (*skewness of glottal pulse*)	H1* - A2*
RC	Schließrate der Glottis (*rate of closure*)	H1* - A3*
CC	Vollständigkeit des Glottisverschlusses (*completeness of closure*)	B1
GO	Grad der Glottisöffnung (*glottal opening*)	H1* - A1
AV	Amplitude des glottalen Anregungssignals (*amplitude of voicing*)	H1

Tabelle 5.1: Glottale Parameter und ihre akustischen Korrelate.

liert die in Tabelle 5.1 wiedergegebenen Zusammenhänge zwischen der Glottisphysiologie und der Akustik. In der Tabelle bezeichnen H1 und H2 die Amplitude der ersten und zweiten Harmonischen, A1, A2 und A3 die Amplitude des ersten, zweiten und dritten Formanten und B1 die Bandbreite des ersten Formanten. Abbildung 5.2 zeigt Formantenlagen, Formantenamplituden und Amplituden der Harmonischen an einem Beispiel. Bei den in der Tabelle 5.1 mit einem Stern versehenen Größen wurde die Beeinflussung durch umgebende Formanten durch Filterung herausgerechnet. Claßen [Cla96] zeigt für das Deutsche (analog zu [SH94, Slu95] für das Holländische und Englische), daß die beiden glottalen Parameter SK und RC am besten mit der Wortbetonung korrelieren. Beide Parameter werden in der Akustik aus der Stärke des spektralen Abfalls geschätzt.

In Forschung und Anwendung der automatischen Spracherkennung werden Merkmale aus dem Sprachsignal extrahiert, die die spektrale Charakteristik der Sprachlaute beschreiben sollen. Eine weit verbreitete Kodierung des Spektrums sind die Mel-Frequenz-skalierten Kepstralkoeffizienten (MFCC). Das Kepstrum, gewonnen aus einer Cosinustransformation des logarithmierten Spektrums, beschreibt die Form des Spektrums. Eigentlich müßte sich der spektrale Abfall in den MFCC niederschlagen. Möglicherweise korrelieren die ersten MFCC, die die mehr globale Form des Spektrums beschreiben, mit der Steilheit des spektralen Abfalls. Eine solche Korrelation würde für die Erkennung von prosodischen Ereignissen vorteilhaft sein, da sich die MFCC leicht aus dem vom Mikrophon aufgefangenen Sprachschall berechnen lassen und damit als leicht bestimmbare Merkmale für die meßtechnisch unzugänglichen glottalen Parameter bereitstehen würden, die ihrerseits mit der Betonung korrelieren. Ob dies zutrifft oder nicht, wird in diesem Kapitel untersucht.

Ein weiteres Korrelat der Betonung, die Ausgeprägtheit der Vokalqualität, ist ebenfalls in der Spektralstruktur enthalten. Vokale betonter Silben sind deutlicher artikuliert und besitzen deshalb eine Formantenstruktur, die im Vokalviereck „mehr in den Ecken“ liegt, oder umgekehrt ausgedrückt: unbetonte Vokale ha-

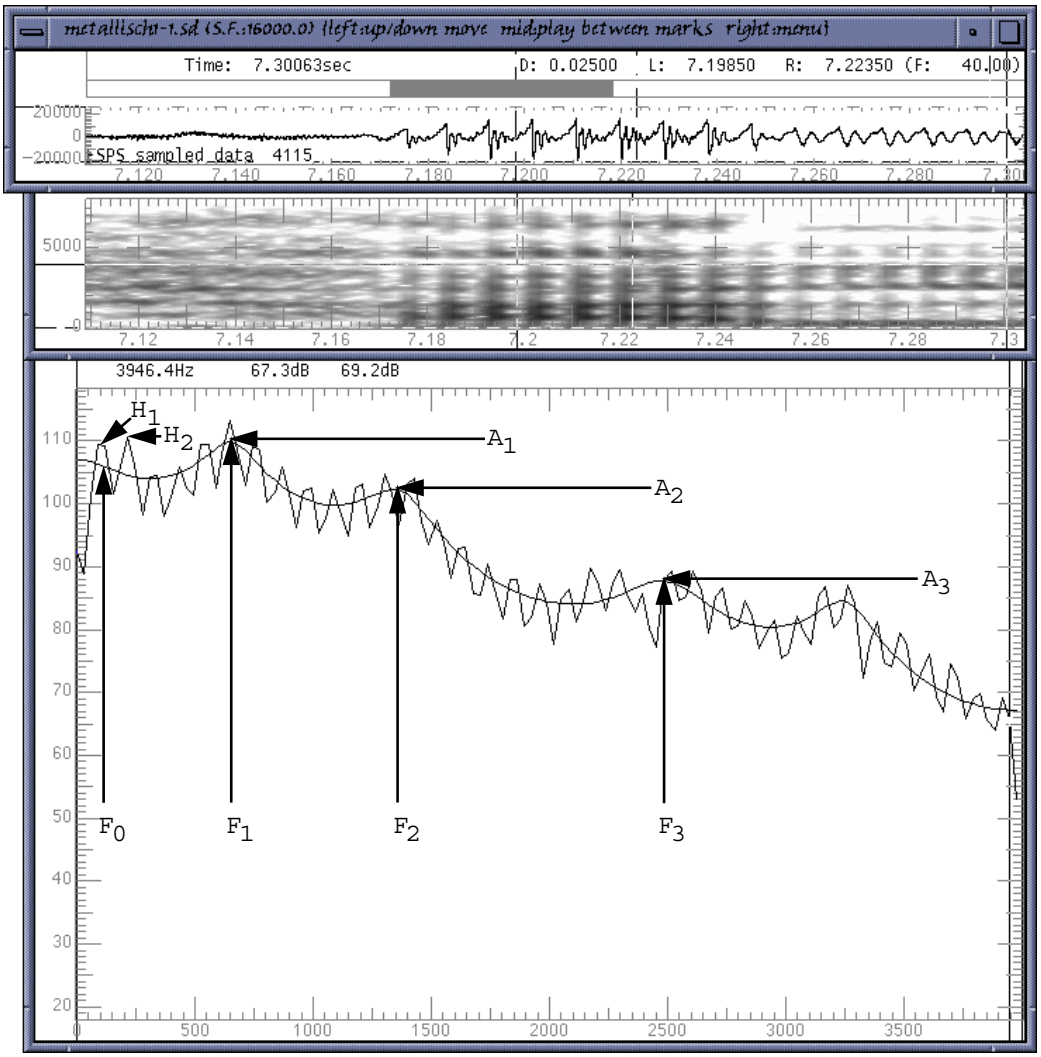


Abbildung 5.2: Die Abbildung zeigt, wie die Formantenlage, Formantenamplitude und Amplitude der Harmonischen gemessen werden. Oben: Zeitsignal der Silbe /tal/ aus „metallisch“. Mitte: Zeitsynchron dargestellte Frequenzanalyse (Spektrogramm) aus dem Zeitsignal. Unten: Spektralanalysen der in den oberen beiden Fenstern mit senkrechten gestrichelten Linien markierten Umgebung des Zeitpunkts der vollständigen Ausbildung des Vokals. Die wellenförmige Linie stellt ein DFT-Spektrum dar und zeigt die Lage sowie die Amplituden H_1 und H_2 der ersten und zweiten Harmonischen. Die Lage der Harmonischen ist charakterisiert durch die erste Harmonische, die mit der Grundfrequenz F_0 zusammenfällt, die weiteren Harmonischen sind ganzzahlige Vielfache davon. Die weniger wellige Linie zeigt eine LPC-Analyse, an der die Lage der Formanten (F_1, F_2, F_3) und die Formantenamplituden (A_1, A_2, A_3) abgelesen werden kann.

ben die Tendenz, sich dem Zentralvokal oder Schwa anzunähern. Zur Illustration zeigt Abb. 5.3 die Lage betonter und unbetonter Vokale des Deutschen im F_1-F_2-Raum. Für die Vorhersage des Betonungsgrades aus der Spektralstruktur sollte die zusätzliche Beeinflussung nicht nachteilig sein, möglicherweise unterstützt sie die Vorhersagegüte sogar.

Zur Untersuchung der Zusammenhänge zwischen den spektralen Merkmalen und der Wortbetonung gibt es mindestens zwei Möglichkeiten:

- Man kann versuchen, den phonetisch gemessenen spektralen Abfall direkt aus den MFCC vorherzusagen, also eine Abbildung von den die Spektralstruktur beschreibenden MFCC auf einen den spektralen Abfall beschreibenden Wert zu finden (MFCC $\mapsto$ SK oder MFCC $\mapsto$ RC).

- Indirekt durch Vergleich zweier Klassifikationsaufgaben: In der einen Klassifikationsaufgabe soll aufgrund der phonetisch gemessenen Werte für den spektralen Abfall entschieden werden, ob es sich um eine betonte oder unbetonte Silbe handelt (SK$\times$RC $\mapsto$ $\{$betont,unbetont$\}$). In der anderen Klassifikationsaufgabe soll die Entscheidung nur aufgrund der automatisch bestimmten MFCC getroffen werden (MFCC $\mapsto$ $\{$betont,unbetont$\}$). Wenn die Klassifikation aus den MFCC sehr viel schlechter gelingt als die Klassifikation aus den glottalen Parametern, kann man annehmen, daß der spektrale Abfall in den MFCC nicht sehr gut repräsentiert ist. Wenn die Klassifikation gleich oder nahezu gleich gut ist, kann man statt der aufwendig manuell zu ermittelnden glottalen Parameter die leicht automatisch bestimmbaren MFCC für die Vorhersage von Betonung verwenden.

Da mit dem Untersuchungsmaterial von Claßen [Cla96] sowohl phonetisch gemessene Werte des spektralen Abfalls als auch zugehörige Sprachsignale und dadurch extrahierbare MFCC zugänglich sind, können im Prinzip beide Ansätze verfolgt werden. Da die in [Cla96] untersuchte Datenbasis mit 400 Fällen jedoch relativ klein ist, bietet sich die zweite Möglichkeit an. Stellt sich heraus, daß die beiden Klassifikationsaufgaben ähnlich gut gelöst werden, kann leicht an einer größeren Datenmenge ein Klassifikator trainiert werden, da für die große Datenmenge keine aufwendige Extraktion der glottalen Parameter vorgenommen werden muß, sondern die aus dem Lexikon entnommene Wortbetonungsinformation ausreicht. Die Realisierung der zweiten Möglichkeit wird im folgenden beschrieben.

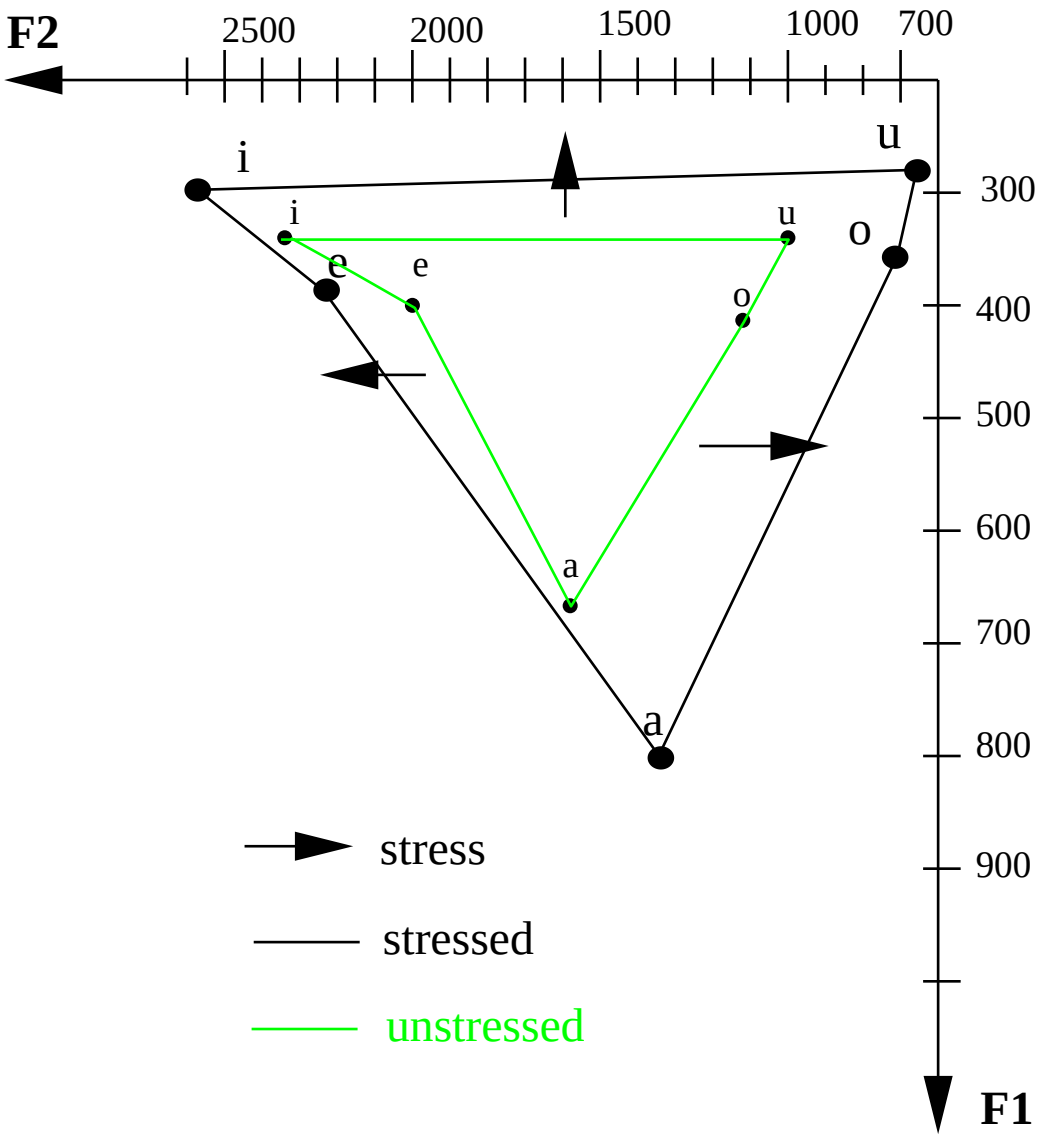


Abbildung 5.3: Die Lage der Vokale des Deutschen im F1-F2 Raum.

	betont	unbetont
aː	Spitalisch	Spitalist
a	Metallisch	Metallist
eː	Klientelisch	Klientelist
ɛ	Kartellisch	Kartellist
iː	Ventilisch	Ventilist
ɪ	Tormentillisch	Tormentillist
oː	Anatolisch	Anatolist
ɔ	Ayatollisch	Ayatollist
uː	Thulisch	Thulist
ʊ	Schatullisch	Schatullist

Tabelle 5.2: Stimuli für die Wortbetonungsuntersuchung.

5.2 Untersuchungsmaterial

Die in [Cla96] untersuchten Sprachdaten entstammen einer Untersuchung der Korrelate der Wortbetonung im Deutschen [JMSC95]. 10 Sprecher äußerten 20 Wörter in 2 Wiederholungen. Es ergeben sich damit knapp 400 Items für die Analyse. (Fehlproduktionen wurden für die weitere Untersuchung ausgeschieden.) Die Wörter sind größtenteils Kunstwörter, die jedoch an lexikalische Formen des Deutschen angelehnt sind. Sie wurden so ausgewählt, daß die untersuchten Vokale in einheitlichem /tVl/ Lautkontext auftreten. Der Betonungsstatus wurde durch Anhängen der Suffixe {-isch} und {-ist} systematisch variiert. Die orthographisch dargebotenen Stimuli sind in Tabelle 5.2 angegeben. Analog zum Vorgehen in [Slu95] wurden in [Cla96] die glottalen Parameter Schiefe und Schließrate aus dem akustischen Signal bestimmt. Für 5 Items war die Berechnung von Schiefe und/oder Schließrate nicht möglich, diese Fälle wurden ausgesondert.

5.3 Schiefe und Schließrate als Korrelate der Betonung

Im ersten Experiment werden die in [Cla96] erhobenen Meßwerte zur Betonungsvorhersage verwendet. Aus den beiden Merkmalen $\mathrm{Skew}^*_{1,3}$ und $\mathrm{RClos}^*_{1,2}$ soll vorhergesagt werden, ob die betrachtete Silbe betont oder unbetont ist. $\mathrm{Skew}^*_{1,3}$ ist der um den Einfluß des ersten und dritten Formanten bereinigte Schätzwert für die Schiefe, in Tabelle 5.1 als SK bezeichnet. $\mathrm{RClos}^*_{1,2}$ ist der um den Einfluß

DataLogic/R		C4.5		Lernsystem
Skew*1,3 RClos*1,2		Skew*1,3 RClos*1,2		Merkmale
A 0.005	68.35	A 0.01	65.8	= richtig klassifiziert
B -	54.43	B 0.005	67.1	= keine Entscheidung
C 0.001	67.09	C 0.001	81.0	= falsch klassifiziert
D 0.005	68.35	D 0.1	59.5	
E 0.025	58.23	E 0.001	68.4	
Σ 0.001	63.29	Σ 0.001	68.36	

Tabelle 5.3: Erzielbare Erkennungsraten aus den phonetisch gemessenen Werten für Schiefe und Schließrate.

der ersten beiden Formanten bereinigte Schätzwert für die Schließrate, in Tabelle 5.1 als RC bezeichnet. Zur Abschätzung der mit diesen Messungen erzielbaren Vorhersagegenauigkeit wurde – wie auch in den folgenden Experimenten – eine 5fache Kreuzvalidierung durchgeführt. Dazu wurden die Daten in fünf Teile geteilt, jeweils vier Teile wurden zum Training verwendet, der verbleibende fünfte diente zum Ermitteln der Erkennungsrate.

Die Ergebnisse der einzelnen Läufe sowie die über die 5 Läufe gemittelten Raten sind in Tabelle 5.3 zusammengefaßt. Der linke Teil zeigt die Ergebnisse, die mit dem auf der Rough-Set-Theorie [Paw91] basierenden Lernsystem DataLogic/R [Dat] erzielt wurden. Im rechten Teil stehen die Resultate, die mit dem System C4.5 erzielt wurden, einem Entscheidungsbaum konstruierenden Verfahren [Qui92], das in Abschnitt 6.2.4 näher beschrieben ist. Beide Lernsysteme arbeiten symbolisch. Sie generieren und/oder-Regeln bzw. Entscheidungsbäume. Man kann also die Plausibilität der gelernten Regeln überprüfen.

Beide maschinellen Lernprogramme erzeugen Klassifikatoren mit signifikanten Erkennungsraten auf den unabhängigen Testdaten. Für jeden Testlauf ist das erreichte Signifikanzniveau in der Tabelle ganz links neben der Laufbezeichnung wiedergegeben. Das Signifikanznizeau bezieht sich auf einen Chi-Quadrat-Unabhängigkeitstest und zeigt, daß die von den Regeln als betont klassifizierten Fälle mit der angegebenen Irrtumswahrscheinlichkeit nicht derselben betont/unbetont-Verteilung unterliegen als die als unbetont klassifizierten. Bei den Testläufen mit den von DataLogic/R konstruierten Klassifikatoren wurden für die Signifikanzuntersuchung die nicht entschiedenen Fälle pessimistisch der „falsch klassifiziert“-Klasse zugeordnet, um die Vorbedingungen des Chi-Quadrat-Tests zu erfüllen. Zur Berechnung des in der untersten Reihe wiedergegebenen Gesamtsignifikanzniveaus wurden die Kontingenztafeln der einzelnen Läufe kumuliert.

Wie Tabelle 5.4 zeigt, sind die glottalen Parameter $\text{Skew}^*_{1,3}$ und $\text{RClos}^*_{1,2}$ allerdings nicht die besten Korrelate der Betonung, sondern die Länge. Die Nukleus-

C4.5		C4.5		Lernsystem
`length` `vowel`		`length`		Merkmale
A 0.001	80.2	A 0.001	76.5	= richtig klassifiziert
B 0.001	75.3	B 0.001	77.8	= keine Entscheidung
C 0.001	76.5	C 0.001	71.6	= falsch klassifiziert
D 0.001	80.2	D 0.001	75.3	
E 0.001	80.5	E 0.001	80.5	
Σ 0.001	78.54	Σ 0.001	76.34	

Tabelle 5.4: Die Länge als wichtigeres Korrelat der Betonung. Erzielbare Erkennungsraten.

vokallängen für die Läufe in Tabelle 5.4 wurden mit dem in Kapitel 3 beschriebenen Verfahren automatisch bestimmt. Für die Ergebnisse auf der linken Seite der Tabelle stand dem Klassifikator sowohl die gemessene Länge als auch die Art des Nukleusvokals zur Verfügung, auf der rechten Seite nur die gemessene Nukleusvokallänge.[1] Wie man sieht, ist die Länge ein besserer Anhaltspunkt zur Vorhersage der Betonung als die ermittelten glottalen Parameter, unabhängig davon, ob man dem Algorithmus die Möglichkeit gibt, intrinsische Faktoren durch Offenlegung der Vokalidentität zu berücksichtigen. Gibt man die nur in kontrollierten Laborexperimenten durchführbare strikte Festlegung des Kontextes auf, so ist zu erwarten, daß die Länge an Nützlichkeit für die Betonungsvorhersage verliert. Länge korreliert neben der Betonung auch mit Gespanntheit (tenseness), Silbenstruktur (die Vokallänge in CV-Silben ist größer als die Vokallänge in CVC-Silben und diese größer als die Vokallänge in CVCC-Silben) und Position in der Äußerung (pre-final lengthening).

5.4 MFCC als Korrelat der Betonung

MFCC werden aus dem Sprachsignal in festen Zeitabschnitten berechnet, typischerweise alle 10 ms. Da die Betonung ein Attribut der Silbe ist und sich die Beeinflussung der glottalen Parameter bei Vokalen besonders stark bemerkbar macht, wird für die Untersuchung der MFCC als Korrelat der Betonung ein Merkmalbündel verwendet, das der Mitte[2] des durch automatische Segmentierung lokalisierten Nukleusvokals entnommen ist. Dabei wird angenommen, daß

[1] Hier und bei den meisten der folgenden Untersuchungen wird wegen der geringen Unterschiede beim ersten Experiment nur noch eines der Lernprogramme eingesetzt, und zwar das dort geringfügig besser abschneidende C4.5

[2] Mitte wird hier als zeitliche Mitte aufgefaßt, bezeichnet also den Zeitpunkt, der vom Beginn und vom Ende des Vokals gleich weit entfernt ist. In der Phonetik wird die Vokalmitte auch als der Zeitpunkt definiert, bei dem F_1 maximal ist, da dort die größte Öffnung des Vokaltrakts (und

DataLogic/R		C4.5		Lernsystem
MFCC1-12		MFCC1-12		Merkmale
A 0.005	66.67	A 0.001	75.3	= richtig klassifiziert
B -	50.62	B 0.1	60.5	= keine Entscheidung
C 0.005	62.96	C -	56.8	= falsch klassifiziert
D 0.025	62.96	D 0.025	63.0	
E 0.005	66.23	E 0.01	64.9	
Σ 0.001	61.888	Σ 0.001	64.1	

Tabelle 5.5: Erzielbare Erkennungsraten für die betont/unbetont Entscheidung aus normalisierten MFCC reichen nahezu an die Resultate der phonetisch gemessenen Werte heran.

die Segmentierung in der Mitte des automatisch gefundenen Vokalsegments am zuverlässigsten und die Beeinflussung durch benachbarte Phoneme am geringsten ist. Die MFCC werden nicht ausschließlich vom Betonungsstatus beeinflußt, sondern in weitaus stärkerem Maße von der Formantenstruktur und Anregungsart des zugrundeliegenden Phonems – schließlich werden die MFCC ja in der Spracherkennung zur Differenzierung der Phoneme verwendet. Die Ausprägung der MFCC bei den verschiedenen Vokalen interessiert hier nicht, es wird jedoch unterstellt, daß die Prominenz die MFCC unabhängig von der Vokalart in gleicher Weise beeinflußt. Um die Beeinflussung durch die Vokalart auszugleichen oder wenigstens zu mindern, wird ein vokalabhängiger Erwartungswert (unabhängig vom Sprecher und der Betonung) von den MFCC abgezogen, bevor sie in das Lernsystem eingegeben werden. Unterstellt man, daß das der Mitte des Vokals entnommene Merkmalbündel vom mittleren Zustand der Hidden-Markov-Modelle aus Abschnitt 3 ausgegeben wird und sich die anderen von diesem Zustand abgedeckten Merkmalbündel nicht sehr vom herausgenommenen unterscheiden, so muß der Erwartungswert für die MFCC nicht erst bestimmt werden, da er als Mittelwert der (einfachen) Normalverteilung der Wahrscheinlichkeitsdichtefunktion in den kontinuierlichen HMM enthalten ist.[3] Die Erwartungswerte wurden beim HMM-Training an den Sprachdaten von 12 verschiedenen Sprechern abgeschätzt.[4] Durch dieses Vorgehen erfolgt eine Normierung auf ein sprecherunabhängiges, typisches a, eɪ, ɪ, etc. Die resultierenden 12 Merkmale je Silbe ergeben die in Tabelle 5.5 wiedergegebenen Erkennungsraten. Die vorherzusagenden Klassen waren wieder „betont“ und „unbetont“.

Aus einem Vergleich zwischen den Erkennungsraten von Tabelle 5.5 und Tabel-

somit die höchste Sonorität) erreicht ist. Da die Umsetzung dieser Definition zwar machbar, aber sehr aufwendig ist, wird sie hier nicht verwendet.

[3]Von einer Berücksichtigung der Varianz bei der Normierung wurde in dieser Untersuchung abgesehen.

[4]Sprecher des HMM-Trainings sind im Untersuchungsmaterial nicht vertreten.

C4.5		Lernsystem
MFCC1-8		Merkmale
A 0.001	69.1	= richtig klassifiziert
B 0.005	66.7	= keine Entscheidung
C 0.05	61.7	= falsch klassifiziert
D 0.005	67.9	
E 0.1	61.0	
Σ 0.001	65.28	

Tabelle 5.6: Die Beschränkung auf 8 MFCC niedriger Ordnung verbessert die Vorhersagegenauigkeit leicht.

le 5.3 geht hervor, daß die automatisch gefundenen spektralen Merkmale nahezu genauso aussagekräftig für die Betonungsvorhersage sind wie die aus dem akustischen Signal manuell gemessenen glottalen Parameter. Betonung ist also tatsächlich auch in den MFCC codiert. Da die Kepstralkoeffizienten höherer Ordnung weniger die globale Form des Spektrums, sondern eher die „Welligkeit“ des Spektrums beschreiben, ist zu vermuten, daß ein Verzicht auf die höherwertigen Koeffizienten die Erkennungsraten nicht verschlechtert und sogar – wegen der nur begrenzt verfügbaren Trainingsmenge – verbessern könnte. Nimmt man dem Lernprogramm nach und nach höherwertige Koeffizienten weg, kann man in etwa gleichbleibende Erkennungsraten beobachten, bis noch fünf Koeffizienten verbleiben. Nimmt man von unten beginnend sukzessive Merkmale weg, so kann man etwa sechs Merkmale entfernen, ohne die Erkennungsraten deutlich zu verschlechtern. Das beste Ergebnis wird mit einem moderaten Kürzen am oberen Ende erzielt. Die Erkennungsraten für die MFCC 1-8 liegen, wie Tabelle 5.6 zeigt, leicht über den Ergebnissen mit allen 12 Merkmalen.

Diese Ergebnisse deuten darauf hin, daß man den spektralen Abfall nicht wie zunächst angenommen, an wenigen MFCC (niedriger Ordnung) festmachen kann. Auch eine Inspektion der produzierten Regeln sowohl von C4.5 als auch von DataLogic/R widerspricht der anfänglichen Vermutung. Beispielsweise wählte DataLogic/R die Merkmale MFCC4, MFCC6, und MFCC12 in jedem der fünf Läufe aus, MFCC7 und MFCC3 in zweien sowie MFCC5 und MFCC8 je einmal.

Um ein wenig näher zu beleuchten, welche MFCC geeignete Kandidaten sein könnten, wird als nächstes die Vorhersagegenauigkeit jeweils eines einzelnen Koeffizienten berechnet. In den in Tabelle 5.7 zusammengefassten Einzeltabellen ist in der linken Spalte der Anteil korrekter Vorhersagen aufgeführt, in der rechten der Mißklassifikationsanteil. Alle Experimente wurden mit C4.5 durchgeführt. Auch diese Lernläufe machen deutlich, daß die Betonung nicht aus einem einzelnen MFCC abgeleitet werden kann. Die Merkmale MFCC3, MFCC4, MFCC5, MFCC11 und MFCC12 schneiden dabei noch am besten ab.

MFCC1		MFCC2		MFCC3		MFCC4	
48.1	51.9	48.1	51.9	54.3	45.7	44.4	55.6
45.7	54.3	53.1	46.9	55.6	44.4	51.9	48.1
54.3	45.7	55.6	44.4	49.4	50.6	63.0	37.0
54.3	45.7	49.4	50.6	55.6	44.4	60.5	39.5
48.1	51.9	49.4	50.6	55.8	44.2	62.3	37.7
50.1	49.9	51.1	48.9	54.1	45.9	56.4	43.6
MFCC5		MFCC6		MFCC7		MFCC8	
56.8	43.2	54.3	45.7	65.4	34.6	63.0	37.0
53.1	46.9	49.4	50.6	55.6	44.4	45.7	54.3
56.8	43.2	54.3	45.7	40.7	59.3	50.6	49.4
46.9	53.1	44.4	55.6	54.3	45.7	54.3	45.7
57.1	42.9	57.1	42.9	44.2	55.8	48.1	51.9
54.1	45.9	51.9	48.1	52.0	48.0	52.3	47.7
MFCC9		MFCC10		MFCC11		MFCC12	
53.1	46.9	53.1	46.9	53.1	46.9	54.3	45.7
42.0	58.0	45.7	54.3	48.1	51.9	48.1	51.9
53.1	46.9	54.3	45.7	53.1	46.9	59.3	40.7
40.7	59.3	51.9	48.1	64.2	35.8	59.3	40.7
51.9	48.1	55.8	44.2	50.6	49.4	54.5	45.5
48.1	51.9	52.2	47.8	53.8	46.2	55.1	44.9

Tabelle 5.7: Die Erkennungsraten der einzelnen Koeffizienten.

C4.5		Lernsystem
MFCC 3,4,5,11,12		Merkmale
A 0.025	64.2	= richtig klassifiziert
B 0.05	63.0	= keine Entscheidung
C -	54.3	= falsch klassifiziert
D -	56.8	
E 0.001	71.4	
Σ 0.001	61.94	

Tabelle 5.8: Die Beschränkung auf die 5 einzeln am besten abschneidenden MFCC verbessert die Vorhersagegenauigkeit nicht.

Da für die Experimente nur sehr wenige Trainingsbeispiele verwendet wurden, kann die Vorhersagegenauigkeit möglicherweise durch die gezielte Auswahl weniger, möglichst aussagekräftiger Merkmale verbessert werden. Daher wurde ein Lauf mit den fünf genannten Merkmalen durchgeführt. Wie Tabelle 5.8 zeigt, ist diese Beschränkung jedoch nicht hilfreich und erbringt keine besseren Resultate. Die Auswahl geeigneter Attribute ist natürlich methodologisch nicht unproblematisch, da für den Trainingsablauf an der Teststichprobe gemessene Ergebnisse entscheidend sind. Dieses Vorgehen birgt die Gefahr in sich, lediglich speziell für diese Teststichprobe geeignete Merkmale aufzuzeigen. In diesem Fall ist es insofern gerechtfertigt, als es zum besseren Verständnis der Beeinflussung einzelner Merkmalsgruppen zur Klassifikationsaufgabe hätte beitragen können. Die Ergebnisse zeigen jedoch, daß es nicht einzelne Bereiche der MFCC sind (also beispielsweise die Koeffizienten 3-5, die mit der Ausprägung der Vokalqualität zusammenhängen könnten, und die Koeffizienten 11 und 12, die von der Struktur der Harmonischen beeinflußt sein könnten), die das Hören einer Betonung (mit)verursachen.

Unproblematisch aus methodologischer Sicht ist eine Transformation des Merkmalsraumes, die auch zur Reduktion der Dimensionalität verwendet werden kann. Ein häufig angewandtes Verfahren dazu ist die Hauptachsentransformation (Principal component analysis). Sie liefert zunächst eine gleich große Anzahl an transformierten Merkmalen, aber auch eine Gewichtung dieser transformierten Merkmale (durch die im Verlauf des Verfahrens berechneten Eigenwerte). Das Standardverfahren zur Reduktion der Dimensionalität besteht darin, die transformierten Merkmale mit kleinem Eigenwert wegzulassen, da sie weniger zur Verteilung der Fälle im Merkmalsraum beitragen. Die in Tabelle 5.9 dargestellten Ergebnisse von derart durchgeführten Experimenten[5] zeigt, daß eine (automatisch bestimmte) Merkmalsraumtransformation die Vorhersagegenauigkeit verbessert. Die Merk-

[5] Zur Berechnung der Hauptachsentransformation wurde das Programm pca von Yoshiro Miyata und Andreas Stolcke verwendet.

C4.5		C4.5		Lernsystem
PC1-8(MFCC1-12)		PC1-12(MFCC1-12)		Merkmale
A 0.001	75.3	A 0.001	72.8	= richtig klassifiziert
B 0.1	59.3	B 0.025	63.0	= keine Entscheidung
C 0.05	61.7	C 0.1	59.3	= falsch klassifiziert
D 0.025	64.2	D 0.1	59.3	
E 0.001	68.8	E 0.001	70.1	
Σ 0.001	65.86	Σ 0.001	64.9	

Tabelle 5.9: Sowohl eine Hauptachsentransformation mit als auch ohne Dimensionalitätsreduktion bringt eine leichte Verbesserung der Vorhersagegenauigkeit.

malstransformation ohne Reduktion der Dimensionalität, in Tabelle 5.9 rechts wiedergegeben, verbessert die Vorhersagegenauigkeit für einen 12-dimensionalen Merkmalsraum leicht von 64.1% auf 64.9%, nach Reduktion auf 8 Merkmale im Vergleich zu den 8 niederwertigen MFCC von 65.28% auf 65.86%.

5.5 Training und Test an einem größeren Korpus

Die Ergebnisse des vorherigen Abschnitts lassen sich nicht direkt in ein System übertragen, das beliebige Daten verarbeiten kann, da in den Daten aus [Cla96] nicht alle Vokale des Deutschen vertreten sind und zudem sämtliche Daten auf einen einzigen allophonischen Kontext eingeschränkt sind. Man kann deshalb die an der „Laborsprache“ gelernten Klassifikatoren nicht unmittelbar auf frei gesprochene Sprache übertragen. Man kann jedoch in diesem Fall das ganze Lernverfahren auf ein größeres Korpus übertragen, da als zu lernende Kategorie der lexikalische Wortakzent und nicht die extrahierten Parameter Schiefe und Schließrate verwendet wurden, um dann die Nützlichkeit des Verfahrens für Sprache in einer natürlicheren Kommunikationssituation zu bewerten. Es werden zwei verschiedene Korpora verwendet, die im Anhang eingehender beschrieben sind. Ein ca. vier Stunden Sprache umfassendes Korpus eines einzelnen Sprechers mit einer Lesung von Goethes „Die Leiden des jungen Werther“ [Goe95, Rap96], sowie ein Korpus, das verschiedene Sprecher und Textarten umfaßt (Radionachrichten im Deutschlandfunk vom 28.07.95 und vom 21.11.95, jeweils zwei Sprecher; „Das Dicke Kind“, sechs Sprecher; „Die Buttergeschichte“, drei Sprecher). Aufgrund der Aufbereitungsweise der Daten und der Textarten sind die Verteilungen der als lexikalisch wortbetont und unbetont gekennzeichneten Silben in den beiden Korpora unterschiedlich. Bei der Literaturlesung tragen etwa 58.6%, bei dem gemischten Korpus etwa 42.8% der Silben eine Markierung für lexikalischen Wortakzent. Speziell ist zu beachten, daß nicht alle Silben, die einen lexikalischen

Gemischtes Korpus		Literaturlesung		Korpus
MFCC1-12		MFCC1-12		Merkmale
A 0.001	55.2	A 0.001	69.3	= richtig klassifiziert
B 0.001	61.0	B 0.001	69.2	= keine Entscheidung
C 0.001	64.4	C 0.001	69.4	= falsch klassifiziert
D 0.001	64.7	D 0.001	68.9	
E 0.001	61.0	E 0.001	69.1	
Σ 0.001	61.26	Σ 0.001	69.18	

Tabelle 5.10: Vorhersage der Wortakzentannotation aus dem Mel-frequenz-skalierten Kepstrum.

Gemischtes Korpus		Literaturlesung		Lernsystem
MFCC1-8		MFCC1-8		Merkmale
A 0.001	56.5	A 0.001	69.4	= richtig klassifiziert
B 0.001	62.2	B 0.001	69.0	= keine Entscheidung
C 0.001	62.5	C 0.001	68.4	= falsch klassifiziert
D 0.001	63.9	D 0.001	68.1	
E 0.001	59.7	E 0.001	69.3	
Σ 0.001	60.96	Σ 0.001	68.84	

Tabelle 5.11: Die Beschränkung auf 8 MFCC niedriger Ordnung verbessert an den Korpusdaten die Vorhersagegenauigkeit nicht.

Wortakzent tragen müßten, auch durch einen solchen gekennzeichnet sind, da nicht im Lexikon aufgefundene Wörter keine Wortakzentannotierung tragen. Bei der Literaturlesung sind etwa 5.2% der Wörter nicht annotiert, bei dem gemischten Korpus sind dies 8.5%, jeweils bezogen auf die Token-Häufigkeit (bezogen auf die lexikalische Häufigkeit 15.4% bzw. 16.6%). Die Textart beeinflußt die Verteilung insofern, als speziell bei den Nachrichten sehr viele Eigennamen und mitunter sehr lange Komposita auftreten, die entweder gar keine Wortakzentannotierung tragen, da sie nicht im Lexikon stehen, oder lediglich aufgrund ihrer überdurchschnittlichen Länge das Verhältnis zwischen unbetont und betont markierten Silben beeinflussen.

Die Erkennungsraten, die unter Verwendung aller MFCC an den Korpora erzielt werden, sind in Tabelle 5.10 dargestellt. Da bei den Korpora mit 63743 bzw. 23307 Silben ausreichend Trainingsmaterial vorhanden ist, erbringt die Reduzierung der Merkmale auf ausgesuchte MFCC keine Verbesserung (Tabelle 5.11, Tabelle 5.12). Eine analog zum Vorgehen oben durchgeführte Principal component analysis erhöht die Vorhersagegenauigkeit wie bei den Labordaten wieder leicht, wie Tabelle 5.13 zeigt. Gegenüber den Labordaten bringt eine Reduktion der Dimensionalität (wie bei den anderen Läufen an den Korpora) aufgrund der großen Datenanzahl keine Verbesserungen.

Gemischtes Korpus		Literaturlesung		Lernsystem
MFCC 3,4,5,11,12		MFCC 3,4,5,11,12		Merkmale
A 0.001	58.4	A 0.001	63.5	= richtig klassifiziert
B 0.005	56.1	B 0.001	63.6	= keine Entscheidung
C 0.001	59.8	C 0.001	63.1	= falsch klassifiziert
D 0.001	60.7	D 0.001	62.8	
E 0.001	58.0	E 0.001	62.5	
Σ 0.001	58.6	Σ 0.001	63.1	

Tabelle 5.12: Die Beschränkung auf ausgesuchte MFCC verbessert die Vorhersagegenauigkeit ebenfalls nicht.

Gemischtes Korpus		Literaturlesung		Lernsystem
PC1-12(MFCC 1-12)		PC1-12(MFCC 1-12)		Merkmale
A 0.001	59.1	A 0.001	70.7	= richtig klassifiziert
B 0.001	63.1	B 0.001	70.1	= keine Entscheidung
C 0.001	64.5	C 0.001	70.3	= falsch klassifiziert
D 0.001	64.3	D 0.001	69.8	
E 0.001	60.7	E 0.001	71.9	
Σ 0.001	62.34	Σ 0.001	70.56	
Gemischtes Korpus		Literaturlesung		Lernsystem
PC1-8(MFCC 1-12)		PC1-8(MFCC 1-12)		Merkmale
A 0.001	58.4	A 0.001	70.7	= richtig klassifiziert
B 0.001	62.0	B 0.001	70.4	= keine Entscheidung
C 0.001	63.8	C 0.001	70.3	= falsch klassifiziert
D 0.001	63.1	D 0.001	69.8	
E 0.001	60.9	E 0.001	70.8	
Σ 0.001	61.64	Σ 0.001	70.40	

Tabelle 5.13: Die Ergebnisse unter Verwendung der Principal component analysis.

Labordaten manuell		Literaturlesung	
Skew*1,3 RClos*1,2		PC1-12(MFCC1-12)	
Σ 0.001	68.36	Σ 0.001	70.56
Verteilung	50.00	Verteilung	58.64
Labordaten automatisch		**Gemischtes Korpus**	
PC1-8(MFCC1-12)		PC1-12(MFCC1-12)	
Σ 0.001	65.86	Σ 0.001	62.34
Verteilung	50.00	Verteilung	57.21

Tabelle 5.14: Zusammenfassung der jeweils besten Ergebnisse für die verschiedenen Korpora.

5.6 Zusammenfassung

Die jeweils besten Vorhersagegenauigkeiten für die unterschiedlichen Datensätze sind in Tabelle 5.14 dargestellt. Da das Verhältnis zwischen betont und unbetont markierten Silben bei den beiden Korpora sowie den Labordaten aus Abschnitt 5.3 unterschiedlich ist, kann man die Erkennungsraten nicht unmittelbar miteinander vergleichen. Beim Betrachten der Ergebnisse fällt jedoch auf, daß die Ergebnisse an dem nur einen Sprecher umfassenden Werther-Korpus deutlich besser ausfallen als bei dem Korpus mit mehreren Sprechern. Dies gilt auch noch, wenn man die nicht balancierte Verteilung der Korpora berücksichtigt. Der triviale Klassifikator „jede Silbe ist unbetont“ liegt beim Mehrsprecherkorpus nicht wesentlich unter den erzielten Erkennungsraten (62.34 vs. 57.21%). Beim Einzelsprecherkorpus, der Werther-Lesung, liegen die Erkennungsraten deutlich über denen des dort geeigneteren trivialen Klassifikator „jede Silbe ist betont“ (70.56 vs. 58.64%). Für die Labordaten sind die besten Ergebnisse für MFCC-basierte Merkmale mit 65.86% deutlich von den mit den trivialen Klassifikatoren erzielbaren 50% abgesetzt und nahe an den mit 68.36 erzielten Erkennungsraten für manuell aus dem Sprachsignal extrahierte glottale Parameter.

Der unterschiedliche Erfolg der Wortakzentvorhersage aus den spektralen Merkmalen bei den Korpora kann verschiedene Ursachen haben. Zum einen könnte es sein, daß der den Werther vortragende Schauspieler, ein geübter Sprecher, deutlicher zwischen betonten und unbetonten Silben bei der Produktion unterscheidet, daß also dieses Korpus leichter zu erkennen ist. In gleicher Weise wirkt sich auch die geringere OOV-Rate beim Werther-Korpus aus; da dadurch weniger Silben fälschlich nicht als wortbetont markiert sind, müßte die wortbetont/unbetont-Unterscheidung leichter gelingen. Besonders nachteilig für das hauptsächlich aus Radionachrichten zusammengesetzte Gemischtsprecherkorpus müßte sich auch auswirken, daß gerade Eigennamen nicht im Lexikon auftreten, somit also gerade Wortakzentsilben mit großer Prominenz (Eigennamen tragen sehr häufig

einen Satzakzent) nicht mit einem Wortakzent markiert sind. Zudem beinhalten die aufgenommenen Radionachrichten eine große Anzahl bosnischer Ortsnamen, die von den Nachrichtensprechern sehr unterschiedlich und unregelmäßig betont werden. Als weitere Ursache könnte angeführt werden, daß die Normierung der MFCC, die vom Vokal abstrahieren soll, für den Einzelsprecher wahrscheinlich überzeugender gelingt. Die Statistiken über die MFCC–Ausprägungen der einzelnen Vokale wurden beim Werther-Korpus den HMMs entnommen, die auf diesen Sprecher nachtrainiert waren [Rap96]. Für das gemischte Korpus wurden sprecherunabhängig trainierte HMMs als Grundlage der Statistik verwendet, denn das vorhandene Sprachmaterial war für manche der Sprecher zu gering, um ein sinnvolles Nachtraining durchzuführen. Auch bei den Untersuchungen der Labordaten in den vorangehenden Abschnitten wurde auf sprecherunabhängige HMM-Emissions-Verteilungen normalisiert. Möglicherweise hat hier jedoch der eingeschränkte phonetische Kontext bessere Ergebnisse ermöglicht.

Wenngleich die Erkennungsraten nicht ganz an die Erkennungsraten für die aufwendig gemessenen und berechneten Werte für Schiefe und Schließrate heranreichen, so erzielt das in diesem Kapitel beschriebene Verfahren eine *automatische* Klassifikation der (Wort-)Betonung, die durch die spektrale Struktur ausgedrückt wird. Gegenüber der lexikalischen Wortakzentmarkierung besitzt sie den Vorteil, daß sie auch bei nicht im Lexikon angetroffenen Einträgen genauso funktioniert wie bei Lexikoneinträgen. Bessere Vorhersagen werden erzielt, wenn die Statistik über die Vokale aus auf den Sprecher adaptierten HMMs gezogen wird, wie das Einzelsprecherkorpus zeigt. Da zum Lernen nur die lexikalische Wortakzentinformation nötig ist, ist der Aufwand dafür gering, setzt aber einen gewissen Mindestdatenumfang voraus.

Zwei Ansätze zur Verbesserung der Klassifikation wurden bewußt nicht verfolgt. Erstens könnten weitere Merkmale, etwa Silben- oder Nukleuslänge, die Vorhersagegenauigkeit verbessern. Dieser Ansatz wurde nicht verfolgt, um eine saubere Trennung der einzelnen Faktoren bei der im nächsten Kapitel beschriebenen Vorhersage der Prosodiebeschreibung zu erreichen. Zweitens würde die Beachtung der Randbedingung „jedes Wort hat genau eine wortbetonte Silbe“ dramatische Verbesserungen erbringen, da dadurch z.B. alle einsilbigen Wörter gezwungenermaßen korrekt vorhergesagt würden. Auch bei Wörtern, bei denen der Klassifikator gar keine oder mehrere Silben als wortbetont markieren möchte, sind Verbesserungen zu erwarten. Auch könnte hier erfolgversprechend phonologisches Wissen bei unklaren Fällen weiterhelfen, etwa daß in monomorphemischen Wörtern der Wortakzent auf eine der letzten drei Silben fällt [Jes94, Rap95], bei Komposita in der Regel der vordere Teil die Hauptbetonung trägt [Jes94], oder die Wortbetonung zum Großteil aus der Silbenstruktur vorhergesagt werden kann [Jes94, Rap95]. Diese denkbaren Verbesserungen wurden ebenfalls nicht verfolgt,

da ein Merkmal gefunden werden sollte, das lediglich die durch Intensität und Stimmqualität hervorgerufene phonetische Prominenz beschreiben soll, die durch spektrale Charakteristik ausgedrückt wird, und von phonologischen und lexikalischen Faktoren frei sein soll.

Kapitel 6

Automatische Bestimmung der phonologischen Intonationsbeschreibung

In diesem Kapitel wird beschrieben, wie aus automatisch gewonnenen phonetischen Merkmalen (Kapitel 3, 4, 5) eine phonologische Intonationsbeschreibung maschinell bestimmt werden kann. Für den Übergang von der Phonetik zur Phonologie der Intonation werden Korpus-basierte Verfahren eingesetzt, das heißt, aus einem Korpus von Sprachdaten, für die sowohl die phonetische als auch die phonologische Beschreibung bekannt ist, werden Verfahren abgeleitet, die aus der phonetischen Beschreibung die phonologische Beschreibung bestimmen können. Diese an den Korpusdaten gelernten Verfahren lassen sich dann auf weitere Daten mit noch unbekannter phonologischer Beschreibung anwenden. Da die phonetische Beschreibung mit den in den Kapiteln 3 bis 5 beschriebenen Verfahren ebenfalls automatisch hergeleitet werden kann, erhält man so eine vollautomatisch gewonnene phonologische Beschreibung der Intonation aus dem Sprachsignal.

In der Statistik, dem maschinellen Lernen und in der Mustererkennung sind viele Lernverfahren entwickelt worden, die für diese Aufgabe prinzipiell in Frage kommen. Die einzelnen Verfahren sind meist theoretisch gut erforscht und haben zum Teil auch mehr oder weniger große Erfolge in praktischen Anwendungen aufzuweisen. Es ist zumeist nicht von vornherein eindeutig klar, welches Lernverfahren für eine bestimmte Lernaufgabe die besten Resultate erbringt. Ein Aspekt, der für die Auswahl eines Lernverfahrens wichtig ist, ist die Größe der verfügbaren Trainingsdatenmenge, da die einzelnen Verfahren unterschiedliche Anforderungen an Umfang und Beschaffenheit der Lerndaten stellen. In der hier vorliegenden Aufgabenstellung erfordert die Bereitstellung geeigneter Trainingsdaten die manuelle

prosodische Annotation umfangreicher Sprachdaten und ist damit ein sehr zeitaufwendiger und kostspieliger Prozeß. Das manuelle prosodische Etikettieren von etwas über einer Stunde Radionachrichten nimmt beispielsweise mehrere Wochen in Anspruch und setzt beim Annotierer Erfahrung mit dem verwendeten Etikettierungssystem voraus. Es können also nicht beliebig große Datenmengen gesammelt werden, nur weil das intendierte Lernverfahren mit wenigen Daten nicht die gewünschten Ergebnisse bringt.

Eng verknüpft mit der verfügbaren Menge an Trainingsdaten ist die Entscheidung, ob ein sprecherabhängiges oder ein sprecherunabhängiges Prosodieerkennungssystem entwickelt werden soll. Die sprecherabhängige Erkennung, der zumeist einfachere Fall, verlangt nach Trainingsdaten des Sprechers, dessen Äußerungen vom System etikettiert werden sollen. Für eine sprecherunabhängige Erkennung sollten dem Lernverfahren Trainingsdaten von möglichst vielen Sprechern und Sprecherinnen zugänglich sein. Generell kann man sagen, daß für eine sprecherunabhängige Etikettierung deutlich mehr Sprachmaterial bereitgestellt werden muß. Es ist aus diesem Grunde ökonomisch, sich zunächst an der sprecherabhängigen Erkennung zu versuchen und anschließend die Erweiterung zur universell einsetzbaren sprecherunabhängigen Etikettierung vorzunehmen. Somit können während der Beschränkung auf sprecherabhängige Erkennung auch sprecherspezifische Charakteristika wie Stimmlage und Sprechtempo als nahezu konstant und damit vernachlässigbar angesehen werden.

Der erste Abschnitt des Kapitels enthält Vorüberlegungen zur Lösung des Problems und diskutiert die bestehenden Wahlmöglichkeiten. Zunächst werden mögliche Lernverfahren zur Bestimmung der Klassifikatoren aus den Trainingsdaten vorgestellt. Anschließend werden Merkmale diskutiert, die sich automatisch extrahieren lassen und somit den Lernverfahren als Eingabe dienen können. Die Merkmale sind dabei in der Reihenfolge der in Kapitel 2 genannten Korrelaten der Betonung, Grundfrequenz, zeitliche Strukturierung, Intensität und Stimmqualität, abgehandelt. Schließlich werden zwei Arbeiten zu einer ähnlichen Aufgabenstellung betrachtet. Im zweiten Abschnitt werden die Resultate von Voruntersuchungen zur Auswahl eines geeigneten Lernverfahrens und zur Auswahl geeigneter Merkmalssätze, also geeigneter phonetischer Beschreibungsformen, vorgestellt. Vier verschiedene Ansätze werden an einem Korpus, das aus Sprachmaterial eines einzigen Sprechers besteht, auf ihre praktische Anwendbarkeit und die mit ihnen erzielbare Vorhersagegüte hin untersucht. Im dritten Abschnitt wird für den besten der vier untersuchten Ansätze ermittelt, welche Kombination von automatisch bestimmten, phonetischen Merkmalen bei der Vorhersage der phonologischen Intonationsbeschreibung die besten Ergebnisse liefert. Auch diese Untersuchungen werden an dem Einzelsprecherkorpus durchgeführt. Der vierte Abschnitt schließlich befaßt sich mit der Erweiterung des Prosodieerkennungssystems zur

sprecherunabhängigen Erkennung. Weiterhin werden Untersuchungen angestellt, wie die dazu notwendigen weiteren Daten zusammengesetzt sein sollten.

6.1 Vorüberlegungen

6.1.1 Überblick über Lernverfahren

Für Klassifikationsaufgaben bieten sich eine Vielzahl von Lernverfahren an. Diese lassen sich nach [LS95] in verschiedene Klassen oder Paradigmen einteilen, wobei die Übergänge durchaus fließend sein können. Die Einteilung geschieht hier aus Sicht des maschinellen Lernens, eines Teilgebiets der Künstlichen Intelligenz (KI). Andere Disziplinen wie Mustererkennung oder Statistik, die sich ebenfalls mit dieser Thematik befassen, würden möglicherweise andere Kriterien und eine teils abweichende Terminologie verwenden.

Im Grenzgebiet zwischen Künstlicher Intelligenz, Kognitionswissenschaft und Mustererkennung haben sich die Künstlichen Neuronalen Netze (*artificial neural networks*, ANN) etabliert. Künstliche Neuronale Netze simulieren Verschaltungen von Nervenzellen (Neuronen). Die künstlichen Neuronen besitzen gewichtete Eingänge und einen Ausgang; eine Aktivierungsfunktion bestimmt, wieviel Aktivierung von den Eingängen zum Ausgang weitergeleitet wird. Der Ausgang eines Neurons kann auf den Eingang eines weiteren Neuron geschaltet sein oder einen Ausgang des gesamten Netzwerks bilden. Klassifikation erfolgt mit ANN, indem die Merkmalswerte des zu klassifizierenden Objekts an die Eingänge des Netzes angelegt werden; die Aktivierung an den Ausgängen legt dann die Klasse des Objekts fest. Dazu müssen sowohl die Merkmalswerte als auch die verschiedenen Klassen numerisch kodiert werden. ANN werden für eine bestimmte Klassifikationsaufgabe trainiert, indem Beispiel-Eingabemuster an die Eingänge des ANN angelegt und die Gewichte der Neuronen dann algorithmisch so modifiziert werden, daß vom ANN erzeugte Ausgabemuster möglichst gut mit der korrekten Klasse der Beispiele übereinstimmen. Als Vorteile der Anwendung von Neuronalen Netzen werden häufig bessere Erkennungsraten sowie Robustheit gegen den Ausfall einzelner Merkmale angegeben. Als Nachteile gelten die schwere Verständlichkeit der gelernten Klassifikatoren, die durch die Gewichte der Neuronen repräsentiert sind, die Notwendigkeit großer Trainingsdatenmengen sowie lange Trainingszeiten.

Bei den genetischen Algorithmen wird der Lernprozeß in Analogie zur biologischen Evolution beschrieben. Eine sogenannte Population von Regeln, die als Ketten (Strings) von binären oder booleschen Merkmalen repräsentiert sind, wird in

wiederholten Evaluationszyklen durch Anwendung genetischer Operatoren wie Crossover und Mutation auf ausgewählte Eltern-Regeln weiterentwickelt. Die Auswahl der Regeln, die fortgepflanzt werden, basiert auf einer Funktion die die „fitness“, also den Erfolg der Regeln in der gestellten Klassifikationsaufgabe, bewerten. Zur Klassifikation eines Objekts werden dessen Merkmale mit den Regeln verglichen. In der Forschung zu den genetischen Algorithmen gibt es zwei Hauptströmungen. Im ersten Ansatz kodiert jedes Individuum der Population eine ganze Lösung, also einen Klassifikator, im anderen Ansatz stellt ein Individuum lediglich eine Teillösung dar; die Gesamtlösung ergibt sich durch Kooperation und Wettbewerb der Individuen aus der ganzen Population.

Für die instanzen- oder fallbasierten Methoden ist charakteristisch, daß sie keine übergeordneten Repräsentationen für das zur Klassifikation nötige Wissen erzeugen, sondern spezifische Beispielfälle abspeichern. Die Klassifikation neuer Fälle erfolgt dann durch Auffinden geeigneter Beispielfälle, die sich auf neue Fälle übertragen lassen. Ein verbreitetes, einfaches Verfahren, „k nearest neighbours“ speichert die Trainingsbeispiele mit ihrer Klassifikation ab. Ein neues Objekt wird klassifiziert, indem die k Beispielfälle bestimmt werden, die ihm am ähnlichsten sind; dem neuen Objekt wird die Klasse zugeordnet, die unter den k ähnlichsten Beispielen am häufigsten auftritt. Für den Erfolg dieser Methoden ist die Qualität der verwendeten Ähnlichkeitsmaße und der Zugriffsmechanismen auf die gespeicherten Beispielfälle ausschlaggebend.

In einem vierten Lernparadigma, das man als Regellernen bezeichnen kann, dienen Entscheidungsbäume, Wenn-Dann-Regeln oder ähnliche Repräsentationsstrukturen zur Beschreibung der Zusammenhänge zwischen Merkmalsausprägungen und Klassenzugehörigkeit. Die Klassifikation erfolgt hier, indem die Merkmalsausprägungen eines zu klassifizierenden Objekts verglichen werden mit den Wenn-Seiten der Regeln oder mit Bedingungen, die in den inneren Knoten des Entscheidungsbaumes gespeichert sind. Die Klasseninformation ist in den Dann-Seiten der Regeln oder in den Blättern des Entscheidungsbaumes gespeichert. Gelernt wird üblicherweise durch die Suche im Raum der darstellbaren Repräsentationsstrukturen (also Entscheidungsbäume oder Regelmengen) typischerweise mit einer heuristisch gesteuerten „gierigen“ Suchstrategie. Regellernverfahren lassen sich nach der Ausdrucksfähigkeit der verwendeten Repräsentationsformalismen in zwei Klassen einteilen. Verfahren, die die Ausdrucksfähigkeit von Aussagenlogik bieten, werden als Attribut-Wert-Lerner bezeichnet. Zu diesen gehören Entscheidungsbaumlerner wie C4.5 [Qui92] und das Rough-Set-basierte Verfahren DataLogic [Dat]. Unter dem Begriff Induktive Logikprogrammierung (ILP) faßt man Verfahren zusammen, deren Ausdrucksfähigkeit die Prädikatenlogik oder Teilmengen derselben erreicht. Das Ziel dieser Verfahren ist, aus einer vorgegebenen Menge an positiven und negativen Beispielen unter Verwendung von Hin-

tergrundwissen eine Theorie in Form eines Logikprogramms zu lernen, aus dem die positiven, jedoch nicht die negativen Beispiele abgeleitet werden können. (Daher auch der Name, es werden Logikprogramme aus Beispielen induziert.) Aufgrund ihrer ausdrucksfähigeren und flexibleren Repräsentationsformen sind diese Verfahren den Attribut-Wert-Lernern bei manchen Anwendungen überlegen, etwa wenn strukturierte Objekte wie zum Beispiel chemische Formeln beschrieben werden sollen. Auch die aus Phonemen zusammengesetzten Silben sind als strukturierte Objekte zu betrachten. Der Nachteil dieser Ansätze gegenüber Attribut-Wert-Lernern ist, daß die größeren Suchräume im allgemeinen auch zu wesentlich längeren Lernzeiten führen.

Als weitere Gruppe von Ansätzen zum Klassifikatorlernen kann man die Verfahren anführen, die aus der Wahrscheinlichkeitsrechnung und beschreibenden Statistik abstammen und probabilistische Repräsentationen verwenden. Dazu gehört beispielsweise der Normalverteilungsklassifikator (Gauß-Klassifikator). Dieser Ansatz schätzt für jede Klasse eine Verteilungsfunktion im Merkmalsraum. Durch Anwendung der Bayes'schen Regel wird bei der Klassifikation den zu entscheidenden Fällen dann die gemäß dieser Modellierung wahrscheinlichste Klasse zugeordnet. Auch die bei der Spracherkennung so erfolgreiche statistische Modellierung mit Hidden-Markov-Modellen gehört zu dieser Gruppe von Ansätzen. Ähnlich wie beim Gauß-Klassifikator wird für jede Klasse getrennt ein statistisches Modell bestimmt, das angibt, wie wahrscheinlich eine gegebene Merkmalsraumausprägung für eine bestimmte Klasse ist. Im Gegensatz zum Gaußklassifikator werden bei den Hidden-Markov-Modellen jedoch Folgen von Merkmalsraumausprägungen betrachtet. Zur Klassenbeschreibung dient statt der Normalverteilung beim Gaußklassifikator ein endlicher Automat, dessen Übergänge und Emissionen statistisch beschrieben sind. Beiden Verfahren gemein ist, daß die statistischen Parameter an einer Trainingsstichprobe, also der Beispiels- oder Lernmenge, abgeschätzt werden müssen.

6.1.2 Automatisch extrahierbare Merkmale für die Vorhersage der Intonationsbeschreibung

Sprachgrundfrequenz

Für die Beschreibung der Sprachgrundfrequenz stehen zunächst die nahezu kontinuierlich anfallenden Werte der Sprachgrundfrequenzschätzung zur Verfügung.[1]

[1] Üblicherweise werden Schätzwerte für die Sprachgrundfrequenz zu festen Zeitabständen ausgegeben, typischerweise etwa alle 0.01 Sekunden. Da für jeden Wert der Schätzung üblicherwei-

Die Ausgabe des Grundfrequenzschätzers (Pitch tracker) kann einer Glättung unterzogen werden, verbreitet ist hier besonders die Glättung durch Medianfilterung, da diese einzelne Ausreißer wirkungsvoll unterdrückt [Nö91]. Andere Autoren schlagen eine Tiefpaßfilterung oder eine Beschränkung durch Bandbeschränkte Fourier-Transformationen vor [Leh95].

Zusätzlich zu den kontinuierlich anfallenden Grundfrequenzmeßwerten und den daraus durch Filterung gewonnenen Wertefolgen sind auch an linguistischen Einheiten orientierte Merkmale denkbar. Diese können aus der Grundfrequenzschätzung (auch mit nachfolgender Glättung) durch Mittelwert-, Maximalwert- oder Minimalwertbildung über die zeitliche Ausdehnung einer linguistischen Einheit, vor allem Phoneme und Silben, gewonnen werden. Auch die in Kapitel 4 beschriebene, automatisch erzeugte phonetische Beschreibung des Grundfrequenzverlaufs für eine Silbe ist ein glättendes, auf der Ausgabe des Grundfrequenzschätzers beruhendes, silbensynchrones Merkmalbündel.

Aufgrund der Tatsache, daß unterschiedliche Sprecher wegen der unterschiedlichen physiologischen Konfiguration ihrer an der Sprachproduktion beteiligten Organe sehr unterschiedliche Stimmlagen besitzen können, ergeben Messungen der Grundfrequenz über verschiedene Sprecher hinweg eine — von forensischen Anwendungen abgesehen — ungewünschte Sprecherabhängigkeit. Um diese Abhängigkeit zu verringern, kann man Grundfrequenzmeßwerte relativ zur Stimmlage des Sprechers angeben. Neben der direkten Angabe der Meßwerte in Hz können auch Mel- oder Erb-Skalierungen sowie Halbtonskalen verwendet werden [Nö91], um physiologische Effekte des perzeptiven Systems zu berücksichtigen.

Auch die Phonemzugehörigkeit beeinflußt die Grundfrequenz, und man kann versuchen, diese intrinsischen Faktoren bei der Merkmalsaufbereitung auszugleichen. Der emotionale Zustand eines Sprechers beeinflußt die Stimmlage, ebenso wie die Diskursstruktur [May97] die Höhe von Satzakzenten beeinflußt. Möglich und sinnvoll ist deshalb auch eine Normierung auf das für jede Intonationsphrase unter Umständen unterschiedliche Register [May95, May97, Lad96]. Eine solche Normierung setzt jedoch im allgemeinen die Kenntnis des verwendeten Registers und der Phrasengrenzen voraus.

---

se ein Signalfenster zur Berechnung herangezogen wird, kann prinzipiell zur Umgebung jedes Zeitpunkts ein Grundfrequenzwert berechnet werden. Auch wenn Grundfrequenzmeßwerte nur in diskreten Abständen gewonnen werden, kann man sie sich als kontinuierlich anfallend vorstellen. Zumindest richtet sich der Abstand nicht nach linguistischen oder physiologischen Kategorien, sondern ist allein von der Implementierung und Parametrisierung des Grundfrequenzbestimmungsalgorithmus abhängig.

Insgesamt ergibt sich für die Grundfrequenzmerkmale eine in systematischen Untersuchungen kaum beherrschbare kombinatorische Vielfalt von Merkmalen, die eine möglicherweise nicht optimale Beschränkung erzwingt.

Zeitliche Strukturierung

Unter zeitlicher Strukturierung werden Merkmale zusammengefaßt, die die Dauer linguistischer Einheiten, Dauerverhältnisse und rhythmische Aspekte beschreiben. Mit dem in Kapitel 3 vorgestellten System Alphons können die Dauer von Phonemen und den phonologischen Kategorien Silbe, Silbenonset, Silbennucleus, Silbencoda und Silbenrhyme sowie anderer denkbarer Einheiten automatisch gemessen werden. Da die Phoneme sich in ihrer intrinsischen Länge unterscheiden, ist neben dem absoluten Meßwert unter Umständen auch eine Normalisierung sinnvoll. Meist wird zur Normalisierung das Verhältnis zwischen gemessener und erwarteter Phonem- (Silben-, Onset-, . . .)länge gebildet. Die erwartete Dauer kann unterschiedlich berechnet sein. Im einfachsten Fall wird der statistische Erwartungswert verwendet, bei unterstellter Normalverteilung[2] also die mittlere Länge des Phonems im Korpus (bzw. für die hierarchisch höher stehenden Einheiten z.B. die Summe der mittleren Längen der beteiligten Phoneme). Es gibt jedoch auch Normalisierungen, die die Umgebung einer linguistischen Einheit stärker berücksichtigen, um Beeinflussungen durch Sprechgeschwindigkeit und Dehnung zum Phrasenende hin (final lengthening) zu kompensieren. Neben einer multiplikativen Normalisierung sind auch additive Normalisierungen denkbar, also beispielsweise die Zeitdauer, die ein Phonem länger oder kürzer ist als die erwartete Länge.

Weiterhin können benachbarte linguistische Einheiten zueinander in Beziehung gesetzt werden durch Bildung von Dauerverhältnissen. Denkbar sind u.a. die Anteile von Onset, Nucleus, Rhyme, Coda an der gesamten Silbe. Aber auch feinere Einheiten können verwendet werden, etwa das Verhältnis zwischen Verschluß- und Aspirationsphase von Plosiven im Onset einer Silbe [JMSC95]. Solche feinen Messungen setzen eine Erweiterung der in Kapitel 3 beschriebenen Modellierung um subphonemische Einheiten voraus.

Schließlich fallen unter die Merkmale der zeitlichen Strukturierung rhythmische Merkmale. Geeignete Theorien über die zu erwartenden Momente vorausgesetzt, kann man den Abstand eines tatsächlich eintretenden phonetischen Ereignisses

[2] Man kann für die Dauerverteilung von Phonemen, Silben etc. auch eine Gammaverteilung anstelle einer Normalverteilung annehmen, da diese auch für Verteilungen mit nahe am Nullpunkt liegenden Erwartungswerten nur für positive Werte eine Dichte aufweist. Sie ergibt eine bessere Modellierung besonders kurzer Segmente.

von seinem erwarteten Moment berechnen und diesen Abstand als Merkmal bereitstellen. Als phonetische Ereignisse sind hier Pivots (CV-Übergänge) [DB88], Avents (acoustic events) [MBGH94, MWB95] oder P-Centers (perceptual centers) [Mar81, PM91] zu nennen. Obwohl man für alle Ereignisse geeignete meßbare Korrelate entwickeln kann, ist die Bereitstellung eines darauf basierenden Merkmals schwierig, da Theorien, die einen erwarteten Zeitpunkt angeben, bisher wenig entwickelt sind.

Intensität

Für die Intensität, genauer für deren meßbares Korrelat, die Lautstärke, kann man, wie bei der Grundfrequenzmessung, verschiedene Typen von Merkmalen berechnen, und zwar sowohl kontinuierliche als auch an linguistischen Kategorien orientierte. Als kontinuierlicher Meßwert kann zu jedem Zeitpunkt die momentane Energie geschätzt werden, indem in einer kleinen Umgebung die Absolutwerte der Zeitsignalamplitude integriert und durch die Länge des Fensters geteilt werden. Da die Amplitude des Zeitsignals von äußeren Faktoren wie Mikrophonabstand und Aussteuerung während der Aufnahme abhängig ist, müssen Intensitätsmerkmale generell normalisiert werden. Da man für Lautstärkemessungen eine logarithmische Skalierung verwendet, ist eine solche Normierung additiv vorzunehmen. Wenn man als Bereich für die Lautstärkemessung linguistische Kategorien zugrundelegt, also vor allem Silbe und Phonem, kann man aus den kontinuierlich anfallenden, aus Ausschnitten berechneten Momentanenergiewerten charakterisierende Merkmale berechnen (durch Summation zur Gesamtenergie oder Mittelwertbildung zur Bestimmung der mittleren Energie im Segment). Als weitere Möglichkeit kann man zur Gewinnung der Gesamtenergie einer linguistischen Kategorie auch direkt über die Gesamtlänge der Kategorie integrieren und zur Berechnung der mittleren Energie die Gesamtenergie durch die Länge des betrachteten Segments dividieren. Merkmale für die Gesamtenergie sind problematisch, da sie direkt von der Länge des Segmentes abhängig sind. Korreliert beispielsweise ein bestimmtes linguistisches Phänomen mit der Längung von Segmenten, kann man eine Korrelation ebenfalls mit Gesamtenergiemerkmalen feststellen, auch wenn das betroffene Segment überhaupt nicht lauter produziert wurde. Eine solche Vermischung von Merkmalen ist für die Theoriebildung nicht wünschenswert, kann aber in der praktischen Anwendung unter Umständen nützlicher sein als getrennte Merkmale.

Alle bisher angesprochenen Energiemaße müssen nicht notwendigerweise aus dem gesamten Frequenzbereich des Sprachsignals berechnet sein, sondern können auch für beliebige Ausschnitte des Frequenzspektrums ermittelt werden. Da-

zu wird das Sprachsignal vor der Intensitätsberechnung zum Beispiel mit einem Bandpaßfilter gefiltert. Beliebig viele weitere kombinatorische Möglichkeiten ergeben sich, wenn man aus diesen Energieanteilen in verschiedenen Frequenzbändern Verhältnisse berechnet.

Generell unterliegen Intensitätsmerkmale zahlreichen Einflüssen, die sie schwer verwendbar machen. Zum einen beeinflussen die beteiligten Phoneme die Merkmale drastisch; so ist selbstverständlich ein stimmloser Plosiv mit einer Verschlußphase, in der nahezu Stille herrscht, sowohl in der Gesamt- als auch der mittleren Energie wesentlich schwächer als beispielsweise ein Vokal. Aber auch bei den Vokalen gibt es beträchtliche Lautstärkeunterschiede. Durch unterschiedliche Formantenstruktur werden die Einflüsse bei Merkmalen für eingeschränkte Frequenzbereiche noch stärker. Auch bei der Verwendung von Sprachmaterial, bei dem sich ein Teil des Signalweges der direkten Einflußnahme entzieht, z.B. Radioaufnahmen, sind Intensitätsmerkmale problematisch, da in den Studios häufig sogenannte Kompressoren oder Limiter eingesetzt werden, die den Dynamikbereich der Sprache verzerren.

Stimmqualität

Auch die Stimmqualität ist, wie bereits in Kapitel 5 erwähnt, ein Korrelat der Betonung. Leider sind glottale Parameter nur relativ aufwendig aus dem Sprachsignal zu extrahieren [Slu95]. Campbell [Cam95] beschreibt die Verwendung eines einfachen Merkmals zur Approximation des Spectral Tilt, den Quotienten aus Energie im Frequenzbereich von über 2 kHz und der Gesamtenergie. In Kapitel 5 wurden Ansätze für ein alternatives automatisches Verfahren aufgezeigt, das den Tilt aus den bei der Spracherkennung verwendeten Mel-Frequenz-skalierten Kepstralkoeffizienten abzuleiten versucht.

6.1.3 Weitere Aspekte

Kontinuierlich vs. diskret. Aus der vorangegangenen Aufzählung der Merkmale wird deutlich, daß die Merkmale, die zur Beschreibung der Prosodie tauglich scheinen, mit unterschiedlichen Zeitpunkten oder Zeitspannen assoziiert sind. Zum einen gibt es die zu feststehenden Zeitpunkten anfallenden Merkmale Grundfrequenzschätzung und Momentanenergie und die glottalen Parameter. Andererseits gibt es Parameter, die an linguistischen Kategorien, etwa der Silbe, orientiert sind. Diese Orientierung ist zwar bei einigen der aufgezählten Merkmale künstlich, da sie aus einem kontinuierlich anfallenden Merkmal abgeleitet sind, dies gilt

jedoch nicht für die Zeitmerkmale; für diese ist eine kontinuierliche Beschreibung sicher nicht die naheliegendste Form.

Zusätzlich zu den oben beschriebenen Merkmalen, die durch die akustisch-physiologischen Korrelate der Betonung motiviert sind, gibt es auch linguistische Merkmale, die für die Vorhersage nützlich sein können. Wie in Kapitel 2 beschrieben, treten im Regelfall Satzakzente nur bei Silben auf, die den lexikalisch bestimmten Wortakzent tragen. Folglich kann ein aus dem Lexikon abgeleitetes Merkmal, das aussagt, ob eine Silbe den Wortakzent trägt oder nicht, für die Vorhersage sinnvoll Verwendung finden, da Silben ohne Wortakzent a priori sehr viel unwahrscheinlicher Träger eines Satzakzentes sind als wortbetonte Silben.[3] Weitere linguistisch motivierte Merkmale, die Einfluß auf die (statistische) Verteilung der phonologischen Prosodiebeschreibung haben, sind Merkmale, die die Silbenstruktur oder die linguistische Funktion eines Wortes beschreiben. Die linguistische Funktion ist (zumindest teilweise) in den Part-of-Speech-Tags enthalten, die, wie in Abschnitt 6.3.2 beschrieben, mit akzeptabler Genauigkeit automatisch bestimmt werden können.

Für die Lernverfahren haben die verschiedenen Zeitbezüge unmittelbare Bedeutung, da bei ihrer Anwendung zu entscheiden ist, ob ein Lernverfahren kontinuierlich anfallende Merkmalbündel, phonemsynchrone Merkmalbündel, silbensynchrone Merkmalbündel oder eine Mischung daraus erhalten soll.

Größe des betrachteten Kontexts. Neben der Entscheidung für oder gegen kontinuierliche oder an linguistischen Kategorien orientierte Merkmale muß auch über die Größe des betrachteten Kontextes nachgedacht werden. Sowohl bei den kontinuierlichen als auch bei den zu linguistischen Einheiten anfallenden Merkmalen, kann man Werte aufeinanderfolgender Merkmale geeignet duplizieren und gruppieren, so daß etwa neben den Merkmalen der gerade betrachteten Silbe die Merkmale der vorhergehenden und nachfolgenden Silbe zur Verfügung stehen. In der Maschinelles Lernen-Forschung wird dieses Vorgehen als Berücksichtigung von Kontext bezeichnet, in der Spracherkennung bezeichnet man dieses Vorgehen gerne als „stacked features“-Anordnung. Lediglich für ein ILP-Verfahren ist eine solche direkte Kodierung der Umgebung nicht nötig, da dieses aufeinanderfolgende Einheiten bei Bedarf über eine Nachbarschaftsbeziehung selbständig verknüpfen kann. Die Wahl der Kontextgröße ist stark abhängig von der verfügbaren Trainingsmenge. Je mehr Daten zur Verfügung stehen, um so mehr Kontext kann

[3] Nicht lexikalisch wortbetonte Silben können nur bei der selten auftretenden kontrastiven Betonung einen Satzakzent tragen, wie etwa in „Du hast GutenBURG geschrieben, und nicht GutenBERG.“

man betrachten. Da in nahezu allen realistischen Lernanwendungen nur eine beschränkte Trainingsdatenmenge bereitgestellt werden kann, muß jeweils sorgfältig abgewogen werden, wieviele Kontextmerkmale beim Lernen mitbetrachtet werden können. Bietet man einem Lernverfahren zu viele Merkmale an, besteht die Gefahr, daß das Verfahren nicht die hinter den Beispielen liegenden allgemeinen Konzepte lernen kann, sondern die einzelnen Instanzen der Trainingsmenge erinnert (man nennt das overfitting), oder im Falle der statistischen Modellierung, daß für die Abschätzung der vielen durch eine zu große Anzahl an Merkmalen entstandenen Parameter zuwenig Trainingsmaterial bereitsteht.

Klassifikationsstrategie. Auch bei der Festlegung der prinzipiellen Klassifikationsstrategie gibt es Optionen. Das in Kapitel 2 vorgestellte Prosodiebeschreibungssystem besitzt zwei grundlegende Kategorien, nämlich Pitchakzente und Phrasentöne, die beide als Attribute einer Silbe aufgefaßt werden können. Die Vorhersage der Pitchakzente kann nun völlig unabhängig von der Entscheidung über die Phrasengrenzen betrieben werden. Dies bedeutet, daß man zwei Klassifikatoren einsetzen kann: Ein Klassifikator entscheidet aus den Merkmalen über die Pitchakzentausprägung der Silben; ein vom ersten unabhängiger zweiter Klassifikator entscheidet über die Ausprägung der Silben bezüglich der Phrasengrenzen. Weiterhin sind kaskadierte Anordnungen denkbar: Zunächst entscheidet ein Klassifikator über die Phrasierung einer gegebenen Äußerung; in einem zweiten Schritt, und unter Zuhilfenahme von Information, die der erste Schritt produziert, entscheidet ein weiterer Klassifikator über die Verteilung der Satzakzente in den Phrasen. Auch eine umgekehrte Anordnung ist denkbar: Zunächst wird über die Akzente entschieden, erst dann über Phrasierung. Als letzte mögliche Anordnung ist ein integriertes Vorgehen möglich: Ein Klassifikator entscheidet gleichzeitig über Phrasierung und Akzentuierung. Diese Anordnung kann man dadurch erreichen, daß das Klassifikationsverfahren einen zweidimensionalen Vektor für jede Silbe ausgibt, dessen erste Komponente die möglichen Pitchakzente und ein Symbol für das Fehlen eines solchen als Wertebereich besitzt, und dessen zweite Komponente die Ausprägung der mit der Silbe assoziierten Phrasengrenze bzw. das Fehlen einer solchen angibt.

Lokal vs. global. Weiterhin gibt es Optionen für die Art, wie Entscheidungen gefällt werden. Im wesentlichen gibt es hier die beiden Möglichkeiten, entweder lokal, das heißt für jede Silbe unabhängig, aufgrund der Ausprägung der Merkmale über die Akzent- oder Phrasentonklasse zu entscheiden, oder die Abfolge der prosodischen Etiketten anhand einer Zielfunktion global zu optimieren.

Die lokal entscheidenden Verfahren bestimmen für das zu einer Silbe gehörende Merkmalbündel über deren Ausprägung. Auch eine Aufnahme von Merkmalen der Umgebung der untersuchten Silbe ist möglich. Ausschließlich lokal entscheidende Verfahren besitzen den Nachteil, daß Symbolfolgen generiert werden können, die nicht der Intonationsgrammatik genügen. Ein Vorteil der lokal entscheidenden Herangehensweise ist, daß auf eine breite Basis erprobter Klassifikationsverfahren aus den Bereichen Mustererkennung, Maschinelles Lernen sowie Konnektionistische Systeme zurückgegriffen werden kann.

6.1.4 Verwandte Arbeiten

Zwei verwandte Arbeiten, die globale Ansätze verwenden, sind in der Literatur für das Amerikanische und das Japanische beschrieben. Ross [Ros95] beschreibt ein Verfahren, das Intensitäts- und Grundfrequenzverlauf mit einem dynamischen System für die Synthese generiert. Durch Anwendung der Bayesschen Regel kann das (Synthese-)Verfahren auch für die Erkennung eingesetzt werden. Ross berichtet für ein amerikanisches Radionachrichtenkorpus eine Erkennungsrate von 88.7% bei der Entscheidung, ob ein Pitchakzent vorhanden ist oder nicht. Für das Verfahren müssen die Phrasengrenzen jedoch bekannt sein, Ross setzt also die Klassifikation über die Phrasengrenzen in einem vorherigen Schritt voraus. Die Unterscheidung in verschiedene Pitchakzente ist ebenfalls nicht sehr fein. Es gibt bei den Pitchakzenten 4 Klassen (unbetont, high, downstepped high und low) und bei den Phrasentönen 5 Klassen. Das System kann natürlich über die Ausprägung des Phrasentones entscheiden, auch wenn bekannt sein muß, an welcher Stelle dieser auftritt. Wenn bekannt ist, daß eine Phrasengrenze vorliegt, entscheidet das Erkennungssystem in 59.0% der Fälle auf den richtigen Typ (eine der Klassen L-L%, L-H%, L-, H-, !H-).

Campbell [Cam96] benutzt ebenfalls ein Generierungssystem für die Erkennung von japanischer Intonation. Sein Ansatz besteht darin, explizit mehrere Versionen für eine Äußerung zu generieren, und diejenige synthetisierte Version anzunehmen, die gemäß einer Abstandsfunktion minimal von der tatsächlichen Äußerung entfernt liegt. Da das Japanische gegenüber dem Englischen und Deutschen prosodisch stark eingeschränkt ist, ist dort die explizite Generierung aller denkbaren Varianten vom rechnerischen Aufwand her machbar (es gibt nur einen Akzenttyp, 2 Anfangsgrenztöne und 3 Endgrenztöne). Die in [Cam96] abgedruckten Verwechslungstafeln für maschinelle vs. manuelle Annotierung (66.5% richtig vorhergesagt) gegenüber der Leistung menschlicher Experten, also manuelle vs. manuelle Annotierung (94.4% richtig vorhergesagt) sprechen nicht für das Verfahren.

6.2 Voruntersuchungen zur Auswahl eines geeigneten Lernansatzes

Aus den vielen möglichen Ansätzen zum Lernen eines automatischen Verfahrens zur Erstellung einer phonologischen Intonationsbeschreibung wurden vier Konstellationen ausgearbeitet und in Voruntersuchungen auf ihre Eignung für die Aufgabenstellung getestet.

1. Der erste Ansatz benutzt kontinuierlich anfallenden Merkmale. Die silben- und phonemsynchronen Merkmale der Dauer und das lexikalische Merkmal Wortakzent werden dazu in kontinuierliche Merkmale umgesetzt. Als Lernverfahren dient die Methode der statistischen Modellierung durch Hidden-Markov-Modelle (HMM).

2. Der zweite Ansatz verwendet ebenfalls die statistische Modellierung mit HMM als Lernmethode, als Merkmale (Observationsfolge) werden jedoch silbensynchrone Merkmalbündel bereitgestellt.

3. Der dritte Ansatz verwendet Wort-, Silben- und phonemsynchrone Merkmale; kontinuierliche Merkmale werden nicht benutzt. Das ausgewählte Lernverfahren ist das ILP-System Foil.

4. Im letzten Ansatz wird ein rein silbenbasiertes Merkmalbündel durch den symbolischen Attribut-Wert-Lerner C4.5 ausgewertet.

Die vier Lernansätze werden zunächst am selben Einzelsprecherkorpus getestet. Für das Training der Lernverfahren werden etwa 4/5 der Nachrichtenmeldungen des männlichen Nachrichtensprechers verwendet, da für diesen Sprecher am meisten manuell prosodisch annotiertes Sprachmaterial vorliegt. Getestet werden alle Ansätze an dem verbleibenden Fünftel. Damit den Verfahren eine möglichst einheitliche Textart vorliegt, sind Wetterberichte und Übersichten nicht in den Trainings- und Testmengen eingeschlossen. Die Trainingsmenge umfaßt 56 Nachrichtenmeldungen (4530 Wörter, 10445 Silben); 14 Nachrichtenmeldungen (1066 Wörter, 2436 Silben) bilden das Testset.

6.2.1 Statistische Modellierung mit HMM

Als erstes wird ein Ansatz skizziert, der mit der Methode der statistischen Modellierung auf kontinuierlichen Merkmalen arbeitet. Als kontinuierliche Merkmale

stehen zunächst die geschätzte Sprachgrundfrequenz und die Momentanenergie zur Verfügung. Für diese beiden Merkmale werden Differenzmerkmale gebildet, die die erste Ableitung von Grundfrequenz und Intensität annähern. Als Merkmal zur Beschreibung des lexikalischen Wortakzents wird innerhalb wortbetonter Silben der Wert 1, ansonsten der Wert 0 ausgegeben. Die synchron zu linguistischen Einheiten anfallenden Dauermerkmale werden ebenfalls in kontinuierliche Merkmale umgewandelt, hier sind das die Merkmale der Phonem-, Silben- und Pausendauer. Die mit dem automatischen Segmentierungssystem Alphons berechneten Dauerwerte für Phoneme und Silben werden wie folgt umgewandelt. Zunächst wird als Erwartungswert für die Dauer eines Phonems die durchschnittliche Dauer aller Vorkommen dieses Phonems für den zu modellierenden Sprecher im Korpus berechnet. Für jedes im Korpus angetroffenene Phonem wird dann das Verhältnis aus gemessener Dauer und durchschnittlicher Dauer des Phonems gebildet. Dieses Verhältnis wird über die Dauer des Lautsegments als kontinuierliches Phonemdauermerkmal ausgegeben. Entsprechende Merkmale werden für die Silbendauer generiert, indem die erwartete Silbendauer aus der Summe der erwarteten Längen der enthaltenen Phoneme gebildet wird.

Eine spezielle Behandlung erhalten die Pausendauern, die den vorangehenden Silben zugeordnet werden. Ein Merkmal kodiert die Länge einer auf die aktuelle Silbe folgenden Pause als Verhältnis zwischen gemessener und erwarteter (mittlerer) Pausenlänge. Folgt keine Pause auf die aktuelle Silbe, so erhält das Merkmal den Wert 0.

In der folgenden Tabelle sind die verwendeten kontinuierlichen Merkmale noch einmal zusammengefaßt:

Grundfrequenz
erste Ableitung der Grundfrequenz
Momentanenergie
erste Ableitung der Momentanenergie
Phonemdauer (Quotient aus gemessener und durchschnittlicher Phonemdauer)
Silbendauer (Quotient aus gemessener Silbendauer und Summe der durchschnittlichen Dauerwerte der in der Silbe enthaltenen Phoneme)
Pausendauer (Quotient aus gemessener und durchschnittlicher Pausendauer, über die Dauer der vorhergehenden Silbe ausgegeben)
lexikalischer Wortakzent

Für alle Pitchakzente und Phrasentongrenzen sowie für unbetonte Silben wurden HMM-Modelle abgeschätzt, die diese Merkmale emittieren. Als HMM-Modelltopologie wurden einfache links-rechts-Modelle gewählt, die 3 emittierende Zustände besitzen. Die Ergebnisse waren so schlecht, daß eine Auswertung der

Fehlerrate nicht angezeigt erschien. Ein Grund für das schlechte Abschneiden lag in der ungenügenden Synchronisation der Modelle mit den Silben. Der Versuch, die Synchronisierung auf Silben durch ein zusätzliches, sägezahnförmiges Merkmal zu verbessern, das am Anfang jeder Silbe 0 und am Ende 1 annimmt, hat das Resultat nur geringfügig verbessert. Aufgrund der schlechten Ergebnisse wurde dieser Ansatz nicht weiterverfolgt.

6.2.2 Silbensynchrone statistische Modellierung mit HMM

Da der soeben beschriebene, auf der statistischen Modellierung beruhende Ansatz Probleme beim Synchronisieren auf einzelne Silben hatte, wird das HMM-Verfahren im nächsten Versuch auf silbensynchronen Merkmalbündeln als Observationsfolge angewendet.

Kodierung der Observationsfolge

Bei der hier verwendeten Kodierung stehen für ein Sprachsignal somit nicht Merkmalsvektoren bereit, die jeweils einen kurzen, 10 ms großen Ausschnitt des Sprachsignals repräsentieren, sondern Merkmalsvektoren, die jeweils eine ganze Silbe repräsentieren; die Observationsfolge umfaßt also genauso viele Vektoren, wie die Äußerung Silben besitzt. Die Lokalisation der Silben im Sprachsignal erfolgt mit dem in Kapitel 3 beschriebenen System Alphons. Als Merkmale stehen dem Verfahren zur Verfügung:

`Tonaldiff`	Parameter α aus der F_0-Parametrisierung
`Tonalsteep`	Parameter β aus der F_0-Parametrisierung
`Tonalalign`	Parameter γ aus der F_0-Parametrisierung
`Peakheight`	Parameter δ aus der F_0-Parametrisierung
`Peaksteep`	Parameter ϵ aus der F_0-Parametrisierung
`Peakalign`	Parameter ζ aus der F_0-Parametrisierung
`Level`	Parameter η aus der F_0-Parametrisierung
`Lenexpected`	Summe der erwarteten Phonemlängen der Silbe
`Lenmeasure`	Gemessene Silbenlänge
`Lenrelative`	Quotient der vorhergehenden beiden Merkmale
`Wordstress`	0 für nicht wortbetonte, 1 für wortbetonte Silben
`Syls2wordend`	Anzahl der Silben zum Wortende
`Syls2nextp`	Anzahl der Silben bis zur nächsten Pause
`Disttonextp`	Abstand zur nächsten Pause in Sekunden
`Lenofnextp`	Länge der nächsten Pause in Sekunden

Modellierung

Für jede der 24 in den Daten auftretenden Pitchakzent-Grenzton-Kombinationen wurde ein HMM abgeschätzt. Es gibt also ein Modell für eine Silbe mit Pitchakzent H*L und ohne Grenzton, ein Modell für eine Silbe mit Pitchakzent L*H und Grenzton % usw. Silben, die keine Intonationsmarkierung tragen, werden von dem Modell der vorausgehenden Silbe mit Intonationsmarkierung mitmodelliert, im allgemeinen also vom vorhergehenden Pitchakzent. Dies ermöglicht eine einfache Beschreibung der Abfolgewahrscheinlichkeiten der Intonationssymbole durch Bigrammstatistiken direkt bei der Erkennung. Dieses Vorgehen steht zudem auch im Einklang mit dem „Spreading“ der phonologischen autosegmentalen Theorie.

Abbildung 6.1 zeigt die Topologien der eingesetzten Modelle für Silben mit Grenztonmarkierung (oben rechts) und markierte Silben ohne Grenztonmarkierung zusammen mit nachfolgenden unmarkierten Silben (oben links). Da auf Silben, die mit einem abgetrennten Ton assoziiert sind (..L und ..H), keine unbetonten Silben folgen dürfen, werden die Merkmale so annotierter Silben von einem Modell mit schleifenloser Topologie (oben rechts) emittiert. Silben, die vor der ersten mit einem prosodischen Etikett assoziierten Silbe auftreten, werden mit einem Modell der oben links abgebildeten Topologie modelliert. In den beiden Kästen neben der Modelltopologiedarstellung sind jeweils die entsprechend aufgebauten Modelle aufgeführt. Der im Modellnamen enthaltene Doppelpunkt trennt die Bezeichnung des Pitchakzents von der Bezeichnung des Grenztones, * tritt in den Modellnamen als s, % als p und . sowie − als d auf. Mit Downstep versehene Akzente werden mit ihren zugrundeliegenden Akzenten in eine Klasse zusammengelegt, um eine größere Anzahl an Exemplaren je Klasse zur Verfügung zu haben. Im unteren Bereich von Abbildung 6.1 ist eine Beispielabfolge der verwendeten Modelle wiedergegeben, zusammen mit der orthographischen Näherung der Silbenfolge und den assoziierten phonologischen Intonationsetiketten (zwei steigende Pitchakzente und eine intermediäre Phrasengrenze). Die Merkmale der Silbenkette werden von 4 Modellen emittiert: das erste, mit 0:0 bezeichnete Modell (kein Akzent, kein Grenzton) emittiert die vor dem ersten Akzent liegenden Silbenmerkmale, das zweite, mit LsH:0 bezeichnete Modell (L*H-Akzent, kein Grenzton) umfaßt den ersten Pitchakzent und die darauf folgenden unbetonten Silben, das dritte Modell, hier wiederum das LsH:0-Modell, emittiert die Merkmale der zweiten satzbetonten Silbe (da keine unbetonten Silben folgen, wird der zweite emittierende Zustand übersprungen), und schließlich modelliert das vierte, mit 0:d bezeichnete Modell (kein Pitchakzent, intermediäre Phrasengrenze −) die Merkmale der mit der intermediären Phrasengrenze assoziierten Silbe.

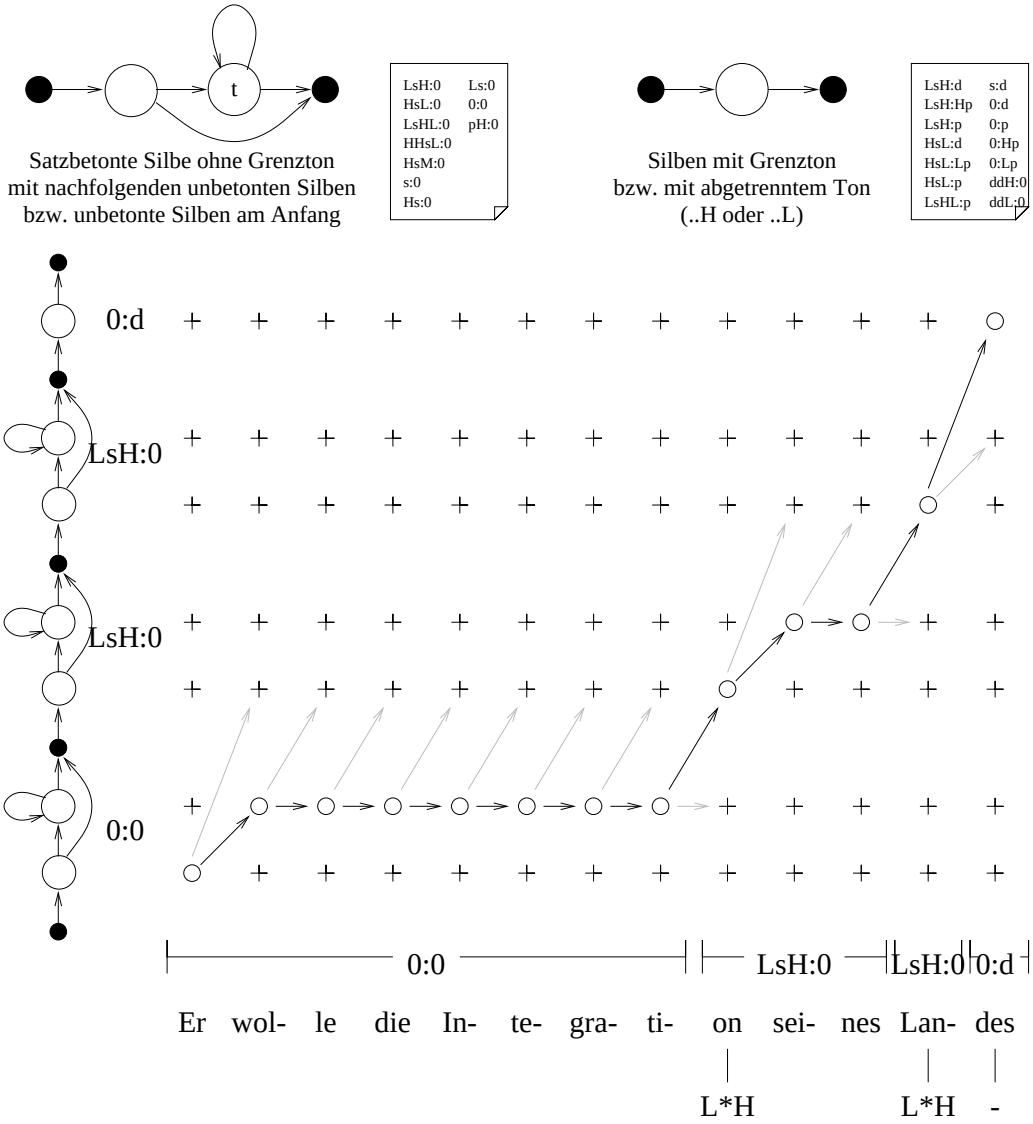


Abbildung 6.1: Modelltopologien und ein Beispiel für die Modellierung bei der zweiten, Silbenmerkmale emittierenden Markov-Modellierung.

Um zu verhindern, daß die unbetonten Silben die Entscheidung zwischen den verschiedenen Pitchakzenttypen beeinflussen, wurde für alle Emissionen der wiederholbaren Zustände „tying“ angewandt – somit also festgelegt, daß die Mittelwerte und Varianzen für die unbetonten Silben stets gleich sind. Der betreffende Zustand ist in Abb. 6.1 mit einem t gekennzeichnet. Da sehr viele unbetonte Silben vorhanden sind, wurden die Emissionen dieses zusammengefaßten Zustands mit zwei gewichtet überlagerten Gaußverteilungen abgeschätzt. Die Merkmale anderer Silben werden von Zuständen mit einer Gaußverteilung je Merkmal emittiert.

Training

Zur Abschätzung der Emissions- und Transitionsparameter wurde das „Hidden Markov Model Toolkit“ (HTK), Version 2.0 verwendet. Initialisierung und Neuabschätzung (HInit und HRest) wurden für jedes Modell getrennt durchgeführt anhand von Labelfiles, die die geänderte zeitliche Abdeckung der Modellsymbolfolge widerspiegeln. Nach der Neuabschätzung wurden die wiederholbaren Zustände zusammengelegt und die Wahrscheinlichkeitsdichtefunktionen für diesen zusammengelegten Zustand von einer auf zwei Normalverteilungen umgestellt. Es folgten mehrere Trainingszyklen mit HERest, bei denen alle Modelle einer Nachrichtenmeldung aneinandergehängt und eine Neuabschätzung äußerungweise statt modellweise durchgeführt wird. Bei der äußerungsweisen Neuabschätzung mit HERest werden die in den Labelfiles festgehaltenen Zeitpunkte, bei denen ein Modellwechsel zu erfolgen hat, nicht berücksichtigt. Für die hier vorgenommene Modellierung ist diese Freiheit nachteilig, da aufgrund der Modellierung in Silbeneinheiten und der manuellen prosodischen Etikettierung ja genau bekannt ist, für welche Silben ein Modell Merkmale emittieren soll. Allerdings ist es mit HRest nicht möglich, zusammengefaßte Zustände abzuschätzen, was die Verwendung von HERest nötig gemacht hat.

Ergebnisse

Mit den abgeschätzten Modellen wurde die prosodische Etikettierung der Nachrichtenmeldungen der Testmenge erzeugt und damit die Erkennungsraten ermittelt. Mit der in Abbildung 6.2 wiedergegebenen Finite-State-Grammar wurde die Erkennung auf erlaubte Symbolfolgen eingeschränkt[4] ohne Anwendung eines

[4]Gemäß der Grammatik allerdings nicht verboten ist das Folgen eines L%-Grenztones auf einen L*H-Akzent und das Folgen eines ..L auf einen L*-Akzent bzw. das Folgen eines ..H auf einen H*-Akzent.

Rest -2.5	53.6	ERest.13 -5.0	58.7
Tied -3.5	54.4	ERest.14 -5.0	58.8
ERest.1 -4.0	55.0	ERest.15 -5.0	58.8
ERest.2 -4.5	56.2	ERest.20 -5.0	57.9
ERest.3 -4.5	55.8	ERest.25 -5.5	60.1
ERest.4 -5.0	58.8	ERest.30 -5.5	59.6
ERest.5 -5.0	59.4	ERest.40 -5.5	59.6
ERest.6 -5.0	59.6	ERest.50 -5.5	59.7
ERest.7 -5.0	59.5	ERest.60 -5.5	59.8
ERest.8 -5.0	59.5	ERest.70 -5.5	59.9
ERest.9 -5.0	59.7	ERest.80 -5.5	59.9
ERest.10 -5.0	59.6	ERest.90 -5.5	59.9
ERest.11 -5.0	59.4	ERest.100 -5.5	60.2
ERest.12 -5.0	58.9		

Tabelle 6.1: Erkennungsraten mit dem silbensynchronen HMM-Verfahren. Die Einstellung für den Parameter p ist jeweils nach der Bezeichnung des Trainingsschritts angegeben.

Bigramm-Symbolabfolgemodells. Da bei der Erkennung viel zu viele Symbole ausgegeben wurden, wurde analog zur gängigen Praxis in der Spracherkennung [HTK] bei der Erkennung jeder Modellwechsel bestraft durch Abzug einer Konstanten p von der logarithmierten Wahrscheinlichkeit. Diese Konstante wurde grob so eingestellt, daß ungefähr gleich viele Symbolfolgen bei der Erkennung produziert werden wie insgesamt in der Teststichprobe gelabelt sind.[5] Eine Einstellung der Konstanten über die Erkennungsrate anstatt über die Anzahl der ausgegebenen Intonationssymbole war aufgrund der schlechten Erkennung nicht möglich. Ein solches Vorgehen hätte zur Folge gehabt, einen Modellwechsel möglichst stark zu bestrafen, also möglichst wenige Symbole auszugeben. Da etwa 2/3 der Silben keine prosodische Etikettierung tragen, werden bessere Ergebnisse (im Sinne von weniger Verwechslungsfehlern) erzielt, wenn fast immer unbetonte Silben ausgegeben werden, ein Verhalten, das für die prosodische Etikettierung nachteilig ist.

Ausgewertet wurden Verwechslungsfehler auf Silbenebene. Dazu mußte die durch die Modellierung verursachte Abbildung wieder umgedreht werden, beispielsweise erhält die erste von einem HsL:0-Modell abgedeckte Silbe eine H*L-Markierung, alle anderen abgedeckten Silben erhalten keine Markierung. Die prosodische Etikettierung einer Silbe wurde als richtig erkannt angesehen, wenn sowohl die Pitchakzent- als auch die Grenztonmarkierung übereinstimmten. In Tabelle 6.1 sind die Erkennungsraten nach unterschiedlichen Trainingsschritten abgebildet.

[5] Die Konstante p wurde schrittweise um 0.5 erniedrigt, bis die Anzahl erkannter Symbole unmittelbar unter die Anzahl gelabelter Symbole fiel.

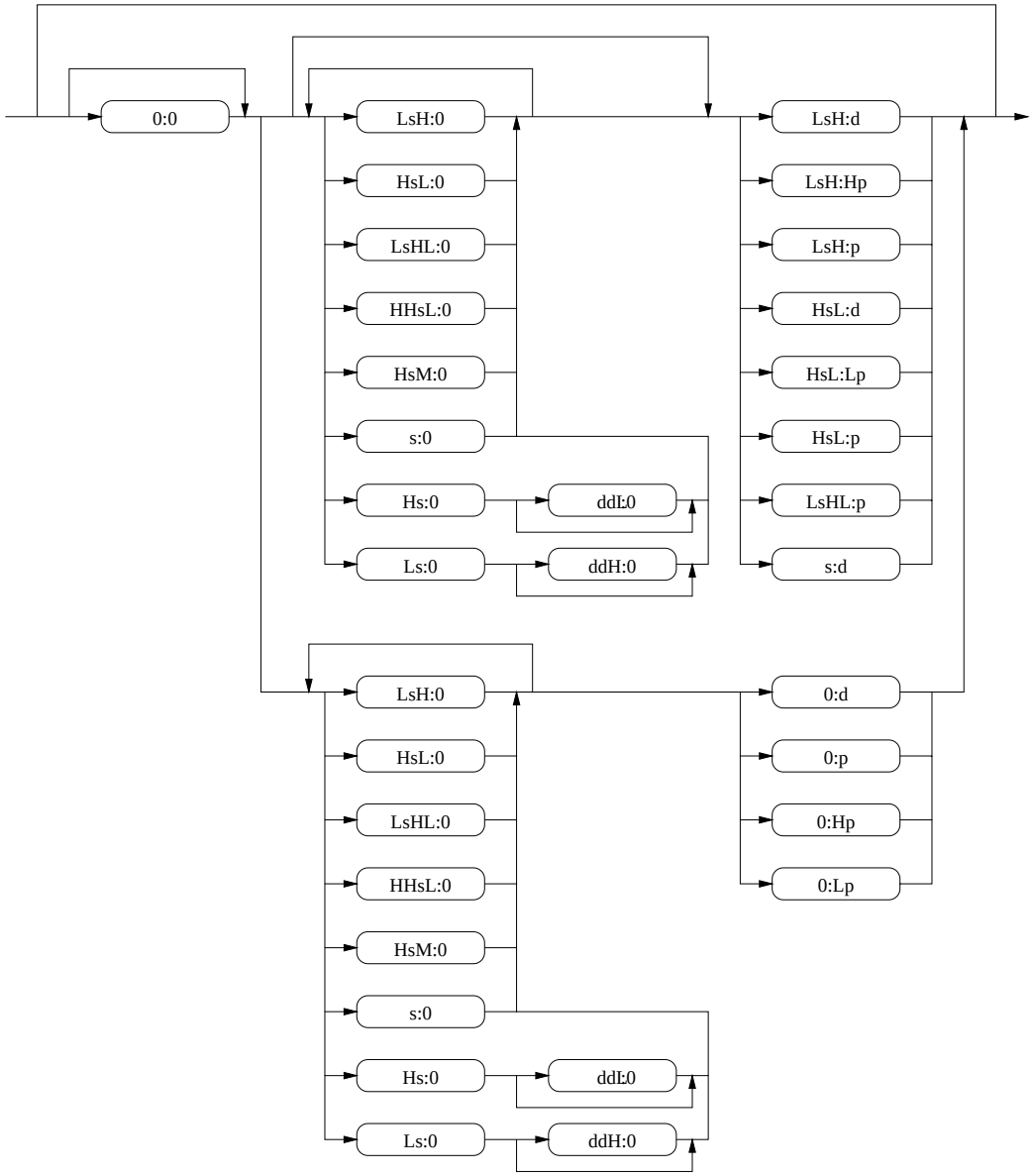


Abbildung 6.2: Grammatik zur Einschränkung erlaubter Intonationssymbolfolgen.

Bewertung

Die mit dem abgewandelten HMM-Verfahren erzielbaren Erkennungsraten sind zwar besser als mit dem ersten Verfahren, sie liegen jedoch deutlich unter dem trivialen Klassifikator „jede Silbe trägt kein prosodisches Etikett“. Mit diesem trivialen Klassifikator würden immerhin 72.70% der Silben richtig klassifiziert werden. Das vorgestellte Verfahren müßte deutlich verbessert werden, um überhaupt einen Nutzen für die Erstellung von Prosodiekorpora zu erbringen. Es gibt jedoch vielfältige Verbesserungsmöglichkeiten, sowohl auf Modellierungsebene, als auch beim Trainingsverfahren sowie bei der Erkennung.

Ansätze zur Verbesserung

Bei der Modellierung sollte für ausgewählte Merkmale eine Modellierung mit mehreren Normalverteilungen versucht werden. Auch scheint die naheliegende Verwendung einer vollen Kovarianzmatrix eine sinnvolle Erweiterung zu sein, da die Merkmale nicht voneinander statistisch unabhängig sind. Die Zustände, die unbetonte Silben repräsentieren, könnten, statt über alle Modelle zusammengefaßt zu werden, getrennt für die Akzente mit hohem Trailton und die Akzente mit tiefem Trailton zusammengefaßt werden. Auch die Berücksichtigung von Kontext könnte möglicherweise eine Verbesserung erbringen, beispielsweise wäre dadurch die Berücksichtigung des Grundfrequenzniveaus nach einem Pitchakzent möglich. Alle diese Verbesserungen gehen einher mit mehr Parametern, so daß gegebenenfalls mehr Daten benötigt würden. Möglicherweise kann man auch nur die Parametersätze der häufig auftretenden Klassen (z.B. 0:0, LsH:0) erweitern und die selten auftretenden Klassen mit wenigen Parametern abschätzen.

Das Zusammenfassen von Parametern wie bei den Zuständen, die unbetonte Silben repräsentieren, kann ebenfalls eine Verbesserung erbringen. Beispielsweise könnten die Verteilungsfunktionen aller Merkmale außer den Grundfrequenzmerkmalen von allen Zuständen, die Pitchakzentsilben vertreten, zusammengefaßt werden. Dadurch wäre das Erkennen eines Pitchakzentes noch von allen Merkmalen abhängig, die Entscheidung, welcher der Pitchakzente ausgewählt wird, jedoch nur von der Grundfrequenz beeinflußt.

Eine weitere, das Training betreffende Verbesserungsmöglichkeit ist, beim eingebetteten Neuabschätzen mit HERest statt ganzer Nachrichtenmeldungen nur einzelne Phrasen zu verwenden. Dadurch könnten sich die Modelle für die Abschätzung besser an die beabsichtigten Lernmuster annähern. Aufgrund der schlechten Erkennungsraten ist zu vermuten, daß eine halbwegs gute Vorbelegung

Tied	-3.5	IG	54.4
Tied	+1.5	3	62.2
Tied	+3.0	5	64.1
Tied	+7.0	10	61.0
ERest.100	-5.5	IG	60.2
ERest.100	+0.0	3	60.7
ERest.100	+2.0	5	61.5
ERest.100	+7.0	10	61.3

Tabelle 6.2: Erkennungsraten mit dem silbensynchronen HMM-Verfahren bei Verwendung eines Symbolabfolgemodells (IG=Intonationsgrammatik (Baseline aus Tabelle 6.1) und Gewichtung des Symbolabfolgemodells mit 3, 5 oder 10).

der Modellparameter für die äußerungsweise Neuabschätzung noch gar nicht vorliegt und deshalb die Erkennung auf diesem niedrigen Niveau bleibt.

Bei der Erkennung schließlich könnten Bigrammstatistiken die Möglichkeiten bei der Abfolge von Intonationssymbolen einschränken. Für diese letzte Verbesserung, also nur die Verbesserung der Erkennung durch ein Bigramm-Symbolabfolgemodell, wurden Versuche durchgeführt. Wie Tabelle 6.2 zeigt, verbessern sie die Ergebnisse nur marginal. Der Wert in der zweiten Spalte gibt wieder die Belegung der Konstanten p an. Der Wert in der dritten Spalte der Tabelle steht für die Gewichtung der Grammatik. IG bezeichnet das seitherige System als Vergleichssystem, das statt der Bigrammstatistik die Intonationsgrammatik verwendet.

6.2.3 Prädikatenlogisches Lernverfahren

Im dritten Versuchsansatz wurde begonnen, Klassifikatoren mit einem Verfahren der induktiven Logikprogrammierung (ILP) zu lernen. Der Vorteil, den ILP-Verfahren gegenüber den anderen Verfahren für die Prosodieerkennung haben, liegt darin, daß der Algorithmus über die Verkettung von Relationen Merkmale benachbarter Silben, Wörter, und anderer strukturell in der Nähe liegender Einheiten für die Klassifikation verwenden kann. Dadurch kann es zumindest teilweise vermieden werden, die Klassifikationsebenen (kontinuierlich/Phoneme/Silben) von vornherein festzulegen; stattdessen bleibt die Auswahl dem Lernverfahren überlassen. Dafür müssen die Daten in eine relationale Repräsentation transformiert werden. Zunächst werden dazu Relationen erstellt, die die Silben oder Phoneme mit ihren Merkmalsausprägungen verknüpfen, beispielsweise die Relation lengthsylmeasure(x,y) die eine Silbe x mit ihrer gemessenen Länge y in Beziehung setzt und natürlichsprachlich mit „x besitzt die Länge y“ beschrieben werden könnte. Würde sich die Repräsentation darauf beschränken, wäre sie äquivalent

zu einer von Attribut-Wert-Lernern verwendbaren Repräsentation. Zusätzlich dazu werden weitere Relationen eingeführt, die Beziehungen zwischen den Einheiten einer Ebene (z. B. Nachfolgebeziehungen zwischen Silben oder Phonemen) und zwischen den Ebenen definieren (z. B. Enthaltenseinsbeziehungen zwischen Silben und Wörtern oder zwischen Phonemen und Silben). Die Daten bilden dadurch ein Netz von Relationen, die an die Strukturen der Autosegmentalen Phonologie erinnern. Abb. 6.3 zeigt die strukturelle Kodierung der Daten an einem Beispiel. Die dem Verfahren zugänglich gemachten Relationen sind in Tabelle 6.3 zusammengestellt. Da durch die Verkettung prinzipiell beliebig weit entfernte Information verwendet werden kann, läßt sich eine solche Datenrepräsentation nicht mehr direkt in eine einzige Tabelle umwandeln, wie sie von einem Attribut-Wert-Lernsystem verarbeitet werden kann.

Aus den in Frage kommenden ILP-Verfahren wurde das System Foil [Qui90] für die Anwendung ausgewählt. Foil gilt als eines der bekanntesten und erfolgreichsten ILP-Verfahren [LWZ+96] und ist für Forschungszwecke frei verfügbar. Als Trainingsdaten dienten die aus den Daten des männlichen Nachrichtensprechers erstellten Relationen. Foil erzeugte daraus Hornklauseln, also wenn-dann-Regeln mit einem Literal im dann-Teil und einer Konjunktion von negierten oder nicht negierten Literalen im wenn-Teil der Regel. Zunächst wurde dem Algorithmus die Aufgabe gestellt, eine Theorie für das Prädikat accent(A) zu bestimmen, das für alle Silben A gelten soll, die mit einem Pitchakzent beliebigen Typs assoziiert sind.

Die Lösung dieser Aufgabe beschäftigte eine SUN-Workstation 27.6 Tage. Es wurden 79 Regeln mit insgesamt 1199 Literalen erzeugt. Die erste der gelernten Regeln ist als Beispiel in Abbildung 6.4 dargestellt. Neben Literalen, die Variablenbelegungen mit Schwellwerten oder mit anderen Variablen vergleichen, treten auch die bereitgestellten Relationen als Literale auf. Deren Auftretenshäufigkeit, die ebenfalls in Tabelle 6.3 angegeben ist, kann man als Hinweis auf die Wichtigkeit der Relation für die Vorhersage ansehen. Da das Auftreten einer Relation als Literal jedoch nichts über die Anzahl der von der Regel abgedeckten Fälle aussagt, dürfen diese Zahlen nicht überinterpretiert werden. Man kann jedoch ganz klar sehen, daß Informationen über die in einer Silbe enthaltenen Phoneme oder die Silbenstruktur nicht verwendet werden. Diese Situation wäre möglicherweise nicht eingetroffen, wenn mehr akustische Informationen an den Phonemen statt an den Silben angelagert gewesen wäre. Die Verwechslungstafeln für die Lern- und Teststichprobe sind in Tabelle 6.4 abgedruckt.

Drei Fünftel der Pitchakzente des menschlichen Labellers wurden auch von dem Klassifikator gesetzt, allerdings labelte der Klassifikator zusätzliche Pitchakzente, die in ihrer Anzahl etwa einem Viertel der vom Experten etikettierten Pitchak-

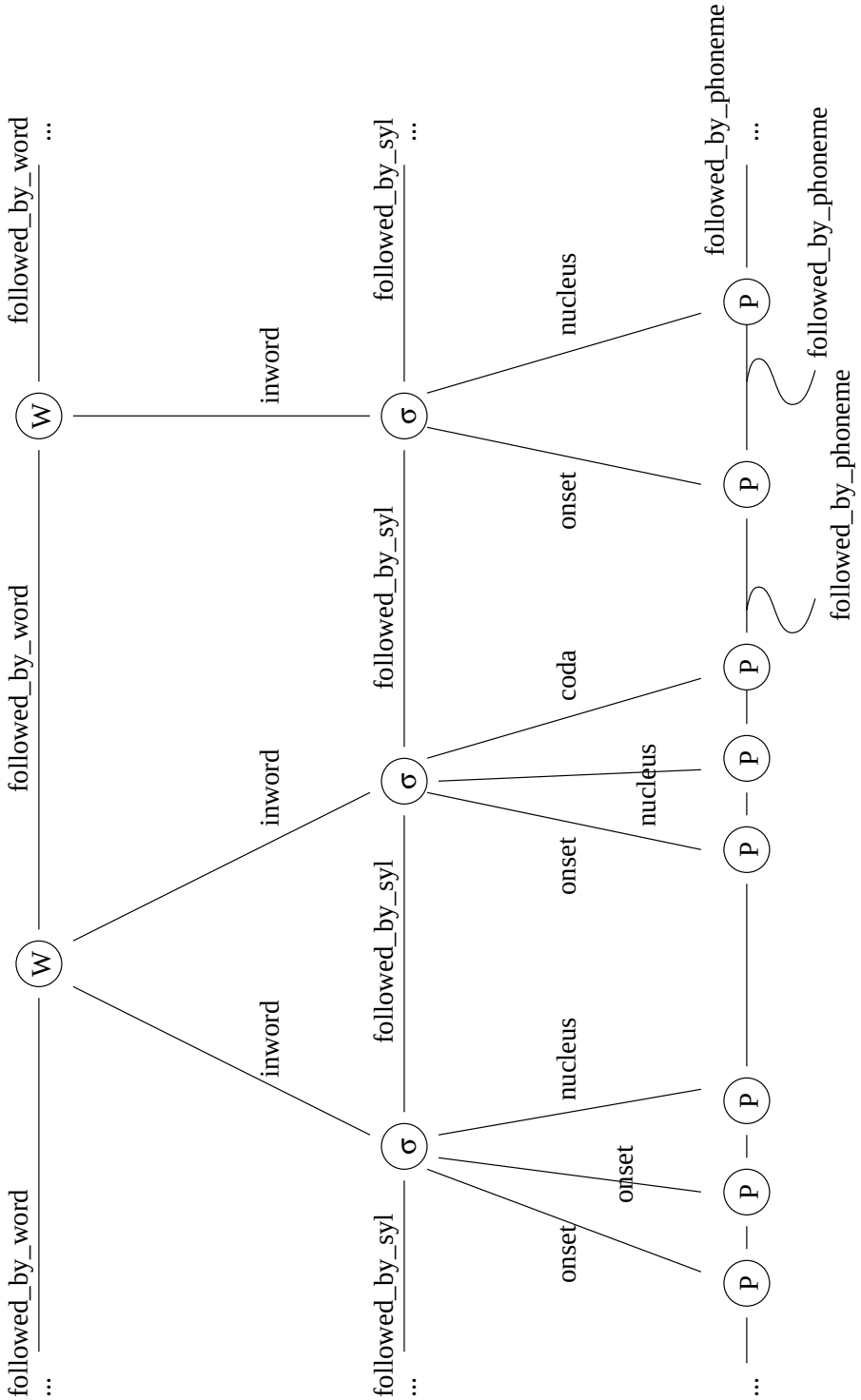


Abbildung 6.3: Strukturelle Relationen zur Beschreibung der Daten für Foil.

Relation	verwendet	negiert verwendet
inword(Syllable,Word)	62	0
insyl(Phoneme,Syllable)	0	0
followed_by_word(Word1, Word2)	2	0
followed_by_syl(Syllable1, Syllable2)	74	0
followed_by_phoneme(Phoneme1, Phoneme2)	0	0
syls2wordbegin(Syllable, Number)	63	0
syls2wordend(Syllable, Number)	34	0
onset(Phoneme)	0	0
nucleus(Phoneme)	0	0
coda(Phoneme)	0	0
dist_to_next_pause(Syllable, Length)	11	0
dist_to_prev_pause(Syllable, Length)	7	0
len_of_next_pause(Syllable, Length)	7	0
len_of_prev_pause(Syllable, Length)	8	0
phoneme(Phoneme, Typename)	0	0
vowel(Typename)	0	0
consonant(Typename)	0	0
diphvow(Typename)	0	0
shortvow(Typename)	0	0
longvow(Typename)	0	0
schwavow(Typename)	0	0
liquid(Typename)	0	0
frikative(Typename)	0	0
plosive(Typename)	0	0
nasal(Typename)	0	0
lengthsylexpected(Syllable, Length)	25	0
lengthsylmeasure(Syllable, Length)	44	0
lengthsylrelative(Syllable, Length)	23	0
lexicalwordstress(Syllable)	57	5
inlexikon(Word)	17	0
tag(Word, Tag)	15	0
tonaldiff(Syllable, Parameter)	42	1
tonalalign(Syllable, Parameter)	21	0
tonalsteep(Syllable, Parameter)	11	0
peakheight(Syllable, Parameter)	29	0
peakalign(Syllable, Parameter)	20	0
peaksteep(Syllable, Parameter)	13	0
level(Syllable, Parameter)	11	0
rms0_500(Syllable, Loudness)	21	0
rms500_1k(Syllable, Loudness)	19	0
rms1k_2k(Syllable, Loudness)	21	0
rms2k_4k(Syllable, Loudness)	25	0
rms4k_8k(Syllable, Loudness)	25	0
rms0_8k(Syllable, Loudness)	29	0
rms2_8k(Syllable, Loudness)	15	0
Summe	751	6

Tabelle 6.3: Relationen, die die Daten für das ILP-Verfahren beschreiben, sowie ihre Verwendungshäufigkeit bei der Akzentvorhersage.

```
accent(A) ← rms500_1k(A,B),
            inword(A,C),
            lengthsylmeasure(A,D),
            tonaldiff(A,E),
            tonalalign(A,F),
            peakalign(A,G),
            peaksteep(A,H),
            tag(C,nn),
            lexical_wordstress(A),
            followed_by_syl(A,J),
            dist_to_next_pause(J,K),
            D>186.292,
            E>H,
            K>1.26,
            B<=1.35568,
            F<=G.
```

Abbildung 6.4: Eine der 79 Regeln, die Foil zur Vorhersage von Pitchakzenten generiert.

Training:		vorhergesagt		
		ja	nein	gesamt
tatsächlich	ja	1677	302	1979
	nein	2	8464	8466
	gesamt	1679	8766	10445

Test:		vorhergesagt		
		ja	nein	gesamt
tatsächlich	ja	273	177	450
	nein	115	1871	1986
	gesamt	388	2048	2436

Tabelle 6.4: Verwechslungstafeln für die von Foil gelernten Regeln zur Unterscheidung zwischen Silben mit und ohne Pitchakzent.

zente entsprechen. Für die Teststichprobe ergibt sich eine Erkennungsrate von 88.0% bezüglich der Entscheidung, ob ein Pitchakzent vorliegt oder nicht. Auf der Lernmenge beträgt die Erkennungsrate 97.1%. Man entnimmt dem Mißverhältnis der Erkennungsraten zwischen Lern- und Testmenge, daß die gelernten Regeln zu speziell sind (Overfitting). Durch Generalisierungen ließen sich die Resultate möglicherweise noch leicht verbessern.

In einem weiteren Versuch sollte Foil Beschreibungen für einzelne Pitchakzenttypen lernen. Bei den seltener auftretenden Akzenten wurden die Regeln in weniger als einem Tag gelernt, für die häufigeren Akzente wurden mehr als 20 Tage benötigt. Beispielhaft sind in Tabelle 6.5 die Verwechslungstafeln für einige der Akzente wiedergegeben. Besonders bei den weniger häufig auftretenden Akzenten ist die Erkennung nicht zufriedenstellend. Diese Akzente sind wegen ihres seltenen Auftretens besonders schwer zu lernen, und auch die Evaluierung ist wenig aussagekräftig: in der Teststichprobe ist beispielsweise ein einziger HH*L-Akzent vertreten, das ist eindeutig zu wenig für eine sinnvolle Bewertung. Bei den H*L-Akzenten kann man aufgrund des häufigeren Auftretens gesichertere Aussagen machen: von den 176 vom Experten ettikettierten H*L-Akzenten wurden vom Algorithmus nur etwa ein Drittel (58) ebenfalls gelabelt. Außerdem ettikettierte der Algorithmus 38 Silben mit H*L, die der Experte nicht markiert hat. Die vom Algorithmus fälschlich gelabelten Fälle entsprechen etwa einem Fünftel der vom Experten insgesamt mit H*L gelabelten Fälle.

Für ein funktionierendes Erkennungssystem auf Basis von mit Foil gelernten Regeln wird noch eine weitere Komponente benötigt, die ebenfalls von Foil gelernt werden kann: Die Einzelergebnisse müssen zusammengefaßt und die dabei sehr wahrscheinlich auftretenden widersprüchlichen Entscheidungen aufgelöst werden, etwa wenn die Regeln für eine Silbe gleichzeitig einen HH*L- und einen H*L-Akzent ergeben. Anstatt diese Konfliktauflösung als Lernproblem anzusehen, kann man auch heuristisch vorgehen und beispielsweise stets das häufigere Konzept ausgeben, wenn mehrere Klassifikatoren einen Akzent vorschlagen. Für ein funktionsfähiges Gesamtsystem, das auch Phrasengrenzen behandeln kann, sind weitere Lernläufe notwendig. Dem stehen jedoch die Lernzeiten von bis zu knapp einem Monat entgegen. Diese Lernzeiten sind ein nicht zu unterschätzendes Moment bei der Systementwicklung, da die Lernläufe zur Optimierung von Parametern und Datenrepräsentation mehrfach ausgeführt werden müssen.

6.2.4 Attribut-Wert-Lernverfahren

Beim letzten Ansatz wird ein Attribut-Wert-Lernverfahren eingesetzt, das für jede Silbe lokal über die mit dieser Silbe assoziierte phonologische Intonations-

H*L				
Trainingsdauer:				1772796.80 s
Anzahl erzeugter Regeln:				40
Test:		vorhergesagt		
		ja	nein	gesamt
tatsächlich	ja	58	118	176
	nein	38	2222	2260
	gesamt	96	2340	2436

L*HL				
Trainingsdauer:				75079.30 s
Anzahl erzeugter Regeln:				4
Test:		vorhergesagt		
		ja	nein	gesamt
tatsächlich	ja	3	7	10
	nein	5	2421	2426
	gesamt	8	2428	2436

HH*L				
Trainingsdauer:				36998.60 s
Anzahl erzeugter Regeln:				1
Test:		vorhergesagt		
		ja	nein	gesamt
tatsächlich	ja	0	1	1
	nein	2	2433	2435
	gesamt	2	2434	2436

Tabelle 6.5: Ergebnisse der von Foil bestimmten Klassifikatoren für H*L, L*HL und HH*L.

beschreibung entscheidet. Die Entscheidungen werden aufgrund von Merkmalsausprägungen getroffen, die die Silbe charakterisieren. Das eingesetzte Lernverfahren ist C4.5 [Qui92], das als ein Standardverfahren zum Klassifikatorlernen im maschinellen Lernen betrachtet werden kann.[6] Wie sich im folgenden zeigen wird, ist C4.5 von den hier untersuchten Verfahren für die vorliegende Aufgabenstellung am besten geeignet und wird daher für weitergehende Untersuchungen verwendet. Aus diesem Grund wird es hier ausführlicher vorgestellt.

Konstruktion von Entscheidungsbäumen mit C4.5

C4.5 generiert aus einer vorgegebenen Menge von Lernbeispielen Entscheidungsbäume. Die Beispiele sind durch eine feste Anzahl von Merkmalen beschrieben. Dabei können sowohl Merkmale mit kontinuierlichem Wertebereich verarbeitet werden als auch solche mit diskretem Wertebereich, das heißt Merkmale, deren Wertebereich aus einer vorgegebenen, endlichen Menge von Bezeichnern besteht. Aus der Lernbeispielmenge erzeugt das Verfahren eine Abbildung vom Merkmalsraum auf Klassenzugehörigkeiten, indem es einen Entscheidungsbaum konstruiert, dessen innere Knoten mit Tests zur Ausprägung eines Merkmals dekoriert sind und dessen Blätter Klassenverweise tragen.

C4.5 baut einen Entscheidungsbaum rekursiv von der Wurzel ausgehend auf, indem es einen Merkmalstest auswählt, der die vorhandenen Beispiele möglichst gut in die verschiedenen Klassen einteilt. Im Wurzelknoten des Entscheidungsbaums wird der gewählte Test auf die Ausprägung eines Attributs eingetragen, und für jeden möglichen Ausgang des Merkmaltests wird ein Zweig an den Knoten angefügt. Die Trainingsbeispiele werden dann nach ihrer Ausprägung des getesteten Attributs in Teilmengen unterteilt, von denen jede in den entsprechenden Zweig des Entscheidungsbaumes weitergeführt wird. Dort wird ein weiterer Knoten erzeugt, dessen Merkmalstest nun anhand der lokalen Beispielmenge ausgewählt wird usw. Wenn in einer Beispielmenge schließlich nur noch Beispiele derselben Klasse enthalten sind, wird ein Blatt mit einem Verweis auf diese Klasse erzeugt. Die Auswahl des jeweils besten Merkmaltests erfolgt anhand des „Information gain“, eines informationstheoretischen Maßes für den Beitrag eines Merkmaltests zur Unterscheidung der verschiedenen Klassen. C4.5 verfolgt beim Aufbau des Entscheidungsbaums also eine „gierige“ Suchstrategie, die zu jedem Zeitpunkt die lokal beste Auswahl trifft und dafür auf eine globale Optimierung des gesamten gelernten Entscheidungsbaums zunächst verzichtet.

[6] Für die Experimente in diesem Kapitel wurde die Programmversion C4.5 Release 8 verwendet, die gegenüber der in Kapitel 5 eingesetzten Release 6 bessere Fähigkeiten beim Umgang mit kontinuierlichen Werten besitzt [Qui96].

Ein neues Beispiel wird mit Hilfe des Entscheidungsbaums klassifiziert, indem von der Wurzel ausgehend die Fragen zur Merkmalsausprägung des Beispiels beantwortet und der der Antwort entsprechende Folgeknoten aufgesucht wird, bis man an ein Blatt gelangt, das dann die Klasse angibt. Neben der einfachen Berechenbarkeit der Entscheidungen, und somit dem effizienten Auffinden einer vorhergesagten Klasse für neue Fälle, besitzen Entscheidungsbäume den Vorteil, auch für Forscher nachvollziehbar zu sein. Der gelernte Klassifikator läßt sich dadurch auf Plausibilität überprüfen und bietet Einsichten in die Struktur des behandelten Klassifikationsproblems. Dies trifft jedoch nur für Bäume bis zu einer bestimmten Größe zu, werden sie zu komplex, so sind auch sie wie etwa die Gewichte in einem künstlichen neuronalen Netz lediglich theoretisch nachvollziehbar. Aus Entscheidungsbäumen läßt sich algorithmisch eine weitere Klassifikatordarstellung ableiten, indem aus den Bäumen Wenn-Dann-Regeln gebildet werden. Manchmal sind diese adäquater für das Untersuchungsgebiet, beispielsweise wenn diese Beschreibungsform näher an der traditionell verwendeten Beschreibung des Untersuchungsgegenstands liegt. Auch für die Regeln gilt, daß sie nur bis zu einer gewissen Komplexität tatsächlich nachvollziehbar sind.

Es existieren „Pruning“-Techniken, um die Komplexität der gelernten Bäume bzw. Regeln zu beschränken, die nicht nur im Interesse einer besseren Verständlichkeit angewendet werden sollten, sondern auch meist zu einer Verbesserung der Vorhersagegenauigkeit auf einer unabhängigen Teststichprobe führen. Ein großer Teil der Komplexität eines Entscheidungsbaumes oder der Regeln wird durch die Beschreibung von Einzelfällen in der Trainingsmenge verursacht, die für die Vorhersage einer unabhängigen Teststichprobe keine Relevanz besitzen. Für die Vorhersage ist es im allgemeinen sinnvoll, von solchen Einzelfällen zu abstrahieren und dafür weniger spezifische Klassifikatoren zu generieren. Das Pruning führt zu Klassifikatoren, die zwar auf den Trainingsdaten eine geringere Genauigkeit, dafür aber häufig auf den Testdaten eine höhere Vorhersagegenauigkeit aufweisen. Die kleineren und damit generalisierten Entscheidungsbäume erreicht man entweder durch das sogenannte Prepruning, bei dem bereits in der Konstruktionsphase des Algorithmus Zweige nicht expandiert werden, oder durch das Postpruning, bei dem nach der Konstruktion Teilbäume bewertet und gegebenenfalls durch ein Blatt ersetzt werden.

C4.5 stellt beide Pruningtechniken zur Verfügung. Das Prepruning ist dadurch realisiert, daß der Anwender eine Mindestanzahl von Fällen vorgibt, die für einen sinnvollen Test und damit für eine weitere Expandierung erforderlich sind. Stehen zu wenige Fälle bereit, wird der Baum statt um einen weiteren Knoten um ein Blatt mit der häufigsten Klasse erweitert. Postpruning realisiert C4.5 dadurch, daß es für die Teilbäume des Entscheidungsbaums prüft, ob sich die erwartete Vorhersagegenauigkeit eines Teilbaums statistisch signifikant von der Vorhersagegenauigkeit

einer einzigen Klasse unterscheidet. Trifft dies zu, wird der Teilbaum belassen, andernfalls durch einen Blattknoten ersetzt.

Die Einstellungsmöglichkeiten des Algorithmus, vor allem die das Pruning betreffenden, haben einen deutlichen Einfluß auf die Klassifikationsgenauigkeit. Sie lassen sich an die jeweilige Klassifikationsaufgabe anpassen. Für die jeweilige Lernaufgabe optimale Parametereinstellungen lassen sich zwar bestimmen, indem man mehrere Lernläufe mit variierten Parametereinstellungen durchführt und dann die Einstellung mit den besten Resultaten auswählt. Eine Versuchsreihe, die alle kombinatorisch möglichen verschiedenen Parametereinstellungen vollständig durchtestet, ist jedoch wegen deren großen Anzahl aus ökonomischen Gründen nicht durchzuführen, so daß es in der Praxis doch unumgänglich ist, die Parametereinstellungen a priori festzulegen. In den in diesem Kapitel durchgeführten Versuchen wurden alle Einstellungen, die die Testauswahl und das Postpruning betreffen, auf ihren Defaultwerten belassen. Variiert wurde jedoch der Parameter m für das Prepruning, der die für einen Merkmalstest erforderliche Mindestanzahl von Beispielen festlegt. So wurden alle Lernläufe sowohl mit der Standardeinstellung $m = 2$ als auch mit dem erhöhten Wert $m = 5$ durchgeführt, der in Vorexperimenten bestimmt wurde.

Verwendete Merkmale

Für diesen ersten Test wurden die gleichen Merkmale benutzt, wie in dem auf Silbenmerkmalen arbeitenden HMM-Verfahren mit Ausnahme des Attributs `Sylls2nextp`, das hier nicht verwendet wurde. Es standen also 14 Merkmale zur Verfügung, nämlich die 7 den Grundfrequenzverlauf beschreibende Attribute (die 7 Parameter aus Kapitel 4), 3 Silbenlängen beschreibende Attribute (erwartete und gemessene Silbenlänge sowie das Verhältnis daraus), 2 lexikalische Attribute (Wortakzent und Anzahl der Silben zum Wortende) sowie 2 Pausen beschreibende Attribute (Abstand und Länge der nächsten Pause). Diese sind nochmals in Tabelle 6.6 zusammengefaßt.

Training

Aus den gleichen Daten wie bei den beiden vorangegangenen Verfahren wurden Entscheidungsbäume induziert mit den Standardeinstellungen des Algorithmus ($m = 2$). Anders als bei dem silbenorientierten HMM-Verfahren wurden hier die Downstep-Akzente in getrennten Klassen geführt.

`Tonaldiff`	Parameter α aus der F_0-Parametrisierung
`Tonalsteep`	Parameter β aus der F_0-Parametrisierung
`Tonalalign`	Parameter γ aus der F_0-Parametrisierung
`Peakheight`	Parameter δ aus der F_0-Parametrisierung
`Peaksteep`	Parameter ϵ aus der F_0-Parametrisierung
`Peakalign`	Parameter ζ aus der F_0-Parametrisierung
`Level`	Parameter η aus der F_0-Parametrisierung
`Lenexpected`	Summe der erwarteten Phonemlängen der Silbe
`Lenmeasure`	Gemessene Silbenlänge
`Lenrelative`	Quotient der vorhergehenden beiden Merkmale
`Wordstress`	0 für nicht wortbetonte, 1 für wortbetonte Silben
`Syls2wordend`	Anzahl der Silben zum Wortende
`Disttonextp`	Abstand zur nächsten Pause in Sekunden
`Lenofnextp`	Länge der nächsten Pause in Sekunden

Tabelle 6.6: Merkmale für das Attribut-Wert-Lernverfahren.

Ergebnisse

Der gelernte Entscheidungsbaum wurde auf die gleichen Testdaten wie bei den beiden vorangegangenen Verfahren angewandt und die Erkennungsrate ermittelt. Bei der Auswertung wurde eine Silbe als korrekt klassifiziert betrachtet, wenn sowohl für den Pitchakzent als auch für den Grenzton die korrekte Entscheidung getroffen war. Die Ergebnisse sind wie folgt:

Auswertung auf Trainingsdaten (10445 Fälle)					
Vor Postpruning			Nach Postpruning		
Größe	Fehler		Größe	Fehler	
1957	776	(7.43%)	1351	991	(9.49%)

Auswertung auf Testdaten (2436 Fälle)					
Vor Postpruning			Nach Postpruning		
Größe	Fehler		Größe	Fehler	
1957	631	(25.90%)	1351	564	(23.15%)

Der gelernte Entscheidungsbaum klassifiziert 76.85% der Testdaten korrekt. Die Fehlerrate auf dem Testkorpus beträgt 23.15%, auf den Trainingsdaten nur 9.49%. Der Vergleich dieser beiden Zahlen läßt vermuten, daß der Algorithmus überspezialisierte Klassifikatoren erzeugt, und daß sich die Vorhersagen verbessern, wenn man stärkeres Generalisieren erzwingt. Am einfachsten läßt sich dies erreichen

durch Erhöhung des Parameters m, der das Prepruning steuert. Mit $m = 5$ steigt die Vorhersagegenauigkeit auf 78.65%:

Auswertung auf Trainingsdaten (10445 Fälle)			
Vor Postpruning		Nach Postpruning	
Größe	Fehler	Größe	Fehler
1011	1388 (13.28%)	609	1515 (14.50%)

Auswertung auf Testdaten (2436 Fälle)			
Vor Postpruning		Nach Postpruning	
Größe	Fehler	Größe	Fehler
1011	559 (22.95%)	609	520 (21.35%)

Durch das moderate Generalisieren (mit $m = 5$) werden die selten vertretenen Klassen nicht allzu sehr unterrepräsentiert, die Erkennungsrate wird jedoch etwas besser.

Wenn man, wie beim silbenmerkmalsbasierten HMM-Verfahren, die Downstep-Annotierung ignoriert, also etwa H*L und !H*L als eine Klasse auffaßt und dadurch den Downstep als globales Phänomen einstuft, das von einer nachfolgenden Stufe entschieden werden soll, erreicht man eine Vorhersagegenauigkeit von 76.85% für $m = 2$ und 78.49% für $m = 5$, also annähernd gleiche Werte.

6.2.5 Bewertung

Das silbenbasierte Attribut-Wert-Lernverfahren C4.5 erbringt gegenüber den HMM-Ansätzen eine deutlich bessere Vorhersagegenauigkeit. Gegenüber dem ILP-Verfahren besitzt es den wichtigen Vorteil, auch bei größeren Datenmengen in sehr kurzer Zeit Klassifikatoren zu konstruieren. Die Konstruktion eines Entscheidungsbaumes mit C4.5 benötigt auf einer gewöhnlichen RISC-Workstation einige Minuten, während das prädikatenlogische Verfahren mehrere Tage für die Konstruktion eines Klassifikators benötigt. Da die mit Foil erreichten Erkennungsraten nicht wesentlich von den mit C4.5 erzielten abweichen, wird für die folgenden Untersuchungen ausschließlich C4.5 eingesetzt. In der gleichen Zeit hätte mit dem ILP-Verfahren nur ein Bruchteil der im folgenden beschriebenen Untersuchungen durchgeführt werden können. Es wird vermutet, daß der Vorteil der symbolischen KI-Verfahren gegenüber den statistischen Modellierungsverfahren dadurch zustande kommt, daß diese Verfahren mit den mitunter sehr wenigen zur

Verfügung stehenden Lerndaten besser zurechtkommen als die statistischen Verfahren, die zur präzisen Parameterabschätzung mehr Daten benötigen als vorhanden sind.

6.3 Auswahl von Merkmalen

6.3.1 Berücksichtigung von Kontext

Im ersten Experiment wird untersucht, ob die Bereitstellung von Merkmalen der benachbarten Silben die Vorhersagegenauigkeit bei dem silbenmerkmalsbasierten Attribut-Wert-Lernverfahren steigern kann. In Kapitel 4 wurde schon angedeutet, daß sich die Grundfrequenzparametrisierung zwischen H*L- und HH*L-Akzenten möglicherweise zu wenig unterscheidet. Die Differenzierung zwischen diesen beiden Akzenten läßt sich eventuell dadurch verbessern, daß die Parametrisierung der vor der untersuchten Silbe liegenden Silbe ebenfalls für eine Klassifikation zugänglich gemacht wird. Auch das Grundfrequenzniveau von nachfolgenden Silben kann für die Unterscheidung z.B. zwischen H*L und L*H sinnvoll sein, ebenso kann die Parametrisierung der nachfolgenden Silbe zur Unterscheidung zwischen L*H- und L*HL-Akzenten beitragen. Die Längenmerkmale benachbarter Silben können die Vorhersage unterstützen, da mit ihnen Entscheidungen über die Silbenlänge an die lokale Sprechgeschwindigkeit angepaßt werden können. Auch der Wortakzentstatus der vorangehenden Silbe kann eventuell zur Unterscheidung zwischen H*L- und HH*L-Akzenten beitragen.

In einer ersten Reihe von Lernläufen mit anschließender Klassifikation der Testmenge werden alle im vorhergehenden Abschnitt dem Lernverfahren angebotenen Merkmale und zusätzlich die entsprechenden Merkmale der jeweils vorhergehenden und nachfolgenden Silbe einbezogen. Statt 14 stehen dem Verfahren nunmehr also $3 * 14 = 42$ Merkmale pro Silbe zur Verfügung. Um eventuelle Unterschiede des Einflußes der vorhergehenden gegenüber den nachfolgenden Silbenmerkmalen zu entdecken, werden Lernläufe durchgeführt, bei denen nur die linken beziehungsweise nur die rechten Kontextmerkmale zugefügt werden. Tabelle 6.7 zeigt die Ergebnisse der Lernläufe.

Man entnimmt den Resultaten, daß der Lernalgorithmus aus den zusätzlich angebotenen Merkmalen keinen Nutzen ziehen kann. Vermutlich sind für diese große Anzahl von Merkmalen nicht mehr ausreichend viele charakterisierende Fälle in den Daten enthalten, und der Algorithmus verläßt sich mitunter auf wenig relevante Merkmalsausprägungen, die zwar zufällig die Lernstichprobe charakterisieren,

Merkmale der vorhergehenden Silbe	aktuellen Silbe	nachfolgenden Silbe	Prepruning m=2	m=5
	14		76.85	78.65
14	14		75.33	76.89
	14	14	75.45	76.60
14	14	14	75.62	76.68

Tabelle 6.7: Erkennungsraten nach Hinzufügen der Merkmale der vorangehenden Silbe, nachfolgenden Silbe, sowie beider umgebender Silben.

Merkmale der vorhergehenden Silbe	aktuellen Silbe	nachfolgenden Silbe	Prepruning m=2	m=5
	14		76.85	78.65
7 F_0	14		75.53	77.13
	14	7 F_0	73.97	78.04
7 F_0	14	7 F_0	74.22	76.48

Tabelle 6.8: Erkennungsraten nach Hinzufügen der Grundfrequenzmerkmale der vorangehenden Silbe, nachfolgenden Silbe, sowie beider umgebender Silben.

aber für die Vorhersage nicht von Belang sind. Zur Untersuchung, ob eine Beschränkung der zusätzlich angebotenen Merkmale eine Verbesserung erbringen kann, dienen Lernläufe, bei denen lediglich die 7 Grundfrequenzmerkmale der umgebenden Silben hinzugefügt wurden. Tabelle 6.8 zeigt, daß auch diese Beschränkung keine Verbesserung hervorrufen kann.

Auch andere Merkmale benachbarter Silben erbringen keine Verbesserung, wie den Tabellen 6.9 (Silbendauermerkmale), 6.10 (Pausendauer und -abstand) und 6.11 (lexikalische Merkmale) zu entnehmen ist. Eine mögliche Erklärung für dieses Verhalten ist, daß für die Unterscheidung der Pitchakzente durch Verwendung eines Zweisilbenfensters bei der F_0-Parametrisierung bereits genügend Information über die nachfolgende Silbe vorhanden ist, und daß die für die Differenzierung zwischen H*L und HH*L wichtige Information aus der vorangehenden Silbe sich nicht in Erkennungsraten auswirkt, da der HH*L-Akzent so selten auftritt (22 HH*L-Akzente vs. 649 H*L-Akzente in den untersuchten Daten).

6.3.2 Berücksichtigung von Tags

Eine in der theoretischen Linguistik weitverbreitete Annahme ist, daß die Wahrnehmung von Pitchakzenten und Phrasengrenzen durch die syntaktische Struktur sowie die linguistische Funktion von Wörtern beeinflußt wird. Man vermutet, daß eine Phrasengrenze an einer erwarteten Stelle viel weniger stark pros-

Merkmale der			Prepruning	
vorhergehenden Silbe	aktuellen Silbe	nachfolgenden Silbe	m=2	m=5
	14		76.85	78.65
3 Dauer	14		75.99	77.30
	14	3 Dauer	75.21	77.05
3 Dauer	14	3 Dauer	74.59	75.74

Tabelle 6.9: Erkennungsraten nach Hinzufügen der Silbendauermerkmale der vorangehenden Silbe, nachfolgenden Silbe, sowie beider umgebender Silben.

Merkmale der			Prepruning	
vorhergehenden Silbe	aktuellen Silbe	nachfolgenden Silbe	m=2	m=5
	14		76.85	78.65
2 Pausenm.	14		76.23	77.71
	14	2 Pausenm.	76.15	77.91
2 Pausenm.	14	2 Pausenm.	75.90	77.67

Tabelle 6.10: Erkennungsraten nach Hinzufügen der Pausenmerkmale der vorangehenden Silbe, nachfolgenden Silbe, sowie beider umgebender Silben.

Merkmale der			Prepruning	
vorhergehenden Silbe	aktuellen Silbe	nachfolgenden Silbe	m=2	m=5
	14		76.85	78.65
2 lex.	14		75.99	78.37
	14	2 lex.	76.31	77.26
2 lex.	14	2 lex.	76.19	78.12

Tabelle 6.11: Erkennungsraten nach Hinzufügen der lexikalischen Merkmale der vorangehenden Silbe, nachfolgenden Silbe, sowie beider umgebender Silben.

odisch ausgeprägt werden muß, um sie beim Hörer als solche kenntlich zu machen, und daß etwa ein Pitchakzent an einer eigentlich unerwarteten Stelle besonders deutlich ausgeprägt sein muß, um vom Hörer erkannt zu werden. Auch bei der automatischen Erkennung der Prosodiebeschreibung kann deshalb die Information, aus der sich diese Erwartungen ableiten, hilfreich sein. Andererseits muß für ein automatisches Verfahren, mit dem beliebiges Sprachmaterial verarbeitet werden soll, diese Information robust automatisch ableitbar sein. Eine für die untersuchte Textart Radionachrichten robust arbeitende syntaktische Analyse stand für die Untersuchungen im Rahmen dieser Arbeit nicht zur Verfügung, so daß auf Part-of-Speech-Tags (POS-Tags) zurückgegriffen werden mußte, um die linguistische Funktion und die syntaktische Gliederung anzunähern.

Unter Tagging versteht man im wesentlichen eine Abbildung der in ihrer Anzahl unbeschränkten Wörter einer Sprache in eine kleine, endliche und festgelegte Anzahl von Kategorien. Die Zuordnung der Kategorie wird anhand der linguistischen Funktion des Wortes im aktuellen Kontext getroffen; das Wort bekommt ein Etikett, das seine Funktion im Kontext beschreibt, den sogenannten Part-of-Speech-Tag (abgekürzt als POS-Tag). Das in diesen Untersuchungen verwendete Inventar an Etiketten ist das Stuttgart-Tübingen-Tag-Set (STTS), das die 54 in Tabelle 6.12 dargestellten Kategorien umfaßt.

Zur Erzeugung der Tags für das Untersuchungsmaterial wurde der auf HMM und Entscheidungsbäumen basierende, in Stuttgart entwickelte POS-Tagger von Helmut Schmid verwendet [Sch95]. Für diesen Tagger berichtet Schmid eine Tagginggenauigkeit von 97,5% auf unabhängigem Testmaterial.

Für die praktische Untersuchung an Sprachmaterial wurde ein Wrapper entwickelt, der aus den von Alphons erzeugten Labelfiles der Wortsegmentierung Zeitpunkte und Wörter entnimmt, die Interpunktion aus der zugehörigen Textdatei (falls vorhanden) hinzumischt und die POS-Tags bestimmt. Ergebnis sind POS-Tags mit Zeitmarkierungen, die eine eindeutige Zuordnung auch zu unter der Wortebene liegenden Einheiten gestatten. Für die Untersuchung wurde zu jeder Silbe der Lern- und Testmenge der POS-Tag des Wortes, in dem die jeweilige Silbe auftritt, der davor stehende POS-Tag, sowie der darauf folgende POS-Tag bereitgestellt und in verschiedenen Lernreihen schrittweise den 14 oben erwähnten Attributen zugefügt. Tabelle 6.13 gibt Aufschluß über die erzielten Erkennungsraten. Man entnimmt, daß die POS-Tags keinen nennenswerten Beitrag zur Erhöhung der Erkennungsraten leisten.

Verläßt man das Erkennungsparadigma und wendet sich der Problemstellung der Sprachgenerierung bzw. Sprachsynthese zu, so kann man unter Mißachtung al-

ADJA	attributives Adjektiv
ADJD	adverbiales oder prädikatives Adjektiv
ADV	Adverb
APPR	Präposition; Zirkumposition links
APPRART	Präposition mit Artikel
APPO	Postposition
APZR	Zirkumposition rechts
ART	bestimmter oder unbestimmter Artikel
CARD	Kardinalzahl
FM	fremdsprachliches Material
ITJ	Interjektion
KOUI	unterordnende Konjunktion mit „zu“ und Infinitiv
KOUS	unterordnende Konjunktion mit Satz
KON	nebenordnende Konjunktion
KOKOM	Vergleichskonjunktion
NN	normales Nomen
NE	Eigennamen
PDS	substituierendes Demonstrativpronomen
PDAT	attribuierendes Demonstrativpronomen
PIS	substituierendes Indefinitpronomen
PIAT	attribuierendes Indefinitpronomen ohne Determiner
PIDAT	attribuierendes Indefinitpronomen mit Determiner
PPER	irreflexives Personalpronomen
PPOSS	substituierendes Possessivpronomen
PPOSAT	attribuierendes Possessivpronomen
PRELS	substituierendes Relativpronomen
PRELAT	attribuierendes Relativpronomen

PRF	reflexives Personalpronomen
PWS	substituierendes Interrogativpronomen
PWAT	attribuierendes Interrogativpronomen
PWAV	adverbiales Interrogativ- oder Relativpronomen
PROAV	Pronomialadverb
PTKZU	„zu“ vor Infinitiv
PTKNEG	Negationspartikel
PTKVZ	abgetrennter Verbzusatz
PTKANT	Antwortpartikel
PTKA	Partikel bei Adjektiv oder Adverb
TRUNC	Kompositions-Erstglied
VVFIN	finites Verb, voll
VVIMP	Imperativ, voll
VVINF	Infinitiv, voll
VVIZU	Infinitiv mit „zu“, voll
VVPP	Partizip Perfekt, voll
VAFIN	finites Verb, aux
VAIMP	Imperativ, aux
VAINF	Infinitiv, aux
VAPP	Partizip Perfekt, aux
VMFIN	finites Verb, modal
VMINF	Infinitiv, modal
VMPP	Partizip Perfekt, modal
XY	Nichtwort, enthält Sonderzeichen
$,	Komma
$.	satzbeendende Interpunktion
$(	sonstige Satzzeichen; satzintern

Tabelle 6.12: Das Stuttgart-Tübingen-Tag-Set (STTS).

Merkmale der			Prepruning	
vorhergeh. Wortes	aktuellen Silbe	nachfolg. Wortes	m=2	m=5
	14		76.85	78.65
	14+POS		77.26	78.49
POS	14+POS		75.90	78.00
	14+POS	POS	76.77	76.77
POS	14+POS	POS	75.86	77.09

Tabelle 6.13: Erkennungsraten nach Hinzufügen der POS-Tags des vorangehenden, nachfolgenden, sowie beider umgebender Wörter (POS=POS-Tag).

Merkmale der aktuellen Silbe	Merkmale der beiden umgebenden Silben/Wörter	Prepruning m=2	Prepruning m=5
2 lex., 1 Lm., POS	2 lex., 1 Lm., POS	73.77	74.96
2 lex., 1 Lm., POS	POS	75.12	73.77
2 lex., 1 Lm., POS		73.89	73.60
2 lex., 1 Lm.		72.66	73.32
Wordstress, POS	POS	73.15	73.73
Syls2wordend, POS	POS	73.36	73.40

Tabelle 6.14: Erkennungsraten mit nur aus dem Text und dem Lexikon (ohne gründliche syntaktische oder semantische Analyse) gewinnbaren Merkmalen (lex.=lexikalische Merkmale, Lm.=Längenmerkmal, POS=POS-Tag).

ler am Sprachsignal gemessener Werte und bei ausschließlicher Beachtung auch aus dem Textmaterial hervorgehender Werte ein Prosodiegenerierungsmodell konstruieren. Von den oben aufgeführten können in einem solchen Szenario folgende Merkmale Verwendung finden: die POS-Tag-Merkmale, die lexikalischen Merkmale Wortakzent und Anzahl der Silben bis zum Wortende und die erwartete Silbendauer, gegebenenfalls auch von der näheren Umgebung der zu entscheidenden Silbe. Die damit erzielbaren Resultate liegen bei bis zu 75% (siehe Tabelle 6.14).

6.3.3 Intensitätsmaße

Die Intensität bzw. deren meßbares Korrelat, die Lautstärke, wurde in den bisherigen Untersuchungen nicht miteinbezogen. Da die Intensität als Faktor bei der Wahrnehmung von Prominenz gilt, wird die Merkmalsmenge in den folgenden Untersuchungen um die Intensität beschreibende Merkmale erweitert. Als erstes wird zu den Silbenmerkmalen die Lautstärke des Nukleus hinzugenommen, die wie folgt erhoben wurde.

Über die Dauer des Nukleusvokals wurde die Intensität als Root-Mean-Square-Wert (abgekürzt RMS-Wert) berechnet, Start- und Endpunkte entstammen dem automatischen Segmentierungssystem Alphons aus Kapitel 3. Um die unterschiedlichen intrinsischen Lautstärken der Vokale auszugleichen, wurden die Mittelwerte der RMS-Werte für jeden Vokal ausgerechnet. Der jeweilige Mittelwert wurde dann zur Normalisierung des Merkmals bei allen Silben abgezogen. Es erfolgte keine Normalisierung auf das Lautstärkeniveau der Äußerung.

In Kapitel 5 wurde bereits festgestellt, daß besonders die Intensität im hohen Frequenzbereich (der sogenannte Spectral Tilt) mit der Prominenz korreliert. Aus

Merkmale	Prepruning m=2	Prepruning m=5
14	76.85	78.65
14 + rms0_8k	76.89	78.28
14 + rms2_8k	76.68	78.04
14 + rms4_8k	76.23	77.67
14 + 7 rms-M.	74.75	76.97

Tabelle 6.15: Erkennungsraten nach Hinzufügen von verschiedenen Intensitätsmerkmalen.

diesem Grund wurde ein weiteres Merkmal berechnet, das die Intensität im oberen Frequenzbereich beschreibt. Dazu wurde das Sprachsignal mit einem Bandpaßfilter mit den Eckfrequenzen 2 kHz und 8 kHz gefiltert und anschließend die Intensität (wieder über die Länge des Nukleusvokals) berechnet. Von der so ermittelten Energie in diesem Frequenzband wurde die Gesamtenergie abgezogen (wegen der logarithmischen Skala wurde nicht der Quotient, sondern die Differenz gebildet). Dies ergibt ein Maß für den Energieanteil im hohen Frequenzbereich. Die so ermittelten Werte wurden wieder auf den Nukleusvokaltyp normalisiert, indem der Mittelwert abgezogen wurde. Weitere Merkmale ergaben sich durch analoge Behandlung mit anderen Unterbändern.

Tabelle 6.15 gibt einen Überblick über die Erkennungsraten in Lernläufen bei Verwendung der Gesamtenergie, der Energie im Bereich 2 kHz–8 kHz, der Energie im Bereich 4 kHz–8 kHz, und der Hinzufügung von 7 die Energie beschreibenden Merkmale (die in den vorhergehenden Läufen verwendeten rms0_8k, rms2k_8k, rms4k_8k sowie zusätzlich rms0_500, rms500_1k, rms1k_2k und rms2k_4k). Offenbar können die angebotenen Intensitätsmerkmale die Ergebnisse nicht verbessern, das Hinzufügen einzelner Merkmale scheint jedoch auch die Ergebnisse nicht nennenswert zu verschlechtern. In einer analog durchgeführten Reihe von Lernläufen wurde ermittelt, ob die Intensitätsmaße aussagekräftig genug sind, um die Längenmerkmale zu ersetzen. Dazu wurden die drei Längenmerkmale aus der dem Lernverfahren angebotenen Merkmalsmenge entfernt. Die Ergebnisse dieser Lernläufe sind in Tabelle 6.16 zusammengefaßt. Man sieht, daß in diesem Fall die Hinzufügung der Intensitätsmerkmale sich eher nachteilig auswirkt.

Als letztes wird untersucht, ob der Betonungsstatus aus dem in Kapitel 5 beschriebenen Verfahren zur Schätzung der Prominenz aus der spektralen Charakteristik ausgenutzt werden kann, um die Erkennung zu verbessern. Wie man in Tabelle 6.17 sieht, erweist sich auch diese Art, Prominenz zu beschreiben, als wenig hilfreich und sehr ähnlich zu den anderen Intensitätsmerkmalen.

Das Ausbleiben positiver Auswirkungen der Intensitäts- bzw. der spektralen Merkmale in dieser Untersuchung muß nicht bedeuten, daß die Rolle der In-

Merkmale	Prepruning m=2	m=5
11 (14 - 3 Dauerm.)	75.29	77.30
11 + rms0_8k	74.18	76.77
11 + rms2_8k	75.00	76.97
11 + rms4_8k	74.10	76.35
11 + 7 rms-M.	73.69	75.25

Tabelle 6.16: Erkennungsraten nach Hinzufügen von verschiedenen Intensitätsmerkmalen ohne Berücksichtigung von Dauermerkmalen.

Merkmale	Prepruning m=2	m=5
14	76.85	78.61
14+wsguess	76.52	78.12
11 (14 - 3 Dauerm.)	75.21	77.30
11+wsguess	75.41	76.89

Tabelle 6.17: Erkennungsraten nach Hinzufügen der Prominenzschätzung aus der spektralen Charakteristik, mit und ohne Berücksichtigung von Dauermerkmalen.

tensität bei der Wahrnehmung von Prosodie neu überdacht werden muß. Da das untersuchte Sprachmaterial nicht vollständig unter kontrollierten Laborbedingungen, sondern vom Radio aufgenommen wurde, ist nicht auszuschließen, daß durch Einsatz von Kompressoren und anderer studiotechnischer Geräte die möglicherweise vom Sprecher produzierten Intensitätsunterschiede zwischen betonten und unbetonten Silben ausgeglichen waren, und somit nicht mehr gut erkannt wurden. Man kann jedoch sicher sagen, daß die Merkmale in der beschriebenen Form bei dem vorliegenden Sprachmaterial das Prosodieerkennungsverfahren nicht verbessern.

6.3.4 Silbendauer vs. Nucleusdauer

In Untersuchungen wurde nachgewiesen, daß die Längung von betonten Silben hauptsächlich im Nukleus vorgenommen wird, während insbesondere Onset aber auch Rhyme gleichbleiben oder sogar verkürzt werden [GvH95]. Aus diesem Grund könnte die Verwendung von Merkmalen, die die Länge des Nukleus statt der Länge der Silbe beschreiben, die Erkennungsraten verbessern. Eine zusätzliche Hilfe könnte die Bereitstellung des Nukleustyps erbringen, da möglicherweise die Längenänderung bei gespannten und ungespannten Vokalen unterschiedlich realisiert wird. In dieser Untersuchung wurde bei der Einteilung der Vokale in Klassen unterschieden in Langvokale (iː, eː, ɛː, aː, oː, uː, yː, øː), Kurzvokale (ɪ, ɛ, a, ɔ, ʊ, ʏ, œ), Diphtonge (aɪ, aʊ, ɔʏ) und Schwas (ə, ɐ). Tabelle 6.18 faßt die Lernläufe zusammen. Man erkennt, daß die Silbenmerkmale besser als die Nu-

Merkmale	Prepruning	
	m=2	m=5
11 + 3 Silbendauerm.	76.85	78.61
11 + 3 Nucleusdauerm.	75.78	77.55
11 + 3 Silbendauerm. + NucType	76.23	78.41
11 + 3 Nucleusdauerm. + NucType	76.23	76.68
11 + 3 Silbendauerm. + Nucleusvok.	77.46	78.37
11 + 3 Nucleusdauerm. + Nucleusvok.	77.01	76.89

Tabelle 6.18: Vergleich der Erkennungsraten mit Silbendauermerkmalen vs. mit Nucleusdauermerkmalen, mit und ohne Berücksichtigung des Nukleustyps (Langvokal, Kurzvokal, Diphtong, Schwa) bzw. des Nukleusvokals.

kleusmerkmale abschneiden, und daß die Bereitstellung der Nukleustypinformation keine Vorteile bringt. Genauso bietet auch die Bereitstellung der Vokalqualität keine Vorteile.

6.4 Von sprecherabhängiger Erkennung zu sprecherunabhängiger Erkennung

In den Untersuchungen, die bis jetzt in diesem Kapitel vorgestellt wurden, stammten sowohl die Sprachdaten, die zum Training des Klassifikators dienten, als auch die Daten, mit denen der Klassifikator getestet wurde, jeweils vom selben Sprecher. Ein Klassifikator, der mit Trainingsmaterial eines Sprechers gelernt wurde, läßt sich nicht ohne weiteres sinnvoll auf die Sprachdaten eines anderen Sprechers anwenden. Die offensichtlichsten systematischen Unterschiede zwischen verschiedenen Sprechern liegen in der persönlichen Stimmlage und der persönlichen Sprechgeschwindigkeit, aber auch darin, wie lange die verschiedenen Phoneme ausgesprochen werden. Wenn beispielsweise ein Entscheidungsbaum mit Trainingsdaten eines Sprechers gelernt wird, sind etwa möglicherweise darin vorkommende Schwellwerte für die Höhe eines Gipfels für Sprecher mit stark abweichender Stimmlage nicht angemessen. Besonders wichtig ist die Berücksichtigung der Stimmlage, wenn in den Trainings- oder Teststichproben sowohl Männer- als auch Frauenstimmen enthalten sind. Die Längenmessungen von Phonemen und Silben sind direkt von der Sprechgeschwindigkeit abhängig, die von Sprecher zu Sprecher sehr unterschiedlich ausfallen kann. Damit sind auch Schwellwerte für die Dauermerkmale möglicherweise nicht adäquat. Die Phonemdauern variieren zum Beispiel mit der Dialektregion, aus der die äußernde Person stammt. Auch ist es plausibel anzunehmen, daß sich sprecher- oder dialektgruppenspezifische, systematische Aussprachevarianten auf die Silbendauermessung auswirken, da die

hier eingesetzte Berechnung der Silbendauer auf der phonemischen Segmentierung von Alphons beruht, und diese die kanonische Standardaussprache unterstellt.

6.4.1 Sprechereinflüsse auf die Dauermerkmale

Für die Dauer wurden in den Untersuchungen seither drei Merkmale erhoben, nämlich die gemessene Silbendauer (Lenmeasure), die Silbenlänge, die aufgrund der in der Silbe enthaltenen Phoneme zu erwarten ist (Lenexpected), und das Verhältnis zwischen beiden (Lenrelative). Wenn ein Sprecher systematisch einzelne Phonemlängen verändert, wirkt sich dies direkt auf das Merkmal Lenmeasure aus. Das zweite Längenmerkmal, die erwartete Silbenlänge, unterliegt einem Sprechereinfluß nur dann, wenn für die Berechnung der Erwartungswerte der Phonemlängen, aus denen die erwartete Silbenlänge abgeleitet wird, Material dieses Sprechers verwendet wird. Da die Radionachrichten Datenmaterial in ausreichendem Umfang lieferten, konnten daraus die erwarteten Phonemlängen für die Sprecher individuell bestimmt werden.[7] Damit unterliegt in dieser Anwendung auch das Merkmal Lenexpected einem Sprechereinfluß. Für das letzte Dauermerkmal Lenrelative, das als Quotient aus gemessener und erwarteter Dauer berechnet wird, ist folglich der geringste Sprechereinfluß zu erwarten. Aus diesen Überlegungen geht hervor, daß für eine sprecherunabhängige Erkennung ausschließlich das Merkmal Lenrelative herangezogen werden sollte. In Anwendungen, bei denen der Umfang des Sprachmaterials es nicht erlaubt, die Erwartungswerte der Phonemlängen für jeden Sprecher individuell zu bestimmen,[8] muß bei der Berechnung des Merkmals Lenexpected auf Statistiken von anderen Sprechern zurückgegiffen werden. In diesem Fall wirkt sich der Sprechereinfluß auf das Merkmal Lenrelative und nicht auf Lenexpected aus.

6.4.2 Sprechereinflüsse auf die Grundfrequenzmerkmale

Eine Möglichkeit, die Beeinflussung der Grundfrequenzmerkmale durch die Stimmlage zu kompensieren, besteht darin, die Grundfrequenzmessungen nicht

[7] Van Santen et al. [vSSM+97] berichten von einer Methode, die für eine stabile Abschätzung von Längenstatistiken mit sehr wenig Sprachmaterial auskommt, indem phonologisch begründbare starke Modellannahmen getroffen werden. Ein solcher Ansatz wurde in der vorliegenden Arbeit nicht verfolgt.

[8] Für die in [vSSM+97] vorgeschlagene Methode reichen unter Umständen bereits wenige Sätze aus zur Abschätzung einer Phonemlängenstatistik [vSB97].

direkt durch Absolutwerte in Hertz auszudrücken, sondern sie als Relativwerte mit Bezug zur Stimmlage darzustellen.

Berechnung der Stimmlage

Die persönliche Stimmlage eines Sprechers kann theoretisch ganz einfach durch den höchsten und den niedrigsten vom Pitch tracker ausgegebenen Grundfrequenzwert angegeben werden. Praktisch produziert der Pitch tracker jedoch gelegentlich viel zu hohe oder viel zu niedrige Grundfrequenzwerte, so daß die vom Pitch tracker ausgegebenen Extremwerte nicht geeignet sind zur Beschreibung der Stimmlage. Um die Bestimmung der Stimmlage gegen die Fehler bei der Sprachgrundfrequenzschätzung robuster zu machen, wurde sie in dieser Arbeit wie folgt ermittelt.

Aus den rohen Grundfrequenzdaten der stimmhaften Segmente wurden das 5%-Quantil, $q_{0,05}$, und das 98%-Quantil, $q_{0,98}$, ermittelt, also die Grundfrequenzwerte, die nur von 5% der Grundfrequenzmessungen unterschritten, beziehungsweise nur von 2% der Messungen überschritten werden. Von diesen wurde wieder auf den gesamten Stimmumfang extrapoliert durch Berechnung der beiden Werte $l = q_{0,05} - \frac{5}{93} \times (q_{0,98} - q_{0,05})$ und $h = q_{0,98} + \frac{2}{93} \times (q_{0,98} - q_{0,05})$.

Bereich für die Stimmlagenberechnung

Die Sprachabschnitte, über die die Stimmlage auf diese Weise geschätzt wird, dürfen nicht zu klein gewählt werden, da sich sonst das Auftreten einzelner Akzente und Grenztöne zu stark auf die Schätzung auswirkt. Als Sprachabschnitte kommen Intonationsphrasen, Signalabschnitte zwischen längeren Sprechpausen, Redebeiträge oder die Gesamtheit aller Sprachdaten des jeweiligen Sprechers in Betracht. Die Intonationsphrase ist als Bereich für die Berechnung interessant, da dadurch neben dem Sprechereinfluß auch der Einfluß durch Register ausgeglichen würde. Gegen die Verwendung der Intonationsphrase als Bereich spricht, daß dafür die Phrasengrenzen und (wegen den mitunter sehr kurzen Phrasen) möglichst auch die auftretenden Akzente schon bekannt sein sollten, um gute Abschätzungen der Stimmlage zu erhalten. Da die Bestimmung der Stimmlage aber eine Vorstufe für die Erkennung von Phrasengrenzen und Akzenten sein soll, kommt die Intonationsphrase als Bereich zunächst nicht in Betracht.

Die nächste Annäherung an die Intonationsphrase, die sich leicht aus dem Signal ableiten läßt, ist die Segmentierung der Sprache durch Pausen. Obwohl zwar

Intonationsphrasen vorkommen können, auf die nicht unmittelbar eine Sprechpause folgt oder die (zumindest kurze) Sprechpausen überspannen, ist doch das Zusammenfallen von Intonationsphrasengrenze und Sprechpause die Regel. Die Sprechpausen werden vom automatischen Segmentierungsverfahren Alphons aus Kapitel 3 mit ausreichender Genauigkeit erkannt, oder sind alternativ durch Auswertung von Energiemessungen in einem über das Sprachsignal verschobenen Fenster mit Signalverarbeitungsmethoden einfach zu bestimmen. Wenn man sehr kurze Sprachstücke, deren Länge einen bestimmten Mindestwert unterschreitet, mit benachbarten Sprachstücken zusammenfaßt, ergeben sich ausreichend lange, automatisch bestimmbare Bereiche für die Berechnung der Stimmlage.

Redebeiträge, also größere, mehrere Intonationsphrasen umfassende, inhaltlich zusammenhängende Äußerungen, können ebenfalls als Bereich für die Stimmlagenberechnung Verwendung finden. Es ist zu erwarten, daß ein solcher Bereich nur wenig von der Verteilung der auftretenden Akzente und Grenztöne und der Registerwahl abhängig ist. Am stärksten wird von der Register- und Akzent/Grenzton-Verteilung abstrahiert, wenn das gesamte verfügbare Material des jeweiligen Sprechers zur Abschätzung der Stimmlage verwendet. Nachteilig könnte es sich dabei jedoch auswirken, wenn von manchen Sprechern sehr wenig Sprachmaterial vorhanden ist, da dann ein Teil der Daten durch Register oder Intonation beeinflußt bleibt, während der andere Teil nahezu gänzlich von diesem Einfluß bereinigt ist.

Aus Gründen der einfachen Realisierung wurde in dieser Arbeit ein künstliches Kriterium zur Festlegung des Bereiches gewählt, nämlich das gesamte, aktuell bearbeitete Sprachsignal. Im Falle der Nachrichten ist jeweils eine Einzelnachricht in einer Datei gespeichert, bestehend aus durchschnittlich 73.6 Wörtern bei einer Standardabweichung von 23.7 Wörtern. Auf eine Dialogsituation übertragen entspricht die Länge einer Einzelnachricht damit in etwa einem längeren Redebeitrag. Zumindest für die Nachrichten hat aufgrund der Bereichsgröße die Wahl von Registern bei der Stimmlagenberechnung keinen nennenswerten Einfluß. Für das Klassifikatorenlernen zur Vorhersage der phonologischen Intonationsbeschreibung bringt diese Festlegung der Bereichsgöße den Nachteil mit sich, daß gleiche Akzente im Merkmalsraum potentiell weit voneinander entfernt liegen können, wenn sie aus Intonationsphrasen mit unterschiedlichen Registern stammen.

Stimmlagenberechnung für die Radionachrichten

Die Wirksamkeit der beschriebenen Stimmlagenberechnung soll am Beispiel der Radionachrichtenmeldungen gezeigt werden. In Abb. 6.5 sind die ermittelten mi-

nimalen und maximalen F_0-Werte l und h sowie der Mittelwert der Grundfrequenz für einzelne Nachrichtenmeldungen abgetragen. Die ersten 3 Nachrichtensendungen sind von einer Frau verlesen, deren Stimmlage etwa zwischen 100 und 260 Hz liegt; der männliche Ansager der restlichen Nachrichten spricht mit einer Sprachgrundfrequenz zwischen etwa 70 und 140 Hz. Die Messungen der Einzelnachrichten einer Nachrichtensendung sind mit einer Linie verbunden, die Messung des Wetterberichts ist stets der am weitesten rechts stehende Punkt einer durch Linien verbundenen Punktmenge. Die Abbildung zeigt, daß die Bestimmung der Stimmlage für die einzelnen Nachrichten ungefähr übereinstimmende Werte ergibt. Lediglich die untere Grenze der Stimmlage bei der Frauenstimme ist etwas uneinheitlich von Nachrichtenmeldung zu Nachrichtenmeldung, die Bestimmung könnte in diesem Punkt noch verbessert werden. Die drei am meisten abweichenden Werte werden bei den Wettermeldungen ermittelt. Interessant ist der Trend, daß sowohl bei der Frauen- als auch bei der Männerstimme die Wetterberichte mit einer Stimmlage mit größerem Umfang verlesen werden als zumindest die vorhergehende Nachrichtenmeldung. Dies könnte durch einen veränderten prosodischen Stil verursacht oder auch ein Artefakt der Stimmlagenberechnung sein.

Normalisierung der Grundfrequenzmerkmale auf die Stimmlage

Nicht alle der Grundfrequenzmerkmale müssen an die Stimmlage angepaßt werden. Die Parameter Tonalalign und Peakalign beziehen sich auf abstrakte Silbeneinheiten, also Zeiten, und sind daher nicht von der Stimmlage abhängig, obwohl auch sie selbstverständlich einem Sprechereinfluß außer der Stimmlage unterliegen können. Dagegen sind die Parameter Tonaldiff, Peakheight und Level zweifellos von der Stimmlage abhängig. Sie wurden wie folgt normiert: Tonaldiff und Peakheight wurden durch den Range $r = h - l$ geteilt. Level wurde um l vermindert und durch den Range geteilt. Durch diese Normierungen bewegt sich Tonaldiff statt im Bereich zwischen etwa $-\frac{r}{2}$ und $+\frac{r}{2}$ nun zwischen $-0,5$ und $+0,5$. Peakheight liegt damit statt zwischen $-r$ und $+r$ nun zwischen -1 und $+1$, und Level statt im Bereich zwischen l und h zwischen 0 und 1. Eine Normierung der verbleibenden beiden Parameter Tonalsteep und Peaksteep wurde nicht vorgenommen.

6.4.3 Experimentelle Untersuchungen zur Kompensation von Sprechereinflüssen

Die diskutierten Methoden zur Kompensation des Sprechereinflusses werden in verschiedenen Untersuchungen experimentell ausgewertet.

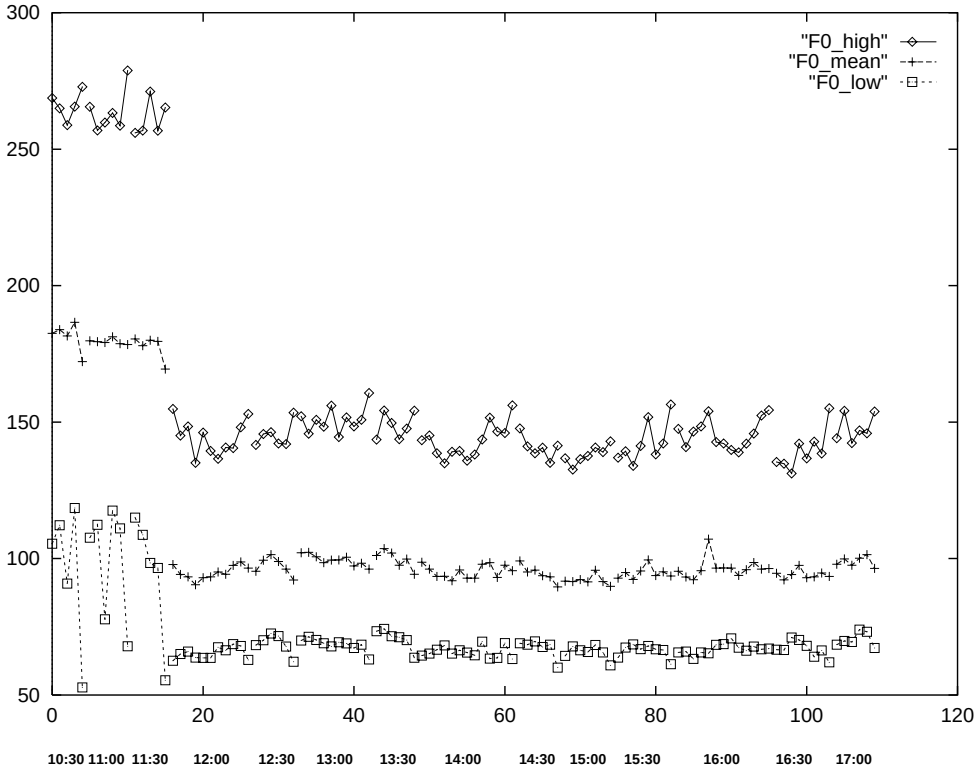


Abbildung 6.5: Automatisch ermittelte Stimmlagen bei den Nachrichtenmeldungen des Deutschlandfunks vom 21.11.1997.

Merkmale	Prepruning m=2	Prepruning m=5
7+7 F_0	76.85	78.61
7+7 F_0 norm.	75.57	77.13

Tabelle 6.19: Vergleich der Erkennungsraten mit Grundfrequenzmerkmalen ohne und mit Normalisierung auf den pro Nachricht detektierten Grundfrequenzbereich (gleicher Nachrichtensprecher).

Merkmale	Prepruning m=2	Prepruning m=5
6 + 7 F_0	76.07	77.59
6 + 7 F_0 norm.	75.86	77.59

Tabelle 6.20: Vergleich der beiden unterschiedlichen Grundfrequenzmerkmalssätze nach Entfernen des nicht normierten Längenmerkmals (gleicher Nachrichtensprecher).

Verluste durch die Normierung

Um zu untersuchen, welche Nachteile die vorgeschlagene Normierung der Grundfrequenzmerkmale mit sich bringt, wurden Lernläufe mit denselben Trainings- und Testdaten wie in Abschnitt 6.3 durchgeführt, wobei die Grundfrequenzmerkmale durch die normierten Merkmale ersetzt wurden. Alle anderen Merkmale blieben unverändert. In Tabelle 6.19 sind die Ergebnisse mit den normierten Daten den entsprechenden Ergebnissen mit nicht normierten Daten gegenübergestellt. Da dabei nur Sprachmaterial von einem einzigen Sprecher verwendet wird, sollte sich die Normierung im Idealfall weder positiv noch negativ auswirken. Es stellt sich jedoch eine leichte Verschlechterung von ungefähr einem Prozent bei beiden Lernläufen ein.

In Tabelle 6.20 sind die gleichen Messungen nach Entfernen des Merkmals Lenmeasure erneut durchgeführt worden. Der Wegfall dieses Merkmals kann recht gut ausgeglichen werden, es werden ungefähr gleich gute Ergebnisse erzielt.

6.4.4 Vorteile durch die Normierung

Nach den Untersuchungen an Sprachdaten eines einzigen Sprechers, die zeigten, daß die Normierung keine starken Nachteile mit sich bringt, werden im nächsten Experiment Daten untersucht, bei denen die Trainings- und die Testdaten von unterschiedlichen Sprechern stammen. Damit soll überprüft werden, ob bei verschiedenen Sprechern die Normierung gegenüber der Verwendung nicht

Merkmale	Prepruning m=2	Prepruning m=5
7+7 F_0	70.32	72.04
7+7 F_0 norm.	71.49	73.55

Tabelle 6.21: Vergleich der Erkennungsraten mit Grundfrequenzmerkmalen ohne und mit Normalisierung auf den pro Nachricht detektierten Grundfrequenzbereich (Training: Männerstimme, Test: Frauenstimme).

Merkmale	Prepruning m=2	Prepruning m=5
6+7 F_0	68.35	71.14
6+7 F_0 norm.	72.14	73.81

Tabelle 6.22: Vergleich der beiden unterschiedlichen Grundfrequenzmerkmalssätze nach Entfernen des nicht normierten Längenmerkmals (Training: Männerstimme, Test: Frauenstimme).

normierter Daten Vorteile erbringt. In den folgenden Läufen ist die in den vorherigen Experimenten verwendete Teststichprobe des männlichen Sprechers, von dem auch die Trainingsdaten stammen, komplett gegen eine Teststichprobe einer weiblichen Sprecherin ausgetauscht. Die neue Teststichprobe wurde aus 32 Einzelnachrichten (einschließlich zweier Wetterberichte) aus 6 Nachrichtensendungen analog zur Trainingsstichprobe erstellt. Die neue Teststichprobe umfaßt 4903 Silben. Die Längenstatistik zur Berechnung des Merkmals Lenexpected wurde für die Frauenstimme getrennt berechnet. Andere sprecherspezifische Faktoren wurden nicht kompensiert. Wenn etwa die Nachrichtensprecherin einen steigenden Akzent konsequent früher realisiert als ihr männlicher Kollege, kann dies zu Mißklassifikationen führen. Aus diesem Grund muß man für die normalisierten Grundfrequenzwerte schlechtere Ergebnisse erwarten als bei dem Lauf mit identischem Sprecher in Lern- und Teststichprobe. Bei den nicht normalisierten Grundfrequenzmerkmalen sind deutlich schlechtere Ergebnisse zu erwarten.

Tabellen 6.21 und 6.22 zeigen, daß die Normierung der Grundfrequenzmerkmale etwa 1.5–2.5% Verbesserung erbringt. Die besten Resultate erhält man, wenn man das Merkmal Lenmeasure ignoriert. Allerdings ist der Vorteil sehr gering und auch nur im Zusammenhang mit den normierten Merkmalen zu beobachten.

Da in der neuen Testmenge mehrere Downstep-Akzente auftraten, für die keine Trainingsbeispiele vorhanden sind, wurden die eben gezeigten Lernläufe wiederholt, nachdem die Downstep-Akzente den zugrundeliegenden Akzentklassen sowohl in der Lern- als auch in der Teststichprobe zugeschlagen wurden, analog zum Vorgehen in Abschnitt 6.2.4. Die Resultate dieser Läufe sind in den Tabel-

Merkmale	Prepruning	
	m=2	m=5
7+F_0	72.02	73.44
7+F_0 norm.	71.85	74.38

Tabelle 6.23: Vergleich der Erkennungsraten mit Grundfrequenzmerkmalen ohne und mit Normalisierung auf den pro Nachricht detektierten Grundfrequenzbereich, ohne Berücksichtigung von Downstep (Training: Männerstimme, Test: Frauenstimme).

Merkmale	Prepruning	
	m=2	m=5
6+7 F_0	69.77	73.89
6+7 F_0 norm.	71.51	73.67

Tabelle 6.24: Vergleich der Erkennungsraten der unterschiedlichen Grundfrequenzmerkmalssätze nach Entfernen des nicht normierten Längenmerkmals und ohne Berücksichtigung von Downstep (Training: Männerstimme, Test: Frauenstimme).

len 6.23 und 6.24 wiedergegeben. Es ergeben sich im Vergleich zu Tabelle 6.21 beziehungsweise 6.22 etwas bessere Ergebnisse. Da die Unterschiede zwischen den Läufen mit und ohne Grundfrequenznormalisierung hier nicht mehr so deutlich sind, ist zu vermuten, daß die Normierung der Grundfrequenz besonders zur Unterscheidung zwischen den Downstep- und den Nicht-Downstep-Varianten der Akzente beiträgt.

6.4.5 Experimente zur Bewertung des Bedarfs an Trainingsmaterial

Die von Sprecher zu Sprecher abweichende phonetische Realisierung der phonologischen Intonationsbeschreibungssymbole, also etwa zu welchem Zeitpunkt ein fallender Akzent produziert oder wie hoch ein hoher Grenzton realisiert wird, dürfte ebenfalls zu den schlechteren Erkennungsraten bei unterschiedlichen Sprechern beitragen. Analog zum Vorgehen in der Spracherkennung sollte auch bei der Erkennung der Prosodiebeschreibung durch Aufnahme möglichst vieler Sprecher und Sprecherinnen in die Lernmenge die Variation in den Entscheidungsbäumen mitgelernt werden, um so eine bessere sprecherunabhängige Erkennung zu ermöglichen. Dazu sind zusätzliche, manuell prosodisch annotierte Daten notwendig.

Da z. Z. am IMS neben den Nachrichtensendungen mit nur etwa 10 Minuten annotiertem Material verschiedener anderer Sprecher zuwenig Daten für eine spre-

cherunabhängige Erkennung verfügbar sind, können keine sehr aussagekräftige Experimente durchgeführt werden. Stattdessen werden für verschiedene Lern- und Testmengen Lernläufe gezeigt, die Anhaltspunkte für die Bewertung eines sprecherunabhängigen Prosodierekennungssystems geben können, auch wenn die Ergebnisse sicher nicht an das erzielbare Optimum heranreichen. An annotierten Sprachdaten stehen am IMS zur Verfügung:

MNSL: die in den obigen Untersuchungen als Lernmenge eingesetzten, von einem männlichen Nachrichtensprecher vorgetragenen Radionachrichten (10445 Silben),

MNST: die in den ersten Experimenten als Testmenge eingesetzten, von einem männlichen Nachrichtensprecher vorgetragenen Radionachrichten (2436 Silben),

WNST: die in den letzten Experimenten als Testmenge eingesetzten, von einer weiblichen Nachrichtensprecherin vorgetragenen Radionachrichten (4903 Silben),

DIKI: ein von 6 Sprechern (4 weiblich, 2 männlich) gelesener Ausschnitt aus der Kurzgeschichte „Das Dicke Kind“ (1175 Silben),

BUTT: eine von 4 Sprechern (2 weiblich, 2 männlich) gelesene kurze Geschichte „Die Buttergeschichte“ (1536 Silben),

WERT: das von einem männlichen Sprecher vorgetragene Vorwort zu „Die Leiden des jungen Werther“ (128 Silben).

Um den experimentellen Aufwand zu begrenzen, werden ausschließlich Daten verwendet, bei denen die Downstep-Varianten ihren jeweiligen Grundklassen zugeschlagen wurden. Weiterhin wurde für das Prepruning die geforderte Mindestzahl vorhandener Fälle stets auf 5 festgelegt ($m = 5$). Von den Transkribenten mit einem Fragezeichen annotierte Unsicherheit wurde unterdrückt.[9]

[9] In Vorexperimenten wurde festgestellt, daß das Entfernen unsicherer Fälle aus der Lernmenge keine wesentlichen Verbesserungen bewirkt.

Training		Test	
DIKI,BUTT,WERT,WNST	(7742 Silben)	MNST(2436 Silben)	75.49
DIKI,BUTT,WERT,MNST	(5275 Silben)	WNST(4903 Silben)	73.44
DIKI,BUTT,WERT,MNST,teils MNSL	(7742 Silben)	WNST(4903 Silben)	74.14
DIKI,BUTT,WERT,MNST,MNSL	(15720 Silben)	WNST(4903 Silben)	75.44

Tabelle 6.25: Erkennungsergebnisse bei kleiner Lernmenge verschiedener Sprecher und Sprecherinnen für männlichen Nachrichtensprecher (oben) und weibliche Nachrichtensprecherin (untere 3 Zeilen).

Vorhersage einer Männerstimme versus einer Frauenstimme

Im ersten Experiment wird aus relativ wenig, von verschiedenen Sprechern und Sprecherinnen stammendem Material (WNST, DIKI, BUTT, WERT) ein Klassifikator gelernt, der dann an einem nicht in der Lernmenge enthaltenen Sprecher (MNST) getestet wird. Direkt anschließend werden die Daten der weiblichen Nachrichtensprecherin und des männlichen Nachrichtensprechers ausgetauscht, also aus MNST, DIKI, BUTT, WERT dann WNST vorhergesagt. Als Merkmale stehen die normierten Grundfrequenzmerkmale, alle 3 Silbendauermerkmale, die beiden lexikalischen und die beiden Pausenmerkmale zur Verfügung.

Tabelle 6.25 zeigt, daß WNST zunächst schlechter vorhergesagt wird, da WNST und MNST unterschiedlich groß sind und somit auch die Lernmenge im zweiten Lauf einen geringeren Umfang besitzt als im ersten Lauf. Wenn man die Trainingsmenge mit Silben vom männlichen Nachrichtensprecher ergänzt, bis sie den gleichen Umfang aufweist wie im ersten Lauf, bleibt die Frauenstimme immer noch etwas schlechter vorhersagbar. Wenn schließlich alle Silben aus MNSL für das Training verwendet werden, erreicht die Erkennungsrate auf WNST die Werte von MNST. Man kann also schließen, daß die männliche Nachrichtenstimme aus den restlichen Daten etwas besser vorhersagbar ist als die weibliche. Wenn man jedoch das Merkmal Lenmeasure nicht verwendet, so läßt sich die Annäherung der Erkennungsraten der weiblichen Stimme an die männliche Stimme mit wachsender Trainingsmenge nicht mehr beobachten. Der Trend, daß die weibliche Stimme schwerer vorherzusagen ist, bleibt jedoch bestehen (s. Tabelle 6.26).

Verbesserung der sprecherunabhängigen Erkennung durch Hinzufügen von umfangreichen Sprachdaten eines einzelnen Sprechers

Die vorhergehende Untersuchungen haben gezeigt, daß die Prosodie verschiedener Sprecher unterschiedlich gut zu erkennen ist und daß sich die Erkennungsraten manchmal dadurch verbessern lassen, daß Daten eines Sprechers in die Trainingsmenge aufgenommen werden, der nicht in der Testmenge vorkommt. Um

Training		Test	
DIKI,BUTT,WERT,WNST	(7742 Silben)	MNST(2436 Silben)	75.45
DIKI,BUTT,WERT,MNST	(5275 Silben)	WNST(4903 Silben)	73.93
DIKI,BUTT,WERT,MNST,teils MNSL	(7742 Silben)	WNST(4903 Silben)	74.63
DIKI,BUTT,WERT,MNST,MNSL	(15720 Silben)	WNST(4903 Silben)	74.42

Tabelle 6.26: Erkennungsergebnisse bei kleiner Lernmenge verschiedener Sprecher und Sprecherinnen für männlichen Nachrichtensprecher (oben) und weibliche Nachrichtensprecherin (untere 3 Zeilen). Bei diesen Läufen wurde das Merkmal Lenmeasure nicht berücksichtigt.

Training		Test	
DIKI,WERT,WNST,MNST	(8642 Silben)	BUTT(1536 Silben)	74.15
DIKI,WERT,WNST,MNST,MNSL	(19087 Silben)	BUTT(1536 Silben)	74.22

Tabelle 6.27: Erkennungsergebnisse bei kleiner Lernmenge verschiedener Sprecher und Sprecherinnen sowie mehreren Sprechern und Sprecherinnnen einer Testmenge. Ein Hinzufügen von viel Material eines einzelnen Sprechers erbringt keine im Verhältnis stehenden Vorteile.

nun den Einfluß der unterschiedlich leichten Vorhersagbarkeit verschiedener Sprecher einzudämmen, wird in der folgenden Untersuchung eine aus mehreren Sprechern stammende Testdatenmenge verwendet. Es wurde der Datensatz BUTT ausgewählt, da er Sprachmaterial von männlichen und weiblichen Sprechern in gleichen Anteilen enthält. Es wird getestet, ob sich die Erkennung auf diesem Testkorpus mit gemischten Sprechern — also die sprecherunabhängige Erkennung — dadurch verbessern läßt, daß die Trainingsmenge um umfangreiches Datenmaterial eines einzelnen Sprechers erweitert wird. Als Trainingsmenge wird zunächst ein recht ausgewogenes, relativ kleines Set von Material verschiedener Sprecher verwendet, das für einen Vergleichslauf um umfangreiches Material eines einzelnen Sprechers erweitert wird. Da die Resultate der seither durchgeführten Läufe bei Ausschluß des Merkmals Lenmeasure im allgemeinen leicht besser ausgefallen sind, wird auch im folgenden der auf 13 Merkmale beschränkte Merkmalssatz verwendet. Wie Tabelle 6.27 zeigt, führt die Hinzunahme umfangreichen Datenmaterials eines einzelnen Sprechers zu keinen wesentlichen Verbesserungen für die sprecherunabhängige Erkennung.

Strategien zur Annotation eines großen Einzelsprecherkorpus

Mit den folgenden Experimenten wird untersucht, ob sich die Erkennung der Prosodie für einen bestimmten Sprecher dadurch verbessern läßt, daß ein Teil des von diesem Sprecher stammenden Materials manuell oder automatisch annotiert und dann in die Trainingsmenge aufgenommen wird. Als realistisches Szenario

kann man sich vorstellen, daß ein großes Korpus eines Einzelsprechers prosodisch annotiert werden soll, von dem zunächst noch kein annotiertes Material zur Verfügung steht.

Die folgenden Strategien werden miteinander verglichen.

1. Ein Teil des zu etikettierenden Korpus wird manuell prosodisch annotiert und zur Trainingsmenge hinzugenommen. Der damit trainierte Klassifikator wird auf den verbleibenden Rest des Korpus angewendet.

2. Der Klassifikator wird nur auf den manuell annotierten Sprachdaten aus dem Zielkorpus trainiert, weitere Trainingsdaten werden nicht hinzugezogen. Der damit trainierte Klassifikator wird auf das verbleibende Restkorpus angewendet.

3. Für eine kostengünstigere Strategie wird die manuelle Labelphase umgangen und der gesamte Zielkorpus mit einem vorläufigen Klassifikator automatisch annotiert. Die so annotierten Daten werden dann ganz oder teilweise zur Trainingsmenge für einen weiteren Lernlauf hinzugefügt, um so eine Anpassung an den Sprecher zu ermöglichen.

Die beschriebenen Strategien werden miteinander verglichen anhand der (fiktiven) Aufgabenstellung, das Korpus mit den Daten des männlichen Nachrichtensprechers automatisch zu annotieren. Die übrigen Daten, also DIKI, BUTT, WERT und WNST dienen als initiale Trainingsmenge. Zum Vergleich wird mit dieser Trainingsmenge zunächst ein Baseline-Klassifikator trainiert, der auf dem Datensatz MNSL, also dem größeren Teil des vom männlichen Nachrichtensprecher stammenden Korpus, getestet wird. Das Ergebnis ist in der ersten Zeile von Tabelle 6.28 wiedergegeben.

Die erste der beschriebenen Strategien wird realisiert, indem die initiale Trainingsmenge um die manuell annotierten Daten MNST des männlichen Nachrichtensprechers erweitert wird.[10] Der damit trainierte Klassifikator wird auf MNSL getestet. Die Erkennungsrate ist in der zweiten Zeile von Tabelle 6.28 abgebildet.

Die zweite der beschriebenen Strategien stellt eine rein sprecherabhängige Erkennung dar, bei der ausschließlich die manuell annotierten Daten MNST des

[10] Die Bezeichnungen MNSL und MNST sind hier etwas verwirrend, da der größere Datensatz des Nachrichtensprechers, MNSL, der seither zum Training verwendet wurde, nunmehr die Testmenge bildet, und die ehemalige Testmenge, MNST, als simulierter manuell annotierter Teil des Sprecherkorpus der Lernmenge zugeschlagen wird.

Training		Test	
DIKI,WERT,BUTT,WNST	(7742 Sil.)	MNSL(10445 Sil.)	74.73
DIKI,WERT,BUTT,WNST,MNST	(10178 Sil.)	MNSL(10445 Sil.)	76.19
MNST	(2436 Sil.)	MNSL(10445 Sil.)	74.82
DIKI,WERT,BUTT,WNST,MNST*	(10178 Sil.)	MNSL(10445 Sil.)	75.42
DIKI,WERT,BUTT,WNST,MNSL*	(18187 Sil.)	MNSL(10445 Sil.)	75.12
DIKI,WERT,BUTT,WNST,MNSL*,MNST*	(20623 Sil.)	MNSL(10445 Sil.)	75.15

Tabelle 6.28: Verschiedene Strategien, um die automatische Erkennung eines neuen Sprechers zu verbessern.
Oben: Baseline-System, nur Trainingsdaten von anderen Sprechern.
2. Zeile: Erste Strategie, 1/5 der Zieldaten wurden manuell annotiert und dem anfänglich vorhandenen Trainingsmaterial von anderen Sprecher hinzugefügt.
3. Zeile: Zweite Strategie, nur das manuell neu gelabelte Material wurde zum Training verwendet.
4. bis 6. Zeile: Dritte Strategie, 1/5, 4/5 bzw. alle Daten des Zielkorpus wurden maschinell annotiert und der anfänglichen Trainingsmenge zugefügt.
Alle Erkennungsraten wurden aufgrund der manuellen Etikettierung von den verbleibenden 4/5 des Korpus ermittelt. Diese standen in keinem der Läufe im Training zur Verfügung.

Zielsprechers als Trainingsmaterial Verwendung finden (siehe 3. Zeile in Tabelle 6.28).

Die dritte Strategie wird in drei Konstellationen untersucht. Auf die manuelle Ettikettierung der Daten des Zielsprechers wird verzichtet; statt dessen werden die Daten durch den am initialen Datenmaterial trainierten Baseline-Klassifikator automatisch annotiert. In einer zweiten Lernphase werden Teile des automatisch annotierten Korpus in verschiedenem Umfang der Ausgangstrainingsmenge hinzugefügt und damit ein zweiter Klassifikator erstellt. Dessen Erkennungsleistung wird ausgewertet auf MNSL durch Vergleich mit der manuell erstellten Etikettierung, die dem Lernverfahren in der Trainingsphase nicht zugänglich war. Dieses iterierte Vorgehen stellt eine Bootstrapping-Strategie dar, da das Gesamtverfahren durch selbstgewonnene Zwischenergebnisse verbessert werden soll.

Aus den Ergebnissen kann man zwei Tendenzen ablesen. Erstens zeigt sich, daß es hilfreich ist, Informationen über den zu etikettierenden Sprecher in die Trainingsmenge aufzunehmen. Die erste Zeile in Tabelle 6.28, bei der beim Training keinerlei Information über den Sprecher vorliegt, enthält das schlechteste Ergebnis. Zweitens beeinflußt die Qualität der Etikettierung die Erkennungsleistung. Die besten Ergebnisse werden erzielt, wenn man manuell annotiertes Material des zu etikettierenden Sprechers der Trainingsmenge hinzufügt. Dies ist wenig überraschend. Interessanter ist, daß auch die zusätzliche Verwendung von automatisch

gelabeltem Material die Erkennung verbessert, wobei jedoch das automatisch gelabelte Material die Trainingsmenge nicht dominieren sollte: bei den Läufen, die auch automatisch gelabeltes Material enthalten, schneidet der Lauf mit einem Fünftel automatisch gelabeltem Zielsprechermaterial am besten ab. Dabei macht das automatisch gelabelte Material etwa ein Viertel der Trainingsmenge aus. Der Verzicht auf die Daten anderer Sprecher bringt keinen Vorteil, wie die Ergebnisse der dritten Zeile zeigen.

6.5 Zusammenfassung

Dieses Kapitel behandelt die Frage, wie sich aus den automatisch aus dem Sprachsignal abgeleiteten phonetischen Beschreibungen der Intonation (Dauer, Grundfrequenzverlauf, Intensität) eine kategorielle, phonologische Beschreibung der Intonation gewinnen läßt. Diesen Übergang von der subsymbolischen zur symbolischen Ebene leisten hier Klassifikatoren, die ein Lernverfahren aus einem Korpus automatisch erzeugt, für das sowohl die phonetische Ausgangs- als auch die phonologische Zielbeschreibung vorliegt. Voruntersuchungen dienten dazu, verschiedene Ansätze zum Lernen der Klassifikatoren zu erproben und miteinander zu vergleichen. Dabei waren mit den beiden statistischen Ansätze mit HM-Modellierung keine zufriedenstellenden Ergebnisse zu erreichen. Das ILP-System Foil erzielte bessere Resultate, benötigte dafür aber sehr lange Trainingszeiten, was seine Verwendung für weitere Experimente ausschloß. Das Attribut-Wert-Lernsystem C4.5 kam zu Ergebnissen von vergleichbarer Qualität wie Foil bei Lernzeiten im Minutenbereich.

Bei den weiteren Experimenten kam daher das Lernsystem C4.5 zum Einsatz. In einer Versuchsreihe wurden die mit verschiedenen Merkmalssätzen erzielbaren Erkennungsraten verglichen, um so die geeignetste Merkmalskombination für die Vorhersage der phonologischen Intonationsbeschreibung zu finden. Die Hauptergebnisse dieser Versuchreihe sind, (1) daß zusätzliche Information über benachbarte Silben keinen Vorteil bei der Erkennung bringt, (2) daß sich die Intensitätskorrelate ebenfalls nicht positiv auf die Erkennungsraten auswirken, (3) daß Silbendauermerkmale Nucleusdauermerkmalen tendenziell überlegen sind, und (4) daß Part-of-Speech-Tags ebenfalls nicht zu wesentlichen Verbesserungen führen. Diese Ergebnisse sind jedoch zunächst nicht als allgemeingültig zu werten, sondern ausschließlich auf das spezifische Korpus und das Vorgehen bei seiner Aufbereitung zu beziehen. Insbesondere bei der Interpretation der Ergebnisse, welche die Intensitätsmerkmale betreffen, ist zu bedenken, daß das untersuchte Sprachmaterial aus Radiosendungen stammt und daß die Sendeanstalten in

der Regel Signalverarbeitungsmethoden einsetzen, die den Dynamikbereich verzerren. Weiterhin kann das Ergebnis, daß die Silbendauer für die Vorhersage der Intonationsbeschreibung aussagekräftiger war als die Nukleusdauer, ein Artefakt der automatischen Segmentierung sein, da diese für größere Einheiten geringere relative Fehler aufweist als für kleinere Einheiten. Schließlich verbieten sich allgemeingültige Schlüsse auch schon deshalb, weil die Untersuchungen über die Merkmalsauswahl nur Sprachmaterial von einem einzigen Sprecher einbezogen.

Ein weiteres Thema dieses Kapitels war die Erweiterung des Verfahrens auf eine sprecherunabhängige Prosodieerkennung. Der vorgeschlagene Ansatz zielt darauf ab, die Sprechereinflüsse auf Grundfrequenzmerkmale durch Normalisierung und die sprecherabhängigen Einflüsse auf die verbleibenden Merkmale durch Diversität im Trainingsmaterial zu kompensieren, da für diese Merkmale eine Normalisierung nicht ohne weiteres möglich ist. Die sprecherunabhängige Erkennung benötigt sehr viel mehr prosodisch annotiertes Sprachmaterial als die sprecherabhängige Erkennung; dieses stand am IMS während der Durchführung dieser Arbeit (noch) nicht zur Verfügung, so daß eine abschließende Bewertung der hier entwickelten sprecherunabhängigen Prosodieerkennung noch aussteht. Mit dem verfügbaren Material erzielt die sprecherunabhängige Prosodieerkennung jedenfalls deutlich schlechtere Ergebnisse als für die sprecherabhängige Erkennung.

Ansatzweise befaßte sich dieses Kapitel mit der Frage, wie zusätzlich erhobenes Spachmaterial zusammengesetzt sein sollte. Ein weiteres Experiment zeigte, daß Bootstrapping im Zusammenhang mit dem symbolischen Lernverfahren C4.5 die prosodische Annotierung eines umfangreichen Einzelsprecherkorpus verbessern kann. Die Annotation großer Einzelsprecherkorpora ist Voraussetzung für die konkatenative Sprachsynthese aus Datenbanken, eine Sprachsynthesetechnik, die Sprache mit natürlicher klingender Prosodie erwarten läßt als die zur Zeit gängige Diphonsynthese.

Insgesamt ist die Erkennung der phonologischen Intonationsbeschreibung noch verbesserungswürdig. Ein Ansatzpunkt für weiterführende Arbeit ist, daß die hier durchgeführten Experimente unterschiedliche Register weitestgehend ignorieren. Da mit der Dissertation von Mayer [May97] nun eine zum hier verwendeten Intonationssystem passende Beschreibung von Registern vorliegt, sind die Voraussetzungen für eine theoretisch fundierte Kategorisierung von Registern bei der Intonationsbeschreibung geschaffen, entweder bei der Normalisierung der Grundfrequenz oder als eigenständiges, vorab klassifiziertes Merkmal, das dem Klassifikator zur Unterscheidung der Intonationssymbole zusätzlich zur Verfügung steht.

Ein weiterer Ansatz zur Verbesserung der Prosodieerkennung besteht darin, die stark unbalancierte Verteilung der auftretenden Klassen zu berücksichtigen. Interessant ist zum Beispiel, eine hierarchische Klassifikation zu versuchen. Dabei ist

zum einen das Ungleichgewicht zwischen unbetonten und betonten Silben, zum anderen auch die extrem ungleiche Verteilung unter den Pitchakzenten zu beachten. Erfolgversprechend erscheint zum Beispiel eine zweistufige Klassifikation: Zunächst wird mit allen verfügbaren Merkmalen über den Betonungsstatus und das Vorhandensein einer Phrasengrenze entschieden, um so die zahlreichen unbetonten Silben von den Silben zu trennen, die Intonationsmarkierungen erhalten sollen. Die zweite Stufe der Klassifikation, die die Ausprägung der Intonationsmarkierung festlegt, verwendet dann zur Unterscheidung der Akzente nur noch die Grundfrequenzmerkmale, die Längenmerkmale sollten etwa nur zur Unterscheidung zwischen intermediären Phrasengrenzen und den Intonationsphrasengrenzen verwendet werden. Möglicherweise ist bei der Klassifikation der Pitchakzente auch eine weitere Kaskadierung zur Abscheidung der sehr häufig auftretenden Konzepte L*H und H*L ein sinnvoller Schritt.

Kapitel 7

Korpuserstellungs- und Korpuszugriffshilfen

Nach der Entwicklung und Diskussion von Verfahren, mit denen aus unaufbereitetem Sprachmaterial eine prosodisch etikettierte Datensammlung gewonnen werden kann, beschreibt dieses Kapitel Verfahren, die bei der Erstellung und Auswertung der Datensammlung hilfreich sind. Im Abschnitt 7.1 wird eine Methode vorgestellt, mit der das Sammeln von Sprachmaterial weiter automatisiert werden kann. Sie erlaubt das Herausfiltern von für die Prosodieforschung interessantem Material aus der Flut von Sprachdaten, die täglich über die Medienkanäle gesendet werden. Abschnitt 7.2 stellt ein Hilfswerkzeug vor, das den Zugriff auf die gesammelten und aufbereiteten Daten erleichtert, indem es übersichtliche, textuelle Darstellungen des Sprachmaterials erzeugt, die die Orthographie und phonologische Intonationsbeschreibung mehrerer sprachlicher Realisierungen miteinander verknüpft. Eine solche Darstellung ist für vergleichende Untersuchungen, etwa typologische oder dialektologische Projekte, von großem Nutzen.

7.1 Selektion wiederkehrender Nachrichten

In diesem Abschnitt werden zwei Methoden vorgestellt, um automatisch gleiche oder nahezu gleiche Nachrichtenmeldungen aus den stündlich gesendeten Radionachrichten herauszufiltern. Die Methoden arbeiten sprecherunabhängig, erlauben es also, neben wiederkehrenden Nachrichten des gleichen Sprechers auch wiederholt verlesene Nachrichten verschiedener Sprecher zu finden. Das grobe Vorgehen

ist dabei wie folgt: Zunächst werden mit HMM-Spracherkennungstechnologie komplette Nachrichtensendungen in Einzelnachrichten aufgespaltet. Das Aufspalten der Nachrichtensendung beruht dabei auf der prosodischen und lexikalischen Markierung der Ansager und der immer gleichen äußeren Form der Nachrichtensendungen. Bei der Aufteilung fallen Wortfolgenschätzungen an, die für das Auffinden gleicher Nachrichten ausgenutzt werden. Die Wortfolgenschätzungen der Einzelnachrichten werden entweder direkt miteinander verglichen (naive Methode) oder aufwendiger mit der sogenannten Realignment-Methode, die die Ähnlichkeit je zweier Einzelnachrichten dadurch quantifiziert, daß mittels der HM-Modellierung die Wahrscheinlichkeit berechnet wird, mit der die Wortfolgenschätzung der einen Nachricht das Sprachsignal der anderen emittiert. Die naive und die Realignmentmethode können verbessert werden, indem man beim Vergleich zweier Nachrichten die jeweiligen Ähnlichkeiten beziehungsweise Unterschiede zu den verbleibenden Nachrichten mitberücksichtigt.

7.1.1 Einleitung

In den stündlich gesendeten Radionachrichten (und besonders bei den in der Nacht gesendeten), werden häufig Einzelnachrichten in aufeinanderfolgenden Sendungen identisch oder nur leicht verändert erneut verlesen. Eine Datenbank solcher Wiederholungen professionell gelesener und aufgenommener Sprache ist interessant für die linguistische Forschung, da sie syntaktische, lexikalische, prosodische und segmentale Variation, aber auch Beständigkeit zwischen einzelnen Realisierungen derselben Nachricht aufzeigen kann. Die Datenbank wird noch interessanter, wenn Einzelnachrichten über einen Sprecherwechsel hinaus in der Nachrichtensendung verbleiben, so daß für eine Nachricht auch Realisierungen von verschiedenen Sprechern bereitstehen.

Aktuelle Modelle der Sprachproduktion [Gar76, Del85, Lev89] benötigen wie alle linguistischen Modelle Daten, um testbar zu sein. Dank der wegweisenden Arbeit von Levelt [Lev89] haben die Modelle einen Komplexitätsgrad erreicht, der für alle Bereiche der Computerlinguistik zur Herausforderung geworden ist. Nach Levelt wird Sprache durch drei Hauptverarbeitungskomponenten erzeugt: KONZEPTUALISIERER, FORMULIERER und ARTIKULIERER. Diese greifen auf zwei Wissensquellen zu, das LEXIKON und das DISKURSMODELL. Abbildung 7.1 [Lev89, S. 9] zeigt den Ablauf und das Zusammenspiel der einzelnen Prozesse bei der Generierung flüssig gesprochener Sprache.

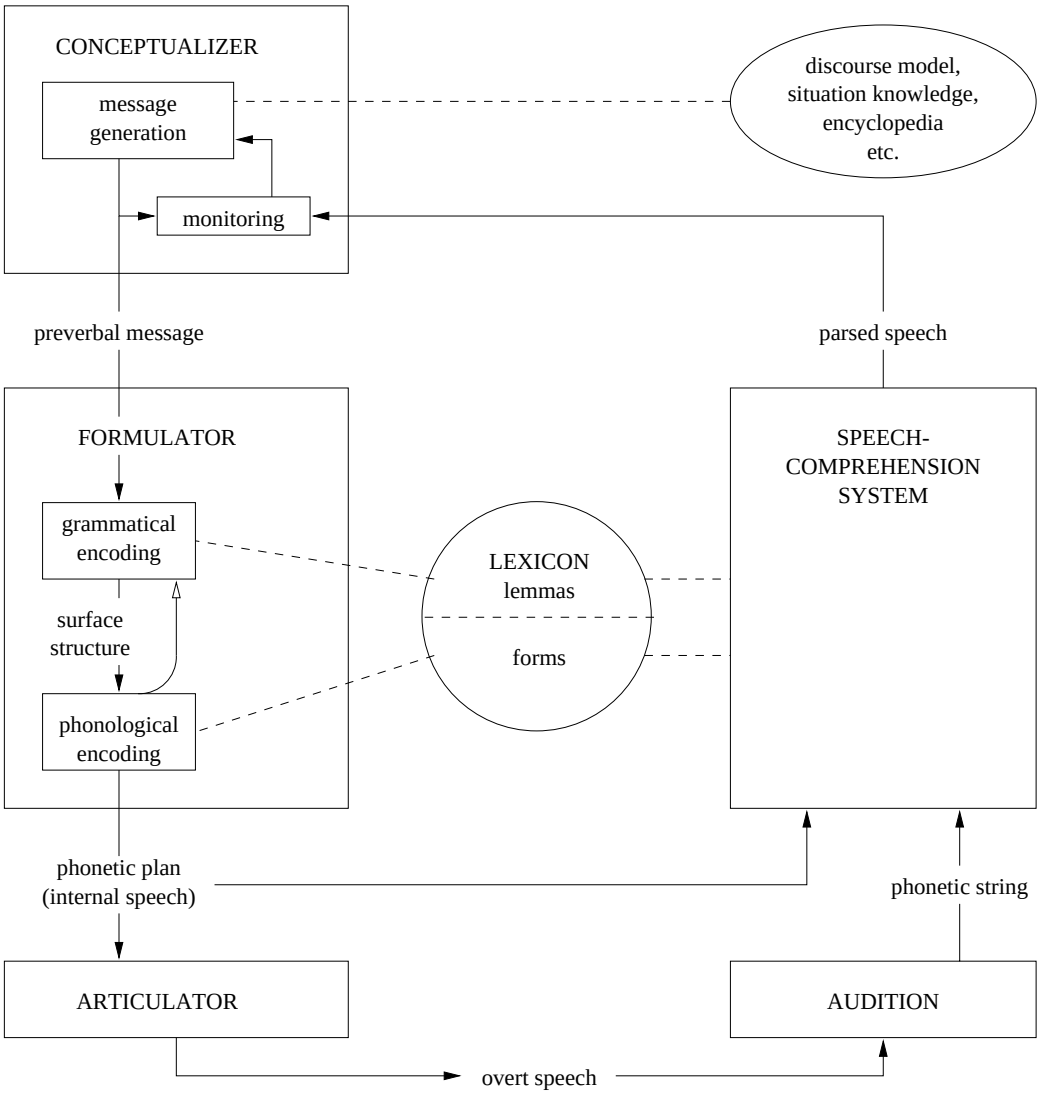


Abbildung 7.1: Ein Modell des Sprechers. Kästen repräsentieren Verarbeitungseinheiten, Kreise und Ellipsen Wissensquellen.
(aus [Lev89, S. 9, Fig. 1.1])

Eine Datenbank wiederkehrender Nachrichten erlaubt es, einzelne Teile des Sprachgenerierungssystems zu untersuchen. Speziell können die Beiträge der Wissensquelle DISKURSMODELL und die Verarbeitungskomponente KONZEPTUALISIERER bei den wiederholten Nachrichten ignoriert werden, was eine isolierte Untersuchung der Teilsysteme LEXIKALISCHE WISSENSQUELLE sowie der Verarbeitungseinheiten FORMULIERER und ARTIKULIERER bei der Generierung flüssig gesprochener Sprache ermöglicht. Da die vorgestellten Selektionsverfahren auch von verschiedenen Sprechern gesprochene, ansonsten aber gleiche Nachrichten erfassen, kann auch die phonetische und phonologische Variabilität bei der Nachrichtenproduktion systematisch untersucht werden.

Die am IMS durchgeführten Untersuchungen zielen primär auf das „phonological encoding"-Modul (siehe Abb. 7.2), und speziell auf dessen als „Prosodiegenerator" bezeichnete Komponente (vgl. [Lev89, Kapitel 10]). An der Datenbank mit wiederkehrenden Radionachrichten können Formulier- und Artikulierstrategien verschiedener Sprecher, also die Encodierung und die phonetische Planung, empirisch untersucht werden. Eine solche Untersuchung ist besonders für die Sprachsyntheseforschung und dort vor allem für die concept-to-speech-Forschung wesentlich, da dabei die Variation der Produktion an verschiedenen Realisierungen der gleichen Konzeptstruktur betrachtet werden kann.

Wäre die Spracherkennung bereits komplett beherrscht, so wäre das Auffinden wiederkehrender Nachrichten sehr einfach. Man müßte lediglich die Nachrichten erkennen lassen und die so gefundenen Wortfolgen der einzelnen Nachrichten miteinander vergleichen. Das Erkennen von Radionachrichten ist jedoch eine recht schwierige Aufgabe. Selbst für die in [ICA97] verzeichneten Systeme, die alle den Stand der Technik repräsentieren, werden Wortfehlerraten beim Erkennen von Radionachrichtensendungen zwischen 20 und 30% berichtet. Die zur Zeit erfolgreichste Erkennungstechnologie, die auch von den in [ICA97] berichteten Systemen eingesetzt wird, verwendet als Erkennungseinheiten Wörter, die ihrerseits aus Untereinheiten zusammengesetzt sind, z.B. aus Phonemen. Das Dilemma eines Erkennungssystems liegt darin, daß das System nur dann die Wortkette einer Nachricht komplett richtig erkennen kann, wenn alle vorkommenden Wörter im Vokabular des Systems enthalten sind. Das Vokabular sollte demnach also mindestens so groß sein, daß es sicher alle Wörter einer Nachricht umfaßt. Da das Erkennungsproblem mit zunehmender Anzahl an Mustern immer schwieriger wird, sollte jedoch andererseits das Vokabular möglichst klein sein, damit gute Erkennungsraten erreicht werden können. Beim Erkennen von Nachrichtensendungen wird man deshalb immer mit nicht im Erkennungsvokabular vorhandenen Wörtern (out of vocabulary words, OOV-words) Probleme haben, da es unmöglich ist, ein ausreichend kleines Vokabular zu definieren, das dennoch alle Wörter der nächsten Nachrichtensendung enthält. Dazu kommt, daß in solchen Nachrichten-

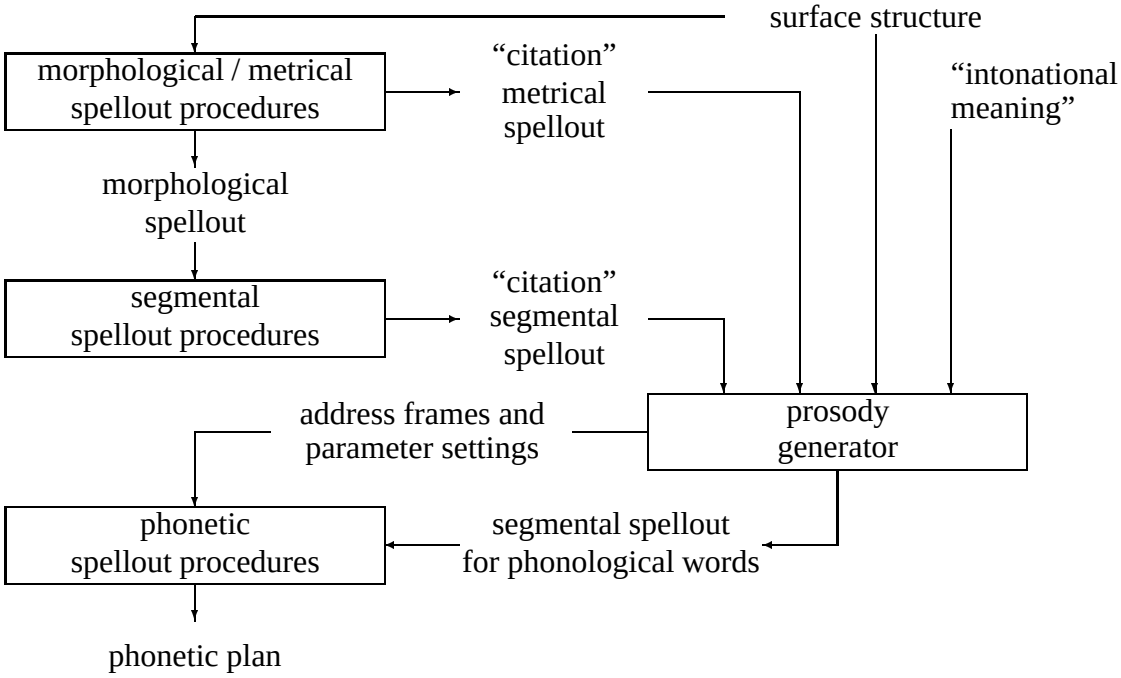


Abbildung 7.2: Übersicht über die Architektur der Phonologischen Encodierung zusammenhängend gesprochener Sprache.

sendungen sehr viele Eigennamen auftreten, und diese Namen wechseln von Tag zu Tag mit dem Fortgang der Weltgeschichte. Selbst wenn es gelänge, alle Wörter der kommenden Nachrichtensendung aufzulisten, wäre eine solche Liste vermutlich zu groß, um mit heutiger Erkennungstechnologie beherrschbar zu sein.

Wenn man sich damit abgefunden hat, daß es mit heutiger Technologie noch nicht möglich ist, Nachrichtensendungen *vollständig richtig* zu erkennen, so kann man darangehen, auch aus einer unzureichenden Erkennung Nutzen für die Erstellung von Korpora für die Prosodieforschung zu ziehen. Eine Möglichkeit dazu zeigt dieser Abschnitt auf, er beschreibt ein vollautomatisches Verfahren zum Aufsammeln wiederholt gleich oder nahezu gleich verlesener Radionachrichten. In den Abschnitten 7.1.3 und 7.1.4 werden zwei Ansätze zur Quantifizierung der Ähnlichkeit zweier Nachrichten vorgestellt. Abschnitt 7.1.5 zeigt eine Verbesserung auf, die beiden Verfahren nachgeschaltet werden kann. Der Rest des Abschnitts 7.1 befaßt sich mit der experimentellen Evaluation der vorgestellten Methoden, durchgeführt an drei Sammlungen von jeweils mehreren vom Radio aufgenommenen aufeinanderfolgenden Nachrichtensendungen.

7.1.2 Aufteilen der Nachrichtensendung in Einzelnachrichten

Der Rundfunksender, dessen Sendungen für die Prosodieuntersuchungen am IMS herangezogen werden, ist der Deutschlandfunk, der am IMS in sehr guter Qualität (32kHz, 16bit linear unkomprimierte, professionelle Studioaufnahmen) über das Digitale Satellitenradio (DSR) empfangen wird. Er sendet in der Zeit zwischen 5 Uhr 30 und 17 Uhr 30 halbstündlich und zwischen 18 und 5 Uhr stündlich Nachrichten. Üblicherweise dauert eine Nachrichtensendung 5 Minuten, einige wenige über den Tag verteilte Nachrichtensendungen dauern 10 Minuten. Eine Nachrichtensendung beinhaltet 5 bis 12 Einzelnachrichten, die immer vom Wetterbericht und einer Verkehrsinformation gefolgt sind. Den zehnminütigen Nachrichten ist zusätzlich eine Zusammenfassung vorausgestellt. Eine Einzelnachricht besteht aus etwa 75 Wörtern, im Bereich zwischen unter 40 und über 100 Wörter. Man beobachtet, daß Sprechpausen zwischen den Einzelnachrichten im allgemeinen länger sind als innerhalb einer Einzelnachricht. Dies gilt für verschiedene Sprecher und Tageszeiten. Die längere Pause zwischen den Einzelnachrichten könnte entweder die Zeit sein, die zum Umdrehen einer Meldungsseite benötigt wird, die Zeit, die der Sprecher benötigt, um die nächste Meldung zu überfliegen, ein Gestaltungsprinzip, dem zu folgen die Nachrichtensprecher angehalten sind, oder schlicht eine prosodische Markierung, die dem Hörer den Anfang einer neuen Nachricht signalisieren soll. Die letzte Annahme wird dadurch gestützt, daß im Falle der Grenze zwischen der letzten Einzelnachricht und dem Wetterbericht die

Pause verhältnismäßig kurz ist. Dort wird die Grenze dem Hörer bereits ausreichend lexikalisch („Das Wetter“) und durch den offensichtlichen Wechsel im Vokabular signalisiert.

Die Einzelnachrichten sind also durch prosodische *und* lexikalische Markierungen voneinander getrennt. Aus diesem Grund bieten sich signalorientierte Methoden für das automatische Auftrennen einer Nachrichtensendung in Einzelnachrichten nicht an. Stattdessen wird hier HMM-Spracherkennungstechnologie eingesetzt, um nach Sprechpausen und bestimmten Schlüsselwörtern zu suchen. Da die Nachrichtensendungen immer nach dem gleichen Schema aufgebaut sind, ist es möglich, eine Finite-State-Grammatik zu entwerfen, die die grobe Struktur einer Nachrichtensendung beschreibt. Natürlich sind die Einzelnachrichten selbst nicht fest vorgegeben, sie müssen durch eine Disjunktion einer beschränkten Wortmenge abgedeckt werden, die beliebig oft wiederholt werden kann. Eine vereinfachte Version der eingesetzten Grammatik ist in Abbildung 7.3 wiedergegeben. Zur Erkennung wird das in Kapitel 3 beschriebene System Alphons verwendet. Nach der Erkennung wird aufgrund der detektierten Pausenlängen und der lexikalischen Markierungen entschieden, welche Segmente der Gesamtnachrichtensendung als Einzelnachrichten anzusehen sind.

7.1.3 Naive Methode

Als Nebenprodukt der Segmentierung einer Nachrichtensendung in Einzelnachrichten fällt durch die HMM-Erkennung eine Wortfolgenschätzung für jede gefundene Einzelnachricht an. Wenn die Erkennung ausreichend genau gelingt, kann man annehmen, daß die Wortfolgenschätzungen für wiederkehrende Nachrichten recht ähnlich und für verschiedene Einzelnachrichten mehr oder weniger unterschiedlich sind. Es bietet sich daher an, diese Wortfolgenschätzungen miteinander zu vergleichen, um herauszufinden, ob zwei Nachrichtensendungen gleich oder verschieden sind. Die häufig zur Evaluation in der Spracherkennung verwendete Bewertungszahl „Akkuratheit“ scheint ein gutes Maß zur Quantifizierung der Ähnlichkeit zu sein. Um sie effizient zu berechnen, wird mittels dynamischer Programmierung zunächst ein optimaler Match zwischen den Wörtern der beiden zu vergleichenden Wortketten gefunden. Dann werden die Ersetzungsfehler S (substitution errors), Einfügefehler I (insertion errors) und Auslassungsfehler D (deletion errors) gezählt und die Akkuratheit wie folgt berechnet:

$$\textit{Akkuratheit} = \frac{N - S - I - D}{N} \cdot 100\%$$

N steht dabei für die Anzahl an Wörtern in der Referenzwortkettenschätzung. Auf einfache Weise kann dann anhand eines festen Schwellwertes über die Ähnlich-

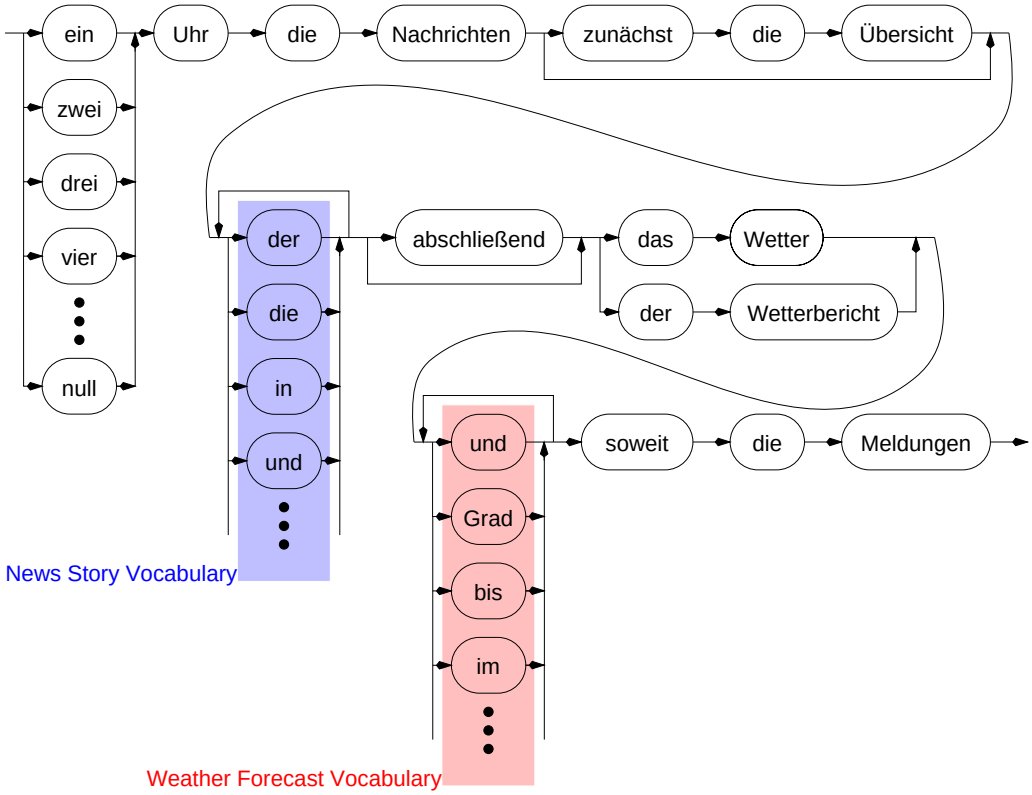


Abbildung 7.3: Vereinfachte Finite-State-Grammatik für das Suchen nach lexikalischen Markierungen zum Aufteilen einer Nachricht in Einzelnachrichten.

keit zweier Nachrichtensendungen entschieden werden. Dazu wird für alle Paare von Einzelnachrichten aus den beiden Wortfolgenschätzungen die Akkuratheit berechnet. Liegt sie über dem Schwellwert, werden die Nachrichten als gleich angesehen, liegt sie darunter, betrachtet man die Nachrichten als verschieden.

7.1.4 Realignment-Methode

Bei der naiven Methode oben wird nicht berücksichtigt, daß manche Wörter leichter miteinander zu verwechseln sind als andere. Ein typischer Erkennungsfehler (sowohl in der maschinellen als auch menschlichen Spracherkennung) ist die Verwechslung eines Wortes mit einem ähnlich klingenden anderen Wort. Ein weiterer häufig auftretender Fehler ist das fälschliche Erkennen mehrerer im Vokabular enthaltener kurzer Wörter anstatt eines (langen) OOV-Wortes. Wegen des offenen Vokabulars sind diese beiden Fehlerarten bei der Erkennung von Nachrichtensendungen besonders häufig zu erwarten. Eine Möglichkeit zur Kompensation dieser Erkennungsfehler bei dem naiven Ansatz besteht darin, ein Maß zu entwickeln, das die Ähnlichkeit zwischen Wörtern abschätzt, und mit diesem Maß die Fehler bei der Berechnung der Akkuratheit zu gewichten. Natürlich ist die Entwicklung eines solchen Maßes mit einem gewissen Aufwand verbunden, der durch die im folgenden beschriebene Methode umgangen wird. Anstatt die Ähnlichkeit zwischen der Wortkette einer Nachricht a mit der Wortkette einer Nachricht b auf abstrakter symbolischer Ebene zu quantifizieren, beruht die Realignment-Methode darauf, daß die Ähnlichkeit der Wortkette der Nachricht a mit dem Sprachsignal der Nachricht b quantifiziert wird. Abbildung 7.4 verdeutlicht die Unterschiede zwischen den beiden Methoden.

Beim zur Berechnung der Ähnlichkeit eingesetzten sogenannten Forced alignment wird, wie in Kapitel 3 beschrieben, eine Kette von Hidden-Markov-Modellen mit einer Beobachtungsfolge in Übereinstimmung gebracht durch Ausnutzen von

1. der Reihenfolgeninformation der Sprachlaute in Einzelnachricht a bzw. deren bei der Segmentierung gefundenen Wortkettenschätzung, welche eine Annäherung an die tatsächliche Abfolge darstellt,
2. phonetischem Wissen, das in den HMM enthalten ist, und
3. der Parametrisierung des Sprachsignals von Einzelnachricht b.

Das Maß, das zur Beschreibung der Übereinstimmungsgüte dient, ist die logarithmierte Wahrscheinlichkeit geteilt durch die Anzahl der Observationen, die

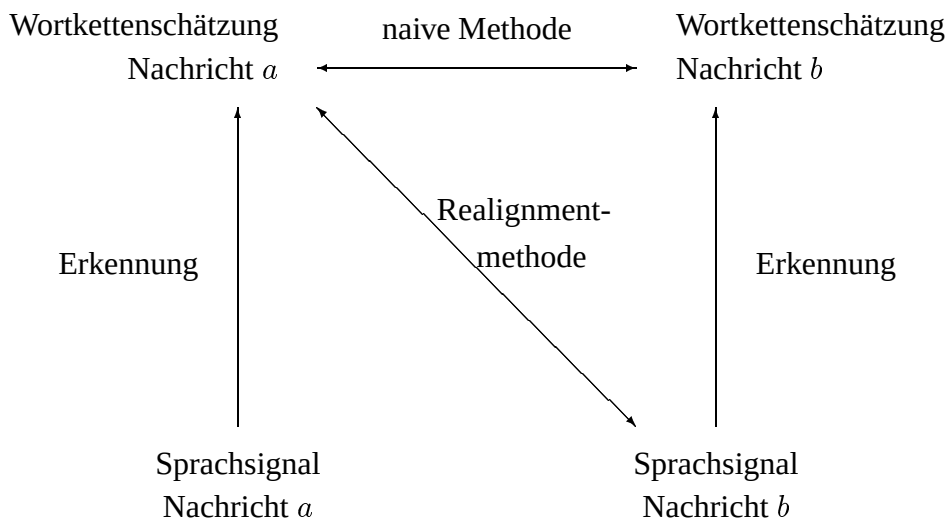


Abbildung 7.4: Die beiden unterschiedlichen Methoden, zwei Einzelnachrichten miteinander zu vergleichen.

während des Alignment berechnet und ausgegeben wird. Das Vorgehen kann wie folgt formalisiert werden.

Sei V das (geschlossene) Vokabular für die Segmentierung in Einzelnachrichten, $\mathcal{W}_i \in V^*$ Wortfolgen davon, $\mathcal{O}_i$ Observationsfolgen von Einzelnachrichten in Form von Parametrisierungen des Sprachsignals. $news$ sei die Indexmenge der untersuchten Einzelnachrichten.

foreach $n \in news$ **do**
  ermittle die Wortfolgenschätzung $\mathcal{W}_n(\mathcal{O}_n, V)$
  durch HMM-Erkennung
od
foreach $i \in news$ **do**
  foreach $j \in news$ **do**
    $A(i,j) = \frac{\log(p(\mathcal{O}_j|\mathcal{W}_i))}{|\mathcal{O}_j|}$
  od
od

Da die aufwendige Berechnung der normalisierten logarithmierten Wahrscheinlichkeit $A(i,j)$ für jedes Nachrichtenpaar durchgeführt werden muß, ist diese

zweite Methode in der Berechnung weit aufwendiger als der direkt symbolisch vergleichende naive Ansatz. Jedoch können die Berechnungen vollständig parallelisiert werden, und etliche Berechnungen kann man einsparen, wenn man einige auf der Hand liegende Heuristiken einsetzt. Beispielsweise kommen Nachrichten in der gleichen Nachrichtensendung nicht noch einmal vor, und sie werden eher in direkt aufeinander vorkommenden Sendungen erneut verlesen als Stunden später. Für die Zwecke der Korpuserschließung sollte es jedoch genügen, wenn beispielsweise eine einzige Menge wiederkehrender Nachrichten pro Tag anfällt, so daß im folgenden auf Effizienzverbesserungen zugunsten von Genauigkeitsverbesserungen verzichtet wird.

Ob zwei Einzelnachrichten als gleich gelten sollen, wird wieder durch Vergleich der normalisierten logarithmierten Wahrscheinlichkeit mit einem Schwellwert t entschieden. Zwei Nachrichten i und j gehören dann zum selben Nachrichtensatz, wenn $A(i,j) > t$ und $A(j,i) > t$.

7.1.5 Ausnutzung der Unterschiede zu anderen Nachrichten

Die im folgenden beschriebene Verbesserung beim Finden wiederkehrender Nachrichten ähnelt dem „nearest neighbour clustering“. Um das Prinzip deutlich zu machen, ist es sinnvoll, zunächst von den beiden obigen Methoden zu abstrahieren. Im Grunde besteht das Ergebnis der beschriebenen Methoden aus einer Abstandsmatrix, die in Zeile i, Spalte j jeweils einen Wert für den Abstand zwischen Nachricht i und Nachricht j enthält. Das Abstandsmaß muß dabei nicht notwendigerweise symmetrisch sein, es kann also gelten $A(i,j) \neq A(j,i)$. Wenn man die Ähnlichkeit zwischen zwei Nachrichten an einem Schwellwert festmacht, entspricht dies dem Vorgehen, jeden über dem Schwellwert liegenden Eintrag der Matrix auf 1, und alle übrigen auf 0 zu setzen. Diese Matrix aus Nullen und Einsen definiert eine Relation $R \in (news \times news)$, und wenn die Abstandsmaße und der Schwellwert gut genug gewählt sind, stellt R idealerweise eine Äquivalenzrelation dar, die die Menge der Einzelnachrichten in Untermengen (nämlich Sätze wiederkehrender Nachrichten, künftig Cluster genannt) partitioniert.

In der Praxis werden jedoch Situationen wie die folgenden auftreten. Nachricht a aus dem Cluster A und Nachricht b aus Cluster B sind ähnlich genug, um über dem Schwellwert zu liegen, also $(a,b) \in R$, obwohl a und b verschieden sind. Nehmen wir der Einfachheit halber an, alle anderen Nachrichtenpaare aus A seien in R enthalten, alle B-Paare ebenfalls und sonst keine. Dann gibt es zwei Optionen. Erstens kann man die Cluster A und B zusammenfassen, da sie durch a und b miteinander verbunden sind. Dieses Vorgehen entspricht dem Überführen von R in

eine Äquivalenzrelation durch Berechnen der transitiven Hülle von R. Die zweite Option ist, aRb als falsch zu betrachten und A und B als verschiedene Cluster weiterzuführen. Dies entspricht dem Überführen von R in eine Äquivalenzrelation durch das Auffinden von Cliquen oder zweifach verbundenen Komponenten im durch R definierten Graphen.

Die notwendigen Korrekturen lassen sich also durch graphentheoretische Algorithmen ausführen, die die Matrix als Ganzes betreffen. Dies zeigt, daß es sinnvoll sein kann, beim Vergleich zweier Nachrichten i und j mehr als nur die zwei Einträge (i, j) und (j, i) in der Matrix A zu betrachten. Angenommen, Nachricht a_1 unterscheidet sich stark von Nachricht b, und Nachricht a_2 ist eine erneut verlesene Variante von a_1. Man würde dann erwarten, daß auch Nachricht a_2 sehr unterschiedlich zu b ist. Und wenn eine Nachricht c zu a_1 ähnlich ist, erwartet man auch a_2 in der Nähe von c. Auf die Abstandsmatrix übertragen bedeutet dies, daß man erwartet, daß die in der a_1-Zeile von A verzeichneten Abstände ähnlich zu denen der a_2-Zeile von A sind, also die Zeilenvektoren $A(a_1, \cdot)$ und $A(a_2, \cdot)$ im $\mathbb{R}^{|news|}$-Raum recht nah beieinander liegen. Der Abstand zwischen Vektoren in $\mathbb{R}^{|news|}$ kann auf verschiedenste Arten gemessen werden. Clustering-Verfahren verwenden häufig den Euklidischen Abstand. Der Euklidische Abstand zweier Vektoren $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ und $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ ist definiert als

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Diese Norm (oder jede andere) kann zur Definition eines neuen Maßes verwendet werden, das die Ähnlichkeit zweier Nachrichten beschreibt und dabei die Ähnlichkeit zu den anderen Nachrichten mitberücksichtigt, sozusagen ein Meta-Abstand. Dieser Meta-Abstand kann zwischen allen Nachrichtenpaaren bestimmt werden, indem er für alle Paare von Zeilenvektoren aus A berechnet wird. Dies ergibt schließlich eine symmetrische Matrix von Zeilenabständen (da d symmetrisch ist). Diese Metadistanzmatrix wird im folgenden als B bezeichnet. Der folgende Algorithmus stellt die Berechnung von B aus A formal dar:

foreach $i \in news$ **do**
  foreach $j \in news$ **do**
    $B(i, j) = d(A(i, \cdot), A(j, \cdot))$
  od
od

Die Berechnung von B befreit nicht von der Notwendigkeit, einen geeigneten Schwellwert t festzulegen, und garantiert auch nicht, daß die aus B und t ableitbare Relation R eine Äquivalenzrelation ist. Aus diesem Grund sollten die graphentheoretischen Reparaturstrategien ebenfalls wieder angewendet werden. Wie

sich jedoch im folgenden Abschnitt zeigen wird, sind diese Berechnungen nunmehr weniger wichtig, da der Schritt von der Matrix A zur Matrix B den Abstand zwischen der Distanz zweier gleicher Einzelnachrichten und der Distanz zweier unterschiedlicher Einzelnachrichten deutlich erhöht.

7.1.6 Experimentelle Auswertung

Für die experimentelle Bewertung der vorgeschlagenen Methoden wurden drei verschiedene Nachrichtensammlungen herangezogen.

Bei den ersten beiden Nachrichtensammlungen wurden die Nachrichtensendungen sorgfältig in Einzelnachrichten aufgeteilt und aufbereitet. Dazu wurden die Einzelnachrichten transliteriert, mit dem System Alphons aus Kapitel 3 automatisch Wort-segmentiert und gemäß den in Kapitel 2 erwähnten Richtlinien manuell prosodisch etikettiert. Da somit für die Nachrichten Transliterationen verfügbar sind, kann man präzise festlegen, ob zwei Nachrichten gleich sind oder nicht, beziehungsweise, in welchem Grad sie ähnlich sind, indem man die Wortfolgen, die aus der nahezu perfekten menschlichen Wahrnehmung resultieren, miteinander vergleicht.[1]

Die erste und kleinste Nachrichtensammlung besteht aus 32 Einzelnachrichten und Wetterberichten, die aus vier aufeinanderfolgenden Nachrichtensendungen des Deutschlandfunks am 28. 07. 1995 (von 10:30 bis 12:00 Uhr) stammen. Die Sammlung enthält 6 Cluster wiederkehrender Nachrichten: drei Cluster mit je 3 gleichen Nachrichten und drei Cluster mit jeweils 2 gleichen Einzelnachrichten. Zwei weitere Nachrichten wurden ebenfalls wiederholt verlesen, dabei aber um einen bzw. zwei Sätze erweitert. Die verlängerten Versionen werden hier als von den ursprünglichen Meldungen verschiedene Nachrichten angesehen. Die verbleibenden 13 Einzelnachrichten wurden nicht wiederholt.

Bei der zweiten Nachrichtensammlung handelt es sich um 14 aus der gleichen Quelle aufgenommene aufeinanderfolgende Nachrichtensendungen, die aufgeteilt in Einzelnachrichten 110 Meldungen und Wetterberichte ergeben (21. 11. 1995, 10:30 bis 17:00 Uhr). Von den 110 Einzelnachrichten treten 49 genau gleich verlesen mehr als einmal auf. Sie sind auf 22 Cluster verteilt, deren größter fünf Realisierungen eines Wetterberichts umfaßt. Wenn man kleinere Umformulierungen beim Vergleich toleriert, können 60 Einzelnachrichten als wiederholt verlesen gelten, die sich auf 23 Cluster verteilen. Ein Cluster enthält 8 Wetterberichte, ein Cluster besteht aus 5 Nachrichtenmeldungen, weiterhin gibt es 5 Cluster, die je

[1] Wie bei allen Datensammlungen haben sich auch hier einige Tippfehler eingeschlichen.

drei Nachrichten umfassen, usw. Toleriert wurden dabei bis zu 10 Wortfehler, die sich durch Einfügen, Auslassen und Ersetzen ergeben haben können und durch Vergleich der manuellen Transliterationen bestimmt wurden. Die Festlegung, bis zu welchem Umformulierungsgrad zwei Nachrichtensendungen noch als gleich betrachtet werden sollen, ist in gewissen Grenzen willkürlich, da die Veränderungen graduell sind. Wenn man etwa nur 5 statt 10 Wortfehler zuläßt, resultieren kleinere Cluster, beispielsweise wird der 5 Nachrichten umfassende Cluster in zwei Cluster aufgespaltet.

Bei der ersten und kleinsten Nachrichtensammlung ist die Bestimmung der Cluster gleicher Nachrichtenmeldungen vergleichsweise einfach, da hier die als identisch geltenden Nachrichtenmeldungen wortgleich wiederholt wurden. Bei der zweiten, größeren Nachrichtensammlung ist das Clustern deshalb schwieriger, weil hier nahezu alle möglichen Grade von Umformulierungen vorkommen — von Ersetzungen einzelner Wörter bis zu fast vollständigen Umformulierungen von inhaltlich gleichen Meldungen. Als Folge dieses graduellen Übergangs sind die Unterschiede zwischen den Clustern weniger deutlich, während die Einzelmeldungen innerhalb der Cluster weniger homogen sind.

Die dritte Nachrichtensammlung besteht aus 17 Nachrichtensendungen, die automatisch über Nacht aufgenommen wurden (Deutschlandfunk, 30./31. 05. 1996, 19:00 bis 8:00 Uhr). Experimente mit dieser Datensammlung dienen dazu, den Gesamtprozeß zu evaluieren, von der Suche nach lexikalischen Markierungen über das Aufteilen anhand lexikalischer und prosodischer Markierungen in Einzelnachrichten bis zum Clustern der Einzelnachrichten mit den beschriebenen Verfahren. Man kann aus den Ergebnissen abschätzen, wieviele wiederkehrende Nachrichten sich aus den Nachrichtensendungen einer Nacht vollautomatisch finden lassen. Naturgemäß gibt es für diese vollständig automatisch erhobenen Daten keine Transliterationen.

Für einen Vergleich der vorgeschlagenen Methoden werden die verschiedenen Verfahren auf alle Nachrichtensammlungen angewendet und die von den verschiedenen Verfahren auf der jeweils gleichen Datenmenge erzielten Ergebnisse miteinander verglichen. Zur Bewertung der beiden Verbesserungsmöglichkeiten, also erstens die aufwendigere Realignment-Methode statt der schneller berechenbaren naiven Methode und zweitens die Metadistanzen statt der direkten Abstände, sind für jeden Datensatz je vier Varianten zu betrachten: zwei Methoden, eine Distanzmatrix zu berechnen (naive versus Realignment-Methode) $\times$ zwei Arten, diese Distanzen auszuwerten (ohne versus mit Berechnung der Metadistanzen).

Für jeden Einzelversuch wird die nach den jeweiligen Verfahren berechnete Abstandsmatrix graphisch dargestellt. Dazu werden die Einträge der Abstandsma-

trix der Größe nach sortiert und so in die Graphik eingezeichnet, daß auf der jeweils linken Seite die Ähnlichkeitswerte der ähnlichsten und auf der rechten Seite die Ähnlichkeitswerte der am wenigsten ähnlichen Meldungspaare zu sehen sind. Wenn die Abstände als direkte Distanzen berechnet werden, entspricht ein hoher Wert einer großen Übereinstimmung des Meldungspaares, bei Verwendung der Metadistanzen bedeutet dagegen ein niedriger Wert eine große Ähnlichkeit. Die linke Graphik stellt immer die Abstände für alle Paare dar, die rechte Graphik ist stets eine Ausschnittsvergrößerung der linken Graphik. Für die erste und zweite Nachrichtensammlung, wo die Gleichheit von Meldungen aus den manuellen Transliterationen bekannt ist, sind gleiche und unterschiedliche Meldungspaare in der Graphik durch verschiedene Symbole dargestellt.

Für jede Einzelnachricht der ersten beiden Nachrichtensammlungen wurde eine Wortfolgenschätzung erstellt mittels HMM-Erkennung mit einem 100 Wörter umfassenden Vokabular. Diese sind die 100 in allen manuellen Transliterationen am häufigsten aufgetretenen Wörter. Die Erkennung mit dem 100-Wörter-Vokabular wurde mit Monophonmodellen mit 3 Zuständen durchgeführt, wie sie in Kapitel 3 beschrieben sind. Diese Modellierung ist zwar weit entfernt vom Stand der Technik in der Spracherkennung, ist aber ausreichend verläßlich, um das Vorgehen zu illustrieren. Speziell bei den Nachrichtensendungen kann man davon ausgehen, fast nie alle auftretenden Wörter im Erkennungsvokabular zu haben, und alle Wörter der Transliterationen zu verwenden wäre sicherlich unangemessen für den allgemeinen Fall. Wie bei den beiden ersten Nachrichtensammlungen beruht die Wortfolgenschätzung bei der dritten Nachrichtensammlung auf einer HMM-Erkennung mit 100 Wörtern Vokabular. Sie muß für die dritte Nachrichtensammlung nicht extra berechnet werden, da sie bei der Suche nach den lexikalischen Markierungen mit der Grammatik aus Abbildung 7.3 als Nebenprodukt abfällt. In allen Versuchen wurde die Bildung der transitiven Hülle als Reparaturstrategie angewendet.

Kleine Nachrichtensammmlung

Die Experimente mit der kleinen Nachrichtensammlung dienen dazu, die verschiedenen Verfahren anhand eines relativ einfachen Clustering-Problems zu evaluieren und zu vergleichen.

Kleine Nachrichtensammlung, naive Methode, direkte Distanz. Die Abbildung 7.5 zeigt die Abstandsmatrix A für die naive Methode, die die Wortfolgeschätzungen jeweils zweier Einzelnachrichten auf symbolischer Ebene miteinander vergleicht. Ganz links in der Graphik sind die Einträge der Hauptdiagonale

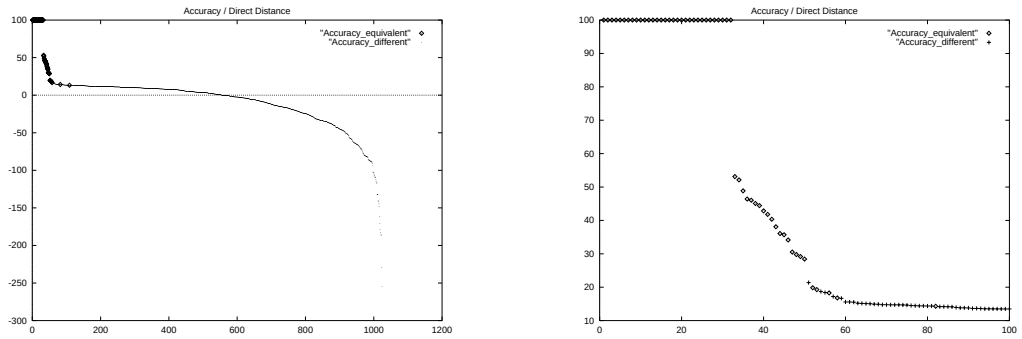


Abbildung 7.5: Die Einträge der Distanzmatrix für die Konstellation kleine Nachrichtenmenge / naiver symbolischer Vergleich / direkte Distanz.

der Matrix A zu sehen, die eine Übreinstimmung von 100% aufweisen. Daran anschließend sind die Akkuratheitswerte aus dem Vergleich verschiedener Einzelnachrichten abgetragen. Man sieht, daß die Akkuratheit erst schnell abnimmt und dann recht lange auf gleichbleibend niedrigem Niveau bleibt, um schließlich gänzlich in den negativen Bereich zu fallen. In der linken Graphik von Abb. 7.5 sind Paare von äquivalenten Einzelnachrichten rautenförmig geplottet, Paare verschiedener Nachrichten sind als Punkte wiedergegeben.

Die Graphiken zeigen, daß sich mit einem gut gewählten Schwellwert, also einem, der mit dem Kniepunkt bei etwa 25 zusammenfällt, recht wenige Klassifikationsfehler ergeben würden, die mit ein wenig Glück von den graphentheoretischen Reparaturstrategien sogar noch korrigiert werden könnten. Der Schwellwert kann besser in der Vergrößerung rechts in Abbildung 7.5 abgelesen werden. Dort sind, um die Sichtbarkeit falsch positiver Fälle zu erhöhen, gute Paare (äquivalente Meldungen) wieder rautenförmig, schlechte (unterschiedliche Meldungen) mit einem Pluszeichen markiert. Wählt man tatsächlich als Schwellwert 25 und berechnet die transitive Hülle zur Erzwingung einer Äquivalenzrelation, so wird ein 2 Meldungen umfassender Cluster verpaßt und ein eigentlich 3 Meldungen umfassender Cluster wird nur teilweise richtig erkannt. In beiden Klassifikationsfehlern ist ein Sprecherwechsel von weiblich zu männlich involviert.

Kleine Nachrichtensammlung, Realignment-Methode, direkte Distanz. In diesem Experiment wird statt der naiven die Realignment-Methode zur Berechnung der Abstände verwendet. Der Abstandswert, der hier für jedes Einzelnachrichtenpaar bestimmt wird, ist die auf die Beobachtungslänge normalisierte logarithmierte Wahrscheinlichkeit, die aus der Anwendung der HMM der Wortfolgenschätzung der ersten Nachricht auf die Parametrisierung des Sprachsi-

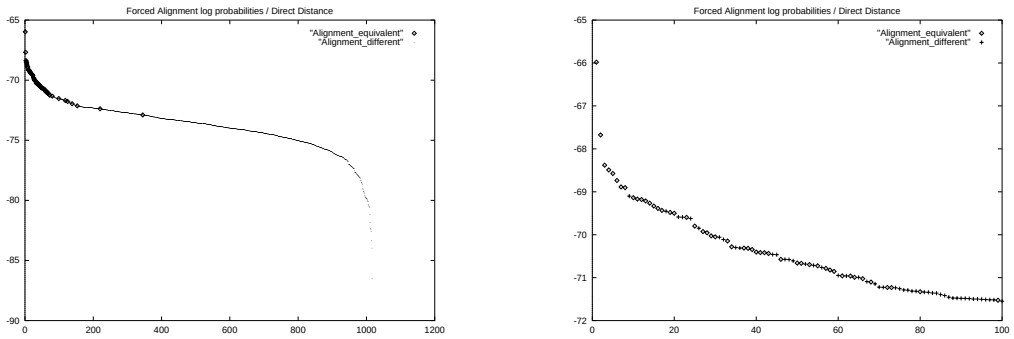


Abbildung 7.6: Die Einträge der Distanzmatrix für die Konstellation kleine Nachrichtenmenge / Realignment-Methode / direkte Distanz.

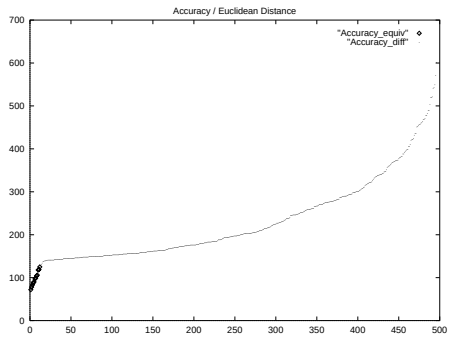
gnals der zweiten Nachricht als Beobachtungsfolge berechnet wurde.[2] Mit anderen Worten: Die damit gefundenen Zahlen geben ein Maß für die Wahrscheinlichkeit, daß die Repräsentation des Sprachsignals der zweiten Nachrichtenmeldung von der Wortfolgenschätzung der ersten Nachricht erzeugt wurde, dividiert durch die Länge der zweiten Nachrichtenmeldung, oder noch einfacher; wie gut das Sprachsignal der zweiten Nachrichtenmeldung zu den gefundenen Wörtern der ersten Nachrichtenmeldung paßt. Die Elemente der mit der Realignment-Methode berechneten Abstandsmatrix sind nach ihrer Größe sortiert in Abb. 7.6 dargestellt. Wieder sind links die Meldungspaare abgebildet, für die die größte Wahrscheinlichkeit berechnet wurde, die also als am ähnlichsten eingestuft wurden, und rechts die Meldungspaare mit den geringsten berechneten Wahrscheinlichkeitswerten, also die als am unterschiedlichsten eingestuften Meldungspaare.

Verglichen mit den Ergebnissen der so einfach und effizient zu berechnenden naiven Methode sind die Resultate zunächst enttäuschend, da Paare äquivalenter Nachrichtenmeldungen schlechter von den unterschiedlichen separiert werden. Die folgenden Überlegungen erklären, warum sich der höhere Berechnungsaufwand erst nicht auszahlt. Zunächst fällt auf, daß die Hauptdiagonalenelemente von A, also die Abstände der Nachrichten zu sich selbst, keineswegs wie bei der naiven Methode deutlich von den anderen abgehoben sind. Sie sind, wie eine nähere Betrachtung ergeben hat, noch nicht einmal die 32 am weitesten links

[2] Das Auffinden wiederkehrender Nachrichten soll es ermöglichen, prosodische Variation von Nachricht zu Nachricht zu untersuchen. Die Nachrichten können sich auch bezüglich der Phrasierung und damit bezüglich der Pausensetzung unterscheiden. Dieser Unterschied soll beim Vergleich nicht stören, deshalb werden zwischen je zwei Wörtern der ersten Nachricht optionale Pausenmodelle eingesetzt, also Pausen modellierende HMM, die komplett übersprungen werden können. Ob und wo in der zweiten Nachricht tatsächlich Pausenmodelle eingesetzt werden, wird im Erkennungsschritt beim Realignment vom Viterbi-Algorithmus entschieden.

stehenden, also die als am ähnlichsten eingestuften Paare. Dies ist dadurch zu erklären, daß das verwendete Vokabular der 100 häufigsten Wörter unterschiedlich gut das annähern kann, was in einer Nachrichtenmeldung verlesen wurde, das heißt, daß das in einer Nachrichtenmeldung verwendete Vokabular unterschiedlich gut mit dem in der HM-Modellierung verwendeten Vokabular übereinstimmt. Eine Meldung mit vielen OOV-Wörtern erhält dadurch beim Vergleich mit sich selbst einen schlechten Wahrscheinlichkeitswert. Eine Meldung mit wenigen OOV-Wörtern erhält beim Vergleich mit sich selbst einen viel höheren Wert. Wenn zwei unterschiedliche Nachrichtenmeldungen verglichen werden, hängt der berechnete Abstandswert sowohl davon ab, wie gut die Wörter der einen zum Sprachsignal der anderen Meldung passen, als auch davon, wie gut beide Nachrichtenmeldungen durch das Vokabular beschreibbar sind. Wenn die Wortfolgenschätzung sehr selektiv ist, wenn beispielsweise viele lange, ausgefallene Wörter gefunden wurden, weil die Meldung gut zum Vokabular paßt, werden äquivalente Meldungen einen sehr guten Ähnlichkeitswert erzielen und nicht äquivalente einen deutlich schlechteren. Wenn andererseits die gefundenen Wörter wenig selektiv sind, beispielsweise ausschließlich gut zur Phonotaktik des Deutschen passende kurze Wörter (weil die Meldung schlecht zum Vokabular paßt), dann werden sich die Ähnlichkeitswerte für äquivalente und für nicht äquivalente Nachrichtenmeldungen nur wenig unterscheiden. Das Problem ist also, daß die Werte in den Reihen der Matrix A auf unterschiedlichem Niveau liegen. Das gleiche kann jedoch auch für die Spalten gezeigt werden: Die HMM, die zur Erkennung verwendet werden, wurden auf Sprachmaterial trainiert, das von Sprechern und Sprecherinnen stammt, die von den Nachrichtensprechern und Nachrichtensprecherinnen verschieden sind. In Abhängigkeit davon, wie gut die Stimme eines Ansagers oder einer Ansagerin zum Trainingsmaterial paßt, können sich unterschiedliche Ähnlichkeitswerteniveaus ergeben. Möglicherweise kann man die erwähnten Probleme durch geschickte Normalisierungen kompensieren. Einige einfache Normalisierungen erbrachten jedenfalls nicht die erwarteten Verbesserungen.

Wenn man einen Schwellwert unmittelbar vor dem ersten nicht äquivalenten Nachrichtenmeldungspaar bei -69 wählt (siehe Abb. 7.6), wird nach Korrektur durch Bildung der transitiven Hülle nur ein einziger korrekter Cluster gefunden. Die nächsten drei Lücken bei $-69,55$, $-69,7$ und $-70,2$ ergeben alle 2 richtige und 1 zu großen Cluster, wobei der zu große Cluster für den höchsten Schwellwert $-70,2$ stärker vergrößert ist als für die beiden niedrigeren Schwellwerte. Beispielsweise fast dieser Cluster zwei richtige Cluster und zwei Meldungen zusammen, die mit beiden nichts zu tun haben. Möglicherweise würde ein relativ niedriger Schwellwert in Verbindung mit dem Finden von zweifach zusammenhängenden Komponenten als Reparaturstrategie zu erfreulicheren Ergebnissen führen.



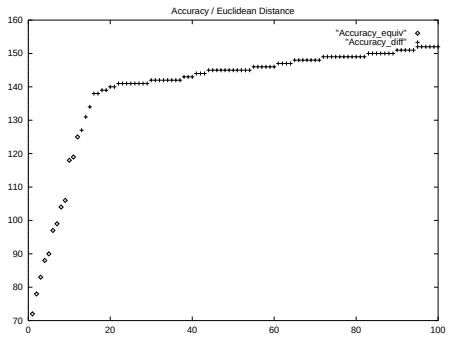


Abbildung 7.7: Die Einträge der Distanzmatrix für die Konstellation kleine Nachrichtenmenge / naiver symbolischer Vergleich / Metadistanz.

Kleine Nachrichtensammlung, naive Methode, Metadistanz. Als nächstes werden die Verbesserungen beschrieben, die die Berechnung der Metadistanz in Verbindung mit der naiven Methode erbringt. Abbildung 7.7 zeigt die Metadistanzmatrix für die naive Methode. Die Graphik wurde auf gleiche Weise wie in den oben beschriebenen Versuchen erstellt mit dem einzigen Unterschied, daß statt der direkten Abstände die Metadistanzen verwendet wurden. Vergleicht man Abb. 7.7 mit Abb. 7.5, so erkennt man, daß die Berechnung der Metadistanz die Ergebnisse der naiven Methode deutlich verbessert. Jetzt sind die beiden Klassen „äquivalente Meldungen“ (rautenförmig) und „verschiedene Meldungen“ (Punkte bzw. Kreuze) eindeutig durch den Schwellwert von 126 voneinander zu trennen, wie man der Ausschnittvergrößerung rechts in Abbildung 7.7 entnehmen kann.[3] Mit dem Schwellwert von 126 werden alle sechs Cluster korrekt gefunden, auch ohne eine graphentheoretische Korrekturstrategie anzuwenden. Problematisch ist es jedoch, diesen Schwellwert zu bestimmen, ohne die vorliegenden Gleichheiten und Ungleichheiten zwischen den Nachrichtenmeldungen zu kennen. Aus der Graphik in Abb. 7.7 lassen sich ohne die Unterscheidung der Punkte in Kreuze und Rauten eher 136 (Kniepunkt) oder 110 (größere Lücke) als Schwellwert ablesen. Mit dem Schwellwert 110 ergeben sich die gleichen Cluster wie ohne Metadistanzberechnung, das heißt, einer von drei Zweierclustern wird übersehen und einer von drei Dreierclustern nur unvollständig entdeckt. Mit dem Schwellwert 136 enthält ein Cluster zusätzlich eine unpassende Nachrichtenmeldung und es werden zwei zusätzliche Cluster gefunden, die je kurze und eine verlängerte Fassung einer Nachrichtenmeldung enthalten (die als verschiedene Meldungen betrachtet werden sollen).

[3]Bei den aus Metadistanzen erzeugten Plots sind nur die Elemente der oberen Dreiecksmatrix gezeigt, da die Metadistanzmatrix symmetrisch ist. (Der Euklidische Abstand ist eine Norm, daher gilt: $d(\mathbf{a},\mathbf{b}) = ||\mathbf{a}-\mathbf{b}|| = ||(-1)\cdot(\mathbf{b}-\mathbf{a})|| = |-1|\cdot||\mathbf{b}-\mathbf{a}|| = ||\mathbf{b}-\mathbf{a}|| = d(\mathbf{b},\mathbf{a})$.) Die Elemente der Hauptdiagonalen sind in diesen Bildern ebenfalls nicht geplottet.

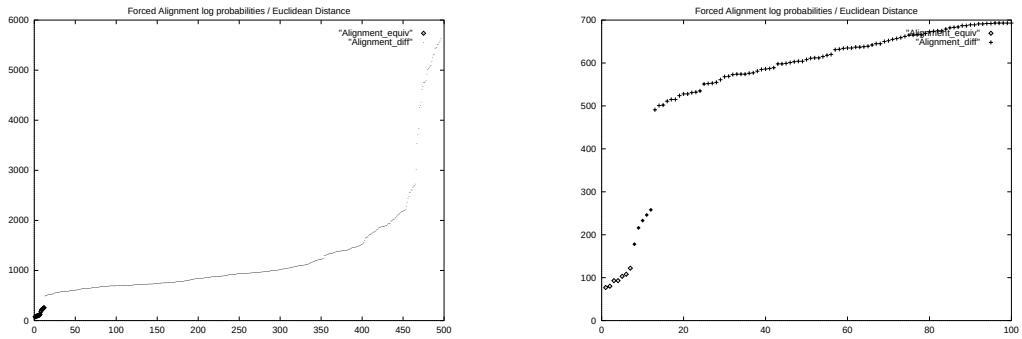


Abbildung 7.8: Die Einträge der Distanzmatrix für die Konstellation kleine Nachrichtenmenge / Realignment-Methode / Metadistanz.

Kleine Nachrichtensammlung, Realignment-Methode, Metadistanz. Die in Abb. 7.8 wiedergegebenen Graphiken zeigen die Metadistanzen für die zur Matrix zusammengefaßten normierten logarithmierten Wahrscheinlichkeiten der Realignment-Methode. Der Vergleich mit den entsprechenden Ergebnissen ohne Metadistanzberechnung (Abb. 7.6) zeigt eine drastische Verbesserung. Äquivalente Meldungen lassen sich jetzt mit dem Schwellwert 400 perfekt von den nicht-äquivalenten Paaren trennen. Der Vergleich mit Abb. 7.7 macht die Vorteile der aufwendigeren Realignment-Methode gegenüber dem naiven Ansatz evident. Zwar war auch dort eine fehlerfreie Separation möglich, doch ist auch die Bestimmung eines geeigneten Schwellwerts mit der Realignment-Methode kein Problem mehr. Wie aus Abbildung 7.8 unmittelbar ersichtlich wird, sind Schwellwerte zwischen 300 und 450 sowohl gleichermaßen geeignet als auch die einzig vernünftigen Möglichkeiten, einen Schwellwert zu bestimmen, da der Plot hier eine deutliche Lücke aufweist. Ein interessantes Detail zeigt die Vergrößerung rechts in Abb. 7.8, nämlich daß ein weiterer, nicht so deutlicher Abstand bei etwa 150 die äquivalenten Meldungspaare trennt in solche, die vom gleichen Sprecher wiederholt wurden, und solche, die von verschiedenen Sprechern verlesen wurden. In der Ausschnittsvergrößerung rechts in Abb. 7.8 sind die von unterschiedlichen Sprechern verlesenen äquivalenten Meldungspaare als gefüllte Rauten repräsentiert.

Größere Nachrichtensammlung

Da die direkte Distanz durchweg schlechtere Ergebnisse liefert, beschränkt sich die Diskussion der mit der größeren Nachrichtenmenge gewonnenen Ergebnisse auf die Experimente, bei denen die Metadistanz verwendet wurde. Um die Untersuchungen vollständig zu dokumentieren, sind die Abstandsmatrizen mit der

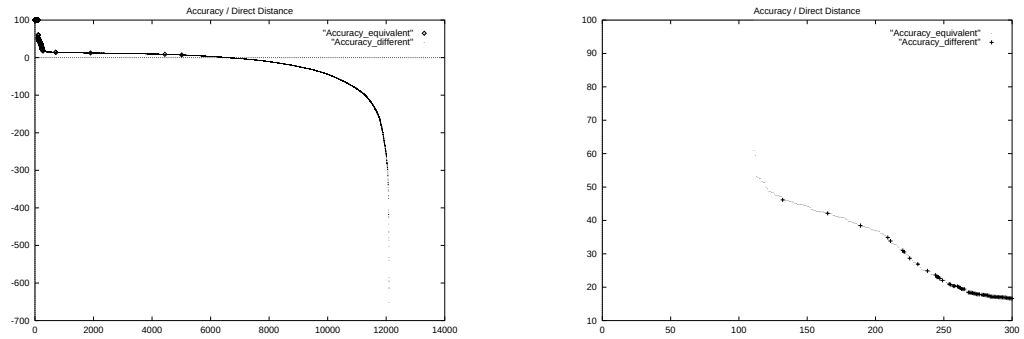


Abbildung 7.9: Die Einträge der Distanzmatrix für die Konstellation größere Nachrichtenmenge / naiver symbolischer Vergleich / direkte Distanz.

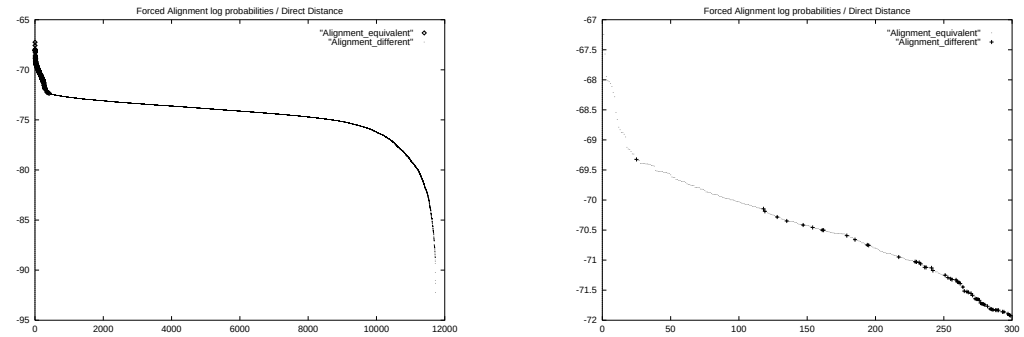


Abbildung 7.10: Die Einträge der Distanzmatrix für die Konstellation größere Nachrichtenmenge / Realignment-Methode / direkte Distanz.

direkten Distanz in den Abbildungen 7.9 (naive Methode/ direkte Distanz) und 7.10 (Realignment-Methode/ direkte Distanz) beigefügt.

Weil für die größere Nachrichtenmenge in den Abbildungen sehr viel mehr Punkte zu plotten sind, stellen die Ausschnittsvergrößerungen rechts in den Abbildungen die äquivalenten Meldungspaare in Form kleiner Punkte statt als Rauten dar. Sonst wären die falsch positiven Fälle (fälschlicherweise als äquivalent erkannte Meldungspaare) nicht mehr zu erkennen. Die falsch positiven Fälle sind in den Ausschnittsvergrößerungen leichter zu unterscheiden, die falsch negativen Fälle (nichterkannte Gleichheiten) in den linken Graphiken.

Größere Nachrichtenmenge, naive Methode, Metadistanz. Mit der naiven Methode und einem Schwellwert von 126 werden in der größeren Nachrichtenmenge insgesamt 22 Cluster gefunden, die 53 Meldungen umfassen. Das größte Cluster beinhaltet fünf Meldungen, die nächstkleineren sind sechs Cluster mit je

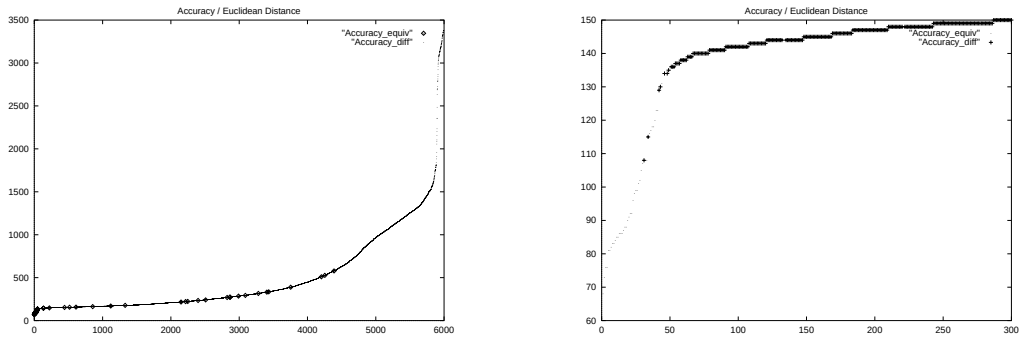


Abbildung 7.11: Die Einträge der Distanzmatrix für die Konstellation größere Nachrichtenmenge / naiver symbolischer Vergleich / Metadistanz.

drei Meldungen. Abb 7.11 zeigt die Abstandsmatrix für diesen Versuch. In der Ausschnittsvergrößerung erkennt man zwei falsch positive Fälle mit einem an der Ordinate abgetragenen Abstand von 108 und 115. Diese stammen von den beiden folgenden Meldungen, die bis auf den jeweils letzten Satz identisch sind.

> Die USA haben der Bosnienkonferenz in Dayton eine neue Frist bis zum Nachmittag gesetzt. Sollte bis sechzehn Uhr Mitteleuropäischer Zeit eine Einigung nicht erzielt sein, würden die Gespräche der Delegationen aus Bosnien, Kroatien und Serbien formell beendet. Weiter hieß es, die amerikanischen Vermittler hätten die Vertreter der Konfliktparteien am Verhandlungsort allein zurückgelassen, in der Hoffnung, daß sie untereinander die letzten strittigen Fragen klären könnten. Der bosnische Außenminister `[SE] [tS@] ["bEj]` sprach bereits von einem Fehlschlag. [dlf951121.1230.n1]

> Die USA haben der Bosnienkonferenz in Dayton eine neue Frist bis zum Nachmittag gesetzt. Sollte bis sechzehn Uhr Mitteleuropäischer Zeit eine Einigung nicht erzielt sein, würden die Gespräche der Delegationen aus Bosnien, Kroatien und Serbien formell beendet. Weiter hieß es, die amerikanischen Vermittler hätten die Vertreter der Konfliktparteien am Verhandlungsort allein zurückgelassen, in der Hoffnung, daß sie untereinander die letzten strittigen Fragen klären könnten. Der Sprecher des US-Außenministeriums Burns machte deutlich, von einem Fehlschlag der Verhandlungen könne nicht die Rede sein. [dlf951121.1430.n2]

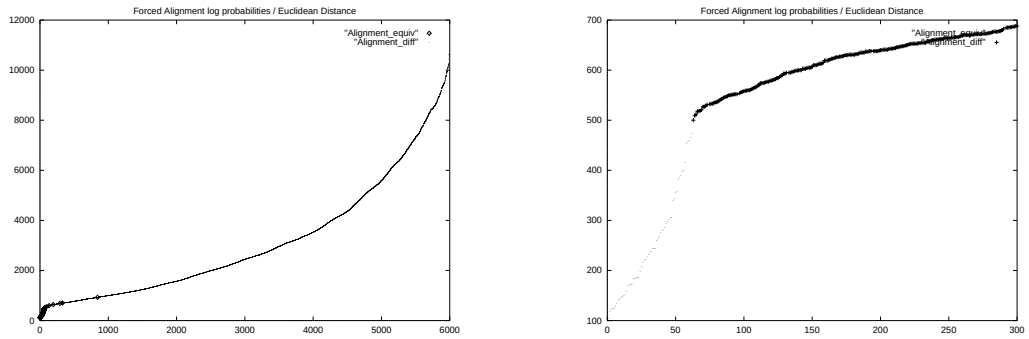


Abbildung 7.12: Die Einträge der Distanzmatrix für die Konstellation größere Nachrichtenmenge / Realignment-Methode / Metadistanz.

Größere Nachrichtenmenge, Realignment-Methode, Metadistanz. Der Vergleich von Abb. 7.11 mit Abb. 7.12 zeigt erneut, daß die aufwendiger zu berechnende Realignment-Methode dem einfacheren naiven Ansatz überlegen ist, da wesentlich weniger falsch negative Fälle auftreten.

Die Realignment-Methode liefert mit einem Schwellwert von 490 wie die naive Methode 22 Cluster, die zusammen 56 Nachrichtenmeldungen beinhalten. Die Realignment-Methode findet hier jedoch größere Cluster: ein Cluster mit acht Meldungen, ein Cluster mit fünf Meldungen und drei Cluster mit je drei Meldungen. Die Abstandsmatrix ist in Abb. 7.12 zu sehen. Sie zeigt, daß es mit der Realignment-Methode keine falsch positiven Fälle unterhalb des Schwellwerts gibt, so daß alle gefundenen Cluster korrekt sind bezüglich der (willkürlichen) Definition der Übereinstimmung, bis zu 10 Wörter betreffende Abweichungen als vernachlässigbare Umformulierungen zu betrachten.

Bemerkenswert ist, daß Paare äquivalenter Meldungen, die von verschiedenen Personen verlesen wurden, tendenziell näher am Schwellwert liegen als vom gleichen Sprecher oder der gleichen Sprecherin verlesene Versionen gleicher Meldungen. Bei dieser umfangreicheren Nachrichtensammlung sind jedoch die Paare gleicher Meldungen von verschiedenen Sprechern und die Paare leicht umformulierter Meldungen eines einzigen Sprechers miteinander vermischt, so daß sich diese Unterklassen hier nicht so gut trennen lassen wie bei der kleineren Nachrichtensammlung. Für die Unterscheidung zwischen Versionen von verschiedenen Sprechern und Versionen des gleichen Sprechers oder der gleichen Sprecherin, versprechen traditionelle Sprechererkennungsmethoden (wie etwa [KMS96]) mehr Erfolg.

Nach der hier verwendeten Definition der Äquivalenz von Nachrichtenmeldungen gibt es in der größeren Nachrichtensammlung 60 mehrfach vorkommende Mel-

dungen, die in 23 Clustern organisiert sind. Die überlegene Realignment-Methode findet also nur vier wiederholte Meldungen nicht, die größten, 8 und 5 Meldungen umfassenden Cluster, die für die prosodische Untersuchung am interessantesten sind, kommen vollzählig vor.

Vollautomatisch erhobene Nachrichtensammlung

Die folgenden Untersuchungen zeigen die Anwendung der effizienteren naiven und der aufwendigeren Realignment-Methode auf automatisch in einzelne Meldungen geschnittene Nachrichtensendungen. Damit ist die Meldungsclustersammlung hier also voll automatisiert. Wie im vorherigen Abschnitt werden für beide Methoden nur die Metadistanzen diskutiert.

Da für die vollautomatisch erhobene Nachrichtensammlung keine Transliterationen vorliegen, läßt sich hier, ohne die Meldungen abzuhören, nicht entscheiden, welche Meldungspaare äquivalent sind und welche nicht. Daher repräsentieren die graphischen Darstellungen der in den Untersuchungen berechneten Abstandsmatrizen alle Meldungspaare mit nur einer Symbolart, in der vollständigen Darstellung links mit Punkten, in der Ausschnittsvergrößerung rechts mit Kreuzen. Zur Auswertung der gewonnenen Ergebnisse werden nur die gefundenen Cluster überprüft, indem die enthaltenen Meldungen verglichen werden. Damit lassen sich also nur die falsch positiven Fälle erkennen, nicht aber die falsch negativen, also vom Verfahren übersehene Übereinstimmungen.

Für die Anwendung beider Methoden ist es wieder nötig, einen Schwellwert festzulegen. Dies erfolgt durch Auswertung der graphischen Darstellungen der Abstandsmatrizen und unter Berücksichtigung der Erfahrungen aus den bisher durchgeführten Untersuchungen. Da die Schwellwerte für die beiden vorangegangenen Nachrichtensammlungen nur unwesentlich voneinander abweichen, läßt jedoch auch ein global einheitlich fixierter Schwellwert vernünftige Ergebnisse erwarten.[4]

Automatisch erhobene Nachrichtensammlung, naive Methode, Metadistanz. Abbildung 7.13 zeigt die sortierten mit der naiven Methode bestimmten Metaabstände. Die Auschnittsvergrößerung legt einen Schwellwert von 140 nahe,

[4]Die Tatsache, daß der Schwellwert für die verschiedenen Nachrichtensammlungen nahezu gleich bleibt, ist etwas überraschend; sie steht im Gegensatz zu der Erwartung, daß der Schwellwert mit der Anzahl der betrachteten Nachrichtenmeldungen wächst. Diese Anzahl entspricht der Dimensionlität des Raums, in dem die Euklidische Distanz berechnet wird, und damit der Anzahl der Summanden unter der Wurzel.

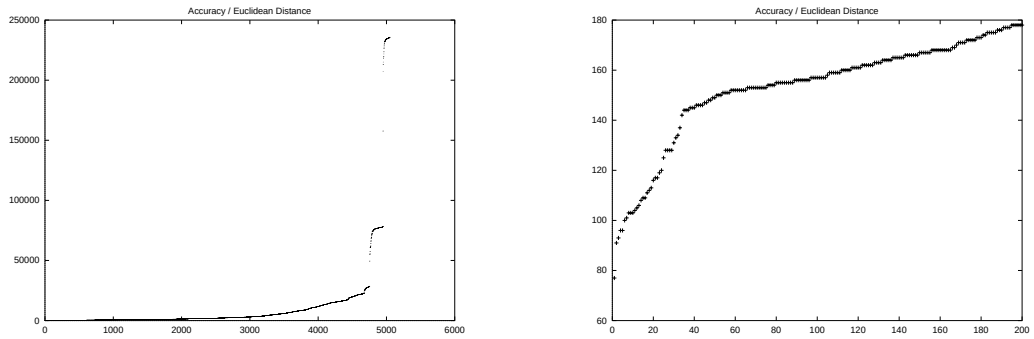


Abbildung 7.13: Die Einträge der Distanzmatrix für die Konstellation automatisch gesammelte Nachrichtenmenge / naiver symbolischer Vergleich / Metadistanz.

oder wenn man die Erfahrungen bei den letzten beiden Nachrichtensammlungen miteinbezieht, einen etwas darunter liegenden Schwellwert. Das Abhören aller Nachrichtenpaare mit einem Abstand zwischen 125 und 140 ergab, daß alle diese Paare äuivalent sind, mit Ausnahme dessen mit dem größten Abstandswert, so daß der vermutlich optimale Schwellwert von 135 gewählt wurde. Damit findet die naive Methode 14 Cluster mit insgesamt 36 mehrfach vorkommenden Meldungen. Die größten so gefundenen Cluster sind zwei Cluster mit je 5 und zwei Cluster mit je drei Meldungen. Einer der beiden Fünfercluster enthält zwar nur Meldungen, die über das gleiche Thema berichten, aber nur zwei bzw. drei Meldungen weisen eine äquivalente Formulierung auf. Nach dem 10-Wörter-Kriterium wäre dieser Cluster also in einen Zweier- und einen Dreiercluster aufzuteilen. Auch einer der Dreiercluster faßt Meldungen zusammen, die in mehr als 10 Wörtern voneinander abweichen. Die Ursache dieses Fehlers ist eine falsche Segmentierung in Einzelnachrichten. Einer der Ansager markiert den Wechsel zur jeweils nächsten Einzelmeldung weniger deutlich durch Pausenlängung als die anderen Sprecher und Sprecherinnen. Zwei der drei „Meldungen“ des fehlerhaften Cluster bestehen deshalb aus beinahe vollständigen Nachrichtensendungen, die dritte Meldung dieses Clusters (die tatsächlich eine Einzelmeldung ist) ist in den beiden anderen enthalten. Insgesamt erhält man mit der naiven Methode einen fünf übereinstimmende Meldungen umfassenden Cluster, zwei Dreiercluster und mehrere Zweiercluster.

Automatisch erhobene Nachrichtensammlung, Realignment-Methode, Metadistanz. Die Abstandsmatrix der Realignment-Methode ist in Abb. 7.14 wiedergegeben. In der Ausschnittsvergrößerung scheint ein Schwellwert von 450 plausibel, wenngleich auch andere Werte als Schwellwerte in Frage kämen.

Mit dem Schwellwert 450 findet die automatische Prozedur 5 Vierercluster, 4 Dreiercluster und 11 Meldungspaare. Insgesamt werden also 54 wiederkehren-

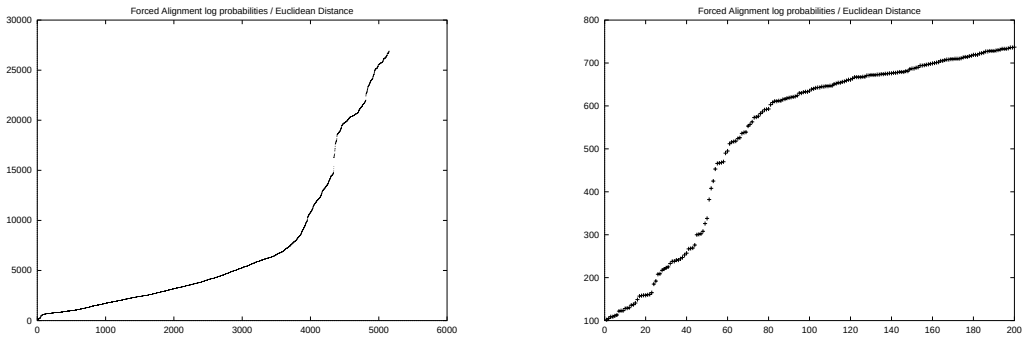


Abbildung 7.14: Die Einträge der Distanzmatrix für die Konstellation automatisch gesammelte Nachrichtenmenge / Realignment-Methode / Metadistanz.

de Meldungen in 20 Clustern gefunden (zu vergleichen mit den 36 Meldungen in 14 Clustern bei der naiven Methode). Die Realignment-Methode übersieht zwar zum Beispiel den korrekt bestimmten Fünfercluster der naiven Methode fast vollständig, ergibt dafür aber wesentlich verläßlicheres Material. Der zweite von der naiven Methode fälschlich erkannte Fünfercluster ist hier korrekt als zwei getrennte Cluster bestimmt, und zwar als ein Vierer- und ein Zweiercluster. Somit ist also noch eine Realisierung der Meldung zusätzlich gefunden und die umfangreiche Umformulierung wird dem Äquivalenzkriterium gemäß richtig behandelt. Eine weitere Illustration der Überlegenheit der Realignment-Methode ist: Zu einem spezifischen Thema findet die naive Methode einen Zweier- und einen Dreiercluster (die mit zwei verschiedenen Formulierungen korrespondieren, die eine, später gesendete, enthält zusätzliche Information), während die Realignment-Methode zum selben Thema ein Meldungspaar, einen Dreier- und einen Viercluster erbringt, die letzten beiden Cluster enthalten nur geringe Unterschiede.

Zur Verifikation wurden alle Nachrichtenmeldungen in jedem der gefundenen Cluster abgehört. Die Überprüfung ergab, daß bis auf eine Ausnahme alle von der Realignment-Methode bestimmten Cluster nur äquivalente Meldungen enthalten. Die einzige Ausnahme ist ein Viercluster mit einer Meldung, die in 13 statt in 10 Wörtern von den übrigen Meldungen des Clusters abweicht und damit das 10-Wörter-Kriterium für eine erlaubte leichte Umformulierung etwas überschreitet.

7.1.7 Bewertung

Dieser Abschnitt hat zwei Methoden vorgestellt, um wiederholt verlesene Nachrichten in Radionachrichten aufzufinden. Die erste, naive Methode ist sehr schnell, da sie lediglich symbolische Vergleiche über die Wortkettenschätzungen vornimmt, die bereits bei der Segmentierung in Einzelmeldungen berechnet werden

müssen. Die zweite Realignment-Methode ist wesentlich aufwendiger zu berechnen, doch ihre Ergebnisse sind denen der naiven Methode klar überlegen. Lediglich bei der kleinen Nachrichtensammlung kann die naive Methode gleich gute Ergebnisse wie die Realignment-Methode erbringen. Für die größere und die vollautomatisch erstellte Nachrichtensammlung jedoch ergibt die Realignment-Methode verläßlichere Cluster und findet mehr Versionen erneut verlesener Meldungen. Die mit der Realignment-Methode gefundenen Meldungscluster sind also sowohl qualitativ als auch quantitativ besser als die der naiven Methode. Methodologisch interessant ist dabei, daß sich der größere Rechenaufwand für die Realignment-Methode erst in Verbindung mit der Berechnung der Metaabstände auszahlt. Diese Verbesserung läßt sich mit der Realignment- und mit der naiven Methode zusammen einsetzen und ergibt stets bessere Ergebnisse als die Verwendung der direkten Abstände. Die Metaabstandsberechnung scheint bei der Realignment-Methode zum Ausgleich der unterschiedlichen Abstandsniveaus beizutragen, die aus den Sprechereinflüssen und der unterschiedlichen Übereinstimmung mit dem Erkennungsvokabular resultieren. Vorteilhaft an der Verbesserung durch Bestimmung der Metadistanz ist außerdem, daß sie keine Parameteranpassung erfordert und schnell zu berechnen ist.

Als Nachteil des Verfahrens mag die Notwendigkeit erscheinen, einen Schwellwert zur Trennung zwischen äquivalenten und nicht-äquivalenten Meldungspaaren bestimmen zu müssen. Dabei ist jedoch zu bedenken, daß auch schon die Definition dessen, was als Äquivalenz von Meldungen gelten soll, durchaus subjektiv ist. Die Erfahrungen in den beschriebenen Untersuchungen haben gezeigt, daß sich ein akzeptabler Schwellwert aus der Form der graphisch dargestellten Abstandsmatrizen ablesen läßt, und daß sogar ein global fixierter Schwellwert zu guten Ergebnissen führen kann. Beim praktischen Einsatz des Verfahrens bieten sich zwei Alternativen an, nämlich entweder einen eher laxen Schwellwert zu wählen und nahe am Schwellwert liegende, zweifelhafte Fälle abhören und gegebenenfalls auszusortieren, oder aber einen rigiden Schwellwert zu wählen und entsprechend mehr Nachrichtenmaterial aufzuzeichnen und zu analysieren. Dies ist deshalb leicht zu realisieren, da die ganze Prozedur automatisch ablaufen kann.

Die Effizienzfrage ist für die Sammlung von Material für die Prosodieforschung nicht dominierend. Damit die vorgestellten Methoden auch für Information-Retrieval-Anwendungen einsetzbar sind, sind jedoch Effizienzverbesserungen sinnvoll. Sowohl die Realignment- als auch die naive Methode setzen die Segmentierung der Nachrichtensendung in einzelne Meldungen voraus, die gleichzeitig die Wortkettenschätzung für jede einzelne Meldung bestimmt. Die Segmentierung der 17 Nachrichtensendungen für die automatisch erstellte Nachrichtensammlung nimmt auf der schnellsten Workstation des Lehrstuhls 41 Stunden

in Anspruch. Durch Beschneidung des Suchraums während der HMM-Erkennung statt der hier verwendeten vollen Suche könnten bei der Segmentierung und Wortkettenschätzung dramatische Einsparungen erzielt werden. Die Cluster-Bildung mit der naiven Methode benötigt nach der Bestimmung der Wortkettenschätzung lediglich noch einige Minuten, während die Realignment-Methode mit einer Ausführungszeit von über einer Woche[5] erheblich länger dauert.

Bisher wurden noch keine Anstrengungen unternommen, um die Kosten der Realignment-Methode zu optimieren. Der weitaus teuerste Teil der Gesamtprozedur ist das Realignment, das mit Kosten von $O(n^2)$ noch deutlich aufwendiger ist als das Finden der Wortketten.[6] Es stehen mehrere Möglichkeiten offen, die Berechnung zu beschleunigen. Zunächst lassen sich die Alignments vollständig parallelisieren und können dann über einen Workstationcluster verteilt ablaufen, z.B. nachts oder über ein Wochenende. Zweitens kann das Alignment selbst wie die Segmentierung und Wortkettenschätzung wesentlich beschleunigt werden durch Beschneidung des Suchraums während der Viterbi-Decodierung. Drittens ist ein hybrider Ansatz möglich, der zunächst mittels der naiven Methode und einem laxen Schwellwert Kandidaten für wiederkehrende Meldungen auswählt und diese dann anhand der Realignment-Methode verifiziert. Allerdings ist bei einer Reduktion der Kandidatenzahl auch zu erwarten, daß sich dann die Vorteile der Metadistanzberechnung weniger stark auswirken. Schließlich besteht noch die Verbesserungsmöglichkeit, intelligente Schlüsselwortsuche auf den Nachrichtenmeldungen selbst anzuwenden. Vorbedingung dafür sind geeignete Schlüsselwörter, die sich idealerweise auch an die veränderten Konstellationen der Weltgeschichte anpassen, beispielsweise indem sie dynamisch aus Nachrichtentexten aus dem WWW als Wissensquellen akquiriert werden. Unabhängig davon, welcher Weg zur Extension in Richtung Information Retrieval aus Radionachrichten eingeschlagen wird, haben sich die in diesem Kapitel vorgeschlagenen Methoden bereits als geeignet erwiesen zur Konstruktion, Sammlung und Klassifikation einer Datenbasis wiederkehrender Radionachrichten, die hochinteressante linguistische Eigenschaften aufweist. Auch ohne die Effizienzverbesserungen sind die Kosten des vorgestellten Verfahrens nicht zu hoch für eine sorgfältig selektierte Datenbank, die für Forschungen in Sprachproduktion, Prosodie- und Sprachgenerierung genutzt werden kann.

---

[5] Von Samstag um Mitternacht bis zum Morgen des übernächsten Sonntags. Während der Bürozeiten hatte die Maschine noch eine geringe zusätzliche Belastung. Alle Zeiten wurden als Gesamtlaufzeit auf einem Einprozessor-Rechner (SGI Indigo^2) mit 200 MHz und MIPS 4400/4010 Prozessor gemessen.

[6] Zu beachten ist allerdings, daß für das Forced Alignment während der Viterbi-Decodierung wesentlich weniger Pfade betrachtet werden müssen als zur Erkennung der Wortkettenschätzungen.

7.2 Erstellung einer Übersicht über die Intonationsvariation auf phonologischer Ebene

Bei mehrfachen Äußerungen, wie den automatisch gefundenen wiederkehrenden Radionachrichtenmeldungen aus dem vorausgehenden Abschnitt, und zugehörigen phonologischen Intonationsbeschreibungen ist für die linguistische Theoriebildung interessant, an welchen Stellen die prosodische Etikettierung über alle Realisierungen hinweg konstant ist und an welchen sie wechselt. Zur Erleichterung solcher Vergleiche wurde ein Werkzeug erstellt, das die unterschiedlichen prosodischen Realisierungen der Varianten übersichtlich darstellt. Dabei wird als unerheblich erachtet, ob der Vergleich zwischen manuellen oder automatisch gefundenen prosodischen Etikettierungen erfolgen soll. Das Werkzeug kann auch dazu dienen, verschiedene Etikettierungen einer einzigen Äußerung zu vergleichen, beispielsweise manuelle und automatisch gefundene Etikettierungen, oder anhand unterschiedlicher Labellingkonventionen bzw. -systeme erstellte Etikettierungen.

Das Werkzeug faßt verschiedene phonologische Ettikettierungen zusammen, indem es statt der Zeit die Wortkette zur Synchronisierung benutzt. Da auch der Einfluß von Umformulierungen auf die intonatorische Realisierung interessant ist, muß das Werkzeug mit leichten Äußerungsunterschieden, also Abweichungen bei den Wortketten der dargestellten Varianten, zurechtkommen. Als Ausgangsdaten stehen Wortetikettierungen und phonologische Intonationsettikettierungen von potentiell verschiedenen, größtenteils jedoch gleichen Äußerungen bereit. Aus den Wortsegmentierungen wird zunächst eine alle Äußerungen repräsentierende Wortkette gebildet, in dem parallele Pfade serialisiert werden. Diese gemeinsame Wortkette wird in der Ausgabe als Referenz jeweils in der obersten Zeile dargestellt. Darunter stehen Wort für Wort die Intonationsbeschreibungen der einzelnen Realisierungen. Durch die Serialisierung bedingte, und in einer Realisierung nicht vorhandene Abschnitte werden dabei mit [_] kenntlich gemacht. Wenn in allen Realisierungen eine Intonationsphrasengrenze auftritt, wird auch schon vor Erreichen des rechten Blattrandes eine neue Zeile begonnen. In der ersten Zeile wird der Dateiname der Variante hinzugefügt, damit deutlich wird, welche Zeile zu welcher Realisierung gehört. Die Ausgabe steht für eine Weiterverarbeitung im Textformat (Wort und zugehörige Intonation der Varianten) und außerdem in LaTeXund Postscript zur Verfügung.

In Abbildung 7.15 ist die Variation der Intonation für eine wiederholt verlesene Radionachricht dargestellt. Die Wortetikettierung wurde dazu automatisch, die phonologische Ettikettierung manuell erstellt.

```
                <P> Die Nato   favorisiert jetzt offenbar den spanischen Außenminister Solana
dlf951121.1100.n2        H*L % L*H                          L*H                          L*HL %
dlf951121.1130.n2        H*L % L*H                          L*!H                         H*L  %
dlf951121.1200.n3        L*H  -            ?                L*H                          L*H  %
                <P> als Nachfolger des zurückgetretenen Generalsekretärs Claes
                [___]    L*H              L*H?              ?                  H*L  %
                [___]    L*H                                                   H*L  %
                         L*H              L*!H                                 !H*L %
                <P> In Madrid erklärte ein Sprecher der regierenden Sozialisten
                        L*H                    L*!H          L*H          L*H      %
                        L*H   -                L*H           L*!H         L*H      %
                        L*H   -                L*H           L*!H         L*H      %
                <P> in der Nato gäbe es einen gewissen Konsens darüber
                           L*H                    L*!H      L*H      L*H   %
                           L*H                              L*HL     H%
                           L*H -                            L*H      L*H  %H
                <P> daß Solana kandidieren solle
                %H       ?       H*L          L%
                [___]    HH*L   !H*L          L%
                         H*L                  %
                <P> Mehrere europäische Länder hätten bereits ihre Unterstützung zugesagt
                     L*H       L*!H         L*H                          H*L              L%
                     L*H       L*!H         L*!H  %                      H*L              L%
                     L*H                    L*!H  -                      H*L?             %
                <P> Auch in den USA   habe es lobende Worte   für den spanischen Außenminister
                                L*H  -          L*H      H*L    -
                                L*H %           L*H      !H*L? %
                     L*H        !H*L -          ?        H*L?   -          ?
                gegeben
                HH*L  %
                HH*L  %
                H*L   %
                <P> Der noch verbliebene einzige offizielle Kandidat <P> der frühere dänische
                           L*H          L*H     L*!H      L*H      - [___]          L*H
                           L*H        - L*H   - L*H      L*!H     %          L*H    L*H
                           L*H                  L*!H?     L*H      % [___]    H*L?
                Außenminister Eleman Jensen
                                     L*H H%
                                     H*L   %
                                     L*H   %
                <P> stößt bei Frankreich auf <P> Vorbehalte
                           H*L              [___] !H*L     L%
                     ?     L*                     H*L      L%
                           H*           ..L       H*L       %
                <P>
```

Abbildung 7.15: Übersicht über die Variation der manuell annotierten phonologischen Intonationsbeschreibung einer wiederholt verlesenen Radionachrichtenmeldung.

Kapitel 8

Schlußbemerkungen

Die in dieser Arbeit vorgestellten und entwickelten Verfahren bilden zusammen ein System, das eine beinahe vollautomatische Erstellung eines Korpus für die Prosodieforschung möglich macht. Die Herstellung einer orthographischen Transliteration ist der einzige Schritt, der menschliche Arbeit erfordert.[1] Dafür sind jedoch keine phonetisch geschulten Experten erforderlich.

Die einzelnen Komponenten spielen wie folgt zusammen: Das Segmentierungssystem Alphons (Kapitel 3) zerlegt das Sprachsignal unter Verwendung der orthographischen Transliteration in Wörter, Silben und Phoneme. Die in den Kapiteln 3, 4 und 5 beschriebenen Verfahren leiten für das segmentierte Sprachsignal Merkmale ab, die die akustische Ausprägung der Silben beschreiben. Die Merkmale bilden eine phonetische Beschreibung der akustischen Prosodiekorrelate Silben- beziehungsweise Pausendauer, Grundfrequenzverlauf und Intensität und stellen außerdem die lexikalischen Informationen Wortakzent und Stellung im Wort bereit. Diese Merkmale bilden die Eingabe für die in Kapitel 6 vorgestellte Systemkomponente, die aus der phonetischen Beschreibung eine phonologische Intonationsbeschreibung erzeugt, unter Verwendung eines auf manuell ettikettierten Sprachdaten traininerten Klassifikators. Diese Systembestandteile realisieren zusammen die automatische Herleitung einer phonologischen Intonationsbeschreibung aus akustischem Signal und orthographischer Transliteration.

Zur Erzeugung und Bearbeitung eines strukturierten Korpus dienen schließlich die in Kapitel 7 beschriebenen Verfahren zum automatischen Herausfiltern und Gruppieren wiederholt verlesener Meldungen aus den stündlichen Radionachrichten

[1] Sofern diese nicht schon vorliegt. Viele Nachrichtensender stellen heute ihre Radionachrichten in Textform im WWW bereit, zum Beispiel der Bayerische Rundfunk und der Westdeutsche Rundfunk. Auch bei Literaturlesungen steht häufig neben dem Sprachsignal die wortgetreue Transliteration zur Verfügung [Goe95, Rap96].

und zur übersichtlichen Darstellung und Zuordnung ihrer Intonationsbeschreibungen zur Orthographie.

Die entwickelten Methoden sind wirksam, aber ungenauer (mitunter deutlich ungenauer) als die Aufbereitung durch phonetische Experten. Andererseits ermöglicht der automatische Ablauf des Verfahrens, wesentlich größere Datenmengen zu analysieren als von menschlichen Experten zu bewältigen wären. Die Zielsetzung der Arbeit war, ein einsatzfähiges Gesamtsystem zu entwickeln. Dieses Ziel wurde erreicht, wenngleich bei den einzelnen Systemkomponenten zum Teil noch Verbesserungen möglich scheinen. Auf deren Ausarbeitung und Implementation wurde zugunsten der Automatisierung des Gesamtablaufs der Korpuserstellung verzichtet. Die einzelnen Verfahren mit möglichen Ansätzen zur Verbesserung und Erweiterung oder alternative Realisierungen wurden bereits in den einzelnen Kapiteln ausführlich diskutiert, so daß sich die Diskussion hier auf einen zusammenfassenden Überblick beschränken kann.

Die Wort- und Silbensegmentierung erreicht eine recht überzeugende Qualität, die Segmentierung in Phoneme ist aufgrund des (bewußten) Ausschlußes von Aussprachevarianten und der Verwendung eines Frame-Abstands von 10 ms in der Genauigkeit eingeschränkt. Die Verwendung von Strahlsuche in der Decodierung kann die Segmentierung noch wesentlich beschleunigen.

Bei der phonetischen Parametrisierung der Grundfrequenz ist der Hauptnachteil der aktuellen Implementierung die aufzuwendende Analysezeit. Auch die Qualität läßt sich in manchen Punkten noch steigern, ein möglicher Ansatz ist die Vorbehandlung der Ausgabe des Pitch trackers durch Unterdrückung von Ausreißern und unsicheren Messungen am Anfang und Ende stimmhafter Abschnitte. Auch sind alternative Parametrisierungsfunktionen, zum Beispiel logistische Funktionen oder die auf den Ergebnissen dieser Arbeit aufbauende polynomielle Funktion, die in [Möh98] beschrieben ist, in Erwägung zu ziehen und auf ihre Eignung für die Erkennung experimentell zu überprüfen. Eventuell lassen sich die vorherzusagenden Klassen im Merkmalsraum deutlicher voneinander trennen und dadurch die Erkennung verbessern, indem das verwendete Register geschätzt und sein Einfluß aus der Parametrisierung herausgerechnet wird.

Bei der Prosodieerkennung haben sich die aus dem Mikrophonsignal abgeleiteten Intensitätsmerkmale als wenig nützlich erwiesen. Dies läßt sich als Indiz dafür werten, daß die Intensität für die Prominenz auf Satzebene im Deutschen weniger wichtig ist als beispielsweise die Silbendauer, unabhängig davon, ob man die Intensität über das gesamte Frequenzspektrum mißt oder nur die hohen Frequenzen berücksichtigt. Eine Wertung der Wichtigkeit der Intensität als prosodisches Korrelat läßt sich aus den vorgelegten Lernläufen jedoch nicht ableiten, da nicht aus-

zuschließen ist, daß die vom Radio aufgenommenen Sprachsignale dynamikverzerrenden Normalisierungen unterliegen. Unabhängig vom theoretischen Status der Intensität als Korrelat der Betonung ist diese Beobachtung jedoch gleichzeitig ein Argument dafür, die Intensitätsmerkmale nicht für die Prosodieerkennung einzusetzen, da prinzipiell mit solchen Beeinflussungen für zu erkennende Sprachsignale zu rechnen ist, wenn sich ein Teil des Signalweges der direkten Kontrolle entzieht.

Die automatische Vorhersage einer phonologischen Intonationsbeschreibung ist zum derzeitigen Entwicklungsstand noch wesentlich unzuverlässiger als die von menschlichen Experten durchgeführten Analysen. Dies bedeutet jedoch nicht, daß man auf den Einsatz eines automatischen Prosodieerkennungsverfahrens verzichten sollte. Da die Erkennung vollautomatisch abläuft (gegebenenfalls abgesehen von der Transliteration), lassen sich wesentlich mehr Daten analysieren als manuell durch menschliche Experten. Wenn das automatische Verfahren für eine bestimmte Entscheidung eine Genauigkeit von ungefähr 75% erreicht, während menschliche Experten die gleiche Entscheidung mit etwa 90-prozentiger Genauigkeit treffen können, so sollte dennoch ein großer Teil der aus dem manuell etikettierten Datenmaterial ableitbaren Folgerungen auch aus den automatisch analysierten Daten gezogen werden können, da wesentlich umfangreicheres Datenmaterial (beispielsweise die hundertfache Menge) bereitsteht. Bei der Auswertung und Interpretation der Ergebnisse muß man jedoch auf alle Fälle diese geringere Genauigkeit berücksichtigen, wie auch weitere Besonderheiten, etwa daß nicht alle Intonationssymbole gleich zuverlässig erkannt werden. Diese Ergebnisse sollten daher in kontrollierten Laborüberprüfungen besonders kritisch verifiziert werden. Wenn ein Ergebnis nur auf einem kleinen Teil der Daten beruht, ist auch eine nachträgliche manuelle Korrektur der Intonationsbeschreibung dieser Daten durchführbar. Da neben der phonologischen Intonationsbeschreibung auch die phonetischen Parameter zur Vefügung stehen, können auch quantitative Untersuchungen auf die automatische Analyse aufsetzen, beispielsweise um eine Modellierung für die Register oder adäquate Parameter für die Synthese zu finden.

Um eine einfache Erweiterung und Anpassung des Systems zu ermöglichen, sind die Strukturen des Systems darauf ausgelegt, weitere Merkmale aufzunehmen. Die Verwaltung der Daten erfolgt stets in einfachen, flexiblen Strukturen in Textdateien, die den leistungsfähigen Unix-Werkzeugen zugänglich sind. Alle vom System erstellten Dateien sind damit offen für komplexe Suchanfragen, für automatische statistische Analysen und für die manuelle, auch auszugsweise Inspektion der Signale durch Experten. Beispielsweise lassen sich aus den aufbereiteten Korpora alle Nomina mit einem vorausgehenden Adjektiv und einer L*H-Markierung herausfiltern und die Höhe des Akzents statistisch analysieren. Oder es lassen sich alle nebenordnenden Konjunktionen mit vorausgehender interme-

diärer Phrasengrenze nach ihrer relativen Länge sortieren und sukzessive einer manuellen Inspektion mit einem Signalvisualisierungsprogramm zuführen.

Mit den in dieser Arbeit entwickelten Verfahren wurden umfangreiche Sprachdaten aus unterschiedlichen Quellen und von verschiedenen Sprechern analysiert. Manche der durchgeführten Datensammlungen sind erst durch den Einsatz der entwickelten Verfahren, besonders des flexiblen Segmentierungssystems Alphons, sinnvoll durchführbar geworden, etwa die fast vollautomatische Sammlung der astronomischen Radio-Sendereihe „Sternzeit“ und die Analyse der 4 Stunden Sprache umfassenden Literaturlesung von „Die Leiden des jungen Werther“ [Goe95, Rap96]. Eine Zusammenstellung der mit den hier entwickelten Verfahren am IMS bisher analysierten Daten ist im Anhang aufgeführt. Derzeit analysiert das IMS mit dem implementierten System weitere Sprachdaten für die eigene Forschung, für den Lehrstuhl für Phonetik an der Universität Konstanz und für das Institut für deutsche Sprache in Mannheim.

Anhang A

Übersicht über aufbereitete Korpora

Im Verlauf der Dissertationsarbeit wurde mit den entwickelten Werkzeugen eine Reihe von Sprachdaten aufbereitet und für Untersuchungen verwendet. Die Sprachdaten entstammen verschiedenen Quellen und sind größtenteils einheitlich analysiert und für automatisierte Untersuchungen zugänglich. Insgesamt stehen etwa 12 Stunden gelesene oder rezitierte Sprache für Untersuchungen zur Verfügung. Davon sind etwa 1,5 Stunden verläßlich manuell prosodisch analysiert. Für etwa 5,5 Stunden stehen automatisch erstellte prosodische Analysen bereit. Alle Sprachdaten erlauben einen direkten Zugriff durch die automatische Wortsegmentierung. Bei Bedarf kann für alle Sprachdaten eine automatische phonemische Segmentierung sowie eine automatische prosodische Analyse erstellt werden.

A.1 Über das digitale Satelliten-Radio aufgenommene und aufbereitete Sprachdaten

Der zentrale Teil dieses Korpus besteht aus Nachrichtensendungen, die manuell transliteriert und manuell prosodisch etikettiert wurden. Dieser Teilkorpus besitzt die verläßlichste prosodische Etikettierung. Daneben gibt es automatisch erstellte Sammlungen wiederholt verlesener Nachrichtenmeldungen. Bei weiterem Bedarf läßt sich mit dem in Abschnitt 7.1 beschriebenen System leicht zusätzliches Nachrichtenmaterial akquirieren und vorselektieren.

A.1.1 Manuell aufbereitete Radionachrichten (zentrales Teilkorpus)

Aufnahme

- vom Deutschlandfunk gesendete Radionachrichten
- digital über das Satellitenradio übermittelt (32 kHz Samplingrate, 16 bit Samplingtiefe linear)
- Samplingrate konvertiert auf 16 kHz

Inhalt

- 4 aufeinanderfolgende Nachrichtensendungen vom 28.07.1995, 10:30 - 12:00 Uhr
  - 17 Meldungen, 3 Wetterberichte einer weiblichen Nachrichtensprecherin, insgesamt 617,272 sec
  - 11 Meldungen, 1 Wetterbericht eines männlichen Nachrichtensprechers, insgesamt 476,553 sec
- 14 aufeinanderfolgende Nachrichtensendungen vom 21.11.1995, 10:30 - 17:00 Uhr
  - 13 Meldungen, 3 Wetterberichte einer weiblichen Nachrichtensprecherin, insgesamt 541,659 sec
  - 94 Meldungen, 12 Wetterberichte eines männlichen Nachrichtensprechers, insgesamt 3551,8 sec

Aufbereitung

- Nachrichtensendungen wurden manuell in einzelne Meldungen segmentiert
- alle Meldungen wurden manuell orthographisch transliteriert
- automatische Wort-, Silben- und Phonemsegmentierung
- automatische POS-Etikettierung
- volle manuelle prosodische Etikettierung nach [May95]

- Trainingsdatenmenge für die automatische prosodische Etikettierung
- Gesamtzeit 1 h 26 min (1 h 07 min ein männlicher Sprecher, 19 min eine weibliche Ansagerin)

A.1.2 Automatisch aufbereitetes Nachrichtenkorpus

Aufnahme

- vom Deutschlandfunk gesendete Radionachrichten
- digital über das Satellitenradio übermittelt (32 kHz Samplingrate, 16 bit Samplingtiefe linear)
- Samplingrate konvertiert auf 16 kHz

Inhalt

- 17 aufeinanderfolgende, automatisch über Nacht aufgenommene Nachrichtensendungen vom 30.05. und 31.05.1996, 21:00 - 8:00 Uhr
- 102 Meldungen, 17 Wetterberichte (gemäß automatischer Segmentierung)
- 2 männliche, 1 weiblicher Sprecher
- Gesamtzeit ca. 1,5 h

Aufbereitung

- vollautomatische Segmentierung in Meldungen
- vollautomatisches Auffinden wiederholter Meldungen
- Transliteration nur der am häufigsten wiederkehrenden Meldungen
- (noch) keine weitere Analyse

A.1.3 Sternzeit

Vom Radio wird täglich automatisch die „Sternzeit“, eine Radio-Sendereihe über astronomische Themen, aufgenommen und vollautomatisch (teil-)analysiert. Zur Zeit steht nur die Wortsegmentierung als Zugriffsmöglichkeit zur Verfügung. Das Material ist relativ reich an Satzpartikeln, Gradadverbien, Quantoren, Entfernungs- und Datumsangaben. Die Verfügbarkeit von fremdsprachlichen Parallelkorpora (Englisch, Spanisch) eröffnet weitere Möglichkeiten für Forschungsaktivitäten.

- deutsche Fassung der Sendung „Stardate“ der Universität von Texas
- englische und spanische Textfassung erhältlich über das Internet, Sprache über amerikanische Radiostationen und über Audio-CD
- beinahe täglich aufgenommen seit 1. August 1996
- ausstrahlender Sender: Deutschlandfunk
- digital über das digitale Satellitenradio verbreitet
- Samplingrate konvertiert auf 16 kHz
- Orthographie akquiriert vom Webserver des Deutschlandfunk
- automatische Wortsegmentierung (mit Kanal/Sprecheradaption)
- Gesamtzeit bisher ca. 6 h, ständig wachsend

A.2 Aufbereitete Fremdaufnahmen

Die automatisch ablaufenden Teile der Korpusaufbereitung, also Segmentierung, Tagging und automatische prosodische Etikettierung, wurden auf eine Literaturlesung angewendet, die einer Multimedia-CD-ROM des Reclam-Verlags entnommen ist [Goe95]. Neben der analysierten CD-ROM sind zur Zeit 14 weitere Werke mit Sprache und Text erhältlich, die in gleicher Weise aufbereitet werden können. Weiterhin sind an anderen Instituten aufgenommene und verarbeitete Sprachdaten im Projekt manuell prosodisch annotiert und automatisch segmentiert worden.

A.2.1 Die Leiden des jungen Werther

Dieses Korpus umfaßt den Text des Romans „Die Leiden des jungen Werther“ von Johann Wolfgang Goethe, der vom Reclam-Verlag auf CD-ROM herausgegeben wird.

- CD-ROM enthält Text und komplette Darbietung durch einen professionellen Sprecher
- automatische Wort-, Silben- und Phonemsegmentierung (mit Sprecheradaption)
- automatische POS-Etikettierung
- automatische phonetische Intonationsbeschreibung
- automatische phonologische Intonationsbeschreibung
- Gesamtzeit 4 h acted speech (ein männlicher Sprecher)

A.2.2 Die Buttergeschichte

- manuelle phonologische Intonationsbeschreibung für „die Buttergeschichte“ des „Kiel Corpus of Read Speech Vol. I“

A.2.3 Saarbrückener Map Task

- manuelle phonologische Intonationsbeschreibung für Teile des Saarbrückener „Map Task“–Korpus

A.2.4 Verbmobil–Dialoge

- manuelle phonologische Intonationsbeschreibung für wenige Verbmobil–Dialoge

A.3 Aufbereitete Eigenaufnahmen

Für einen intensiv semantisch analysierten kurzen Auszug einer Kurzgeschichte wurden Aufnahmen erstellt und manuell prosodisch annotiert. Weiteres Material wurde für verschiedene Untersuchungen im Rahmen des Projekts aufgenommen und analysiert.

A.3.1 Das Dicke Kind

Dieses Korpus enthält den folgenden kurzen Auszug aus Marie Luise Kaschnitz' Kurzgeschichte „Das Dicke Kind“:

> *Es war Ende Januar, bald nach den Weihnachtsferien, als das dicke Kind zu mir kam. Ich hatte in diesem Winter angefangen, an die Kinder aus der Nachbarschaft Bücher auszuleihen, die sie an einem bestimmten Wochentag holen und zurückbringen sollten. Natürlich kannte ich die meisten dieser Kinder, aber es kamen auch manchmal Fremde, die nicht in unserer Straße wohnten. Und wenn auch die Mehrzahl von ihnen gerade nur so lange Zeit blieb, wie der Umtausch in Anspruch nahm, so gab es doch einige, die sich hinsetzten und gleich auf der Stelle zu lesen begannen. Dann saß ich an meinem Schreibtisch und arbeitete, und die Kinder saßen an dem kleinen Tisch bei der Bücherwand, und ihre Gegenwart war mir angenehm und störte mich nicht. Das dicke Kind kam an einem Freitag oder Samstag, jedenfalls nicht an dem zum Ausleihen bestimmten Tag. Ich hatte vor auszugehen...*

Der semantisch dichte Text enthält Quantoren, Ellipsen, anaphorische Pronomina etc. und besitzt eine interessante Diskursstruktur. Eine gründliche semantische Analyse ist verfügbar in [Roß94].

- Umfang: 6 Sätze, 144 Wörter
- Aufnahmen von 6 Sprechern (4 weiblich, 2 männlich)
- 48 kHz DAT Aufnahmen in schallabsorbierender Aufnahmekabine
- Samplingrate konvertiert auf 16 kHz

- automatische Wort-, Silben- und Phonemsegmentierung
- automatische phonetische Intonationsanalyse
- manuelle prosodische Analyse

A.3.2 Anaphernresolution

Hierbei handelt es sich um Daten, die für das Entscheidungsexperiment zur Anaphernresolution in [BDM96] verwendet wurden.

- 130 Versuchstoken bestehend aus je zwei Sätzen
- 7 Situationen, verschiedene Anapher-Referenzen
- verschiedene prosodische Realisierungen generiert aus je einer einzigen Äußerung durch PSOLA-Resynthese [Möh98]
- graphische Darstellung von 4 Situationen für jeden Token, die als nichtsprachliche Situationsbeschreibung der Anaphernreferenz den Versuchspersonen vorgelegt wurden

A.3.3 „Erst“-Daten

Diese Daten stammen aus einem Produktionsexperiment über das semantisch ambige deutsche Wort „erst“ [DKM+97].

- [Ebe96] beschreibt semantisch-pragmatische Analysen der drei Lesarten des deutschen Wortes „erst“
- Aufnahmen der Sätze aus [Ebe96], die jeweils eine der drei Lesarten erzwingen sollten
- 5 Sprecher (3 weiblich, 2 männlich)
- Versuchspersonen wurden angehalten, die Sätze mit der korrespondierenden semantischen Interpretation zu äußern.
- automatische Wort- Silben und Phonemsegmentierung
- manuelle prosodische Analyse

A.3.4 Glottale Parameter von Pitchakzenten

Dieses Material wurde für ein Experiment zur Untersuchung glottaler Parameter von Pitchakzenten zusammengestellt.

- Sprachsignal und Elektroglottogram
- 5 Sprecher (3 weiblich, 2 männlich)
- 48 kHz DAT Aufnahmen in schallabsorbierender Aufnahmekabine
- Samplingrate konvertiert auf 16 kHz

A.3.5 Zeitungsartikel

- gelesene Zeitungsartikel (2 Artikel, 846 Wörter)
- manuell getaggte Part-of-Speech-Etiketten
- eine weibliche Sprecherin
- 48 kHz DAT Aufnahmen in schallabsorbierender Aufnahmekabine
- Samplingrate konvertiert auf 16 kHz

Literaturverzeichnis

[BDM96] Annegret Bender, Grzegorz Dogil, and Jörg Mayer. Prosodic disambiguation of anaphoric pronouns in German discourses. In *Proceedings Discourse Anaphora and Resolution Colloquium (DAARC96)*, Lancaster University, Juli 1996.

[BK94] H.-P. Bernhard and G. Kubin. A fast mutual information calculation algorithm. In *Proceedings 7th European Signal Processing Conference, EUSIPCO*, Edinburgh, 1994.

[BP86] M. E. Beckman and J. B. Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:255–309, 1986.

[BPvR93] R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.

[Cam95] W. N. Campbell. From read speech to real speech. In *Proceedings ICPhS 95 Stockholm*, volume 2, pages 20–27, 1995.

[Cam96] Nick Campbell. Autolabelling Japanese ToBI. In *ICSLP96 Proceedings Fourth International Conference on Spoken Language Processing*, pages 2399–2402, October 1996.

[CDJ+] Katrin Claßen, Grzegorz Dogil, Michael Jessen, Krzysztof Marasek, and Wolfgang Wokurek. Stimmqualität und Wortbetonung im Deutschen. Erscheint in Linguistische Berichte.

[Cla96] Katrin Claßen. Phonetische Korrelate der Wortbetonung im Deutschen – Produktionsexperiment. Magisterarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1996.

[Dat] Reduct Systems Inc., P. O. Box 3570, Regina, Saskatchewan, Canada S4P 3L7. *DataLogic/R*.

[DB88] Grzegorz Dogil and G. Braun. *The PIVOT Model of Speech Parsing*. Verlag der Akademie, Wien, 1988.

[Del85] G. S. Dell. Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science*, 9:3–23, 1985.

[DKM+97] Grzegorz Dogil, Jonas Kuhn, Jörg Mayer, Gregor Möhler, and Stefan Rapp. Prosody and discourse structure: Issues and experiments. In Antonis Botinis, Georgios Kouroupetroglou, and George Carayannis, editors, *Intonation: Theory, Models and Applications*, pages 99–102. European Speech Communication Association, September 1997.

[DW87] J. E. Dennis, Jr. and D. J. Woods. New computing environments: Microcomputers in large-scale computing. pages 116–122. SIAM, 1987.

[Ebe96] Kurt Eberle. Disambiguation by information structure. In *COLING 96*, pages 334–339, 1996.

[Eng95] Fred Englert. *Automatische Segmentation von Sprachsignalen*. Hector, Frankfurt am Main, 1995.

[Fér93] Caroline Féry. *German Intonational Patterns*. Niemeyer, Tübingen, 1993.

[Gar76] Merrill F. Garrett. Syntactic processes in sentence production. In R. J. Wales and E. Walker, editors, *New Approaches to Language Mechanisms*. North-Holland, Amsterdam, 1976.

[Goe95] Johann Wolfgang Goethe. *Die Leiden des jungen Werther*. Philipp Reclam jun., Stuttgart and Silver Spring, Berlin, 1995. Klassiker auf CD-ROM.

[Gol90] John A. Goldsmith. *Autosegmental & Metrical Phonology*. Basil Blackwell, Oxford and Cambridge, 1990.

[GvH95] Rob Goedemans and Vincent J. van Heuven. Duration perception in subsyllabic constituents. In *Proceedings EUROSPEECH '95 Madrid*, pages 1315–1318, Madrid Spain, 1995.

[Hal92] T. A. Hall. *Syllable structure and syllable-related processes in German*. Niemeyer, Tübingen, 1992.

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

[HTK] Entropic Research Laboratory, Inc, 600 Pennsylvania Avenue, Washington DC 20003. *HTK – Hidden Markov Model Toolkit*.

[ICA97] IEEE. *ICASSP 97*, volume II, 10662 Los Vaqueros Circle, Los Alamitos CA 90720-1264, April 1997. IEEE Computer Society Press. S. 711 – 726.

[Jes94] Michael Jessen. A survey of German word stress. Manuskript, Universität Stuttgart, 1994.

[JMSC95] Michael Jessen, Krzysztof Marasek, Katrin Schneider, and Kathrin Claßen. Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German. In *Proceedings ICPhS 95*, volume 4, pages 428–431, Stockholm, August 1995.

[Kam97] Christiane Kamm. Implementierung eines wortbasierten, konkatenativen Sprachsynthesesystems. Diplomarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1997.

[KMS96] Tetsuo Kosaka, Shoichi Matsunega, and Shigeki Sagayama. Speaker-independent speech recognition based on tree-structured speaker clustering. *Computer Speech and Language*, 10(1):55–74, January 1996.

[Koh94] Lexica of the Kiel PHONDAT corpus, read speech, volume I. Arbeitsberichte (AIPUK) 27/28, Institut für Phonetik und digitale Sprachverarbeitung Universität Kiel, 1994.

[Koh95] K. J. Kohler, editor. *The Kiel Corpus of Read Speech (CD-ROM)*. Institut für Phonetik und digitale Sprachverarbeitung, Christian–Albrechts–Universität zu Kiel, D–24098 Kiel, Germany, e–mail: `gph33@rz.uni-kiel.de`, 1995.

[Kva93] Knut Kvale. *Segmentation and labelling of speech*. Dissertation, The department of telecommunications of the Norwegian institute of technology, 1993.

[Lad96] D. R. Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 1996.

[Leh95] Michael Lehning. Bandbegrenzte Funktionen als Hilfsmittel zur Fehlerkorrektur von Grundfrequenzverläufen. Verbmobil Report 72, TU Braunschweig, Juli 1995.

[Lev89] W. J. M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, Cambridge Massachusetts, 1989.

[LS95] Pat Langley and Herbert A. Simon. Applications of machine learning and rule induction. *Communications of the ACM*, 38:55–64, 1995.

[LWZ+96] Nada Lavrač, Irene Weber, Darko Zupanič, Dimitar Kazakov, Olga Štěpánková, and Sašo Džeroski. ILPNET repositories on WWW: Inductive Logic Programming systems, datasets and bibliography. *AI Communications*, 9(4):157–206, 1996.

[Man90] Max Mangold. *Duden Aussprachewörterbuch: Wörterbuch der deutschen Standardaussprache*. Bibliographisches Institut &F. A. Brockhaus AG, 1990. 3. Auflage.

[Mar81] S. Marcus. Acoustic determinants of perceptual center (p-centre) location. *Perception and Psycholinguistics*, 30:122–140, 1981.

[May95] Jörg Mayer. Transcribing German intonation — the Stuttgart system. Manuscript, Universität Stuttgart, 1995.

[May97] Jörg Mayer. *Intonation und Bedeutung. Aspekte der Prosodie-Semantik-Schittstelle im Deutschen*. Dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1997.

[MBGH94] Nelson Morgan, Hervé Bourlard, Steve Greenberg, and Hynek Hermansky. Stochastic perceptual auditory-event-based models for speech recognition. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 1943–1946, Yokohama, Japan, September 1994.

[Möh98] Gregor Möhler. *Eine theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Dissertation, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1998.

[MWB95] Nelson Morgan, Su-Lin Wu, and Hervé Bourlard. Digit recognition with stochastic perceptual speech models. In *Proceedings EUROSPEECH'95*, volume 1, pages 771–774, Madrid, Spain, September 1995.

[NM] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313.

[Nö91] Elmar Nöth. *Prosodische Information in der automatischen Spracherkennung*. Niemeyer, Tübingen, 1991.

[Pap84] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill, New York, 1984.

[Paw91] Zdisław Pawlak. *Rough sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic, Dordrecht, The Netherlands, 1991.

[PB88] J. Pierrehumbert and M. E. Beckman. *Japanese Tone Structure*. MIT Press, Cambridge, 1988.

[Pie80] Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980.

[PM91] B. Pompino-Marschall. The syllable as a prosodic unit and the so-called p-center effect. Technical Report 29, IPSK München, 1991.

[Qui90] J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5:239–266, 1990.

[Qui92] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kauffmann, San Mateo, CA, 1992.

[Qui96] J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[Rap95] Stefan Rapp. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An aligner for German. In *Workshop "Integration of Language and Speech in Academia and Industry"*, Moscow, November 1995. ELSNET goes east and IMACS. http://www.ims.uni-stuttgart.de/˜rapp/aligner.ps.gz.

[Rap96] Stefan Rapp. Goethe for prosody. In *ICSLP96 Proceedings Fourth International Conference on Spoken Language Processing*, pages 1636–1639, October 1996.

[Roß94] Antje Roßdeutscher. Fat child meets DRT. A semantic representation for the opening lines of Kaschnitz' "Das dicke Kind". *Theoretical Linguistics*, 20:237–305, 1994.

[Ros95] Kenneth N. Ross. *Modeling of Intonation for Speech Synthesis*. PhD thesis, Boston University College of Engineering, 1995.

[Roy83] H. Royé. *Segmentierung und Hervorhebung in gesprochener deutscher Standardaussprache*. PHONAI Band 27. Niemeyer, Tübingen, 1983.

[SBP+92] K. E. A. Silverman, M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling english prosody. In *2nd International Conference on Spoken Language Processing*, pages 867–870, Banff Canada, 1992.

[Sch95] H. Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of EACL SIGDAT-Workshop*, Dublin, Ireland, 1995. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.ps.gz.

[SH94] K. N. Stevens and H. M. Hanson. *Classification of Glottal Vibration from Acoustic Meassurements*. 1994.

[Slu95] Agaath M. C. Sluijter. *Phonetic Correlates of Stress and Accent*. PhD thesis, Leiden University, 1995.

[Tab95] Edgar Tabar. Graphem–Phonem Konvertierung und Silbentrennung für die deutsche Sprache. Studienarbeit Nr. 1474, Universität Stuttgart, Fakultät Informatik und Institut für Maschinelle Sprachverarbeitung, 1995.

[Ven90] T. Vennemann. Syllable structure and syllable cut prosodies in modern standard German. In P. M. Bertinetto, M. Kentstowicz, and M. Loporcaro, editors, *Certamen Phonologicum II. Papers from the 1990 Cortona Phonology Meeting*, pages 211–243, Turin, 1990. Rosenberg & Sellier.

[vSB97] Jan P. H. van Santen and Adam L. Buchsbaum. Methods for optimal text selection. In *Proceedings EUROSPEECH '97 Rhodos*, pages 553–557, Rhodos Greece, 1997.

[vSSM+97] Jan van Santen, Chilin Shih, Bernd Möbius, Evelyne Tzoukermann, and Michael Tanenblatt. Muliti-lingual duration modeling. In *Proceedings EUROSPEECH '97 Rhodos*, pages 2651–2654, Rhodos Greece, 1997.

[Wes95] Maria-Barbara Wesenick. Regelsystem zur Generierung von Aussprachevarianten. Verbmobil Memo 96, Ludwig-Maximilians-Universität München, November 1995.