Linguistische Arbeiten

LA

Elmar Nöth

Prosodische Information in der automatischen Spracherkennung

Berechnung und Anwendung

Niemeyer

Linguistische
Arbeiten 259

Herausgegeben von Hans Altmann, Peter Blumenthal, Herbert E. Brekle,
Hans Jürgen Heringer, Heinz Vater und Richard Wiese

Elmar Nöth

Prosodische Information in der automatischen Spracherkennung

Berechnung und Anwendung

Max Niemeyer Verlag
Tübingen 1991

Meiner Frau Jina und meinen Kindern Maria, Eric und Ryan

CIP-Titelaufnahme der Deutschen Bibliothek

Nöth, Elmar : Prosodische Information in der automatischen Spracherkennung : Berechnung und Anwendung / Elmar Nöth. – Tübingen : Niemeyer, 1991
(Linguistische Arbeiten ; 259)
NE: GT

ISBN 3-484-30259-3 ISSN 0344-6727

Druck: Weihert-Druck GmbH, Darmstadt
Einband: Heinr. Koch, Tübingen

Inhaltsverzeichnis

1 Einleitung ... 1
1.1 Motivation ... 1
1.2 Die potentielle Rolle der Prosodie in einem automatischen Spracherkennungssystem ... 5
1.3 Das System EVAR ... 11
1.4 Beitrag dieser Arbeit zur Prosodie-Forschung ... 16

2 Definitionen, Begriffe und sprachwissenschaftliche Grundlagen ... 21
2.1 Prosodie ... 22
2.2 Intonation ... 23
2.3 Akzent ... 26
2.4 Fokus ... 29
2.5 Satzmodus ... 31
2.6 Gliederung von Äußerungen ... 32
2.7 Akzentuierungsmittel ... 33
2.7.1 Intrinsische Eigenschaften und Koartikulation ... 35
2.7.2 Das Akzentuierungsmittel Tonhöhe ... 38
2.7.3 Das Akzentuierungsmittel zeitliche Strukturierung ... 41
2.7.4 Das Akzentuierungsmittel Lautheit ... 43
2.7.5 Das Akzentuierungsmittel Klangfarbe ... 45
2.8 Intonatorische Markierung des Satzmodus ... 46
2.9 Intonatorische Gliederungsmittel ... 48
2.10 Intonation als Komplexphänomen ... 52

3 Stichproben ... 55
3.1 Die EVAR- und die Pragmatik-Stichprobe ... 55
3.2 Die Dialog-Stichprobe ... 57
3.2.1 Erstellung der Dialog-Stichprobe ... 57
3.2.2 Das erweiterte Erlanger Transkriptionssystem ... 58
3.2.3 Erstellung einer Betonungsbeschreibung ... 59
3.3 Das Royé-Korpus ... 64
3.4 Die Modus-Fokus-Korpora ... 65
3.4.1 Das Fokus-Korpus ... 66
3.4.2 Das Leo-Korpus ... 67

4 Akzentuierungsmittel und Merkmalextraktion 69
4.1 Silbenkerndetektion 69
4.1.1 Berechnung der Energie in verschiedenen Frequenzbereichen 70
4.1.2 Lokalisierung von Silbenkernen aufgrund der spektralen Energie 71
4.1.3 Korrektur der Silbenkerngrenzen durch Vergleich mit dem Analyseergebnis des Akustik-Phonetik-Moduls 79
4.2 Tonhöhenmerkmale 85
4.2.1 Ein Modell der Spracherzeugung 85
4.2.2 Grundperiode und Grundfrequenz 88
4.2.3 Berechnung von Grundfrequenzschätzwerten im Zeit- und Frequenzbereich .. 94
4.2.3.1 Bestimmung der stimmhaften Bereiche eines Zeitsignals 95
4.2.3.2 Tiefpaßfilterung 96
4.2.3.3 Center Clipping und dreistufige Quantisierung 97
4.2.3.4 Fensterfunktionen 100
4.2.3.5 Das AMDF-Verfahren 101
4.2.3.6 Das Seneff-Verfahren 102
4.2.4 Berechnung der Grundfrequenzkontur mit der Dynamischen Programmierung . 104
4.2.5 Normierungsoperationen und Extraktion der Tonhöhenmerkmale 110
4.3 Dauermerkmale 116
4.4 Lautheitsmerkmale 117

5 Erstellung einer Betonungsbeschreibung 119
5.1 Bewertung der Merkmale 121
5.2 Bewertung der prosodischen Eigenschaften 121
5.3 Bewertung der Gesamtbetonung 123

6 Experimentelle Untersuchungen zum Einsatz der prosodischen Information in einem sprachverstehenden System 125
6.1 Fehleranalyse für die Grundfrequenzbestimmung 126
6.1.1 SH/SL-Fehler 127
6.1.2 Feinfehler 127
6.1.3 Grobfehler 129

6.2 Prosodische Satzmodus-Bestimmung ... 133
6.2.1 Ein einfaches Modell der intonatorischen Markierung des Satzmodus ... 133
6.2.2 Quantitative Gültigkeit des Modells für die Modus-Fokus-Korpora ... 134
6.2.3 Automatische Merkmalextraktion - Reproduzierbarkeit der Ergebnisse und weitere Merkmale ... 137
6.2.4 Fälle, in denen das einfache Modell nicht zutrifft ... 137
6.3 Datengetriebene Silbenkernbestimmung und Betonungszuweisung ... 140
6.3.1 Ergebnisse zur Silbenkerndetektion ... 140
6.3.2 Ergebnisse zur Betonungsbeschreibung ... 142
6.4 Prosodische Verifikation der Silbenstruktur und des lexikalischen Wortakzents ... 145
6.4.1 Vorbemerkungen zur Leistungsbeurteilung bei der Worthypothesengenerierung und zur Verifikation ... 145
6.4.2 Das DEL-Filter ... 147
6.5 Einschränkung des Lexikons an betonten Stellen ... 151
6.6 Verbesserung der Worterkennung durch die neue Lautsegmentierung - Ein Zwischenbericht ... 157

7 Grundlegende Voruntersuchungen zur prosodischen Verifikation von Satzhypothesen - Ein Ausblick ... 161
7.1 Motivation ... 161
7.2 Vorgehensweise und Untersuchungsmaterial ... 163
7.3 Merkmalberechnung und Bewertung ... 164
7.4 Bestimmung der Prototypen ... 167
7.5 Identifikation der Kern- und Rand-Prototypen ... 169
7.6 Darstellung des erstellten Intonationsmodells in einem maschinellen Wissensrepräsentationsschema ... 171

8 Zusammenfassung ... 175

9 Literaturverzeichnis ... 181

Anhang ... 195

1 Einleitung

1.1 Motivation

Ein sehr bekannter amerikanischer Phonologe soll zu Beginn der sechziger Jahre gegenüber einem deutschen Kollegen bei der Führung durch sein Institut sinngemäß gesagt haben:

> "... und in diesem Zimmer haben wir zwei Jungs - die arbeiten an der automatischen Spracherkennung. Ich nehme an, daß wir das Problem in drei Jahren gelöst haben. ..."

1969 schrieb J. Pierce von den Bell Laboratories zur Frage *"Warum überhaupt Forschungen zur automatischen Spracherkennung":*

> "THE PURPOSE OF THIS LETTER IS TO EXAMINE BOTH MOTIVATIONS and progress in the area of speech recognition (that is, word recognition). ...
>
> It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not a sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamor.
>
> ... When we look further for reasons, we encounter that of communication with computers.
>
> In this general form, the reason is as specious as insisting that an automobile should respond to *gee, haw, giddap, whoa*, and slaps or tugs of the reins. We communicate with children by words, coos, embraces, and slaps. We communicate with people by these means and by nods, winks, and smiles. It is not clear that we should resort to the same means with computers. In fact, we do very well with keyboards, cards, tapes, and cathode-ray tubes. ...
>
> These considerations lead us to believe that a general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of a native speaker of English."
>
> [PIERCE 69, S.1049-1050]

Diese beiden Zitate repräsentieren zwei extreme Standpunkte in einer Diskussion, die - wie das zweite Zitat andeutet - zeitweise außergewöhnlich *emotional* geführt wurde und wird. *Extrem* ist hier im Sinne von *sehr gegensätzlich*, nicht im Sinne von *sehr selten* zu verstehen. Ein Grund dafür, daß die Diskussion um den Sinn und die Machbarkeit von automatischer Spracherkennung (ASE)[1] so sehr emotionsgeladen war, ist sicherlich in der euphorischen Anfangsbegeisterung einiger Forscher zu sehen. Sie gaben ihren Projekten sehr anspruchsvolle, falsche Erwartungen weckende Namen und stellten sehr gewagte Prognosen über zukünftige Erfolge auf. Der Hauptgrund für die von Emotionen geprägte Atmosphäre in dieser Diskussion ist aber sicher darin zu suchen, daß alle Versuche zur ASE eine neue Einschätzung des Systems "Sprache" schlechthin implizieren. Da ja Sprache eine über Jahrtausende erlernte Fähigkeit des Menschen zur Kommunikation darstellt,

---

[1] Im folgenden wird ASE als Oberbegriff für die automatische Auswertung von Sprachsignalen verwendet, vgl. Kap.1.2.

die ihn von allen anderen (bekannten) Lebensformen abhebt, ist eine versuchte "Degradierung" dieser wichtigen und komplexen Fähigkeit zu einer Handvoll *if ... then ... else ...* - *Regeln* sicherlich ein Schritt, der das Bild des Menschen vom Menschen verändert ([TURKLE 84]).

Daher spielen sich ähnlich emotionsgeladene Diskussionen nicht nur in vielen Bereichen der *künstlichen Intelligenz* und der *Mustererkennung/Musteranalyse* ab [DREYFUS 87], zu deren typischen Anwendungsgebieten die ASE zählt, sondern in allen Forschungsgebieten, die vergleichbar grundlegende Themen zum Forschungsgegenstand haben (z.B. die Gen-Manipulation).

Heute - rund zwanzig Jahre nach Fixierung der oben zitierten Standpunkte - kann man sagen, daß die Realität wie so oft den goldenen Mittelweg genommen hat. Auf der einen Seite benötigt man mehr als zwei *"boys"* (s.o.), um die Spracherkennung in den Griff zu bekommen. Auf der anderen Seite ist man der oben erwähnten *universellen phonetischen Schreibmaschine* (*"general phonetic typewriter"*) ein gutes Stück näher gekommen, genauso wie man auf allen Gebieten der ASE beachtliche Fortschritte erzielt hat: Die Frage, ob es sich bei dem IBM-System [JELINEK 85] oder dem Dragon-System [BAKER 87] um den ersten Prototypen einer *universellen phonetischen Schreibmaschine* handelt, hängt nicht nur vom Problem der Erstellung eines *Sprachmodells* für einen neuen Problemkreis ab, sondern auch von der Definition von *universell.* Ist eine *universelle phonetische Schreibmaschine* erst dann realisiert, wenn

- jeder beliebige Text
- über jedes beliebige Thema
- ohne Einschränkung des Vokabulars
- für jeden Sprecher
- fehlerfrei

wiedergegeben wird (dies wäre im übrigen auch für den *Menschen* eine unlösbare Aufgabe), oder ist sie schon realisiert, wenn

- nach einer zumutbaren Trainingsphase
- unter zumutbarer Veränderung der Sprechweise
- für mehrere, nicht zu sehr eingeschränkte Themenbereiche
- unter Benutzung eines nicht zu sehr eingeschränkten, aber (mit zumutbarem Aufwand) erweiterbaren Wortschatzes
- 85% (90%, 95%) der Wörter fehlerfrei

wiedergegeben werden?

Ob eine solche Schreibmaschine in absehbarer Zeit kommerziell verfügbar ist und akzeptiert wird, hängt von den Kosten, den geweckten Erwartungen, dem individuellen Verständnis der Begriffe *zumutbar* und *nicht zu sehr eingeschränkt* und dem Komfort ab, mit dem die vom System abgelieferte "Rohfassung" des zu erstellenden Textes weiterverarbeitet werden kann.

Die Nichtbeachtung der *sozialen Innovation* ist nach [WOHLAND 89] häufig der Grund für das Scheitern von Projekten zur Einführung neuer Techniken (in seinem Fall speziell von CIM-Projekten). Auch im Falle einer phonetischen Schreibmaschine, etwa zum Diktieren von Patientenbefunden, muß der Systementwickler die Arbeitsweise und die Arbeitsumgebung des späteren Benutzers *Arzt* genau untersuchen, damit die beiden notwendigen Stufen

"Abspaltung formaler Anteile der individuellen und der sozialen, geistigen Arbeit und Übertragung auf eine Maschine"

"Entwicklung einer neuen Arbeitsorganisation für die verbleibende menschliche Arbeit" [WOHLAND 89, S.17]

erfolgreich durchgeführt werden können.

Die Ansicht von Pierce (s.o.), daß wir mit *"Tastaturen, Lochkarten, Magnetbändern und Kathodenstrahlröhren"* gut zurechtkommen, erzeugt eher Schmunzeln in einer Zeit, in der kaum noch Lochkarten benutzt werden. Sie verdeutlicht, daß man die Form der **M**ensch-**M**aschine-**I**nteraktion (MMI) nicht als etwas Statisches betrachten sollte.

Das Interesse an der ASE ist vor allem auch deshalb ungebrochen, weil mit zunehmender Leistungsfähigkeit der Hardware immer mehr Anwendungen möglich erscheinen. Das Spektrum der (sich aus den Anwendungen ergebenden) Systemanforderungen reicht von sprecherabhängiger Einzelworterkennung mit kleinem Wortschatz und guten Sprachaufnahmebedingungen (z.B. für Gerätebedienung) bis zum sprecherunabhängigen Verstehen kontinuierlicher Sprache bei großem Wortschatz (z.B. für das Durchführen informationsabfragender Dialoge).

Man kann davon ausgehen, daß ein Spracherkennungssystem umso eher vom Benutzer akzeptiert und damit anderen Interaktionsmitteln wie Tastatur vorgezogen wird, je natürlicher die Interaktion abläuft. So wird eine phonetische Schreibmaschine, bei der die Spracheingabe wie bei einem Diktaphon ablaufen kann, trotz höherer Fehlerrate eher akzeptiert werden als ein System, das monotones Sprechen und Pausen zwischen den Wörtern erfordert.

Selbst wenn also bei vielen Forschungsprojekten zum Zwecke der Erhöhung der Erkennungsraten die zulässige Sprechweise stark eingeschränkt wird, muß eine langfristige Forschungsplanung auf dem Gebiet der ASE auch Forschungen zur Überwindung solcher Beschränkungen vorsehen. Der folgende Ausschnitt aus einem über Telefon geführten Auskunftsdialog (siehe Kap.3.2) enthält mehrere Beispiele für Sprechweisen, die in der Mensch-Mensch-Kommunikation durchaus gebräuchlich sind und die in einem ASE-System explizit verboten sein könnten:

Ja, hier ist Obermayer, Grüß Gott, ich hätt' 'ne Frage - ich möcht' morgen von Nürnberg nach Ulm fahren und möcht' ungefähr um vier Uhr in Ulm sein. Wann muß ich da in Nürnberg wegfahren?

- Die in Telefon-Dialogen (es fehlt der non-verbale Kommunikationskanal) übliche Nennung des Namens zum Zwecke der Dialogeinleitung muß in einem ASE-System besonders behandelt werden, da Eigennamen normalerweise nicht im Lexikon des Systems enthalten sind.
- Dialektale und überregionale Verschleifungsformen wie die Elision unbetonter Silben an "unwichtigen" Stellen der Äußerung (*"hätt' 'ne* statt *"hätte eine"*) sind sehr gebräuchlich und müssen modelliert werden.
- Das Wort *"Ulm"* wurde in diesem Beispiel beim ersten Mal *sehr stark betont* (der Zielort ist für den Auskunftsbeamten besonders wichtig) und beim zweiten Mal *unbetont* ausgesprochen

(Ulm ist bekannt, die Ankunftszeit ist wichtig und wird daher betont). Man erhält also zwei sehr verschiedene Repräsentanten für *ein* Wort von *einem* Sprecher in *einer* Äußerung.

Es bietet sich an, solche Effekte durch Instruktion der Sprecher zu verhindern. Man kann z.B. den Sprecher anweisen, jedes Wort gleich deutlich auszusprechen, also in einem für den Menschen unüblichen Stil zu sprechen. Auf der anderen Seite ist es intuitiv klar, daß die Betonung (Hervorhebung) bestimmter Stellen einer Äußerung gewissen Konventionen unterliegt und vom Sprecher bewußt eingesetzt wird. Zwar würde die Anfrage u.U. auch verstanden werden, falls der Sprecher als einzige Wörter die drei vorkommenden Wörter *"ich"* hervorheben würde, aber der Auskunftsbeamte würde die Sprechweise vermutlich als "komisch" oder "unnatürlich" empfinden. Der Kunde weiß ja, daß der Beamte für eine richtige Auskunft auf die Anfrage vor allem die Wörter *"morgen", "von Nürnberg", "Ulm"* und *"vier Uhr"* verstehen muß und wird diese Stellen normalerweise besonders deutlich aussprechen. Dagegen ist es unwichtig, wer fährt, so daß der Sprecher das Wort *"ich"* undeutlich aussprechen wird.

Das verschliffene (unbetonte) Sprechen der unwichtigen Stellen und das deutliche (betonte) Sprechen der wichtigen Stellen, sind Gegenstand *prosodischer Untersuchungen.* Das Beispiel verdeutlicht, daß das Wissen über betonte und unbetonte Stellen einer Äußerung für den Verstehensprozeß sehr wichtig ist. Das stark angewachsene Interesse an *prosodischer Information* als Wissensquelle für die ASE ist deshalb unter zwei Gesichtspunkten zu sehen:

- Prosodische Information wird vom Menschen sehr stark eingesetzt, sowohl in der Sprachproduktion als auch in der Sprachperzeption. Ein automatisches System sollte *alle* Wissensquellen benutzen, die auch beim menschlichen Perzeptionsprozeß eingesetzt werden. Insbesondere ist zu erwarten, daß ein enger Zusammenhang zwischen den *betonten* Stellen einer Äußerung und den für die richtige Interpretation einer Äußerung *wichtigen* Stellen besteht.
- Systeme, in denen prosodische Information als unerwünschte Variabilität des Sprachsignals betrachtet wird, und in denen der Benutzer folgerichtig zu einer monotonen, ungewohnten Sprechweise gezwungen wird, werden in einer echten Anwendung vermutlich nur begrenzt akzeptiert werden, inbesondere von ungeübten Benutzern.

Die vorliegende Arbeit befaßt sich mit dem Einsatz prosodischer Information im ASE-System *EVAR.* Bevor dieses System sowie die im Rahmen seiner Entwicklung durchgeführten Untersuchungen zur Prosodie vorgestellt werden, soll kurz darauf eingegangen werden, auf welchen Verarbeitungsebenen welche prosodische Information in automatischen Systemen eingesetzt werden kann bzw. eingesetzt wird. Weitere Literatur zum Einsatz prosodischer Information in der ASE findet sich in [LEA 75], [LEA 80b], [VAISSIÈRE 88], [MUDLER 86] und [KORI 87]. Eine Bibliographie über Veröffentlichungen zur ASE ist in [HOUSE 88] zusammengestellt. Übersichten über aktuelle Systeme und Projekte finden sich in [BAKER 87], [MARIANI 87] und [MARIANI 89].

1.2 Die potentielle Rolle der Prosodie in einem automatischen Spracherkennungssystem

In diesem Kapitel sollen Einsatzmöglichkeiten für prosodische Information in der ASE erörtert werden. Auf die sehr wichtigen Aspekte *Sozialverträglichkeit* und *Ethik* beim Einsatz von ASE-Systemen kann im Rahmen dieser Arbeit nicht eingegangen werden. Es soll aber ausdrücklich betont werden, daß die Nennung einer potentiellen ASE-Anwendung nicht gleichzeitig ihre Billigung beinhaltet. Hinweise auf Veröffentlichungen zu diesen Themen für den Bereich Informatik finden sich z.B. in dem Mitteilungsblatt des "Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung e.V. (FIFF)", der "FIFF Kommunikation" (siehe auch [WEIZENBAUM 76], [LÖWE 87], das Oktoberheft 1989 des Informatik-Spektrums, das dem Thema "Aufgabe und Verantwortung des Informatikers" gewidmet ist, sowie das Augustheft 1989 der "Communications of the ACM" über das Thema "Social Responsibility").

Zunächst einmal werden die Begriffe *Prosodie* und *automatische Spracherkennung* vorläufig eingeführt. Unter *Prosodie* sollen alle Gegenstände sprachwissenschaftlicher Untersuchungen zusammengefaßt werden, die sich mit lautübergreifenden sprachlichen Eigenschaften befassen. Die wohl wichtigsten Themen prosodischer Untersuchungen sind die Fragen, wie Teile einer Äußerung *betont* werden, wie der *Satzmodus* (Frage, Aussage, usw.) einer Äußerungen markiert wird und wie Äußerungen *gegliedert* werden. Diese grobe Definition von *Prosodie* sowie die intuitive Vorstellung des Lesers darüber, was eine *betonte Stelle* einer Äußerung ist, sollen an dieser Stelle genügen. Eine genaue Begriffsbestimmung findet sich in Kapitel 2.

Unter *automatischer Spracherkennung* sollen alle Verfahren zusammengefaßt werden, die sich mit der automatischen (d.h. in diesem Zusammenhang mit einem digitalen Rechner durchgeführten) Erkennung eines Aspektes der menschlichen Sprache befassen. Unter diese sehr informelle Definition von *automatischer Spracherkennung* würde ein Projekt zur *Sprecheridentifikation* fallen, während Untersuchungen zur *Natürlichkeit synthetisierter Sprache* hier nicht berücksichtigt werden sollen (in einem solchen Projekt steht ja die Erkennung durch den Menschen im Vordergrund).

Wenn man versucht, das Gebiet der ASE zu gliedern, kann dies unter der Fragestellung tun, *was* erkannt werden soll. Damit ergeben sich mindestens vier Teilgebiete:

- Es soll etwas über den *Sprecher* herausgefunden werden (z.B. *Sprecheridentifikation*).
- Es soll etwas über das Gesprochene *herausgefunden* werden (z.B. *Telefonüberwachungssysteme*).
- Das Gesprochene soll *erkannt* werden (z.B. *Einzelworterkennung zur Gerätesteuerung, phonetische Schreibmaschine*).
- Das Gesprochene soll *verstanden* werden (z.B. *Systeme zum Führen von Informations-Dialogen*).

Auf welche Weise *Prosodie* in diesen vier Teilgebieten eingesetzt werden kann, wird im folgenden diskutiert. Systeme, bei denen etwas über den *Sprecher* herausgefunden werden soll, sind häufig für Zugangskontrollen gedacht, etwa um über Telefon Bankauskünfte erhalten zu können.

Der Benutzer muß eine Testäußerung sprechen. Die Aufnahme bzw. die daraus abgeleiteten Merkmale werden mit Daten über den Sprecher verglichen. Zwar kann man zumindest beim Benutzer, der zum Zugang berechtigt ist, i.a. von einem kooperativen Sprecher ausgehen, aber es gibt trotzdem verschiedene Faktoren, die die *Stimmqualität* eines Sprechers verändern können. Hier könnte prosodische Information dazu eingesetzt werden, um systematische Veränderungen der Stimme zu modellieren, wie sie z.B. bei Heiserkeit auftreten. Untersuchungen zur *Stimmqualität* finden sich z.B. in [KLINGHOLZ 88].

Eine weitere Aufgabe neben der Sprecheridentifikation ergibt sich im Zusammenhang mit kriminalistischen Untersuchungen. Paralinguistische Aspekte der Prosodie, wie etwa das Vermitteln von Gemütszuständen, können in diesem Zusammenhang eine Rolle spielen (z.B. Einschätzung des emotionalen Zustandes eines Entführers). Denkbar ist der Einsatz von Untersuchungen zur Stimmqualität auch in gerichtsmedizinischen und medizinischen Anwendungen, etwa um festzustellen, ob der Sprecher unter Alkoholeinfluß steht. In [MUELLER 81] wurde von einem Ansatz berichtet, aufgrund von Beobachtungen über Veränderungen der Stimmqualität Rückschlüsse auf den Heilungsprozeß bei der Behandlung von Kehlkopfkrebs zu ziehen (siehe auch [HECKER 71]). Eine Übersicht über Systeme und Anwendungen der Sprechererkennung findet sich in [CORSI 82].

Bei den Anwendungsgebieten für Systeme, bei denen etwas über das *Gesprochene herausgefunden* werden soll, handelt es sich vorrangig um militärische und geheimdienstliche Fragen, insbesondere um die automatische Überwachung von Telefongesprächen. Aufgrund der Datenflut ergibt sich die Randbedingung, daß eine Entscheidung über das Aufzeichnen des Gespräches innerhalb weniger Sekunden geschehen muß. Mindestens zwei Vorgehensweisen sind denkbar:

- Aufzeichnen eines Gesprächs, sobald ein gewisses Wort gesprochen wird (*word spotting*). Der Einsatz prosodischer Information ist hier auf zwei Ebenen denkbar:
  * Falls ein Zeitbereich gefunden wird, in dem eines der interessierenden Wörter gesprochenen worden sein könnte, wird überprüft, ob das lexikalische Betonungsmuster des Wortes mit einem automatisch berechneten Betonungsmuster der Worthypothese übereinstimmt (im folgenden als *prosodische Wortakzent-Verifikation* bezeichnet).
  * Geht man davon aus, daß sich die interessierenden Wörter an betonten Stellen des Sprachsignals befinden, kann man mit prosodischer Information betonte Stellen suchen und dort das Signal genauer mit den interessierenden Wörtern vergleichen (im folgenden als *prosodische Satzakzent-Suche* bezeichnet).
- Aufzeichnen eines Gesprächs, sobald eine gewisse Sprache gesprochen wird. Da jede Sprache eine eigene Intonation hat, ist der Einsatz prosodischer Information auch hier vorstellbar (siehe hierzu die Ergebnisse von [BLESSER 69], zitiert in [WAIBEL 86], die belegen, daß sieben von elf untersuchten Sprachen aufgrund prosodischer Information von Hörern mit einer Erkennungsrate identifiziert werden konnten, die deutlich über der Zufallsrate lag).

Systeme zur *Erkennung des Gesprochenen* lassen sich in solche für Einzelworterkennung und solche für kontinuierliche Sprache unterteilen. Bei der Einzelworterkennung kann zwischen

Systemen mit *kleinem* und *großem Wortschatz* unterschieden werden. Bei Anwendungen für Systeme mit kleinem Wortschatz handelt es sich vor allem um Systeme zur Gerätesteuerung und Dateneingabe. Neben einem breiten Spektrum industrieller Anwendungen (z.B. Qualitätskontrolle [NELSON 85]; siehe [BAKER 87] und die Proceedings der jährlichen "Speech Tech"-Konferenz in New York für einen Überblick über den industriellen Einsatz von ASE-Systemen) werden bei Einzelworterkennern medizinische (z.B. Gerätesteuerung für Behinderte [AWAD 86]) und militärische Anwendungen (z.B. Gerätesteuerung zur Unterstützung von Kampfflugzeug-Piloten [NORTH 82]) untersucht. Da es sich um wenige zu unterscheidende Wörter handelt (<100), können die zu erkennenden Wörter als Ganzes betrachtet werden. Üblicherweise wird das zu erkennende Befehls- oder Datenwort mit einem oder mehreren in einer Trainingsphase gesprochenen Repräsentanten aller erkennbaren Wörter verglichen. Das Interesse an prosodischen Untersuchungen konzentriert sich in diesem Zusammenhang vor allem auf paralinguistische Aspekte. So können Erkenntnisse über die Eigenschaften der Sprache von Menschen mit speziellen Behinderungen dazu benutzt werden, die Qualität des Befehlswortes bei Behindertensystemen zu verbessern (siehe z.B. [ALIM 87], [TREHERN 87]). Auch bei dem Beispiel Kampfflugzeug-Pilot gibt es Ansätze zum Einsatz prosodischer Information: Wenn ein Wort z.B. unter mehrfacher Erdbeschleunigung gesprochen wird, verändert sich das Signal im Vergleich zum Referenzsignal sehr stark. Da es in dieser Anwendung verschiedene, bekannte und bis zu einem gewissen Grad modellierbare Streßfaktoren gibt, kann die Veränderung der Stimme systematisch erforscht werden (siehe z.B. [HANSEN 89]).

Bei Einzelworterkennung mit großem Wortschatz (>500) bietet sich vor allem die *prosodische Wortakzent-Verifikation* an. Diese Anwendung hat gegenüber kontinuierlicher Sprache (s.u.) den Vorteil, daß man einerseits in vielen Anwendungen davon ausgehen kann, daß jede Wortbetonung realisiert wird, daß andererseits aber die Markierung der Wortbetonung nicht überlagert wird von anderen prosodischen Ebenen, wie der prosodischen Markierung der syntaktischen Struktur einer Äußerung oder der Markierung des Satzmodus (s.u.). Durch die Bestimmung der Silbenanzahl und des Betonungsmusters kann eine schnelle Präselektion möglicher Kandidaten aus dem Lexikon erreicht werden. Untersuchungen zur *prosodischen Wortakzent-Suche* für Einzelworterkenner mit großem Wortschatz finden sich z.B. in [AULL 84], [WAIBEL 86]. In [HUTTENLOCHER 84] wird gezeigt, daß das Wissen über die *Anzahl der Silben* und die grobe Lautzerlegung der *betonten Silben* ein Lexikon mit 20000 Wörtern nur unwesentlich schlechter zerlegt als das Wissen über die *Anzahl der Silben* und die grobe Lautzerlegung *aller Silben*.

Da bei großem Wortschatz kein direkter Ganzwortvergleich mit dem Sprachsignal mehr durchgeführt werden kann, wird das Signal in Wortuntereinheiten zerlegt, und diese werden mit einer symbolischen Repräsentation jedes Lexikoneintrages verglichen. Prosodische Information kann bei der Segmentierung in Wortuntereinheiten sowie bei ihrer Identifizierung eine Rolle spielen (im folgenden als *prosodische Laut-Suche* bezeichnet). Ein Beispiel ist der Grundfrequenzverlauf zu Beginn des Vokals in Plosiv-Vokal-Folgen, der Rückschlüsse auf die Klassenzugehörigkeit des vorangehenden Plosivs zuläßt ([GARTENBERG 87], [MÖBIUS 87]). Eine typische Anwendung für

Einzelworterkennung mit großem Wortschatz ist ein System medizinischer Fachwörter [AKTAS 86], bei dem eine hohe Zahl phonetisch ähnlicher Wörter zu erwarten ist.

Für Systeme zur *Erkennung von Sätzen* mit Einschränkung der Sprechweise (Pausen zwischen den Wörtern, deutliche Aussprache *aller* Wörter) bieten sich dieselben Anwendungen prosodischer Information an wie für die Einzelworterkennung.

Bei Systemen zur *Erkennung kontinuierlicher Sprache* kommen weitere prosodische Ebenen hinzu. Dies macht die Prosodie als Informationsquelle zwar für den Einsatz bei kontinuierlicher Sprache attraktiver, aber gleichzeitig wird die Extraktion der Information wesentlich schwieriger und *fehlerhafter,* da die verschiedenen Rollen der Prosodie zum Teil mit denselben prosodischen Eigenschaften markiert werden und sich somit Überlagerungen ergeben (siehe z.B. Kap.2.10).

Viele paralinguistische Aspekte der Prosodie spielen hier eher eine untergeordnete Rolle, wenn man von dem Gebiet der schnellen Sprecheradaption absieht (z.B. sprechen manche Menschen ungewöhnlich behaucht, was eine zusätzliche Vorverarbeitung günstig erscheinen läßt).

Da in kontinuierlicher Sprache nicht jedes Wort betont, d.h. nicht jeder lexikalische Wortakzent realisiert wird, ist zu erwarten, daß die Wichtigkeit der *prosodischen Wortakzent-Suche* im Vergleich zur Einzelworterkennung zurückgeht: Das Wortbetonungsmuster läßt sich an betonten Stellen einer Äußerung verifizieren, nicht aber an unbetonten (an unbetonten Stellen ist der Wortakzent nicht realisiert, siehe Kap.2.3). Von Interesse ist hier mehr das Betonungsmuster innerhalb einer Phrase oder Äußerung (*prosodische Phrasenakzent-Verifikation* und *Satzakzent-Verifikation*) sowie das Finden der am stärksten betonten Stellen (*prosodische Satzakzent-Suche*). Das Wissen über betonte Stellen kann dazu benutzt werden, gewisse Wortklassen an diesen Stellen zu verbieten (*prosodische Lexikon-Beschränkung,* [NÖTH 89a]). Schließlich kann das Wissen über betonte Stellen dazu benutzt werden, der Analyse auf der *Wort- und Wortuntereinheiten-Ebene* mehr "zu glauben" ("Islands of Phonetic Reliability" [LEA 80b, S.170]), da z.B. in [LEA 73a] und [NÖTH 88b] gezeigt wurde, daß betonte Stellen besser erkannt werden. Diese Stellen können bei Vergleichsverfahren wie der dynamischen Zeitnormierung (**D**ynamic **T**ime **W**arping, DTW) zwischen Lexikonrepräsentanten und Hypothesen für Wortuntereinheiten zur Pfadbeschränkung benutzt werden. Bei inselgetriebener Syntaxanalyse können die betonten Stellen als Startpunkte benutzt werden. Die Tatsache, daß betonte Stellen besser erkannt werden, ist kein Widerspruch zu der oben genannten Ansicht, daß monoton gesprochene Äußerungen u.U. besser erkannt werden können. Denn zum Zwecke der Hervorhebung der *betonten Stellen* werden nicht nur diese Stellen besonders deutlich artikuliert, sondern es werden auch die *unbetonten Stellen* reduziert artikuliert und damit schlecht erkannt. Die *Gesamterkennungsrate* (z.B. gemessen in Anzahl erkannter Wörter) kann bei monotonem Sprechen steigen.

Je nach Anwendungsgebiet bzw. zulässiger Sprechweise werden syntaktische Strukturgrenzen mehr oder weniger stark mit prosodischen Mitteln markiert (*prosodische Struktur-Suche*). Sowohl in [LEA 75] als auch in [VAISSIÈRE 82] wird von guten Ergebnissen beim Finden syntaktischer Grenzen berichtet.

Zusätzlich zu diesen Einsatzgebieten kommt bei Systemen zum *Verstehen gesprochener Sprache* ein weiteres, wichtiges Einsatzgebiet prosodischer Information hinzu: Die *prosodische Satzmodus-*

Bestimmung kann die Bestimmung der Sprecherintention unterstützen, wobei vor allem die Unterscheidung zwischen *Fragen* einerseits und *Aussagen und Aufforderungen* andererseits von Interesse ist. Das folgende Beispiel aus dem Anwendungsgebiet des Systems EVAR (Zugauskunft) verdeutlicht, daß es Situationen geben kann, in denen ein richtiges Systemverhalten nur aufgrund *prosodischer Satzmodus-Bestimmung* erreicht werden kann:

		das System generiert eine Antwort auf eine Kundenanfrage
System:	Erlangen ab fünfzehn Uhr einundzwanzig Nürnberg an fünfzehn Uhr einundvierzig Nürnberg ab fünfzehn Uhr neunundvierzig ...	
Kunde:	neunundvierzig	*der Kunde signalisiert, daß er die Information soweit notiert hat, indem er das letzte Wort mit einer fallenden Grundfrequenz-Kontur wiederholt. Das System kann mit der Auskunft fortfahren.*
		vs.
Kunde:	neunundvierzig?	*der Kunde signalisiert, daß er sich nicht sicher ist, ob er die Information soweit richtig notiert hat, indem er das letzte Wort mit einer steigenden Grundfrequenz-Kontur wiederholt. Das System muß den letzten Teil der Information wiederholen und kann dann fortfahren.*

Bei *sprachverstehenden Systemen* kann im Vergleich zu *spracherkennenden Systemen* eine Verschiebung in der *Art* des Einsatzes notwendig sein. Eine *phonetische Schreibmaschine* muß in der Äußerung *"Hiermit möchte ich für den 17.1. einen Flug von Frankfurt nach New York buchen"* das Wort *"buchen"* nur *erkennen*, nicht *verstehen*. Ziel ist es vielmehr, möglichst *wenige Wortfehler* zu erhalten. Anders bei einem *automatischen Buchungssystem*, das die Bedeutung der Wörter *"frei sein"* und *"buchen"* verstehen muß, um auf die Anfragen *"Kann ich für den Flug von Frankfurt nach New York am 17.1. einen Platz buchen?"* und *"Ist für den Flug von Frankfurt nach New York am 17.1. noch ein Platz frei?"* adäquat reagieren zu können. Ziel ist es hier, in einer formalen Darstellung der *semantisch/pragmatischen Bedeutung* von Wörtern wie *buchen* notwendige *Ergänzungen (z.B. Zielort)* zu finden. Die Analyse einer Äußerung kann u.U. nach der Analyse der (mit prosodischen Mitteln) als *wichtig* markierten Stellen bereits beendet werden, obwohl noch nicht alle Teile der Äußerung analysiert sind. In [LEA 75] wird sogar ein *prosodisch kontrolliertes Spracherkennungssystem* vorgeschlagen.

In diesem Abschnitt sollte aufgezeigt werden, daß alle Gebiete prosodischer Untersuchungen in der ASE von Interesse sein können. Je nachdem, *was* erkannt werden soll, treten gewisse Verschiebungen in der Wichtigkeit der Teilgebiete auf, bzw. es werden Teilgebiete explizit ausgeschlossen. Tabelle 1.1 zeigt für die Bereiche *Erkennung* und *Verstehen des Gesprochenen* die angesprochenen prosodischen Teilgebiete, ein beispielhaftes Einsatzgebiet in einem ASE-System sowie die linguistische Verarbeitungsebene des Einsatzgebietes.

Im folgenden Abschnitt wird das sprachverstehende System EVAR kurz vorgestellt, im darauffolgenden werden die prosodischen Teilgebiete vorgestellt. Danach werden die Ziele dieser Arbeit beschrieben.

Untersuchungsgegenstand	Einsatz in ASE	Linguistische Ebene
Sprechergruppeneigenschaften	Sprecheradaption (z.B. an Sprecher mit speziellen Spracheigenheiten wie Lispeln, etc.)	Akustik-Phonetik-Ebene
Eigenschaften der Sprache unter verschiedenen äußeren Bedingungen	Normierung des Sprachsignals (z.B. bei Sprechen unter mehrfacher Erdbeschleunigung)	Akustik-Phonetik-Ebene
Satzmodus-Bestimmung	Unterscheidung von Frage/Nicht-Frage	Dialog-Ebene
Laut-Suche	Identifikation von Wortuntereinheiten	Akustik-Phonetik-Ebene
Wortakzent-Verifikation	Verifikation von Worthypothesen	Lexikalische Ebene
Phrasenakzent-Verifikation	Verifikation von Phrasen-Hypothesen	Syntaktisch-Semantische Ebene
Satzakzent-Verifikation	Verifikation von Satz-Hypothesen	Syntaktisch-Semantische Ebene
Satzakzent-Suche	Finden der am stärksten betonten Stellen eines Sprachsignals	Lexikalische Ebene Syntaktisch-Semantische Ebene Dialog-Ebene
Gliederungs-Suche	Finden von Gliederungsgrenzen in Äußerungen	Syntaktisch-Semantische-Ebene

Tabelle 1.1: Teilgebiete prosodischer Untersuchungen, ihr Einsatz in ASE-Systemen sowie die linguistische Ebene, auf der das prosodische Wissen eingesetzt wird.

1.3 Das System EVAR

Am Lehrstuhl für Informatik 5 (IMMD 5) der Universität Erlangen wird seit 1979 am sprachverstehenden Dialogsystem EVAR gearbeitet. Das Akronym EVAR steht für notwendige Teilschritte bei der Durchführung eines informationsabfragenden Dialogs: **E**rkennen, **V**erstehen, **A**ntworten, **R**ückfragen.

Die Systemstruktur lehnt sich an ein geschichtetes linguistisches Modell an. Das System ist in mehrere Module unterteilt, die verschiedene linguistische Ebenen repräsentieren. Anforderungsziele an das zu entwickelnde System sind unter anderem:

- ein Dialog-System zum Zwecke der Informationsabfrage über einen definierten und eingeschränkten Problemkreis
- Erkennung von kontinuierlich gesprochener deutscher Standardsprache unter Zulassung von Verschleifungen und Ausschluß von Dialekten
- sprecherunabhängige Erkennung
- Beschränkung der Qualität des Sprachsignals auf Telefonbandbreite
- Verwendung eines *großen* Lexikons
- Verwendung einer flexiblen Systemstruktur, die für Grundlagenforschung auf dem Gebiet des automatischen Sprachverstehens geeignet ist
- Trennung zwischen *problemabhängigem* und *problemunabhängigem* Wissen zum Zwecke eines möglichst einfachen Austausches des Anwendungsbereichs.

Als Pilot-Diskursbereich wurden Auskunfts-Dialoge über das Inter-City-Zugsystem der Deutschen Bundesbahn gewählt. Bild 1.1 zeigt das geschichtete Modell für ein sprachverstehendes Dialogsystem, auf dem die Architektur des Systems basiert. Der Analyseprozeß ist grundsätzlich flexibel gehalten und soll sich in einer Kombination von daten- und erwartungsgesteuerten Methoden vollziehen. Die verschiedenen Module können über definierte Schnittstellen direkt miteinander kommunizieren. Daten werden auf Anforderung ausgetauscht.

Im folgenden werden die einzelnen Module (bis auf das Prosodie-Modul) kurz vorgestellt. Eine ausführliche Übersicht des Systems findet sich in [NIEMANN 85, 88b].

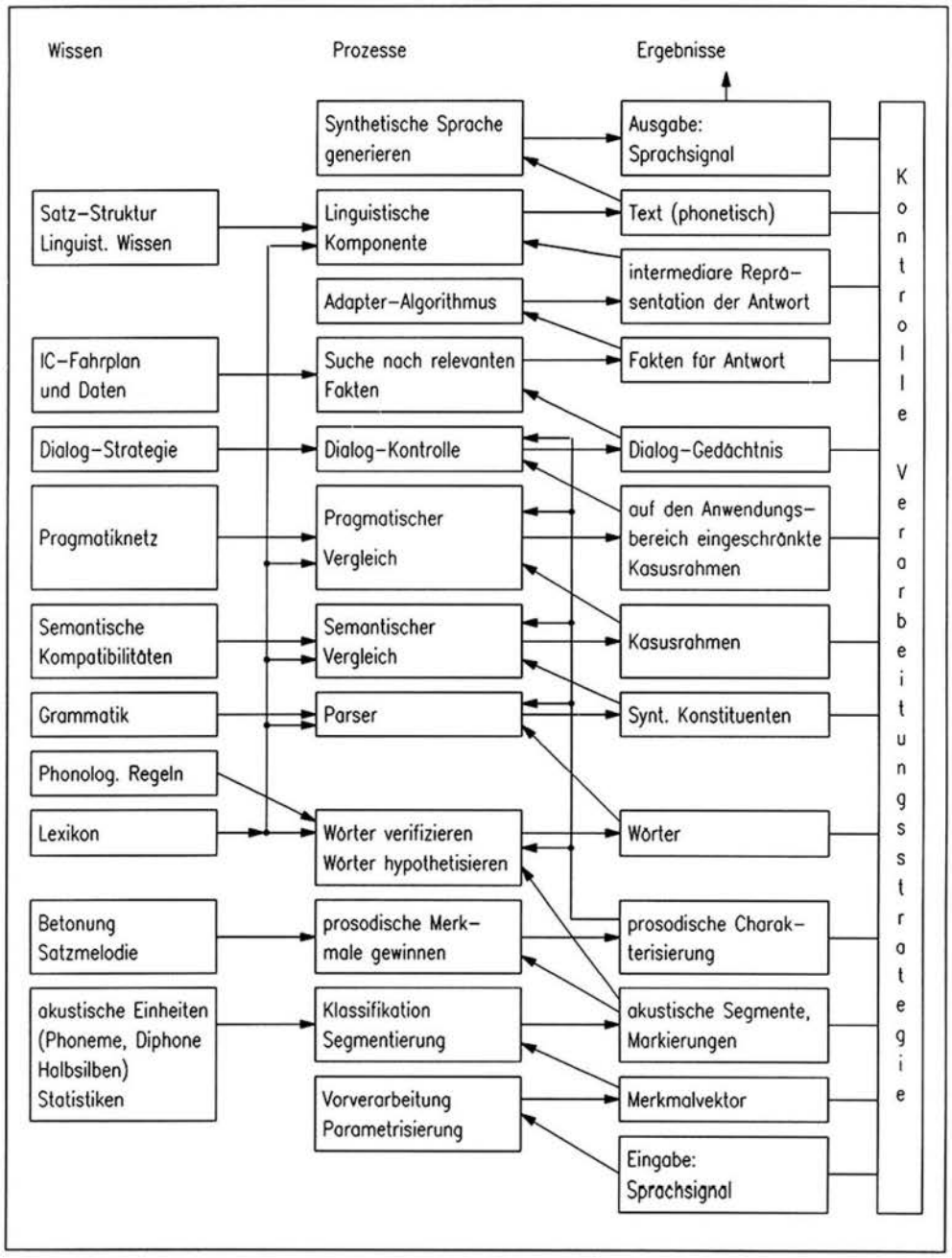


Bild 1.1: Das geschichtete Modell für ein sprachverstehendes Dialogsystem, auf dem die Architektur des Systems EVAR basiert (nach [Niemann 87a]).

Das *Akustik-Phonetik-Modul* ([REGEL 88]) zerlegt das Sprachsignal in äquidistante Zeitscheiben von 12.8 Millisekunden. Diese werden im folgenden als *Frames* bezeichnet. Für jeden Frame werden Merkmale berechnet, die mit einem Bayes-Klassifikator nach Lautkomponenten klassifiziert werden. Die Lautkomponentenhypothesen werden mit Hilfe eines syntaktischen Verfahrens zu *Segmenten (Lauten)* variabler Länge zusammengefaßt. Es werden keine alternativen Segmentgrenzen betrachtet[2].

Das *Worterkennungs-Modul* ([KUNZMANN 90]) erzeugt mit einem statistischen Verfahren (**H**idden-**M**arkov-**M**odell, HMM) *Worthypothesen* durch den Vergleich der Lauthypothesenfolge mit der Standardaussprache eines jeden Wortes eines Lexikons. Eine Worthypothese besteht aus Anfangs- und Endframe des Bereichs, über dem die Hypothese generiert wurde, Wortnummer, Bewertung sowie einem optimalen *Zuordnungspfad* der Standardaussprache zu der Lauthypothesenfolge.

Das *Lexikon* ([EHRLICH 86]) enthält für jedes Wort u.a. eine systemweit eindeutige Wortnummer, die Standardaussprache nach [DUDEN 74] und drei Bitvektoren. Diese repräsentieren die Zugehörigkeit des Wortes zu Wortarten sowie zu semantischen Klassen und pragmatischen Konzepten. Alle (verschiedenen) Wörter mit gleicher orthographischer Umschrift sind unter einem Worteintrag zusammengefaßt. Momentan werden drei Vollformen-Lexika verwendet: *GLEXS* mit ca. 4200 Einträgen, *PLEXS* (ca. 1600 Einträge) und *KLEXS* (ca. 500 Einträge).

Zeitlich aufeinanderfolgende Worthypothesen werden im *Syntax-Modul* auf *syntaktische Konsistenz* überprüft und zu *Syntaxhypothesen* zusammengefaßt [BRIETZMANN 87]. Es werden auch Hypothesen für diskontinuierliche Konstituenten erstellt (z.B. Hypothesen für Verbalgruppen: *"kann ... abfahren"*). Syntaxhypothesen bestehen (analog zu Worthypothesen) aus einer Liste der Wortnummern der enthaltenen Worthypothesen, der Position im Sprachsignal, einer Bewertung sowie einer Darstellung der syntaktischen Struktur. Da zum einen nicht damit gerechnet werden kann, daß alle tatsächlich gesprochenen Wörter hypothetisiert werden, und zum anderen in natürlich-sprachlichen Dialogen Äußerungen teilweise aus unvollständigen Sätzen bestehen, werden nicht nur syntaktisch korrekte Sätze, sondern auch einfachere Konstituenten als Syntaxhypothesen weitergegeben.

Die Syntaxhypothesen werden semantisch verifiziert (*Semantik-Modul*), pragmatisch interpretiert (*Pragmatik-Modul*) und, soweit möglich, zu komplexen Konstituenten und Sätzen zusammengefaßt [EHRLICH 89] (z.B. würden die Hypothesen für die Nominalgruppe *"der nächste Zug"* und die Präpositionalgruppe *"nach Hamburg"* zu der komplexen Hypothese *"der nächste Zug nach Hamburg"* zusammengefaßt, falls die zeitlichen Positionen der beiden Hypothesen gewissen Restriktionen

---

[2] Im folgenden wird für die Lautkomponenten und Laute die Erlanger Kodierung verwendet (siehe Kap.3.2.2). Die Laut- und Lautkomponentensymbole werden durch "/" vom restlichen Text getrennt. Ein Verzeichnis der erkennbaren Klassen sowie ihre Abbildung auf die IPA-Symbole findet sich in Anhang A und B.

genügen). Semantisch verifizierte und komplexe Konstituenten werden als *Semantikhypothesen* bezeichnet.

Unter *pragmatischer Interpretation* ist in EVAR der anwendungsabhängige Teil der semantischen Analyse zu verstehen [BRIETZMANN 84]: Die Hypothese *"mit dem Auto fahren"* würde zwar die semantische Analyse passieren, nicht aber die pragmatische. Zwar handelt es sich bei *Auto* um ein Transportmittel, aber im Anwendungsbereich *Inter-City-Zugsystem* kann ein Auto in dieser Funktion nicht vorkommen. Das Beispiel verdeutlicht, daß pragmatisches Wissen nicht allein durch Ausschluß von Wörtern aus einem anwendungsunabhängigen Lexikon realisiert werden kann. Da man über die Deutsche Bundesbahn an gewissen IC-Bahnhöfen ein Auto mieten kann, gibt es zulässige Fragen, in denen das Wort *Auto* vorkommen kann. Pragmatisch kann jedoch im Anwendungsbereich von EVAR die Klasse *Transportmittel* auf Züge eingeschränkt werden.

Das *Dialog-Modul* ([NIEMANN 88a], siehe auch [BRIETZMANN 84]) hat Aufgaben der Dialogführung; dazu gehört das Reagieren auf Benutzeranfragen mit Antworten und Rückfragen. Hierfür ist unter anderem die Interpretation von metakommunikativen Äußerungen (z.B. *"Grüß Gott, ..."* , *"Ich hätte eine Frage."* , *"Was haben Sie gesagt?"*) notwendig. Diese Äußerungen werden vom Pragmatik-Modul nicht verarbeitet.

Bild 1.2 zeigt für die Module Akustik-Phonetik, Worterkennung und Syntax die Analyse-Ergebnisse (Schnittstellen zu den anderen Modulen):

a) das Sprachsignal für die Wörter *"nach Mainz"* aus der Äußerung *"Wir möchten am Wochenende nach Mainz fahren."*
b) die Handklassifikation auf Lautebene
c) die Handklassifikation auf Wortebene
d) die maximal fünf am besten bewerteten Lautkomponenten-Hypothesen
e) die maximal fünf am besten bewerteten Laut-Hypothesen
f) die am besten bewerteten Wort-Hypothesen (Die richtigen Hypothesen sind mit ### markiert. Das Wort *"fahren"* ist nur partiell im dargestellten Sprachsignal enthalten.)
g) die am besten bewerteten Syntax-Hypothesen (NG ≙ Nominalgruppe, PNG ≙ Präpositionalgruppe und ADJUG ≙ unflektierte Adjektivgruppe).

Bild 1.2 (nächste Seite): Analyse-Ergebnisse für die Module **Akustik-Phonetik, Worterkennung** und **Syntax.**

a)

b)

N	A	XA	M	AI	N	D	S	F	A

c)

noch	Mainz	fahren

d)

c)

N	N	XA	N	A	E	N	XI	S	A
NE	A	F	M	E	ER	NE		Z	N
M	AR	S	NE	AR	AR	NG			M
NG	ER	Z	NG	U	EH	M			
ER	M	SH	U	R	A	ER			

f)

nehmen			Mainz###						o
noch###			neun				ist		o
noch			meine				Hamm		
Nachmittag			um	ein			sein		
wenn	ob		in	eins					w
Moss			nahe			ich	von		
in	auch		da		Anfang				
Minden			im	einer			vom		
im	alle	kleine					Sonne		
Nacht			natuerlich				Zahl		
den	Essen			eine			dann		
macht			mal			mich	fahren###		
koennen			ja		in		kann		
Mark			wollen			bis			l
Doch			Jahren						e
immer			Mann				Dank		
dem	acht		bei			nichts			u
nahe			mon				zum		
Nase				Rad	Im				B
muss			Uhr		welcher				o

g)

NG Nacht		NG Mann	
NG Netz		NG mon	
PNG noch Mainz			
PNG in Mann			
ADJUG gerne		PNG an nichts	
NG Minden		NG Mainz	
PNG in Mainz			
PNG noch Bahn			
PNG noch Jahren			
PNG noch Mann			
PNG noch Rahmen			
PNG noch Wagen			
NG Muenchen	PNG von Bahn		
PNG am Mor	PNG von Jahren		
PNG an Mor	PNG von Mann		
PNG im nae	PNG von Rahmen		
PNG in nae	PNG von Wagen		
PNG in nur		NG nichts	
PNG in acht Jahren			
PNG in acht Rahmen			

1.4 Beitrag dieser Arbeit zur Prosodie-Forschung

In Kap.1.2 wurden verschiedene Möglichkeiten erörtert, prosodische Information in potentiellen ASE-Anwendungen einzusetzen. Ein im zwischenmenschlichen Bereich besonders wichtiger Aspekt der Prosodie, die Übermittlung des eigenen Gemütszustandes, ist für ASE-Systeme so gut wie nicht relevant. Untersuchungen zur Veränderung der Sprache unter verschiedenen äußeren Bedingungen spielen lediglich für sehr spezielle Anwendungen eine Rolle. Sprechergruppeneigenschaften sind für ASE-Systeme mit medizinischen Anwendungen teilweise sehr wichtig (Gerätesteuerung für Behinderte), bei Systemen für einen allgemeinen Benutzerkreis sind diese Eigenschaften jedoch von untergeordneter Bedeutung. Somit verbleiben vier Teilgebiete prosodischer Information für den Einsatz in einem ASE-System: Identifikation der gesprochenen Laute, der Gliederungsgrenzen, des Satzmodus und der betonten Stellen. Bei der Identifikation der betonten Stellen ist zu unterscheiden, ob zusätzliches Wissen über das Gesprochene (die Analyseergebnisse der anderen Module) verwendet wird oder nicht. Der erste Fall wird als *erwartungsgesteuerte* Analyse bezeichnet (die Analyseergebnisse der anderen Module werden verifiziert), der zweite als *datengetriebene* (die interessierenden Stellen werden ohne Berücksichtigung der Ergebnisse anderer Module gesucht).

Diese Arbeit ist ein Beitrag zum Einsatz prosodischer Information im sprachverstehenden System EVAR. Schwerpunkt ist die automatische Erstellung einer Betonungsbeschreibung, also die Identifikation der betonten Stellen. Dabei wird zunächst eine datengetriebene Beschreibung entwickelt. Nach [VAISSIÈRE 88] kann diese Aufgabe in fünf Teilschritte zerlegt werden (in Klammern sind die Kapitel dieser Arbeit angegeben, in denen auf die entsprechenden Teilaufgaben eingegangen wird):

1) Detektion der Silbenstruktur (Kap.4.1 und 6.3.1)
2) Extraktion der prosodischen Merkmale (Kap.4.2 bis 4.4 und 6.1)
3) Normalisierung dieser Merkmale (Kap.4.2.5)
4) Identifikation der betonten Silben (Kap.5 und 6.3.2)
5) Überprüfung des Beitrags der Betonungsbeschreibung zur Gesamtanalyse des ASE-Systems (Kap. 6.4 und 6.5)

Es existieren sehr wenige Ansätze zur Erstellung einer datengetriebenen Betonungsbeschreibung:

- Von den insgesamt 252 bzw. 359 Beiträgen zu der ECST 1987 und ECST 1989 (European Conference on Speech Communication and Technology) beschäftigen sich 24 bzw. 50 (10 bzw. 14 Prozent) im weitesten Sinne mit Prosodie. Davon behandeln jeweils 2 Artikel (3 Prozent) die Erstellung einer vollständigen Intonationsbeschreibung. 72 Prozent (14 bzw. 39 Beiträge), berichten über Untersuchungen zu Teilaspekten (Silbendetektion, F_0-Bestimmung (Pitch), SH/SL-Entscheidung, Dauermerkmale, Bestimmung des Fokus). Lediglich bei jeweils 2 Beiträgen wird die Integration in ein ASE-System berücksichtigt bzw. als Forschungsziel genannt. Beim Anwendungsbereich liegt der Schwerpunkt unter den gesichteten Beiträgen auf

dem Nutzen der Prosodie in der Sprachsynthese, insbesondere in der Text-to-Speech-Konvertierung (8 bzw. 10 Artikel, 24 Prozent).

- In [VAISSIÈRE 88], einem ausgezeichneten Überblick über den Einsatz prosodischer Information in ASE-Systemen, werden für kontinuierliche Sprache fünf Literaturstellen für vier Sprachen genannt: [LEA 73b] und [WAIBEL 86] für das Englische, [MARTIN 79] für das Französische, [PIERACCINI 86] für das Italienische und [HOUSE 87] für das Schwedische.

Nach Wissen des Autors handelt es sich bei der in dieser Arbeit beschriebenen Methode um den ersten Versuch einer automatischen datengetriebenen Betonungsbeschreibung für das Deutsche.

Ein Vergleich mit den oben genannten Ansätzen ist sehr schwierig, da eine Veränderung der Aufnahmebedingungen die Erkennungsraten desselben Algorithmus stark verändern kann, wie in Kap.6.3 gezeigt wird. Es stellt sich hier die Frage, auf welche Art der Beitrag der prosodischen Information zum Analyseprozeß in einem ASE-System überhaupt gemessen werden kann (Teilaufgabe 5 in der obigen Aufzählung).

- Ein automatisches Verfahren, etwa zur Extraktion der Grundfrequenz, ist in der Regel fehlerhaft. Die Fehlerhäufigkeit der Algorithmen ist anhand von Referenzdaten zu beurteilen. Da an dieser Stelle die Robustheit und Genauigkeit sowie der Vergleich verschiedener Algorithmen im Vordergrund stehen, kann auf Sprachmaterial zurückgegriffen werden, für das eine weitere Verarbeitung im System EVAR nicht sinnvoll durchgeführt werden kann (siehe die *Leo-* und *Fokussätze,* Kap.3.4). Diese Vorgehensweise ist angebracht, da für das Sprachmaterial bereits Referenzdaten vorliegen und zusätzlich Test- und Lernstichproben separiert werden.
- Die verschiedenen linguistischen Einheiten (Laute, Silben, Wörter, Phrasen, ...) können nicht fehlerfrei aus dem Sprachsignal extrahiert werden. Auf jeder Analyseebene wird daher mit einer Vielzahl bewerteter Hypothesen gearbeitet. Die Bewertung stellt ein Maß dafür dar, wie sicher sich das generierende Modul über die Richtigkeit der Hypothese ist. Eine Möglichkeit, die Leistungsfähigkeit der einzelnen Verarbeitungsmodule zu beurteilen, ist die Beantwortung der Frage, wieviele der nach ihrer Bewertung sortierten Hypothesen im **Durchschnitt** (gemittelt über alle Äußerungen einer Stichprobe) generiert werden müssen, um *n Prozent* der richtigen Hypothesen zu erhalten. Man erhält somit keine *Erkennungsrate,* sondern eine *Erkennungskurve.* Der Beitrag des Prosodie-Moduls zum Analyseprozeß läßt sich daran messen, ob sich die Erkennungskurve verbessert. Das Prosodie-Modul kann also richtige Hypothesen durchaus verwerfen und die Erkennung im Schnitt trotzdem verbessern. Richtige Hypothesen können aufgrund fehlerhafter Berechnung der Korrelate oder aufgrund von heuristischen Ansätzen verworfen werden, die "meistens" zum Erfolg führen (und deswegen auch prozentual mehr falsche Hypothesen verwerfen).

- Während eines Analyseprozesses ist es häufig notwendig, im Suchraum, der durch die *n besten Hypothesen* aufgespannt wird, nach der richtigen Lösung zu suchen. Geht man davon aus, daß bei einer vollständigen Suche für eine gegebene Hypothesenmenge immer die optimale Lösung gefunden wird, so kann eine Beschränkung des Suchraumes das Analyseergebnis nicht verbessern, sondern nur dafür sorgen, daß die richtige Lösung schneller oder mit geringerem Speicheraufwand gefunden wird. Da auch hierbei damit gerechnet werden muß, daß richtige Lösungen durch die Suchraumbeschränkung nicht gefunden werden, kann man die Aufwandsverringerung einer Suchraumbeschränkung mit Hilfe einer *Reduktionskurve* beurteilen. Für jeden Faktor, um den der Analyseaufwand reduziert wird, kann man feststellen, um wieviel sich die Analyse verschlechtert.

Alle drei Beurteilungskriterien werden im Rahmen dieser Arbeit verwendet.

Die Tatsache, daß bei einer datengetriebenen Betonungsbeschreibung kein weiteres Wissen (z.B. segmentale und syntaktische Struktur) über die Äußerung verwendet werden kann, stellt eine sehr starke Einschränkung dar, die unter der Randbedingung des Einsatzes während der Initialisierung der Erkennungsphase in einem ASE-System zu sehen ist. Diese Einschränkung entfällt z.B. bei Forschungen zum Einsatz prosodischer Information für Sprachsynthese. Auch bei der Erstellung von Intonationsmodellen wird meistens "generativ" gearbeitet, d.h. man geht von einer Äußerung aus und erzeugt hierfür eine Beschreibung der prosodischen Eigenschaften (ein Überblick über die verschiedenen Intonationsmodelle bis zum Erscheinungsjahr findet sich in [LADD 83b]). In einem ASE-System steht Wissen über die Äußerung in Form von Hypothesen anderer Module zur Verfügung. Daher sollte die datengetriebene Betonungsbeschreibung während der Analyse einer Äußerung durch eine erwartungsgesteuerte Analyse verifiziert werden, d.h. nachdem (unter Verwendung der datengetriebenen Betonungsbeschreibung) Phrasen- oder Satzhypothesen erstellt wurden, sollte die Betonungsbeschreibung für den entsprechenden Teil des Sprachsignals überprüft werden.

Leider konnte zu Beginn dieser Arbeit auf kein geeignetes intonatorisches Modell des Deutschen zurückgegriffen werden. In einer Einschätzung der deutschen Intonationsforschung schrieb W. Klein 1980 (siehe auch [KLEIN 82]):

> Der Stand der Forschung zur deutschen Satzintonation gibt zum Rühmen keinen Anlaß. ...
> Es scheint mir kein gutes Zeichen für den Stand der deutschen Phonetik, daß völlige Unklarheit darüber herrscht, was das signalphonetische Gegenstück des „Akzents" ist. ...
>
> [KLEIN 80, S.3 und Anmerkung 22, S.32]

Seitdem sind, insbesondere im Rahmen des von 1984 bis 1988 durchgeführten DFG-Schwerpunkts "Formen und Funktionen der Intonation", mehrere Ansätze untersucht worden. Die Ergebnisse der meisten Projekte sind in [ALTMANN 88] zusammengestellt (für Ergebnisse außerhalb des DFG-Schwerpunkts siehe [ADRIAENS 84], [BANNERT 89] und [MUDLER 86]).

Die Untersuchungen zu einer erwartungsgesteuerten Betonungsbeschreibung, die im Rahmen dieser Arbeit durchgeführt wurden (siehe Kap.7 sowie [NÖTH 87] und [BATLINER 89a, 89b]), erfolgten in enger Kooperation mit den Mitarbeitern des Münchner DFG-Projekts "Modus-Fokus-Intonation" ([ALTMANN 88, 89a]). Neben dem zweifellosen Vorteil der geographischen Nähe erschien der Münchner Ansatz für eine Darstellung in einem maschinellen Wissensrepräsentationsschema am geeignetsten:

- Es werden wenige, einfach extrahierbare Merkmale untersucht.
- Die Merkmale beschränken sich nicht allein auf Tonhöhenkorrelate.
- Die Relevanz der Merkmale wird an einem großen Korpus mit statistischen Methoden überprüft.
- Die im Deutschen möglichen intonatorischen Minimalpaarkonstellationen (siehe Kap.3.4) sind so gut wie vollständig erfaßt.
- Zwar handelt es sich bei dem untersuchten Datenmaterial - ebenso wie beim Material fast aller Intonationsprojekte - um gelesene Äußerungen, eine Erweiterung des Modells auf spontan gesprochene Äußerungen war aber von vorneherein geplant und wird zur Zeit im Rahmen des DFG-Projekts "Intonation - Register - Modus/Fokus" durchgeführt.

Die im Rahmen dieser Zusammenarbeit durchgeführten Untersuchungen in Erlangen konzentrierten sich auf den Aspekt der automatischen Extrahierbarkeit der verwendeten Merkmale und auf Möglichkeiten zur Darstellung des erstellten Modells in einem automatischen Wissensrepräsentationsschema. Die korrekte automatische Bestimmung der Betonung aufgrund eines Intonationsmodells kann als letzter Prüfstein für die Validität des Modells angesehen werden, so daß die im Ausblick aufgezeigte Vorgehensweise nicht nur für die ASE, sondern auch für die Intonationsforschung an sich einen wichtigen Beitrag leistet.

Die weitere Arbeit ist folgendermaßen gegliedert:

In Kapitel 2 werden die für diese Arbeit wichtigen sprachwissenschaftlichen Begriffe eingeführt. Gleichzeitig werden Ergebnisse aus der Literatur vorgestellt, wie sich die drei im Zusammenhang mit ASE-Systemen bedeutendsten Rollen der Prosodie im akustischen Signal ausdrücken. Die verschiedenen Sprachstichproben, mit denen experimentelle Untersuchungen durchgeführt wurden, werden in Kapitel 3 vorgestellt. Auf prosodische Eigenschaften, die bei der Markierung von *Betontheit* eingesetzt werden, wird unter dem Aspekt der Berechnung von korrespondierenden Merkmalen aus dem Sprachsignal in Kapitel 4 eingegangen. Dabei wird ein neuer Grundfrequenz-Algorithmus vorgestellt, der im Rahmen der Arbeiten am Prosodie-Modul entwickelt wurde. In Kapitel 5 wird ein Verfahren vorgestellt, mit dem eine automatische datengetriebene Betonungsbeschreibung aus diesen Merkmalen berechnet wird. Die Ergebnisse der experimentellen Untersuchungen werden in Kapitel 6 präsentiert: Der eigene Grundfrequenz-Algorithmus wird mit zwei aus der Literatur bekannten Algorithmen verglichen. Die automatische Betonungsbeschreibung wird mit dem Ergebnis eines Perzeptionsexperimentes verglichen, in dem 15 Hörer die Silben einer

Sprachstichprobe in betont/unbetont einteilen mußten. Die Auswirkungen auf den Analyse-Prozeß im System EVAR werden diskutiert. Es werden Ergebnisse zu den folgenden Einsatzbereichen prosodischer Information präsentiert: *Satzmodus-Bestimmung, Laut-Suche, Wortakzent-Verifikation* und *Satzakzent-Suche.* Der Erfolg des Einsatzes prosodischer Information wird anhand der oben genannten Auswertungskriterien gemessen. In Kapitel 7 wird ein grundsätzlich neuer Ansatz zur Modellbildung für eine erwartungsgesteuerte automatische Betonungsbeschreibung vorgestellt. In Kapitel 8 schließlich werden die Ergebnisse dieser Arbeit zusammengefaßt.

2 Definitionen, Begriffe und sprachwissenschaftliche Grundlagen

Ziel dieses Kapitels ist es, einige sprachwissenschaftliche Begriffe im Sinne dieser Arbeit zu definieren. Dies ist notwendig, da Begriffe wie *Prosodie, Intonation* oder *Fokus* von Autoren verschiedener sprachwissenschaftlicher Schulen unterschiedlich verwendet werden (siehe z.B. [ARTEMOV 78], der einige der verschiedenen Begriffsbildungen für die beiden Begriffe *Prosodie* und *Intonation* ausführlich behandelt). Dieser "Sprach-" oder besser "Begriffswirrwarr" ist sicherlich auch darauf zurückzuführen, daß die Prosodie ein Bindeglied zwischen verschiedenen sprachwissenschaftlichen Disziplinen darstellt. Somit werden prosodische Themenbereiche von Autoren verschiedener sprachwissenschaftlicher Schulen und Disziplinen behandelt. Weiterhin wird die notwendige Trennung der verschiedenen Abstraktionsebenen nicht immer konsequent eingehalten. So stehen z.B. in [WAIBEL 86, S.6] zu Beginn *"duration"*, *"intensity"*, *"pitch"* und *"stress"* als sogenannte *"prosodic cues"* scheinbar gleichberechtigt nebeneinander, während später [WAIBEL 86, S.105] *"pitch"*, *"duration"* und *"intensity"* als *"acoustic manifestations of stress"* bezeichnet werden.

Die im weiteren vorgestellten Definitionen sollen keinen Beitrag zur Vereinheitlichung der Nomenklatur darstellen. Sie haben sich im Rahmen dieser Arbeit als sinnvoll erwiesen und lehnen sich an die folgenden Literaturstellen an: [ALTMANN 88, 89a], [BUSSMANN 83], [KOHLER 77], [LEHISTE 70], [SCHMÖLZ 87], [BECKMANN 86] und [TILLMANN 80]; für sprachwissenschaftliche Begriffe, die nicht explizit eingeführt werden, sind im Rahmen dieser Arbeit die Definitionen in [BUSSMANN 83] maßgebend. Die Ausführungen beziehen sich, sofern nicht anders angegeben, auf die *deutsche* Sprache bzw. sie sind auf die deutsche Sprache übertragbar, da viele der Aussagen universeller Natur sind oder für ganze Sprachenklassen gelten.

Bei den angeführten Beispielen wurde, soweit möglich, bewußt darauf verzichtet, die Mittel der prosodischen Markierung mit Hilfe einer symbolischen Darstellung anzugeben, da diese oft stark sprecherspezifisch sind. Der Leser möge also, seiner eigenen Intuition folgend, die Beispiele laut sprechen, um sich die *Unterschiede* bzw. die *Funktion* der intonatorischen Markierungen anhand der Beispiele zu veranschaulichen. Betonte Silben werden im folgenden durch Großbuchstaben und Fettdruck gekennzeichnet.

2.1 Prosodie

In [BUSSMANN 83] wird *Prosodie* folgendermaßen definiert:

> "**Prosodie/Prosodik** [engl. *polysystematic phonology*]. Untersuchung sprachlicher Eigenschaften wie →Akzent, →Intonation, Sprechpausen u.a., die sich auf größere Einheiten als einzelne Phoneme beziehen bzw. diese überlagern. Man bezeichnet sie daher auch als →suprasegmentale Merkmale. Die P. kann als Verbindung zwischen →Phonologie und →Syntax bezeichnet werden, insofern →Silben, Wörter und Sätze ihr Untersuchungsgegenstand sind."
> [BUSSMANN 83, S.417]

Diese Definition soll im folgenden übernommen werden. Untersuchungsgegenstand der Prosodie sind die prosodischen Eigenschaften eines lautsprachlichen Ereignisses: *Tonhöhe, Lautheit, zeitliche Strukturierung, Sprechtempo, Stimmlage, Stimmqualität, Klangfarbe, Rhythmus* und auch das Fehlen eines sprachlichen Ereignisses, die *Pause.* Sprachliche Einheiten, denen diese Eigenschaften zuzuordnen sind, umfassen mehr als einen Laut, weshalb die prosodischen Eigenschaften insbesondere in der angelsächsischen Literatur auch als *suprasegmentale Merkmale* bezeichnet werden. Diese sprachlichen Einheiten werden häufig als *Prosodeme* bezeichnet. Es kann sich dabei um *Silben, Wörter, Phrasen, Sätze* oder *Redebeiträge* handeln.

Die oben genannten prosodischen Eigenschaften sind *perzeptive Einheiten.* Den perzeptiven Einheiten entsprechen *akustische Parameter.* So ist z.B. die *Grundfrequenz* des Sprachsignals (siehe Kap.2.7.2 und Kap.4.2) das akustische Korrelat der Tonhöhe. Die Abbildung der akustischen auf die perzeptive Ebene ist i.allg. nicht eindeutig, da auch subjektive Eindrücke des Hörers eine Rolle spielen[1]. Im weitesten Sinne läßt sich dies z.B. mit der Eigenschaft eines Menschen vergleichen, *alt zu sein:* Das eindeutig meßbare Korrelat dieser Eigenschaft, das *Lebensalter,* läßt sich zwar genau angeben, aber ob ein Mensch von jemanden als *alt* angesehen wird, hängt auch von anderen Faktoren wie dem Lebensalter des Beobachters selbst ab.

Eine exakte Definition der prosodischen Eigenschaften ist daher teilweise nicht möglich. Für einige prosodische Eigenschaften, wie z.B. die *Stimmqualität,* sind sogar die akustischen Korrelate noch nicht vollständig bestimmt worden (Warum wird eine Stimme als *rauh* empfunden und wie läßt sich die *Rauhigkeit* einer Stimme im Sprachsignal messen?).

---

[1] Im speziellen Fall der prosodischen Eigenschaft *Tonhöhe* ist die Abbildung wohl noch am besten erforscht.

2.2 Intonation

Der Begriff Intonation soll enger gefaßt werden als in [BUSSMANN 83]:

> "Gesamtheit der prosodischen Eigenschaften von sprachlichen Äußerungen, die nicht an einen Einzellaut gebunden sind. ..."
>
> [BUSSMANN 83, S.219]

Unter *Intonation* wird in dieser Arbeit die ***distinktive*** *Verwendung prosodischer Eigenschaften zur* ***Bedeutungsdifferenzierung*** *ganzer Äußerungen* verstanden, *und zwar sowohl in der Darstellungsfunktion (die intellektuelle Bedeutung) als auch in der Ausdrucks- und Appellfunktion (die emotionale Bedeutung)*. Diese Definition stimmt im wesentlichen mit der von [KOHLER 77, S.126] überein. *Intonation* soll jedoch nicht nur als *distinktive Verwendung der* ***Tonhöhe*** (wie bei Kohler), sondern als *distinktive Verwendung* ***aller*** *prosodischen Eigenschaften* verstanden werden.

Eine Äußerung kann einen oder mehrere Sätze (im grammatikalischen Sinn) umfassen, aber auch nur ein Wort, wenn es sich um elliptische Sätze handelt:

Frage:	*"Wann kommen Sie?"*	Elliptische Antwort: *"Morgen"*
Elliptische Frage:	*"Morgen?"*	Elliptische Antwort: *"Morgen!"*

Die Tatsache, daß die kleinste als selbständige Äußerung verwendbare linguistische Einheit ein Wort, normalerweise sogar ein vollständiger Satz ist, deutet an, daß dem *zeitlichen Verlauf* der prosodischen Eigenschaften eine besondere Bedeutung zukommt.

Da diese Definition von *Intonation* in der Literatur nicht üblich ist, soll im folgenden auf einige wichtige Unterschiede eingegangen werden: Im Gegensatz zu [BUSSMANN 83] trennt diese Definition zwischen dem gezielten Einsetzen einer prosodischen Eigenschaft und artikulatorisch bedingten Variationen der physikalischen Korrelate dieser Eigenschaften. Diese Variationen, die von [LEHISTE 70] als *phonetic conditioning factors* bezeichnet werden, sind Gegenstand *mikroprosodischer* Untersuchungen. Den durch diese Faktoren bedingten Variationen im Verlauf der prosodischen Eigenschaften (bzw. ihrer Korrelate) kann i.allg. keine funktionale Rolle zugeordnet werden.

Die in der Literatur weit verbreitete Beschränkung von Intonation auf Untersuchungen des Tonhöhenverlaufs (siehe u.a. [BANNERT 85], [KOHLER 77], [PIERREHUMBERT 80], [WUNDERLICH 88]) mag für viele intonatorische Untersuchungen ausreichen, kann aber sicher nicht alle prosodischen Aspekte der Bedeutungsdifferenzierung von sprachlichen Äußerungen abdecken. Beschränkt man sich z.B. bei der Untersuchung der Intonation auf Sätze, die anhand einer schriftlichen Vorlage nachgesprochen wurden, so kann man davon ausgehen, daß *Pausen* für die Intonation keine Rolle spielen. Dies ist aber nicht der Fall, wenn man ganze Redebeiträge, z.B. die *freie politische Rede*, zum Untersuchungsgegenstand macht. Hier kann gezielt gesetzten Pausen durchaus eine funktionale Rolle zugewiesen werden, sie können beispielsweise der besonderen Hervorhebung der im Satz davor getroffenen Aussage dienen (siehe das Beispiel in Kap.2.9).

In Diskussionen werden Pausen häufig ebenfalls ganz bewußt gesetzt, insbesondere von geübten Sprechern. Allerdings haben sie hier u.U. eine ganz andere funktionale Rolle, die mit der *Bedeutungsunterscheidung* und somit (im Sinne dieser Arbeit) mit *Intonation* nicht mehr viel zu tun hat: Legt ein Sprecher seine Atempausen in ein Wort bzw. in eine Phrase, so gibt er die Kontrolle über die Diskussion nicht aus der Hand. Sein Diskussionsgegner kann ihm nicht ins Wort fallen (siehe [ROYÉ 83], der für eine Fernsehdiskussion u.a. die Verteilung von Pausen in freier Rede empirisch untersucht). Wenn man die verschiedenen Situationen untersucht, unter denen die sprachliche Kommunikation stattfindet, kann man genaugenommen nicht von ***der** Intonation des Deutschen* sprechen, sondern eher von der *Intonation der .. (öffentlichen Rede, usw.) im Deutschen.* (Es handelt sich also hier um verschiedene Register oder Stilebenen.)

Die Tatsache, daß hier unter Intonation der distinktive Einsatz *aller* prosodischen Eigenschaften verstanden wird, ist dahingehend zu verstehen, daß beim momentanen Stand der Intonationsforschung keine Eigenschaft ausgeschlossen werden kann. Es gibt sicher Kommunikationssituationen, in denen die eine oder andere Eigenschaft keine oder nur eine sehr untergeordnete Rolle spielt oder in denen der Tonhöhenverlauf so stark dominiert, daß man sich auf die Untersuchung dieser Eigenschaft beschränken kann. Eine Verallgemeinerung der Untersuchungsergebnisse auf alle Kommunikationssituationen erscheint dann allerdings nicht gerechtfertigt. In diese Richtung geht die sehr vorsichtige Begriffsklärung in [ALTMANN 89b]:

> "Der Begriff 'Intonation' wird hier im weiteren Sinn gebraucht: Er umfaßt die nichtsegmentalen lautlichen Eigenschaften von Äußerungen, also **mindestens** (*Hervorhebung nicht im Original*) die Grundfrequenz(Fo)/Tonhöhe, die Intensität/Lautstärke und die zeitliche Strukturierung von Äußerungen."
>
> [ALTMANN 89b, S.2]

Die Darstellungsfunktion der Intonation gliedert sich in mindestens drei Bereiche ([KOHLER 77, S.127]):

1) **Markierung des Akzents**

*Der **LEO** säuft.* vs. *Der Leo **SÄUFT.***

Im ersten Satz wird die Tatsache hervorgehoben, daß es sich bei dem Säufer um den *Leo* handelt und nicht um eine andere Person. Im zweiten Satz wird die Tatsache hervorgehoben, daß der Leo *Alkoholprobleme* hat.

2) **Markierung des Satzmodus**

neunundvierzig. vs. *neunundvierzig?*

(siehe das Beispiel in Kap.1.2)

bzw. *Der Zug fährt durch.* vs. *Der Zug fährt durch?*

(Beispiel mit grammatikalisch vollständigen Satz)

3) **Gliederung der Äußerung**

Das folgende Beispiel aus dem Englischen ist ohne Interpunktion (Gliederung in der geschriebenen Form) bzw. Intonation (Gliederung in der gesprochenen Form) nicht verständlich:

John where James had had had had had had had had had had been correct.

Zwei Interpretationen sind möglich:

John, where James had had "had had", had had "had"; "had had" had been correct.

oder

John, where James had had "had", had had "had had"; "had had" had been correct.

Der Rest dieses Kapitels beschäftigt sich mit diesen drei Rollen der Intonation in der zwischenmenschlichen Kommunikation: In Kap.2.3 wird der Begriff *Akzent* eingeführt. Der semantische Begriff *Fokus,* der phonetisch durch den Akzent gekennzeichnet wird, wird in Kap.2.4 behandelt. Kap.2.5 beschäftigt sich mit dem *Satzmodus* und Kap.2.6 mit der *Gliederung von Äußerungen.* Die prosodischen Mittel, mit denen Akzent und Satzmodus markiert und Äußerungen gegliedert werden, werden in Kap.2.7-2.9 näher vorgestellt. Kap.2.10 schließlich geht auf die Interaktionen zwischen den drei Rollen der Intonation ein.

2.3 Akzent

Der Begriff *Akzent* soll im Rahmen dieser Arbeit synonym mit *Betonung* verwendet werden. Abgesehen von der intuitiven Gleichsetzung von *akzentuiert* mit *hervorgehoben* herrscht keine Einigkeit darüber, was Akzent ist und wie Akzent markiert bzw. perzipiert wird. In [SCHMÖLZ 87] wird anhand von sechs Zitaten mit Definitionsversuchen eindrucksvoll demonstriert, daß diese *Intuition* vom Betrachterstandpunkt abhängt. Hervorhebung wird vom Standpunkt der *Produktion* (1, 2), der *Perzeption* (3, 4) und *signalphonetisch* (5, 6) betrachtet:

1) "Akzent (Druck) ist Energie, intensive Muskeltätigkeit ... soll eine starke Silbe ausgesprochen werden, wird in allen Organen die größte Energie aufgewandt."
[JESPERSEN 32, S.119]

2) "When the speaker's activity in producing stressed syllables is in focus, stress may be defined in terms of greater effort that enters into the production of stressed syllables as compared to an unstressed syllable."
[LEHISTE 70, S.106]

3) "In zahlreichen Sprachen ist in jedem Wort ... eine bestimmte Silbe gegenüber allen anderen durch das Vorhandensein eines phonologischen Merkmals 'akzentuiert' ausgezeichnet, das sich in verschiedenen Kontexten in unterschiedlicher Weise durch einen größeren Grad der phonetischen Prominenz, d.h. des perzeptiv stärkeren Hervortretens, manifestiert."
[KOHLER 77, S.122]

4) "Unter Schwere verstehen wir das Gewicht, mit dem eine Silbe ins Ohr fällt."
[WINKLER 73, S.642]

5) "Syllabic prominence is a product of length, loudness, pitch (and pitch movement) and quality (or 'timbre') of a syllabic nucleus. These auditory properties correlate in a complex manner with the acoustic parameters: duration, intensity, frequency (and frequency change) and spectral structure."
[BARRY 81, S.322]

6) "Stress ... is reflected in at least four acoustical parameters: speech power, fundamental voice frequency, phonetic quality, and duration."
[LEHISTE 59, S.428]

zitiert in [SCHMÖLZ 87, S.5-6]

Schmölz selbst versucht, die verschiedenen Teilaspekte zusammenzufassen, indem sie die *Intention des Sprechers* in den Vordergrund stellt:

"Berücksichtigt man die Intention des Sprechers, so läßt sich ein möglicher Zusammenhang zwischen den eben beschriebenen Teilaspekten des Akzents denken: Durch die Akzentuierung will der Sprecher die Aufmerksamkeit des Hörers auf bestimmte Stellen seiner Rede lenken. Zu diesem Zweck gestaltet er diese Stellen perzeptiv auffällig, indem er dort die akustischen Parameter Tonhöhe, Dauer, Intensität und Lautqualität ändert; für diese Veränderungen ist erhöhter Kraftaufwand erforderlich."
[SCHMÖLZ 87, S.6]

Man unterscheidet allgemein zwischen *Wort-, Phrasen-* und *Satzakzent.* Isoliert gesprochen hat jedes Wort eine Silbe, die Träger des Wortakzents ist (zum Begriff der Silbe siehe z.B. [KOHLER 77, S.79ff], [TILLMANN 64] und [LINDNER 81]). Die Position der Wortakzentsilbe ist nicht fixiert

(das würde beispielsweise bedeuten, daß immer die erste Silbe Träger des Wortakzents ist). Sie liegt aber für jedes einzelne Wort fest und läßt sich i. allg. mit einigen Regeln (siehe z.B. [KOHLER 77, S.191-196]) vorhersagen.

Innerhalb eines Satzes wird nicht jede Wortakzentsilbe auch tatsächlich akzentuiert produziert, sondern nur solche, die Träger des Phrasen- bzw. Satzakzents sind. Man spricht daher von der *potentiellen Realisierung* des Wortakzents:

> "It appears probable that word-level stress is in a very real sense an abstract quality: a potential for being stressed. Word-level stress is the capacity of a syllable within a word to receive sentence stress when the word is realized as part of the sentence."
>
> [LEHISTE 70, S.150]

Im Falle von *emphatischem* Akzent (*"das ist einfach **UNERHÖRT**, daß man für diesen Bummelzug auch noch Zuschlag bezahlen muß"*) und *kontrastivem* Akzent (*"Ich möchte nach **HOM**burg fahren, nicht nach **HAM**burg"*), kann auch eine Nicht-Wortakzentsilbe Träger des Phrasen- oder Satzakzents sein.

Einem Satz wird *ein* Satzakzent und, je nach Länge der Äußerung, *ein* bis *mehrere* Phrasenakzente zugeordnet. Jeder Satzakzent ist auch Phrasenakzent, die Umkehrung gilt jedoch nicht. Dem Satzakzent wird eine *höhere Prominenzstufe* zugeordnet, man spricht daher auch von *Primär-* und *Sekundärakzent* (die Zuordnung ist etwas problematisch, da nicht klar ist, wieviele Stufen produziert bzw. perzipiert werden, siehe [LIEBERMAN 65]).

In der Vergangenheit wurden immer wieder Versuche unternommen, die Position des Satzakzents und der Phrasenakzente aufgrund der *syntaktischen Oberflächenstruktur* zu bestimmen (für das Deutsche z.B. in [KIPARSKY 66], [BIERWISCH 66]; für neuere Ansätze, bei denen auch die Ebene der Semantik mit einbezogen wird, siehe z.B. [UHMANN 88]). Diese Akzentstruktur wird *Normalbetonung* genannt. Sie tritt beim Sprechen von Sätzen ohne Kontextwissen auf, etwa wenn man aufgefordert wird, einen isolierten Satz vom Blatt zu lesen. Solche Äußerungen werden in der Literatur auch als *"out of the blue"-Sätze* bezeichnet.

> "Wir berücksichtigen durchaus nur die normale, affektfreie Betonung von Sätzen. Emphatische oder kontrastive Betonung lassen wir konsequent beiseite. Entgegen der Normalbetonung, die strengen Regeln unterliegt, kann diese ein beliebiges Wort des Satzes treffen und stellt daher keine besonderen Probleme."
>
> [KIPARSKY 66, S.79]

In dieser sogenannten *Normalbetonung* liegt z.B. in dem Satz

"Mein Bruder wohnt in München."

aufgrund der syntaktischen Struktur der Satzakzent auf *München* und ein Phrasenakzent auf *Bruder*. In dem Satz

"Es ist mein Bruder, der in München wohnt."

ist die Verteilung der Akzentstellen gerade umgekehrt.

Da die zwischenmenschliche Kommunikation in gesprochener Sprache jedoch so gut wie nie "kontextfrei" abläuft, wird diese *Normalbetonung* auch ohne Einbezug des *kontrastiven* und *emphatischen* Akzents aufgrund eines gegebenen Kontextes so häufig durchbrochen, daß man an sich nicht von einer *Normalbetonung* sprechen kann. Daher sind diese "strengen Regeln" für sich alleine auch unzureichend für einen Einsatz in einem ASE-System.

In [LÖTSCHER 83, S.40-53] wird auf die Problematik der unzureichenden Wiedergabe der Realität durch die Normalbetonung ausführlich eingegangen. Lötscher schlägt den Begriff *neutraler Akzent* vor. Als Musterbeispiele für neutral akzentuierte Sätze führt Lötscher u.a. Sätze an, die Rundfunknachrichten eröffnen können.

In *neutraler Betonung* besteht ein enger Zusammenhang zwischen den Positionen der Phrasenakzente und dem rhythmischen Aufbau der Sprache. Im Deutschen besteht die Tendenz,

> "... die rhythmischen Hebungen, die akzentuierten Silben, in approximativ gleichen Abständen aufeinander folgen zu lassen (vorausgesetzt natürlich, daß der Sprechfluß nicht durch Häsitationen unterbrochen oder das Tempo verändert wird)."
>
> [KOHLER 77, S.123]

Die Länge dieser Einheiten, die Kohler *Takte* nennt (*Phrasierungseinheit* bei [BIERWISCH 66]), hängt sehr stark vom *Sprechtempo* und dem sprecherspezifischen *Sprechstil* ab. Beispielsweise sind in dem Satz ([LÖTSCHER 83, S.21])

Fritz untersuchte Johanns linke Zehe.

zwei intonatorische Zerteilungen in Phrasen (mit "/" markiert) möglich:

Fritz / untersuchte / Johanns / linke Zehe.
Fritz / untersuchte / Johanns linke Zehe.

2.4 Fokus

Die bisherigen Ausführungen haben gezeigt, daß es zwar so etwas wie eine *neutrale Betonung* gibt, für die sich auch die Akzentsteuerung generativ erzeugen läßt. Sie kann jedoch nur einen Teilaspekt der sprachlichen Kommunikation wiedergeben (*Leseintonation*) und ist insbesondere für die Repräsentation von Betonungsmustern in *Dialogsituationen* (und somit für den Einsatz in einem ASE-Dialogsystem) inadäquat. Es gibt daher eine ganze Reihe von Ansätzen, in denen ein Satz nicht nur unter dem Aspekt seiner syntaktischen Struktur, sondern unter dem Aspekt seiner Mitteilungsfunktion gegliedert wird. Hierfür wurde von [MATHESIUS 29] der Begriff *Funktionale Satzperspektive* geprägt. Mit diesem Ansatz ist es möglich, situationellen Kontext für die Vorhersage der Akzentpositionen mit einzubeziehen. Jeder Kommunikationspartner hat ein Modell vom Wissensstand seines Dialogpartners. Dieses Modell wird bestimmt vom bisherigen Verlauf der Kommunikation, von der Beziehung zwischen den Kommunikationsteilnehmern (gut bekannt, befreundet, unbekannt, ...) und den gemeinsam zugänglichen Sinneseindrücken (Hintergrundabläufe, nonverbale Kommunikationsmittel, ...). Aufgrund dieses Modells ordnet der Dialogpartner jedem Teil seiner Äußerung einen *Informationsgehalt* zu, wobei zwischen *alter, d.h. gegebener* und *neuer Information* unterschieden wird:

> "Virtually every sentence a speaker utters is a mixture of what ... I will call GIVEN material, which the speaker assumes is already in the addressee's consciousness, and NEW material, which he assumes is not. As he converts this mixture into sound, the speaker does not treat the given and new material in the same way: typically, he will attenuate the given material in one way or another, e.g. by pronouncing the items that convey such material with **lower pitch and weaker stress** (*Hervorhebung nicht im Original*) or by the attenuated specification or pronominalization of such items."
>
> [CHAFE 74, S.112]

Die *neue Information* kann aus einer Menge von Alternativen ausgewählt werden, die durch den situationellen Kontext bzw. das daraus abgeleitete Modell vorgegeben ist. Die *bekannte Hintergrundinformation* wird auch als *Topik*, die *neue* als *Fokus* bezeichnet. Der *Fokus* kann außer durch den *Akzent* auch (zusätzlich) durch die *Wortstellung* markiert werden:

"Mein Bruder wohnt in München" vs. *"Es ist mein Bruder, der in München wohnt."*

(siehe das Beispiel in Kap.2.3)

Die Konstituente des Fokus ist Träger des Satzakzents. Somit besteht ein sehr enger Zusammenhang zwischen dem *Informationsgehalt* und dem Grad der *Akzentuierung*. Aus diesem Grund wurden in Perzeptionsexperimenten ([LEA 80b]) fast immer Inhaltswörter (Substantive, Verben, Adverbien und Adjektive) und nicht Funktionswörter (Artikel, Präpositionen, Konjunktionen und Pronomina) von Versuchspersonen als *betont* markiert.

Etwas problematisch sind in diesem Zusammenhang *metakommunikative* Äußerungen wie:

*"Müller, Grüß **GOTT**. Ich hätte eine **FRA**ge. ..."*

Das Wort *Frage* ist zweifellos Träger des *Satzakzents,* aber man kann nicht immer davon reden, daß es *fokussiert* ist. Selbst wenn man argumentiert, daß an dieser Stelle die gesamte Äußerung *"Ich hätte eine Frage."* fokussiert ist, kann man ihr häufig nur einen sehr geringen semantischen Informationsgehalt zuweisen. Wie in [HITZENBERGER 86] gezeigt wird, ist diese Floskel eine übliche Dialogeinleitung, selbst wenn, wie in den dort untersuchten Dialogen, ein Dialogpartner ein Bundesbahn-Auskunftsbeamter ist und die Floskel vom Informationsgehalt her absolut überflüssig ist. Diese Frage soll hier nicht weiter verfolgt werden (siehe aber Kap.3.2.3).

2.5 Satzmodus

Der Begriff *Satzmodus* wird im Sinne von [ALTMANN 84] und [ALTMANN 87] verwendet. Die folgende kurze Charakterisierung aus [ALTMANN 89b] ist in diesem Zusammenhang vollkommen ausreichend:

> "Satzmodi werden ... verstanden als komplexe syntaktische Strukturen (Beispiele: Aussagesatz, Entscheidungsfragesatz, Wunschsatz), denen regelhaft bestimmte abstrakte Funktionstypen zugeordnet sind. Die jeweiligen Formtypen ergeben sich aus dem Zusammenspiel von Merkmalen aus vier verschiedenen Merkmalsmengen:
>
> a) Kategoriale Füllung (Beispiel: *w*-Ausdrücke in *w*-Fragesätzen und *w*-Exklamativsätzen)
>
> b) Stellungseigenschaften (Beispiele: Stellung des finiten Verbs an erster, zweiter oder letzter Position, Stellung des *w*-Frage-Ausdrucks in Versicherungsfragen)
>
> c) morphologische Markierung (Beispiel Konjunktiv II in Wunschsätzen)
>
> d) intonatorische Markierung."
>
> [ALTMANN 89b, S.1]

Diese Definition deutet an, daß die intonatorische Markierung mehr oder weniger stark ausgeprägt sein kann. So ist z.B. der Satz

"Wo muß ich umsteigen?"

bereits durch das *w*-Element eindeutig als Frage gekennzeichnet, eine intonatorische Markierung also fakultativ. In anderen Fällen kann der Satzmodus nur durch die intonatorische Markierung bestimmt werden, z.B. bei Versicherungsfragen:

"Da muß ich gar nicht umsteigen?"

Da die Markierung des Satzmodus und des Satzakzents teilweise mit den gleichen prosodischen Eigenschaften erfolgt, muß für eine detaillierte Analyse der intonatorischen Markierung die gegenseitige Beeinflussung berücksichtigt werden (siehe [BATLINER 88b, 89c] und Kap.2.10).

2.6 Gliederung von Äußerungen

Die gliedernde Funktion der Intonation hängt eng mit dem rhythmischen Aufbau der Sprache an sich und mit der syntaktischen Struktur der Äußerung (und somit der neutralen Betonung) zusammen. Die gliedernde Rolle ist vor allem dann wichtig, wenn eine Unterbrechung oder Veränderung angezeigt werden soll, also insbesondere bei Ergänzungen und Einschüben:

"Mein Arbeitskollege, der Gerhard Schuhmann - der kennt Dich übrigens - hat sich letzte Woche einen Porsche gekauft."

Eine weitere sehr wichtige gliedernde Rolle hängt gerade nicht mit der syntaktischen Struktur und dem rhythmischen Aufbau zusammen: Die Markierung eines *gestörten Sprech-Denk-Ablaufs* in der freien Rede. In den Demonstrationsbeispielen des folgenden Zitats aus [ROYÉ 83] bedeutet die Notation "/" eine *Kurzpause* (bis 0,4 Sek.), "//" eine *Normalpause* (bis 1 Sek.) und das Zusatzzeichen "+" eine Pause mit *Atmung*. Das Kapitel, aus dem das Zitat entnommen ist, beschäftigt sich mit der Funktion der Pause zur Formulierungskorrektur und Wortfindung. Daher sind außer der Pause keine prosodischen Eigenschaften markiert. Der Begriff *Proposition (Pp.)* wird in Kap.2.9 erläutert.

"Im gesprochenen Text sind Pausen festzustellen, die durch ihre Länge und ihre Häufigkeit innerhalb der Pp. auffallen. Es handelt sich hier offensichtlich nicht um eine vom Sprecher intendierte Segmentierung, sondern um eine Störung im Sprech-Denkablauf. Indizien dafür sind:
1. Pausen in Verbindung mit Formulierungskorrekturen des Sprechers,
2. auffällige Pausen an nicht sinngemäßen Stellen innerhalb der Proposition.
Demonstrationsbeispiele
Pp 4651 ich stimme Herrn R. /
2 damit // eh
3 durch... durchaus darin zu //+
Nach dem Präpo-Ausdruck "damit" macht der Sprecher eine komb. Pause, gewinnt noch etwas Zeit durch eine Teilwiederholung des folgenden "durchaus" und formuliert dann das passende Wort im Sinne der verbalen Wortkette: Herrn R. darin zustimmen.
Pp 3371 jetzt müssen wir uns mal über die Funk... überhaupt über die /+ eh /
2 meinetwegen Rolle des Richters überhaupt in unserer Rechtsordnung /
3 klar werden
Die ursprünglich geplante verbale Wortkette hieß wahrscheinlich: sich über die Funktion des Richters klar werden. Der Sprecher korrigiert sich in folgender Weise: er unterbricht seine Rede mitten im Wort "Funktion" und setzt mit dem Präpo-Akk., erweitert durch das Situativ "überhaupt", nochmals an, findet aber nicht sofort das Wort, das er anstelle von "Funktion" setzen wollte, macht deshalb eine Pause und schiebt vor dem Ersatzwort "Rolle" noch "meinetwegen" ein, das in diesem Zusammenhang als "Füllwort" gedeutet werden kann."
[ROYÉ 83, S.90-91]

2.7 Akzentuierungsmittel

Im folgenden soll auf einige Untersuchungen zur Bestimmung der Akzentuierungsmittel eingegangen werden. Ein ausgezeichneter Überblick findet sich in [LEHISTE 70], sowie in [SCHMÖLZ 87, Kap. 2] und [BECKMANN 86]. Auf eine Darstellung vom Standpunkt der *Produktion* (erhöhte Muskeltätigkeit, usw.) wird hier verzichtet, da bei den eigenen Untersuchungen nur das Sprachsignal zur Verfügung steht.

Entscheidend für die Bestimmung der Akzentuierungsmittel sollte das Urteil von *phonetisch **untrainierten** Hörern* sein, d.h. die Frage, ob von einer ***Gruppe*** eine Stelle im Sprachsignal als *hervorgehoben* perzipiert wird. Auf eine *Gruppe* von *untrainierten* Hörern sollte zurückgegriffen werden, damit

- die Zahl der Fehlkategorisierungen verringert wird (die Wahrscheinlichkeit, daß *n* Versuchspersonen beim *exakt gleichen* Stimulus eine falsche Zuordnung aufgrund von Unkonzentriertheit usw. treffen, ist gering),
- nicht die *intuitive Vorstellung* von Betonung eines einzelnen Hörers benutzt wird,
- nicht (*unbewußt!*) Kategorisierungen vorgenommen werden, die die eigene Hypothese über die Markierung des Akzents unterstützen,
- nicht Einflußfaktoren als Akzentuierungsmittel interpretiert werden, die zwar von einem trainierten Hörer (z.B. von einem Phonetiker oder von einer Person mit musikalischer Ausbildung) noch wahrgenommen werden, die aber in der zwischenmenschlichen Kommunikation keine Rolle spielen, da von der Mehrheit der Kommunikationspartner die "Hervorhebung" gar nicht perzipiert wird.

Allerdings ist zu bedenken, daß eine ganze Reihe von Faktoren beim Versuchsaufbau das Urteil der Hörer systematisch beeinflussen können. So wird die Mehrheit der Hörer in der Äußerung

"Ich möchte morgen ähh übermorgen nach Hamburg fahren"

den signalphonetisch prominentesten Teil, das *"ähh"*, nicht als betont empfinden. Zwar hat das *"ähh"* in diesem Beispiel sehr wohl eine wichtige kommunikative Rolle, nämlich zu signalisieren

"Einen Moment, ich habe mich versprochen, ich korrigiere mich sofort",

aber in das Hörerurteil geht in diesem Beispiel vor allem die Bedeutung der Äußerung mit ein, so daß die Mehrheit der Hörer dazu tendieren wird, dem Wort *"übermorgen"* den Satzakzent zuzuweisen (siehe Kap.3.3). Dieser Effekt ist auch in der anderen Richtung beobachtbar. Aufgrund der Bedeutung und des Gedächtniseindrucks der zugehörigen Normbetonung weist der Hörer den Akzent einer Stelle zu, die signalphonetisch nicht hervorgehoben ist. Diese Tatsache ist sogar nach Entfernung der Bedeutung zu beobachten:

> "Man muß zweifellos von einem linguistischen Faktum sprechen, wenn beispielsweise im gleichen sinnlosen Zweisilber /apa/ einmal von einem Hörer, in dessen Sprache Anfangsbetonung vorliegt, die erste Silbe betont gehört wird und andererseits von einem Hörer einer Sprache mit Endbetonung die zweite Silbe als betont aufgefaßt wird, obwohl die akustisch meßbare Intensität beider Silben identisch ist."
>
> [HEIKE 69, S.15]

Heikes Ergebnisse zeigen eine Bevorzugung der ersten Silbe in den Betonungsurteilen deutscher Hörer, was sich vermutlich damit erklären läßt, daß Heike mit Vokalpaaren gearbeitet hat und daß im Deutschen die Tendenz besteht,

> "... im nicht-abgeleiteten und unflektierten Wort die Pänultima, also die vorletzte Silbe, zu akzentuieren, ..."
>
> [KOHLER 77, S.191]

Ein weiterer bekannter Einflußfaktor ist der Ordnungseffekt: Die Beurteilung eines Paares von Stimuli kann von seiner Anordnung abhängen (z.B. wird die Frage, ob zwei Laute gleich lang sind, für die Anordnung AB der beiden Stimuli i.allg. anders beantwortet als für die Anordnung BA; siehe [BATLINER 87a], [SCHIEFER 88]).

Auch die Tatsache, daß manipuliertes Testmaterial beim Hörer den Eindruck einer *"Hervorhebung"* hervorrufen kann, und zwar aufgrund des ungewohnten Höreindrucks und nicht aufgrund eines intonatorischen Phänomens, kann bei Perzeptionsexperimenten mit synthetischem Material eine Rolle spielen.

Als Mittel zur Markierung der Betonung wurden vor allem die prosodischen Eigenschaften *Tonhöhe, Lautheit, zeitliche Strukturierung* und *Klangfarbe* systematisch untersucht. Dabei beschränkte man sich häufig auf die Ebene des Wortakzents und untersuchte, wie sich die Veränderung dieser Eigenschaften signalphonetisch manifestiert. Dies hat verschiedene organisatorische und systematische Vorteile:

- Es existiert eine ganze Reihe von Wörtern, bei denen es sich, abgesehen von der Position des Wortakzents, um Homophone handelt (z.B. Substantiv/Verb-Paare im Englischen wie ***PER**mit* vs. *per**MIT***, Wortpaare wie ***UM**fahren* vs. *um**FAH**ren* im Deutschen). Mikroprosodische Einflußfaktoren können somit konstant gehalten werden.
- Eine kontrollierte Stichprobe von Sprachaufnahmen ist einfacher zu erstellen. Man kann die Position des Akzents über vergleichsweise einfache Perzeptionsexperimente bestimmen.
- Die Überlagerung mit den weiteren Rollen der Intonation kann ausgeschlossen werden. (Das soll nicht heißen, daß die Überlagerungen auch tatsächlich immer ausgeschlossen werden.)
- Benutzt man als Testwörter Logatome (Unsinnswörter wie *"sisi", "sasa",* usw.), so kann man das semantische Wissen der Testpersonen ausschalten.

Die Ergebnisse aus der Literatur sind sehr widersprüchlich. Während z.B. in [BOLINGER 58b] die *Grundfrequenz* als wichtigstes Akzentkorrelat bezeichnet wird, die *Dauer* als Kovariable der Grundfrequenz und die *Intensität* als bedeutungslos eingestuft wird, kommt [BECKMANN 86] zu

der Schlußfolgerung, daß die *Intensität* das wichtigste Akzentkorrelat ist, und [ADAMS 78] zu der, daß die *Dauer* das häufigste Korrelat darstellt.

2.7.1 Intrinsische Eigenschaften und Koartikulation

Als kleinste betonbare Einheit gilt die *Silbe.* Veränderungen zum Zwecke der Betonung betreffen insbesondere den *Silbenkern,* so daß hier nur *Sonoranten* (Vokale, Liquide und Nasale) betrachtet werden müssen. Den prosodischen Eigenschaften *Tonhöhe, Lautheit, zeitliche Strukturierung* und *Klangfarbe* auf der perzeptiven Ebene entsprechen die akustischen Parameter *Grundfrequenz (F_0), Intensität, Lautdauer* und *spektrale Struktur.* Bevor das Verhalten dieser Parameter bei der Akzentmarkierung untersucht werden kann, müssen intrinsische (dem Laut eigene) und koartikulatorische Einflüsse bestimmt werden, welche die Parameterwerte beeinflussen. Folgende Einflüsse wurden festgestellt:

- Für Dauer und spektrale Struktur: die Zugehörigkeit zur Klasse der *Langvokale* oder *Kurzvokale.* Außer in der Dauer (Dauerwerte für das Deutsche finden sich z.B. in [MAACK 49a]) unterscheiden sich *kurze Vokale* von *langen* durch eine Verschiebung der ersten beiden Formanten in Richtung Zentralvokal /ER/ ([WODARZ 71]). Bild 2.1 veranschaulicht diese Verschiebung für die Vokale aus der EVAR-Stichprobe (Kap.3.1), sowie die Verschiebung beim Übergang von *isoliert gesprochenen Vokalen* nach solchen aus *kontinuierlich gesprochener Sprache.*
- Für F_0, Dauer und Intensität: die Zungenstellung des Vokals (siehe Bild 2.1). Vokale mit tiefer Zungenlage haben eine niedrigere F_0 ([LEHISTE 70], [MOHR 71], [ANTONIADIS 81]), sind länger ([MOHR 71], [ANTONIADIS 84d]) und haben eine höhere Intensität als solche mit hoher Zungenlage ([LEHISTE 59], siehe auch Bild 2.4 in Kap.2.7.4).
- Für F_0, Dauer und Intensität: die Konsonantenumgebung. Nach *stimmlosen* Konsonanten ist die F_0 höher als nach *stimmhaften,* nach *aspirierten* höher als nach *nicht aspirierten* ([HOUSE 53], [MOHR 71], [THORSEN 79]). Vokale in stimmloser Konsonantenumgebung sind kürzer als in stimmhafter ([BELASCO 53], [KIM 70], [DOMMELEN 82], [KOHLER 82]). Je näher die Artikulationsstellen des Vokals und des folgenden Konsonanten sind, umso kürzer ist der Vokal ([MAACK 53], [FISCHER-JØRGENSEN 64]). Stimmlose Plosive dämpfen die Intensität stärker als stimmlose Frikative und diese stärker als Gleitlaute, Nasale und stimmhafte Frikative. Den geringsten Einfluß auf die Intensität haben Halbvokale und stimmhafte Plosive ([LEHISTE 59]).
- Für F_0 und Dauer: die Silbenstruktur. Vokale in offenen Silben haben eine höhere F_0 und sind länger als in geschlossenen ([MOHR 71]).
- Die Dauer der Vokale sinkt mit zunehmender Anzahl von Silben im Wort ([RIETVELD 75]) und Wörtern im Takt ([LEHISTE 70], [KOHLER 82]) sowie mit zunehmender Sprechgeschwindigkeit ([KOHLER 81]).

- Die F_0, Dauer und Intensität werden von der Stellung der Silbe in der Äußerung beeinflußt: In einer Atmungsgruppe ("breath-group", d.h. einem Äußerungsteil zwischen zwei Atmungsstellen) existiert die Tendenz des F_0-Verlaufs, im Mittel abzufallen ([MAEDA 76], [PIERREHUMBERT 79], [VAISSIÈRRE 83], [ANTONIADIS 84b]). Dies wird mit *Deklination* bezeichnet. Die genaue Form des Abfalls und die Art der Berechnung ist stark umstritten (siehe die sehr emotional geführte Diskussion im JASA [LIEBERMAN 85a, 85b, 86], [REPP 85], ['T HART 86]).

  Am Ende einer Phrase oder einer Äußerung besteht die Tendenz, die Vokale des letzten Wortes und insbesondere den letzten Silbenkern zu dehnen ([LEHISTE 72], [KLATT 76], [KOHLER 82, 83]).

  Die Intensität nimmt zum Ende einer Äußerung hin ab ([ADAMS 78]).
- Die F_0 hängt vom Geschlecht und der Altersstufe des Sprechers ab ([LEHISTE 70]). Der Grund dafür liegt in der unterschiedlichen Länge der Stimmbänder. Daher ist die F_0 bei Kindern am höchsten, niedriger bei Frauen und am niedrigsten bei Männern.

  Es gibt allerdings auch Untersuchungen, die zumindest einen Teil des geschlechtsspezifischen F_0-Unterschieds mit gesellschaftlichen Verhältnissen erklären (siehe die bei [SPENDER 80, S.38-41] zitierten Untersuchungen).

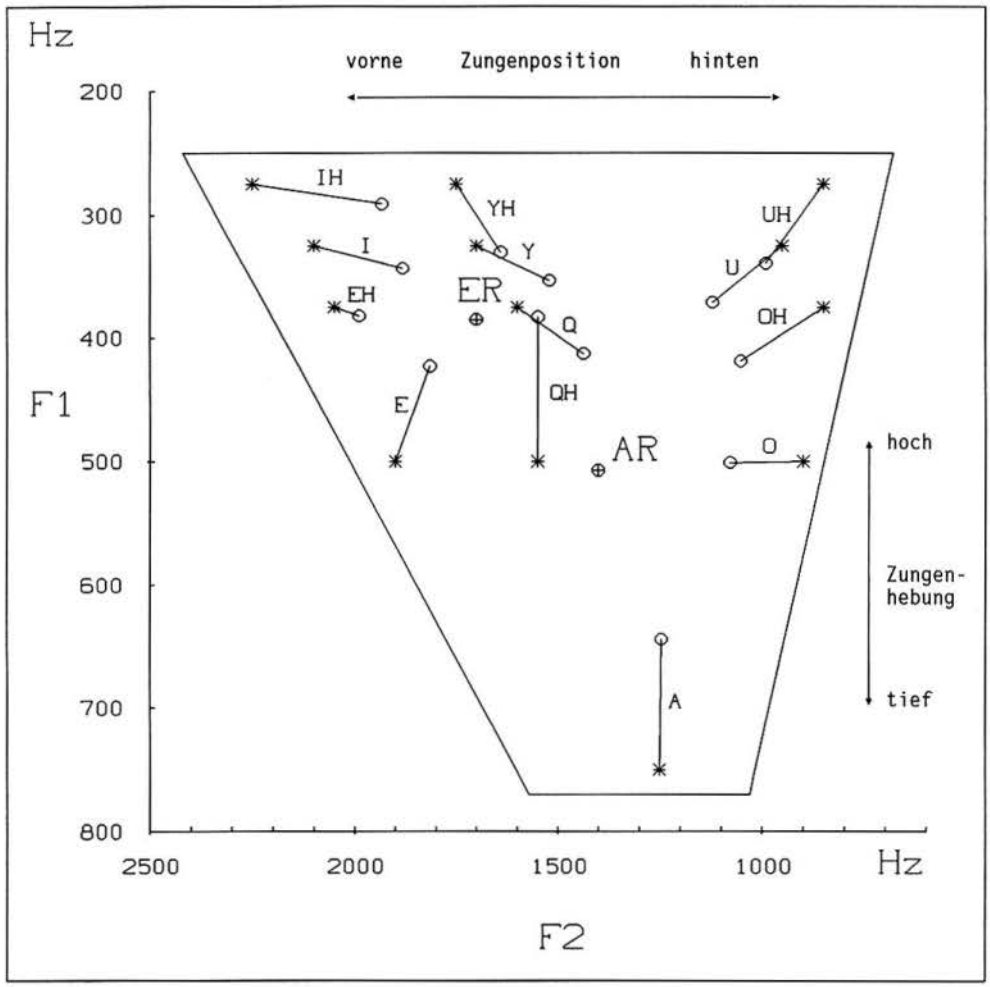


Bild 2.1 **Vokalviereck des Deutschen.** In X-Richtung ist die Position des zweiten Formanten bzw. die horizontale Zungenstellung dargestellt, in Y-Richtung die Position des ersten Formanten sowie die Zungenhöhe.

*: Mittelwerte für isoliert gesprochene Vokale nach [DELATTRE 65]
O: Mittelwerte für Vokale in kontinuierlich gesprochener Sprache nach [NÖTH 85]
⊕: Mittelwerte für Zentralvokal (/ER/) und abgeschwächtes A (/AR/) nach [NÖTH 85]

Man erkennt deutlich die Tendenz der Verschiebung zum Zentralvokal, sowohl für den Übergang von isoliert nach kontinuierlich gesprochenen Vokalen, als auch für den Übergang von geschlossenen (≙ langen) nach offenen (≙ kurzen) Vokalen. (Die geschlossenen Vokale sind durch ein H gekennzeichnet, siehe Anhang A.)

2.7.2 Das Akzentuierungsmittel Tonhöhe

Von vielen Autoren wird die Tonhöhe bzw. das akustische Korrelat Grundfrequenz als wichtigstes Akzentuierungsmittel genannt[2]. Es besteht die Tendenz, daß eine *höhere* Tonhöhe mit Betontheit assoziiert wird. Ist die Tonhöhenveränderung deutlich wahrnehmbar, so scheint eine weitere Vergrößerung des Tonhöhenunterschiedes keine Auswirkungen auf die Perzeption zu haben:

> "Change in fundamental frequency differs from change of duration and intensity in that it tends to produce an all-or-none effect, that is to say the magnitude of the frequency change seems to be relatively unimportant while the fact that a frequency change has taken place is all-important. The experiments ... show that a higher syllable is more likely to be percieved as stressed; the experiments with more complex patterns of fundamental frequency change suggest that sentence intonation is an over-riding factor in determining the perception of stress and that in this sense the fundamental frequency may outweigh the duration cue."
>
> [FRY 58, S.151]

Frys Ergebnisse sind unter anderem konsistent mit [MORTON 65], [ISAČENKO 66] und [LIEBERMAN 60] (beispielsweise wiesen 90 Prozent der als betont perzipierten Silben in zweisilbigen Wörtern bei Lieberman eine höhere Grundfrequenz auf). In [LEHISTE 70] wird der Zusammenhang zwischen höherer F_0 und Betonung vom Standpunkt der Produktion her erklärt:

> "Stress, on the other hand, is frequently associated with higher fundamental frequency. The conditioning seems to proceed from stress to fundamental frequency. As was discussed above, one of the factors causing the rate of vocal fold vibration to increase is increased rate of airflow. Since stress has been shown to be associated with increased subglottal pressure, the increase in vocal fold vibration may be considered automatic."
>
> [LEHISTE 70, S.82]

Dieses Zitat kann aber höchstens erklären, warum an betonten Stellen öfter eine steigende als eine fallende Grundfrequenzbewegung zu beobachten ist (für das Deutsche siehe z.B. [MAACK 37], [BLEAKLEY 73], [WINKLER 73]). Betrachtet man das Untersuchungsmaterial für viele der Untersuchungen zur Betonungswahrnehmung, so stellt man fest, daß nur maximal eine betonte Stelle pro Stimulus vorhanden ist. In einer solchen Situation ist es sicherlich die "natürliche" Art der Produktion, die betonte Stelle durch erhöhte Grundfrequenz hervorzuheben bzw. durch erhöhten subglotttalen Druck, welcher die höhere F_0 nach sich zieht. Folgen mehrere betonte Stellen aufeinander, so kann die F_0 jedoch nicht beliebig weiter steigen. Nun kann die Grundfrequenz an den dazwischen liegenden, unbetonten und mit niedrigerem subglottalen Druck produzierten Stellen wieder langsam absinken, so daß durch die langsame Veränderung über den unbetonten Stellen nicht der Eindruck einer Hervorhebung hervorgerufen wird.

Allerdings kann der Mensch die Grundfrequenz auch unabhängig vom subglottalen Druck kontrollieren (man denke nur an die bekannte Tatsache, daß die meisten Menschen mit Kindern

---

[2] Im Rahmen dieser Arbeit werden die Begriffe *Markierung des Akzents* und *Akzentuierung* synonym verwendet.

in einer höheren Stimmlage sprechen als mit Erwachsenen). In Abhängigkeit von der Sprechgeschwindigkeit, der Stilebene und der persönlichen Sprechweise kann der Sprecher nach *Erhöhung* der Grundfrequenz (zum Zwecke der Markierung einer Stelle der Äußerung) mit der Stimme "oben" bleiben, um an der nächsten zu akzentuierenden Stelle die Hervorhebung durch *Erniedrigung* der Grundfrequenz hervorzurufen (Brückenakzent bei [WUNDERLICH 88], Hutkontur bei [COHEN 67]). In einigen Fällen ist auch eine nochmalige Erhöhung der Grundfrequenz zu beobachten (siehe die Transkription im Anhang von [ROYÉ 83]).

In [ROYÉ 83] werden die Grundfrequenzverläufe von fünf Sprechern aus einer 45-minütigen Fernseh-Diskussion untersucht. Royé unterteilt den Grundfrequenzbereich sprecherabhängig in vier Bereiche (Stufen). Die intonatorisch relevanten Grundfrequenzveränderungen, also u.a. die mit der prosodischen Eigenschaft Tonhöhe markierten Akzentstellen, beschreibt er in Anlehnung an [STOCK 76] durch ein Liniensystem und unterscheidet zwischen Sprüngen (Tonbruch) und gleitenden Übergängen (Schleifton). Bei dieser Vorgehensweise ergeben sich kombinatorisch folgende Beschreibungsmöglichkeiten, die Royé *Stimmführungsfiguren* nennt:

- einstufig vs. mehrstufig
- in eine Richtung verlaufend vs. gegenläufig (steigend / fallend vs. steigend-fallend / fallend-steigend bzw. einfache vs. doppelte Stimmführungsfigur)
- Vorakzent- vs. Nachakzentfigur (bei einfachen Stimmführungsfiguren)
- Schleifton vs. Tonbruch.

Bild 2.2 zeigt die im Royé-Korpus beobachteten Stimmführungsfiguren. Bei den leeren bzw. mit einem Strich markierten Kästchen handelt es sich um Stimmführungsfiguren, die im Untersuchungsmaterial nicht auftreten.

In dem untersuchten Material überwiegt die Akzentmarkierung durch die Tonhöhe (93 Prozent der Fälle, in 70 Prozent ausschließliche Markierung durch die Tonhöhe).

In [BATLINER 89c] wird gezeigt, daß die Tonhöhe das wichtigste Betonungsmittel ist, und daß eine gesonderte Behandlung von Fragen und Nicht-Fragen die (statistisch gemessene) Relevanz der Tonhöhenkorrelate verbessert (siehe Kap.2.10).

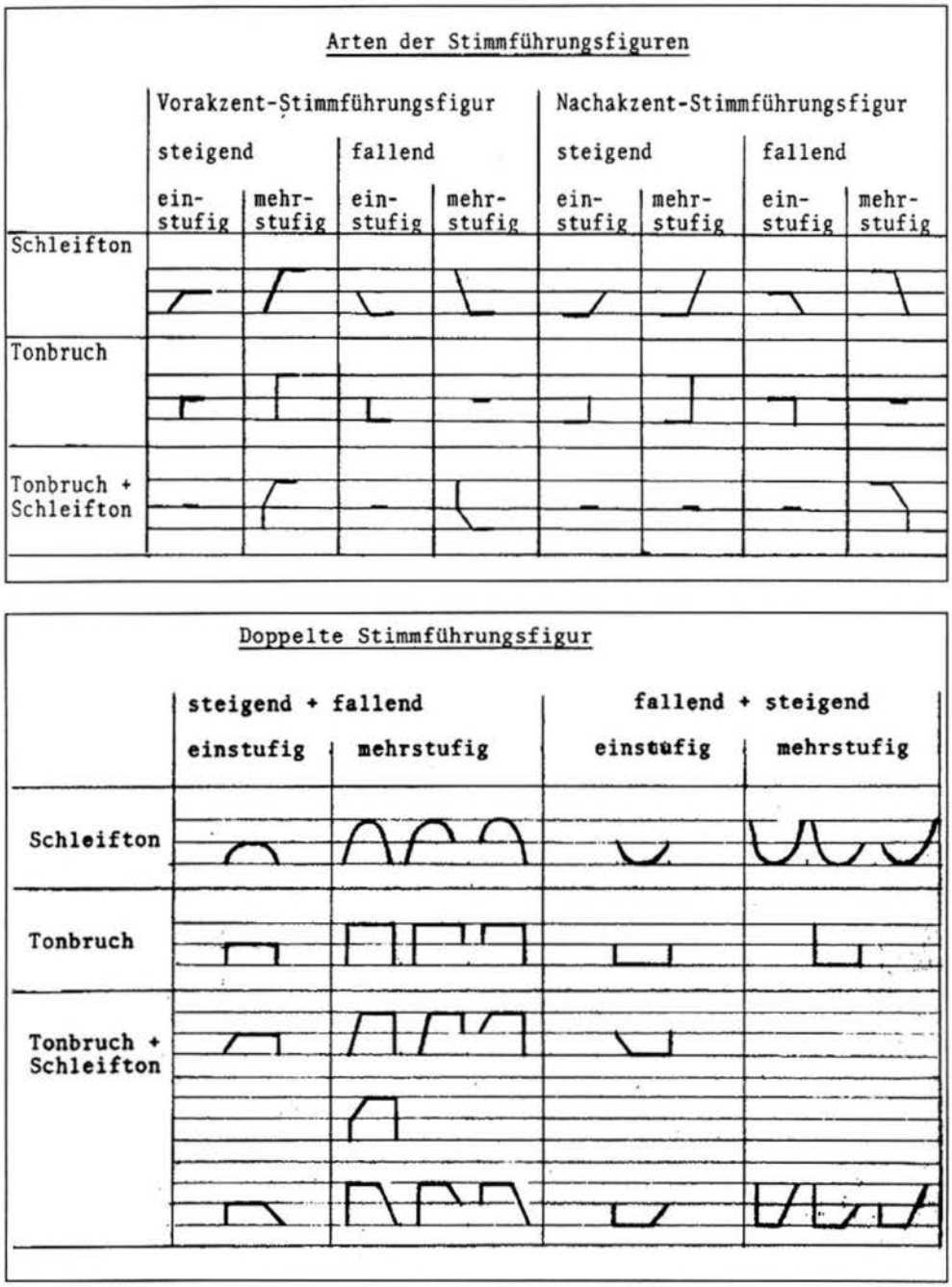


Bild 2.2 Arten der einfachen und doppelten Stimmführungsfiguren (aus [ROYÉ 83, S.144-145])

2.7.3 Das Akzentuierungsmittel zeitliche Strukturierung

Als zweitwichtigstes Akzentuierungsmittel wird in der Regel die zeitliche Strukturierung bzw. das akustische Korrelat Lautdauer angeführt (z.B. [FRY 55, 58], [BATLINER 89c]). In betonter Stellung sind Silben i.allg. länger als in unbetonter, wobei die Akzentuierung hauptsächlich die Dauer des Vokals beeinflußt ([KOHLER 82]). Das temporale intonatorische Akzentuierungsmittel wird hier als *zeitliche Strukturierung* bezeichnet, da eine Verkürzung der unbetonten Umgebung (in Verbindung mit reduzierter Klangfarbe) den Eindruck der Hervorhebung verstärken kann.

In [MAACK 49b] wird für die *"Sonanten"* des Deutschen (Vokale und Diphtonge) ein Lautdauer-Betonungsfaktor B als Maß für ihre relative Dehnung in betonter Stellung angegeben (die Mittelwerte beziehen sich auf das dort untersuchte Datenmaterial):

$$B_S = \frac{L_{S/}}{L_{SU}}$$

mit S : Sonant
$L_{S/}$: mittlere Lautdauer von S in betonter Silbe
L_{SU}: mittlere Lautdauer von S in unbetonter Silbe

Der Betonungsfaktor hängt von der Zungenstellung ab (kleiner bei hoher Zungenstellung) und von der Zugehörigkeit zur Klasse der Lang- und Kurzvokale (kleiner bei Kurzvokalen). Gedehnte Kurzvokale sind i.allg. kürzer als ungedehnte Langvokale. Weiterhin treten im Deutschen mehrsilbige Wörter mit der Silbenstruktur *Silbe mit Langvokal gefolgt von Wortakzentsilbe* so gut wie nie auf ([DUDEN 73, S.36]).

In [LIEBERMAN 60] wird festgestellt, daß 90 Prozent der betonten Silben eine höhere F_0 hatten, 92 Prozent ein höheres Amplituden-Integral, aber nur 66 Prozent eine höhere Dauer. In [AULL 84] wird jedoch darauf hingewiesen, daß Lieberman die *äußerungsfinale Dehnung* (siehe Kap.2.7.1 und Kap.2.9) offensichtlich nicht kompensiert hat, die nach [KLATT 75] für die letzte Silbe einer Äußerung durchschnittlich 30 Prozent beträgt. Aull selbst konnte zeigen, daß in ihrem Korpus nach einer Kompensierung der äußerungsfinalen Dehnung 90 Prozent der betonten Silben eine größere Dauer aufweisen.

In [MORTON 65] wird von einem interessanten Effekt im Zusammenhang mit dem akustischen Korrelat Dauer berichtet: Eine *Verkürzung* eines Vokals in zweisilbigen Logatomen führte bei einem Teil der Testpersonen konsistent zur Zuweisung der Betonung an die kürzere Silbe. Eine mögliche Erklärung könnte sein, daß ein enger Zusammenhang zwischen Dauer und Lautreduktion besteht ([LINDBLOM 63], siehe auch Kap.2.7.4), der vom Hörer bereits "einberechnet" und somit kompensiert wird. Da in [MORTON 65] mit synthetischem Material gearbeitet wurde, kann die Kombination von unreduzierten Vokalen und starker Verkürzung diesen Eindruck ausgelöst haben. Unabhängig davon, ob es sich bei dem beschriebenen Ergebnis um ein Artefakt des Versuchsaufbaus handelt, kann die Kombination von Dauer und Lautreduktion aber in Wörtern mit betonten Kurzvokalen eine Rolle spielen. Dies ist insbesondere dann der Fall, wenn der Kurzvokal

phonematische (bedeutungsunterscheidende) Funktion hat. So kann z.B. die Akzentsilbe des Wortes *bitten* in Satzakzentstellung nur sehr begrenzt gedehnt werden, da sich *bitten* von *bieten* hauptsächlich durch die Vokaldauer unterscheidet.

Die in [SCHMÖLZ 87, S.12, S.18] getroffenen Aussagen bezüglich des begrenzten Gesamteffekts einzelner intrinsischer Einflüsse (siehe auch [LEHISTE 72], [KLATT 73], [PORT 81]) sowie die Aussagen bezüglich der Tatsache, daß betonte Silben i.allg. länger als unbetonte sind (s.o.), treffen unter anderem dann nicht mehr zu, wenn es in fließender Sprache an Wortgrenzen mit Vokal-Vokal-Übergängen zu Silbenverschmelzungen kommt. Bild 2.3 zeigt das Zeitsignal für eine stark verschliffene Äußerung (gesprochen wurde *"Ja, hier ist ..."*) sowie eine enge Transkription nach dem erweiterten Erlanger Transkriptionssystem (siehe Kap.3.2.2 und Anhang C). Die Silbenkerne der beiden unbetonten und einsilbigen Wörter *hier* und *ist* werden zu einem überdurchschnittlich langen Vokal (171 Millisekunden) verschmolzen.

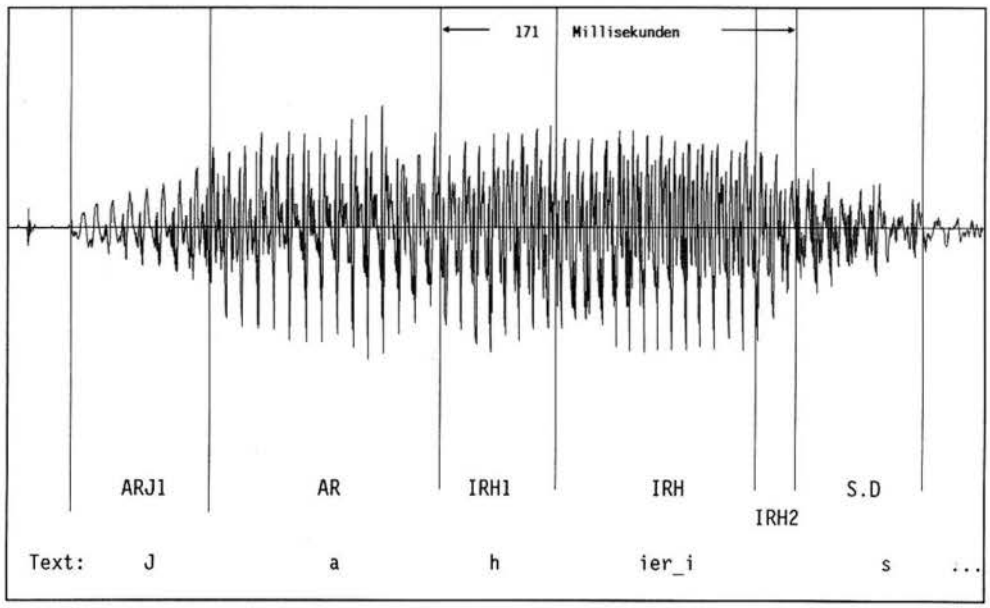


Bild 2.3: Stark verschliffene Realisierung der Äußerung *"Ja, hier ist ..."* mit Verschmelzung der beiden unbetonten Vokale der Wörter *"hier"* und *"ist"*.

2.7.4 Das Akzentuierungsmittel Lautheit

Die Untersuchungsergebnisse zum Akzentuierungsmittel Lautheit sind wohl am widersprüchlichsten (s.S.36). Hierfür können folgende Gründe aufgeführt werden (vergleiche auch Kap.4.):

- Wie bereits oben erwähnt, wird meistens ein aus dem Sprachsignal berechnetes Korrelat mit Hörerurteilen verglichen. Der Einfluß der intrinsischen Eigenschaften (siehe Kap.2.7.1) ist bei der Intensität als Lautheitskorrelat stärker als bei den Tonhöhen- und Dauerkorrelaten, da meßbare Energiewerte stark von den Abstrahlungseigenschaften der für jeden Vokal typischen Vokaltraktform abhängen.
- Im Gegensatz zu Tonhöhe und zeitlicher Strukturierung ist die Abbildung auf das entsprechende Korrelat nicht so eindeutig: Soll das Korrelat im Zeit- oder Frequenzbereich gemessen werden? Ist die Gesamtenergie oder ein Energieband relevant? Ist der Maximalwert, der Durchschnittswert oder der Integralwert zu nehmen?
- Für die Lautheit als Akzentuierungsmittel spricht, daß die automatische Extraktion von Lautheitskorrelaten nicht so fehleranfällig wie für Tonhöhen- und Dauerkorrelate ist. Interessant ist an dieser Stelle, daß gute Ergebnisse für die Bestimmung der richtigen Akzentstelle vor allem auch bei automatischer Parameterextraktion angeführt werden ([AULL 84], [WAIBEL 86]). Von mehreren Autoren (z.B. [LIEBERMAN 60], [BECKMANN 86]) werden für Korrelate, bei denen über den Silbenkernbereich integriert wird, bessere Ergebnisse angeführt als für Durchschnitts- oder Maximalwerte. Da betonte Silben tendenziell länger sind als unbetonte, werden durch die Integralbildung die betonten Silben bevorzugt.

Bild 2.4 zeigt die intrinsischen Energiewerte der Vokale in Abhängigkeit von ihrer Formantlage. Die Pfeile gehen vom Vokal mit der niedrigsten (/IH/) in steigender Reihenfolge zum Vokal mit der höchsten intrinsischen Energie (/O/). Die Graphik kann nur eine grobe Abschätzung der Realität sein, da die Formantlage für deutsche Vokale gilt (aus [NÖTH 85]) und die intrinsischen Energiewerte für die entsprechenden englischen Vokale gelten (die Zahlen hinter den Vokalklassen geben die intrinsische Energie in dB an, entnommen aus [LEHISTE 59]). In derselben Untersuchung, aus der die intrinsischen Werte entnommen sind, wurden Substantiv/Verb-Minimalpaare (z.B. *INcline* vs. *inCLINE)* unter denselben Aufnahmebedingungen und in derselben Satzakzent-Position gesprochen. Der Unterschied zwischen den Energiewerten der identischen Silben in betonter und unbetonter Position betrug im Mittel 2 dB und lag somit deutlich unter den intrinsischen Unterschieden von maximal 5.5 dB. Nach Abzug des jeweiligen intrinsischen Wertes konnte die Position des Wortakzents aufgrund des Energiewertes in allen Fällen richtig bestimmt werden. Lehiste und Peterson kommen zu der Schlußfolgerung, daß der Mensch während der Perzeption die intrinsischen Unterschiede zurückrechnen kann:

"These observations suggest that the listener associates a certain intrinsic relative amplitude (or perhaps average power) with each vowel spectrum, and applies a corresponding "correction factor" to the incoming signal. Assuming that duration and fundamental voice frequency are held constant, this procedure would enable a listener to identify a stressed syllable, even if the average or peak power of that syllable were less than that of an adjacent unstressed syllable containing a more open vowel. If such perceptual corrections are made, an automatic device for identifying linguistically stressed syllables should contain a set of such built-in correction factors."

[LEHISTE 59, S.429]

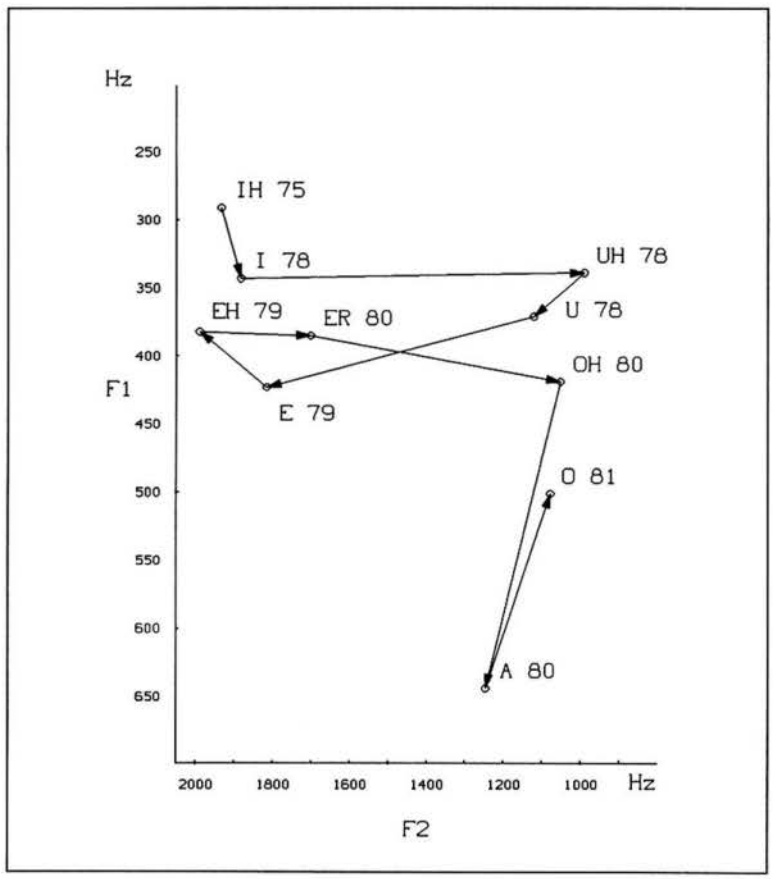


Bild 2.4: Intrinsische Energiewerte für Vokale in Abhängigkeit von ihrer Formantlage bzw. der Zungenposition. In Pfeilrichtung nimmt die Energie nicht ab (Formantwerte aus [NÖTH 85], Energiewerte aus [LEHISTE 59]).

2.7.5 Das Akzentuierungsmittel Klangfarbe

In der soeben erwähnten Untersuchung ([LEHISTE 59], siehe auch [LEHISTE 70, S.120-122]) wurde neben der Vokalenergie auch die *spektrale Struktur* oder *Vokalqualität* als akustisches Korrelat der Klangfarbe untersucht. Unter 15 Substantiv/Verb-Minimalpaaren war nur eines (***PER**vert* vs. *per**VERT***), in dem die Vokalqualität der Silbenkerne beider Silben und beider Produktionen relativ konstant blieb. Dies ist nicht überraschend, da die Tendenz besteht, daß sich die Vokalqualität in unbetonter Stellung der Qualität des Zentral-Vokals (Schwa-Laut) annähert ([LEHISTE 70, S.139], siehe auch [STETSON 51], [TIFFANY 59]). Zu den beiden Zentralisierungstendenzen, die in Bild 2.1 dargestellt wurden (isoliert gesprochen → kontinuierlich gesprochen und Langvokal → Kurzvokal), kommt also noch eine dritte (betont gesprochen → unbetont gesprochen) hinzu. Es besteht ein offensichtlicher Zusammenhang zwischen der Intensität und Dauer, mit der ein Silbenkern produziert wird, und dem Grad der Reduktion ([LINDNER 69, 81]). In [LINDBLOM 63] wird die Reduktion hauptsächlich auf die Dauerverhältnisse zurückgeführt und von der artikulatorischen Seite her erklärt: Die Vokalposition wird als Ziel (target) gesehen, das in unbetonter Position nicht erreicht wird (undershoot).

Über die Zentralisierung hinaus kann es auch zur Elision des Silbenkerns bzw. der Silbe kommen. Vor Nasalen und /L/ kann der Schwa-Laut vollständig elidiert werden:

/R.EH:.D.ERN./ → /R.EH:.D.NE/ (reden)

Insbesondere bei Formwörtern kann, je nach Sprechstil und Sprechsituation, eine fortschreitende Reduktion zur Löschung des Wortes führen. In [TILLMANN 80, S.99-102] werden die fortschreitenden Reduktionsstufen anhand des Wortes *"und"* im Zahlwort *"99"* demonstriert:

Vokal → entrundeter Vokal → Schwa → silbischer Nasal → Silbenelision

/N.OYN.+.U.N.T.N.OYN.TSI.XI/ → ... → /N.OYN.OYN.TSI.XI/ (neunundneunzig)

Die Aussprache der am stärksten reduzierten Form ist mit der Aussprache von 9,90 (z.B. bei Preisangaben) identisch. Die Zweideutigkeit ist nur durch den Kontext auflösbar, bzw. die am stärksten reduzierte Form wird normalerweise nur benutzt, wenn die Interpretation vom Kontext her eindeutig ist. Tillmann bezeichnet die verschiedenen Realisierungsformen als *alphabetisch explizite* und *im Kontext reduzierte Formen* (zur Beschreibung von Reduktionserscheinungen siehe auch [KOHLER 77, Kap.6.3] und [MÜHLFELD 86]).

Die zunehmende Reduktion ist ein Indikator für zunehmend *unbetont* produzierte Silben und führt gleichzeitig zum verstärkten Eindruck der Betonung für die nicht reduziert produzierten Nachbarsilben.

In [AULL 84] wird die Vokalqualität als prosodisches Merkmal für die automatische Akzentberechnung insoweit ausgenutzt, als für einen Silbenkern ein Maß für die Veränderung im Frequenzspektrum berechnet wird (spectral change). Diese Veränderung ist in betonten Silben geringer, da im Sinne von Lindbloms artikulatorischer Theorie eine stabile Zielposition länger konstant eingehalten wird.

2.8 Intonatorische Markierung des Satzmodus

Die intonatorische Markierung des Satzmodus geschieht mit der prosodischen Eigenschaft Tonhöhe, wobei die Eigenschaften zeitliche Strukturierung und Lautheit eine sekundäre Rolle spielen (z.B. bei der Markierung des Exklamativs, siehe [NÖTH 87], [BATLINER 88a]; man denke auch an den Gebrauch von Lautstärke beim Erteilen von Befehlen an Dienstuntergebene im Militär, also der Markierung des Satzmodus *Imperativ* in einer sehr speziellen Kommunikationssituation). Zur Betrachtung der intonatorischen Markierung des Satzmodus ist eine weitere Unterscheidung der Satzmodi *Frage, Aussage, Imperativ, Wunsch* und *Exklamativ* notwendig. Eine solche Differenzierung bietet das Satzmodus-System nach [ALTMANN 84, 87]. In [BATLINER 89g] (siehe auch Kap.3.4) finden sich Beispielsätze für intonatorische Minimalpaare, d.h. Äußerungen, deren Satzmodus nur aufgrund der intonatorischen Markierung bestimmbar ist.

Die intonatorische Markierung des Modus kann geringer ausgeprägt oder fakultativ sein, wenn andere (grammatische) Merkmale wie z.B. die Verb-Position den Modus ganz oder teilweise bestimmen. Die genaue Form der Modus-Markierung kann i.allg. nur unter Einbezug der Position des Fokus beschrieben werden. So kann eine Entscheidungsfrage *nicht, in der fokussierten Phrase und/oder am Satzende* markiert werden (siehe Kap.2.10).

Es existieren sehr viele unterschiedliche Ansätze zur Beschreibung des Tonhöhenverlaufs, die von impressionistischen, schwer nachvollziehbaren perzeptiven bis zu sehr oberflächengetreuen und auf instrumentell meßbaren Kurven beruhenden meßphonetischen Beschreibungsformen reichen. In [OPPENRIEDER 88a] werden satzmodustypische intonatorische *Prototypen* präsentiert (zur Modellierung der Intonation mit dem Prototypenkonzept siehe auch [BATLINER 89c] und Kap.7). Das Modell beruht auf der Annahme, daß für jede mögliche Kombination von Satzmodus und Fokus eine oder mehrere Standard-Realisierungen existieren, die als *Prototypen* bezeichnet werden. Die intonatorische Form einer konkreten Äußerung entspricht dann mehr oder weniger genau einem dieser Prototypen. Die Prototypen werden aus Perzeptionsexperimenten und aus statistischen Größen von akustischen Parameterwerten gewonnen. Diese Parameter (z.B. der F_0-Wert am Ende der Äußerung) sollen den Verlauf der Korrelate beschreiben.

Im Tonsequenzansatz ([PIERREHUMBERT 80], für Übertragungen auf das Deutsche siehe z.B. [FÉRY 88], [UHMANN 88] und [WUNDERLICH 88]) werden lokal bestimmte Töne fortlaufend verkettet. Dabei werden die möglichen Töne *Hochton* und *Tiefton* weiter kategorisiert, z.B. als *Grenzton.*

Ein kritischer historischer Überblick bis zum Erscheinungsjahr über die verschiedenen Beschreibungsformen, speziell für das Deutsche, findet sich in [KLEIN 80, 82].

In [BATLINER 88c, 89d, 89e, 89f] werden unterschiedliche aus der F_0-Kontur abgeleitete Merkmale (z.B. die Position und Höhe des Grundfrequenzgipfels) auf ihre Relevanz für die Satzmodusmarkierung hin untersucht (zur Rolle der Position des F_0-Gipfels siehe auch [KOHLER 87]).

Im Zusammenhang mit der ASE ist insbesondere die Unterscheidung von *Fragen* und *Nicht-Fragen* von Interesse (siehe Kap.1.2). In [BATLINER 89a] (siehe auch Kap.6.2) wird an einem großen Korpus untersucht, inwieweit der F_0*-Wert am Äußerungsende (Offset)* als Frage/Nicht-Frage-Indikator ausreicht, wie dieser Wert aussehen soll und ob es specherspezifische Strategien beim Einsatz dieses Parameters gibt. In dem dort untersuchten Material reicht der Offset in 87 Prozent der Fälle als Frage/Nicht-Frage Indikator aus. [WAIBEL 86] berichtet von einer Erkennungsrate von 78 Prozent.

In Kap.4.2.2 wird auf eine "Tonhöhenbewegung" genauer eingegangen, die u.a. bei der Satzmodus-Markierung eingesetzt werden kann: die *Laryngalisierung* ([LEHISTE 70, S.60]). Es handelt sich dabei um einen Sprung in eine äußerst *tiefe Stimmlage* (vocal fry), der außer an Wortgrenzen vor allem am Ende von Aussagesätzen beobachtet werden kann und dort sowohl als Grenzsignal als auch zur Markierung des Satzmodus *Aussage* benutzt wird.

2.9 Intonatorische Gliederungsmittel

Die gliedernde Rolle der Intonation hängt sehr stark von der Stilebene der Äußerung ab. Als intonatorische Gliederungsmittel gelten insbesondere die Tonhöhe, die zeitliche Strukturierung und die gezielte Pausensetzung. In Kapitel 2.7.1 wurde die Stellung der Silbe in einer Äußerung als Einflußfaktor auf das Akzentkorrelat Silbendauer genannt. Im Grunde genommen ist dieser Einflußfaktor natürlich nichts anderes als eine Überlagerung der beiden Rollen *Akzentuierung* und *Gliederung* der Intonation. Die satz- bzw. phrasenfinale Dehnung markiert ja gerade das Ende einer "Sinneinheit".

In [ROYÉ 83] wird u.a. die intonatorische Gliederung mit den prosodischen Eigenschaften Tonhöhe und Pausensetzung untersucht. Im Sinne von [GLINZ 79] versteht Royé den Begriff *Proposition* als eine *grammatisch-strukturelle (morphosyntaktische) Gliederungseinheit, als*

> "das Stück Text, das von einem Verb aus strukturiert ist (auf einer und nur einer verbalen Wortkette beruht) oder das als eigene Einheit neben solchen vom Verb aus strukturierten Einheiten steht, wie z.B. eine Anrede, eine Grußformel, ein Ausdruck der Bejahung oder Verneinung ohne Verb ("ja - nein - auf keinen Fall - vielleicht") oder auch eine Überschrift"
> [GLINZ 79, S.180-181], zitiert in [ROYÉ, S.43]

Royé unterteilt die von ihm untersuchte Fernsehdiskussion in Propositionen und untersucht u.a. den Einsatz der *phonodischen Parameter Pause* und *Stimmführung. Phonodie* kann im Sinne dieser Arbeit ungefähr als *Intonation oberhalb der Wortakzentebene,* also *Satzakzent, Melodiebildung, Pausenbildung* verstanden werden, siehe [GLINZ 65]. Royés Untersuchungen belegen, daß die isolierte Betrachtung der Pause zum Zwecke der Gliederung in grammatische Einheiten kein einheitliches Bild ergibt und nur unter Einbezug der Stimmführung betrachtet werden sollte (siehe Kap.2.6 und [ROYÉ 83, S.90-93] zur Funktion der Pause bei der Formulierungskorrektur und Wortfindung, die eben nicht an grammatischen Grenzen, sondern u.U. mitten in einem Wort stattfindet). Es werden fünf verschiedene Stimmführungsarten unterschieden, und zwar die *Stimmhebung,* die *Stimmschwebe* und drei Stufen der *Stimmsenkung:*

> "Die Pausen allein sind aber für die phonodische Gliederung nicht aussagekräftig genug, denn an den Pausenstellen entscheidet die Stimmführung durch ihre Grundstrukturen (Stimmhebung bis volle Stimmsenkung), ob eine Pause (etwa mit Stimmschwebe) spannungsvoll eine mittlere Einheit mit der nächsten verbindet oder mit voller Stimmsenkung abschließt und von der nächsten trennt. Im ersten Fall kann eine höhere phonodische Einheit entstehen, indem 2 bis x mittlere Einheiten durch Pausen mit den entsprechenden Grundstrukturen der Stimmführung aneinandergefügt und erst am Ende der letzten mittleren Einheit durch Pause mit voller Stimmsenkung abgeschlossen sind."
> [ROYÉ 83, S.100]

Im Material von Royé ist an jedem dritten Propositionsende eine Pause zu beobachten, wobei diese Zahl auch für Grenzstellen zu eingeschobenen Propositionen gilt. Nur 44 Prozent der Pausen fallen mit einem Propositionsende zusammen. Innerhalb der Propositionen können Pausen an jeder

Stelle beobachtet werden, wobei 73 Prozent der Pausen an Satzglied- oder Ausdrucksgrenzen auftreten, die gliedernde Rolle also überwiegt.

Betrachtet man die Stimmführung an Propositionsenden, so stellt man fest, daß 50 Prozent der Stimmfiguren im oberen Stimmbereich bleiben (Hebung, Schwebe oder Senkung in Stufe 3 und 4), was einer weiterweisenden Funktion entspricht. An Pausenstellen innerhalb von Propositionen bleiben rund 40 Prozent im oberen Stimmbereich, in rund 20 Prozent der Fälle ist eine volle Stimmsenkung zu beobachten. Die Funktion der Lautdehnung zum Zwecke der Gliederung wird von Royé leider nicht untersucht.

Bild 2.5 zeigt zwei Beispiele für Stimmführungsfiguren mit Pausen aus [ROYÉ 83]. Um dem Leser einen Eindruck über die Dauerverhältnisse zu vermitteln, ist auch der dazugehörige Meßstreifen wiedergegeben. Für die beiden Beispiele ist von oben nach unten zu sehen:

- Zeitmarken im 20 Millisekunden-Abstand
- Oszillogramm des Zeitsignals
- Schallpegel (in dB)
- Schalldruck
- Tonhöhenkurve (logarithmisch von 65 Hz bis 522 Hz; die beiden weißen Linien in der Mitte entsprechen 130 und 261 Hz)
- Transkription von Royé mit Stimmführungsfiguren (durchgezogene Linien) und dazwischenliegende neutrale Zonen (gestrichelte Linien), Markierung von Lautdehnung bei allen Silben, die länger als 0,2 Sek sind (:), und Markierung von Pausen und Atmung (/, //, + ; s. S.28).

Das erste Beispiel (die Pause zwischen *DIE ZEIT* und *Rechtsanwalt*) stammt aus einer unterbrochenen Proposition, von der im folgenden der zweite Teil wiedergegeben wird:

"Hans Schue:ler von der Wochenzeitung DIE ZEI:T / Rechtsanwalt und Journalist //+ und // Joachim Sobotta / den Chefredakteur der RHEINISCHEN POST"

Proposition AAA 32b aus [ROYÉ 83, S.186-187]

Es handelt sich um eine Kurzpause mit Stimmschwebe (weiterweisende Funktion) zur Abgrenzung der nachfolgenden Apposition *"Rechtsanwalt und Journalist"*. Man beachte die Dehnung des Silbenkerns von *ZEIT*, der fast doppelt so lang ist wie der Silbenkern in Wochen*zei*tung. Zweifellos handelt es sich hier um eine Dehnung zum Zwecke der Betonung **und** der Gliederung.

Das zweite Beispiel (die Pause zwischen *mutmaßlich* und *negative*) ist ein Beispiel für eine Sprechpause ohne gliedernde Funktion. Sie ist der folgenden Proposition entnommen:

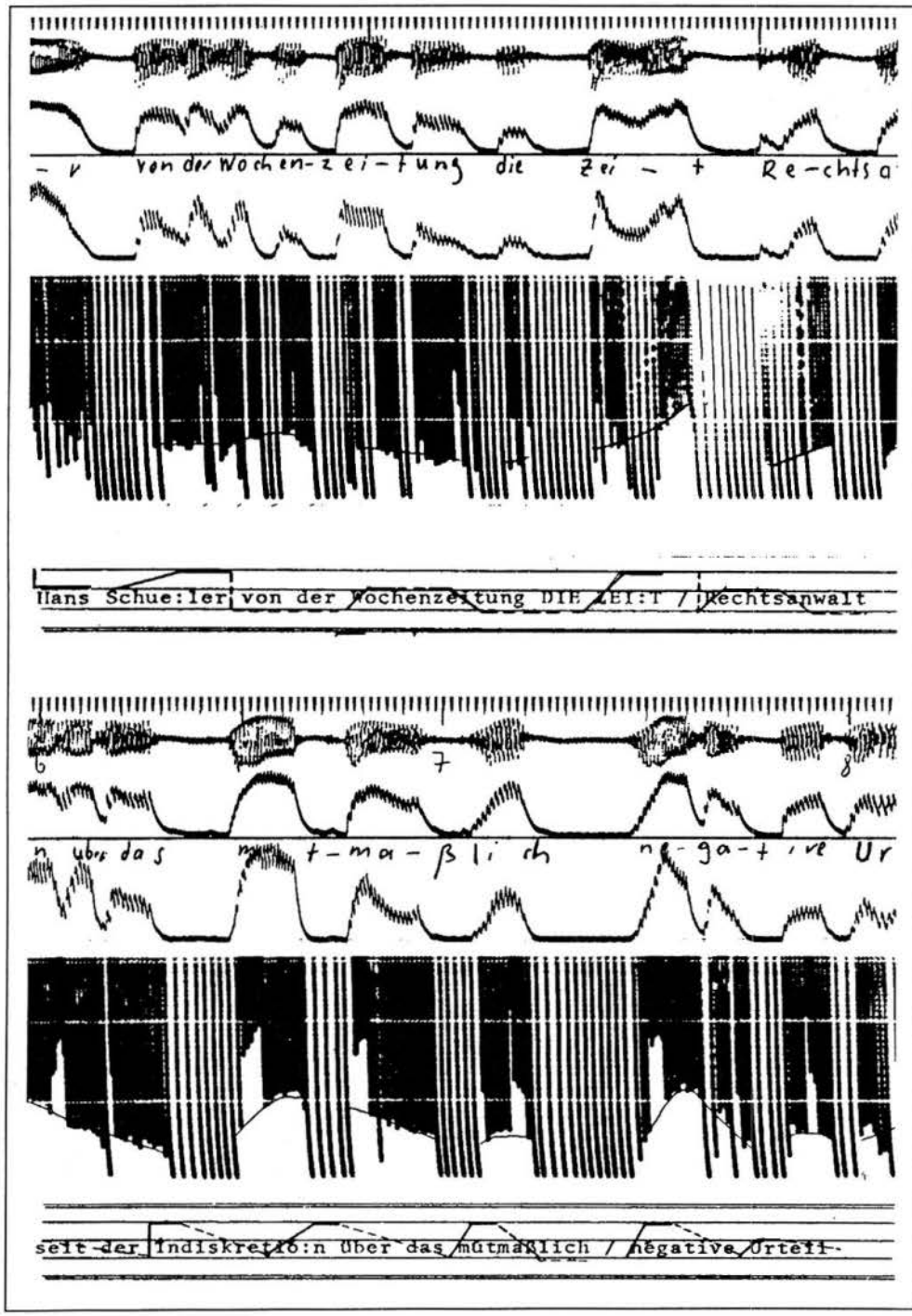
"seit der Indiskretio:n über das mutmaßlich / negative Urteil seines ersten Senats zur Frage der Straffreiheit der Abtreibung in den ersten drei Monaten //+ ist das höchste deutsche Gericht wieder einmal //+ Ziel verhaltener / und offener Kritik auch vor allem von Politikern geworden //"

Proposition AAA 5 aus [ROYÉ 83, S.183]

Obwohl die Pause ca. 40 Prozent länger ist als im ersten Beispiel und mit einer Stimmsenkung zur Stufe zwei verbunden ist, dient sie nicht der Gliederung, sondern dem Spannungsaufbau zum Zwecke der besonderen Hervorhebung des nachfolgenden Wortes *"negative"*.

In [VAISSIÈRE 88] wird ein Überblick über Untersuchungen zum automatischen Finden syntaktischer Grenzen gegeben. Im Gegensatz zu Leas Untersuchungen, in dessen Material 94 Prozent der syntaktischen Grenzen durch eine fallend-steigende F_0-Kontur markiert sind [LEA 72, zitiert in VAISSIÈRE 88], belegen Vaissières eigene Untersuchungen ([VAISSIÈRE 82]), daß die phrasenfinale Lautdehnung der zuverlässigste Indikator für die "Hauptgrenze" (main boundary) in Sätzen ist, und daß die gedehnte Silbe entweder durch eine steigende, weiterweisende F_0-Kontur (continuation rise) oder einen hohen F_0-Wert markiert ist. Eine mögliche Erklärung für die unterschiedlichen Ergebnisse ist die Tatsache, daß Leas Ergebnisse auf englischem Sprachmaterial basieren und Vaissières auf französischem. Hier wurden also eventuell sprachabhängige intonatorische Phänomene betrachtet.

Bild 2.5 (nächste Seite): Zwei Beispiele für Pausen mit Stimmführung innerhalb von Propositionen (aus [ROYÉ 83], Erklärung siehe Text).


- v von der Wochen-zei-tung die zei - t Re-chts a
Hans Schue:ler von der Wochenzeitung DIE ZEI:T / Rechtsanwalt
b 7 8
n übrs das mut-ma-ßli ch ne-ga-ti ve Ur
seit der Indiskretio:n über das mutmaßlich / negative Urteil

2.10 Intonation als Komplexphänomen

Bei der Besprechung der Mittel zur intonatorischen Markierung von Akzent und Satzmodus wurde immer wieder deutlich, daß es sich in allen drei Rollen der Intonation um ein Komplexphänomen handelt. In den häufigsten Fällen spielen mehrere prosodische Eigenschaften zusammen, sei es als unabhängige Variable, die durch ihre gleichzeitige Benutzung die jeweilige funktionale Rolle der Intonation verstärken (z.B. die gleichzeitige Markierung einer Akzentstelle durch die Tonhöhe und die Lautheit), sei es als mehr oder weniger stark abhängige Variable (z.B. die Klangfarbe als Kovariable der zeitlichen Strukturierung), oder sei es, daß erst das gemeinsame Auftreten mehrerer Eigenschaften den vom Sprecher intendierten Eindruck bewirkt (z.B. die Gliederung einer Äußerung durch Pausen **und** Tonhöhenverlauf).

Da der Hörer normalerweise nicht in der Lage ist, zu begründen, warum er an einer gewissen Stelle einer Äußerung eine *Betonung,* eine *Frage-Markierung* oder eine *Phrasengrenze* hört, beruhen Abschätzungen der Wichtigkeit der einzelnen Eigenschaften auf Arbeiten mit synthetischem Material und systematischem Variieren der akustischen Korrelate ([MORTON 65]), auf Häufigkeitszählungen ([ROYÉ 83]) sowie auf statistischen Korrelationsmaßen zwischen Parameterwerten und Hörerurteilen ([BATLINER 89c]).

Im folgenden wird auf eine Untersuchung eingegangen, die beispielhaft belegt, daß zur Untersuchung einer Funktion der Intonation die anderen Funktionen nicht ausgeschlossen werden dürfen, sondern daß gerade die Wechselbeziehungen zwischen den Funktionen mit einbezogen werden müssen.

In [BATLINER 89b, 89c] (siehe auch Kap.7) wird ein Korpus von 355 Äußerungen untersucht, die aufgrund der vorgegebenen Modus/Fokus-Konstellation in vier "Klassen" eingeteilt werden können:

Frage vs. *Nicht-Frage*

und

Fokus auf der letzten Phrase vs. *Fokus auf der vorletzten Phrase.*

Aufgrund der Konstruktion der Testsätze ist eine intonatorische Markierung des Satzmodus *Frage* obligatorisch. Die Bilder 2.6a - 2.6c zeigen die F_0-Kontur der letzten beiden Phrasen für verschiedene Realisierungen derselben Konstellation *Frage mit Fokus auf der vorletzten Phrase* von drei verschiedenen Sprechern. Bild 2.6d zeigt eine Realisierung der Konstellation "Nicht-Frage mit Fokus auf der vorletzten Phrase" vom selben Sprecher wie Bild 2.6a. Bei diesen Äußerungen herrschte in Perzeptionsexperimenten eine starke Übereinstimmung der Hörer in Bezug auf

- Position des Fokus
- Satzmodus
- hohe Natürlichkeit der Produktion.

Es handelt sich hier also im Sinne von [BATLINER 89c] um *Prototypen* der Modus/Fokus-Konstellation *Frage/Fokus auf vorletzter Phrase* bzw. *Nicht-Frage/Fokus auf vorletzter Phrase.* Der größtenteils produzierte Typ 2.6a wird dort als *Kern-Prototyp,* die selteneren, aber akzeptablen Typen 2.6b und 2.6c werden als *Rand-Prototypen* bezeichnet. Typ 2.6d ist der *Kern-Prototyp* der Konstellation *Nicht-Frage/Fokus auf vorletzter Phrase.*

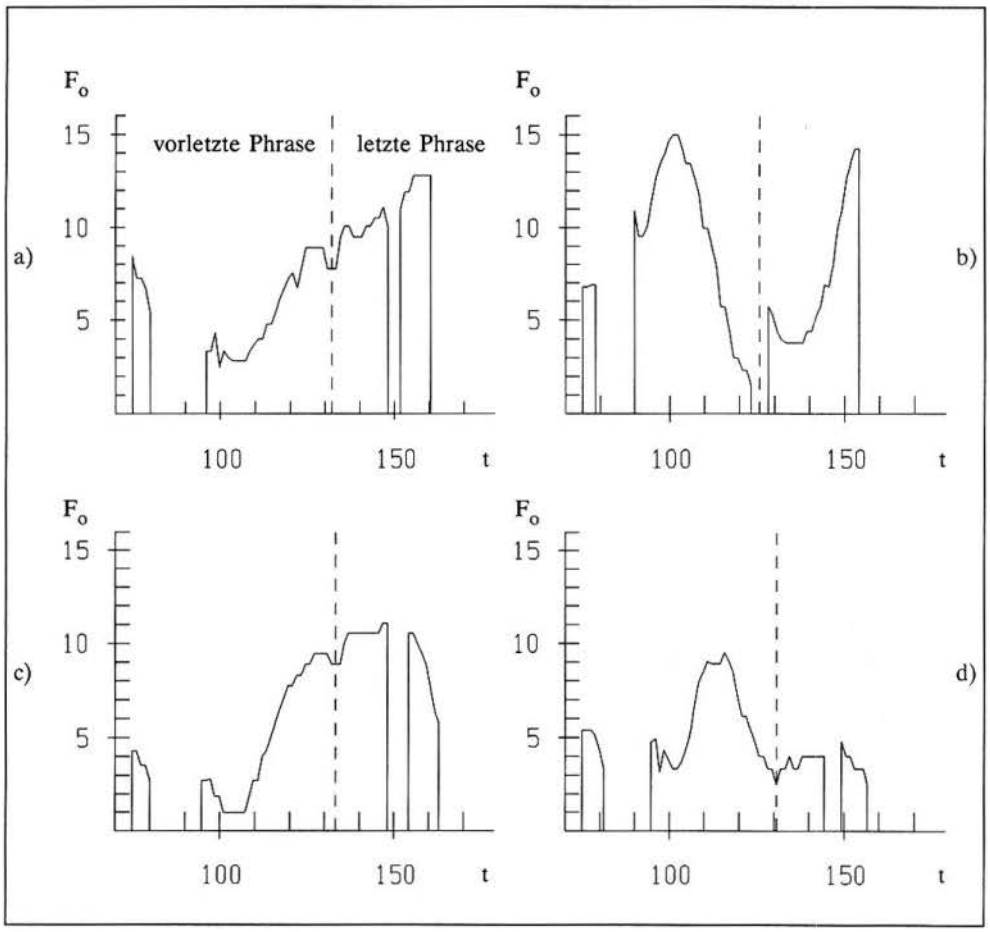


Bild 2.6: Grundfrequenzkonturen für vier Äußerungen mit Fokus auf der vorletzten Phrase. Die X-Achse zeigt die Zeit in Centisekunden vom Beginn der Äußerung, die Y-Achse die Grundfrequenz in Halbtönen über dem sprecherspezifischen Basiswert (siehe Kap.7).

		vorletzte Phrase	letzte Phrase
a)	Kern-Prototyp Frage/Fokus auf vorl. Phrase (Sprecher 5)	*das Leinen*	*weben*
b)	Rand-Prototyp Frage/Fokus auf vorl. Phrase (Sprecher 6)	*die Bohnen*	*schneiden*
c)	Rand-Prototyp Frage/Fokus auf vorl. Phrase (Sprecher 1)	*das Leinen*	*weben*
d)	Kern-Prototyp Nicht-Frage/Fokus auf vorl. Phrase (Sprecher 5)	*das Leinen*	*weben*

Während Sprecher 6 den Fokus auf der vorletzten Phrase bei Fragen (b) ähnlich wie bei Nicht-Fragen markiert (d), wenden die anderen Sprecher für die Markierung des Fokus bei Fragen (a, c) eine andere Strategie an als bei Nicht-Fragen (d). Die zusätzliche Markierung des Satzmodus in der letzten Phrase ist zwar üblich, aber fakultativ, so daß Sprecher 1 fast immer den Modus lediglich auf der fokussierten Phrase markiert (c).

Benutzt man aus dem F_0-Verlauf abgeleitete Merkmale wie *F_0-Maximum der i-ten Phrase und seine zeitliche Position* als Eingangsgrößen für einen statistischen Klassifikator zur Prädiktion der fokussierten Phrase, so verbessern sich die Ergebnisse, wenn man zuerst den Modus bestimmt und dann den Fokus.

Die Prototypen aus Bild 2.6 zeigen beispielhaft für die prosodische Eigenschaft *Tonhöhe* die Überlagerungen, die dadurch entstehen, daß sowohl der *Satzmodus* als auch der *Fokus* mit dieser prosodischen Eigenschaft markiert werden. Ähnlich wie die Beispiele, die in [KLEIN 82] angeführt werden, zeigen sie, daß bei isolierter Betrachtung *einer* funktionalen Rolle der Intonation *"fast alles geht"* ([KLEIN 82, S.295]). Die Tatsache, daß es sich um "prototypische" Realisierungen handelt, deutet allerdings auch an, daß in gewissen Kontext-Situationen (d.h. unter anderem bei Betrachtung der anderen Rollen der Intonation) *"nur manches geht"*. In dieser Arbeit soll versucht werden, sowohl das *"alles"* als auch das *"manche"* zu berücksichtigen:

- Bei der *datengetriebenen Betonungsbeschreibung* (Kap.5) muß davon ausgegangen werden, daß über das Gesprochene noch nichts bekannt ist (Existenz eines W-Fragewortes, syntaktische Struktur, usw.). Daher wird die Betonungsbeschreibung aufgrund von Merkmalen erstellt, welche in bezug auf die prosodische Eigenschaft *Tonhöhe* sowohl eine Erhöhung als auch eine Senkung der Grundfrequenz als Indiz für Betonung zulassen (das Merkmal `STG`, siehe Kap.4.2.5). Gleichzeitig wird die Tatsache berücksichtigt, daß die Markierung einer Betonung häufiger eine Erhöhung der Grundfrequenz zur Folge hat (das Merkmal `NIVFO`, siehe Kap.4.2.5). Durch den geplanten Einsatz der Betonungsbeschreibung während der Erkennungsphase eines sprachverstehenden Systems ist vorgegeben, daß weitere Wissensqellen für die Betonungsbeschreibung nicht verwendet werden können. Deren Nichtberücksichtigung führt zwangsläufig zu einer höheren Fehlerrate. Es wird sich aber zeigen (Kap.6.3 bis Kap.6.5), daß die Betonungsbeschreibung trotzdem zu einer Verbesserung der Worterkennungsleistung eingesetzt werden kann.
- In Kap.7 wird gezeigt, wie die intonatorische Markierung des Fokus unter Einbeziehung der hier erläuterten Überlagerungen durch die Erweiterung des in [BATLINER 89b,89c] entwickelten *intonatorischen Modells des Deutschen* in der Verstehensphase eines sprachverstehenden Modells eingesetzt werden kann.

3 Stichproben

Da bei der Beschreibung der Algorithmen in Kapitel 4 und 5 an verschiedenen Stellen Abbildungen von Äußerungen gezeigt werden, werden bereits hier die Sprachstichproben beschrieben, die für die durchgeführten Experimente (Kapitel 6) verwendet wurden. Neben Einzelheiten über die Aufnahmesituation für jedes Korpus wird auf die Motivation für ihre Erstellung, ihre Größen sowie die vorhandenen Handklassifikationen bezüglich der verschiedenen linguistischen Ebenen eingegangen.

3.1 Die EVAR- und die Pragmatik-Stichprobe

Im Rahmen der Erstellung des Akustik-Phonetik-Moduls [REGEL 88] wurden von 12 Sprechern (6 männlich, 6 weiblich) Sprachaufnahmen erstellt. Die Aufnahmen wurden in Büroräumen des IMMD 5 auf Band gesprochen (Philips N4415, 19 m/Sek.). Die Sprecher waren "naive" Versuchspersonen (Studenten der Informatik und Mitarbeiter des IMMD 5). Die Äußerungen wurden vom Blatt gelesen, wobei es sich fast ausschließlich um mögliche Fragen an ein Zugauskunftssystem und um Texte aus Prospekten der Deutschen Bundesbahn handelt. Zusätzlich wurden die Namen der Städte mit IC-Anschluß isoliert gesprochen. 250 dieser Äußerungen mit durchschnittlich 6 Wörtern liegen in digitalisierter Form vor. Die Äußerungen haben eine durchschnittliche Dauer von 2.5 Sekunden. Die Abtastrate beträgt 10 kHz. Das Sprachsignal ist bandpaßgefiltert (0.1-3.4 kHz; 48 dB/Oktave Flankensteilheit). Dieses Korpus wird im folgenden als *EVAR-Stichprobe* bezeichnet. Tabelle 3.1 zeigt einige typische Beispiele.

BD2121	"Wo muß ich umsteigen?"
BR2066	"Wechselstuben finden Sie in größeren Bahnhöfen."
BR2072	"Das GKA für Großkunden, Behörden und Verbände."
CI0018	"Bremen" "Hamburg-Altona" "Osnabrück" "Münster" "Essen" "Bochum" "Dortmund" "Hamm" "Bielefeld" "Minden"
EM5520	"Wir möchten am Wochenende nach Mainz fahren."
JA245F	"Guten Morgen, am Donnerstag Vormittag um sieben Uhr muß ich in München sein."
UT2001	"Abfahren, entspannen, ankommen, ausgeruht sein ..."
UT2019	"Die Welt ist schöner ohne Scheibenwischer."
UT2026	"Ihr Büro auf Schienen!"

Tab. 3.1: Beispielsätze aus der EVAR-Stichprobe.

Die zu den Äußerungen existierenden Handklassifikationen auf der Ebene der Lautkomponenten, Laute und Wörter sind in [REGEL 88] beschrieben. Die Länge einer Lautkomponente (Frame) wurde so festgelegt, daß sie einerseits kurz genug ist, um das Signal als quasistationär betrachten zu können, andererseits lang genug, um genügend lauttypische Information zu enthalten. Sie ist mit

12.8 Millisekunden deutlich kürzer als ein Laut, dessen durchschnittliche Länge in der EVAR-Stichprobe 83 Millisekunden beträgt. Bis auf die diskontinuierlichen Laute (Plosive und Affrikate) und die Gleitlaute (Diphthonge) stimmen die Lautkomponentenklassen mit Lautklassen überein. Die Plosive werden auf Lautkomponentenebene nach ihren Phasen *Verschluß*, *Burst* und (bei den Fortis-Plosiven) *Behauchung* transkribiert (z.B. zerfällt der Laut /P/ in die Lautkomponenten /PP/, /PB/ und /PH/). Die Affrikate werden als Plosiv und Frikativ transkribiert (z.B. zerfällt das /PF/ in die Komponenten /PP/, /PB/ und /F/). Bei den Gleitlauten existiert auf Lautkomponentenebene nur dann ein Label, wenn der Frame entweder dem Start- oder Ziellaut zugeordnet werden kann (z.B. /A/,/E/ bei /AI/). Ein Verzeichnis der verwendeten Laute und Lautkomponenten findet sich in Anhang A und B.

Zum Zeitpunkt der Digitalisierung der EVAR-Stichprobe war die Leistungsfähigkeit und Speicherkapazität des Rechners (nach heutigem Maßstab) sehr beschränkt. Somit kam es mehrfach vor, daß die auf Analogband gesprochene Äußerung länger war als der bereitgestellte A/D-Puffer. Da die Stichprobe in erster Linie für die Dimensionierung des Akustik-Phonetik-Moduls und des Worthypothesen-Moduls digitalisiert wurde, wurden auch diese unvollständigen Aufnahmen verwendet. Im Rahmen eines Inter-City-Auskunftssystems können allerdings die meisten der unvollständigen Äußerungen und einige der DB-Werbetexte aus der EVAR-Stichprobe (z.B. UT2019) pragmatisch nicht analysiert werden. Daher wurden im Rahmen der Arbeiten am Pragmatik-Modul aus der EVAR-Stichprobe diejenigen Äußerungen ausgewählt, für die eine Pragmatikanalyse sinnvoll erscheint. Diese Stichprobe umfaßt 62 Äußerungen mit insgesamt 354 Wörtern und wird im folgenden als *Pragmatik-Stichprobe* bezeichnet.

3.2 Die Dialog-Stichprobe

3.2.1 Erstellung der Dialog-Stichprobe

Die Äußerungen der EVAR-Stichprobe sind für Untersuchungen in bezug auf Intonation nur bedingt geeignet. Da die Sätze vom Blatt abgelesen wurden, können gewisse für freie Rede typische Effekte, wie gefüllte Pausen ("... *ääh* ..."), systematisch nicht beobachtet werden (siehe Kap.2 und [ROYÉ 83]). Aus diesem Grund wurde im Rahmen der vorliegenden Arbeit eine neue Sprachstichprobe erstellt. Das neue Material sollte verschiedene Bedingungen erfüllen:

1) Es sollte sich um freie Rede handeln.
2) Die Aufnahmen sollten in Dialogsituationen gemacht werden.
3) Das Material sollte aus dem Anwendungsbereich (Intercity-Zugauskunft) stammen.
4) Die Aufnahmen sollten so gut sein, daß sie für akustisch-phonetische sowie für prosodische Analysen geeignet sind. (Das bedeutet u.a., daß der Testperson aus aufnahmetechnischen Gründen die Aufzeichnung des Dialogs bewußt wird.)

Es wurden Aufnahmen mit vier Sprechern (zwei männlich, zwei weiblich) erstellt. Die Sprecher waren wiederum "naive" Versuchspersonen. Zwei der Personen waren auch für die EVAR-Stichprobe Testsprecher. Es wurden Dialograhmen in drei Schwierigkeitsstufen etwa der folgenden Art entworfen:

1. Leichte Dialoge
   Von Nürnberg nach Ulm
   nachmittag ankommen, morgen

2. Mittelschwere Dialoge
   Von Nürnberg nach Hamm
   nächster Sonntag
   Zusatzinformation: Direktverbindung?

3. Schwere Dialoge
   Von Nürnberg nach Dortmund mit Fahrtunterbrechung in Köln
   Köln Ankunft: Nachmittag
   Zusatzinformationen: Gibt es früheren/späteren Zug?
   Wie lange kann Fahrtunterbrechung dauern?

Die Testpersonen wurden aufgefordert, bei der Zugauskunft der DB oder bei einem Reisebüro die gewünschte Information telefonisch einzuholen. Die Dialograhmen wurden *absichtlich* ungenau gehalten, da sich gezeigt hatte, daß sonst die Anrufer die Fragen exakt in der vorgegebenen Form stellten. Fehlende Information bei Rückfragen des Auskunftspersonals konnte der Anrufer nach eigenem Gutdünken einsetzen. Die Dialoge wurden doppelt aufgenommen, zum einen der gesamte Dialog mit einem Gerät zur Aufzeichnung von Telefongesprächen (sehr schlechte akustische

Qualität; diese Aufnahme wurde lediglich zur Transkription des Dialoges verwendet), zum anderen der Teil des Anrufenden mit einem handelsüblichen Mikrofon auf Analogband (relativ gute akustische Qualität; diese Aufnahme wurde später teilweise digitalisiert, s.u.). Die Aufnahmen fanden in Büroumgebung statt.

Es wurden insgesamt 19 Dialoge aufgenommen, die Dauer der Dialoge lag zwischen 40 Sekunden und 4 Minuten (je nach Komplexität des vorgegebenen Dialograhmens).

Die Dialoge wurden orthographisch transkribiert, wobei auffällige Verschleifungen markiert wurden, wie z.B. in folgendem Dialogteil:

Auskunftsbeamter:
"Also da könnt'mer fahr'n ab Erlangen um 10 Uhr 2 ..."

Im Rahmen einer Zusammenarbeit mit der Universität Regensburg wurden die in Erlangen aufgenommenen Dialoge in das dort erstellte Korpus informationsabfragender Dialoge (FACID) aufgenommen. Die Transkripte von sämtlichen in Erlangen erstellten Dialogen finden sich in [HITZENBERGER 86].

Die Dialog-Teile der Anrufenden von acht Dialogen (je zwei pro Sprecher) wurden digitalisiert. Die Abtastrate beträgt 16 kHz, das Sprachsignal ist bandpaßgefiltert (0.1-6.4 kHz; 48 dB/Oktave Flankensteilheit). Ingesamt handelt es sich um 17 Passagen mit einer Gesamtdauer von ca. 100 Sekunden. Zu den digitalisierten Dialogteilen wurde eine Handklassifikation auf der Lautkomponenten-Ebene (Kap.3.2.2) sowie eine Betonungs-Bewertung (Kap.3.2.3) erstellt.

3.2.2 Das erweiterte Erlanger Transkriptionssystem

In [FISCHER 88] wird die Fortschaltzeit und Breite des Analysefensters bei der akustisch-phonetischen Analyse, die bei [REGEL 88] jeweils 12.8 Millisekunden betragen, variiert. Eine genaue Analyse der Klassifikationsergebnisse in [FISCHER 88] zeigt, daß bei anderen Analysebedingungen (veränderte Fortschalt- und Analysefensterraster) die Transkription der Lautkomponenten mit angepaßt werden muß[1]. Daher und um bei der Weiterentwicklung des Akustik-Phonetik-Moduls *koartikulatorische Effekte* besser untersuchen zu können, wurde das Erlanger Transkriptionssystem für die *enge* Transkription der Dialog-Stichproben in einigen Punkten erweitert (enge Transkription bedeutet, daß zusätzlich zu der Umschrift nach Lauten auch Details der Artikulation beschrieben werden, wie z.B. */AR/* vs. *behauchtes /AR/*, siehe [TILLMANN 80, S.102ff):

- Die Transkription erfolgt punktgenau, d.h. nicht in einem festen Fortschaltraster.
- Die ersten beiden Symbole dienen der weiten Transkription, danach können beliebig viele Zeichen folgen, die der engeren Transkription dienen.

---

[1] Problematisch kann z.B. der Burst eines Plosivs sein, der oft nur zwei bis drei Millisekunden lang ist.

Beispiele: EH1 *linksrandige (→ 1)* Transition eines *geschlossenen E (→ EH)*
A.C *laryngal (creaky, → C)* artikuliertes *helles A (→ A.)*
ERT2 *rechtsrandige (→ 2)* Transition eines *entstimmten (→ T) Schwa-Lautes (→ ER)*

- Zusätzlich zu den in Anhang B angegebenen Symbolen zur weiten Transkription (1. und 2. Stelle) werden einige weitere Symbole vereinbart:

Beispiele: IR *hoher Schwa*
RF *Flap*

Eine komplette Liste der zusätzlich vereinbarten Symbole für die enge und weite Transkription findet sich in Anhang C.

Die Dialog-Stichprobe wurde von einer Phonetikerin des Instituts für Phonetik und Sprachliche Kommunikation der Universität München transkribiert.

3.2.3 Erstellung einer Betonungsbeschreibung

Um die automatische Intonationsbeschreibung beurteilen zu können (Kap.6), mußte eine Referenzkontur erstellt werden. Für diesen Zweck wurden die digitalisierten Dialog-Teile 15 Versuchspersonen zusammen mit der orthographischen Mitschrift zur Beurteilung vorgelegt. Das ca. 100 Sekunden umfassende Sprachmaterial enthält 307 Wörter mit 458 Sprechsilben[2]. Die Hörer konnten sich die Sätze beliebig oft anhören. Es handelte sich wiederum um "naive" Versuchspersonen (Informatik-Studenten bzw. Mitarbeiter des IMMD 5).

Die Testpersonen hatten die Anweisung, die Silben in *unbetont* und *betont* einzuteilen. Auf weitere Angaben (z.B. wieviele Silben pro Satz zu betonen sind) wurde mit Absicht verzichtet, da die intuitive Vorstellung der Testperson von "betont/unbetont" erfaßt werden sollte. Eine differenziertere Beurteilung des Betonungsgrades einer Silbe sollte sich dann aus der Summe der Beurteilungen ergeben (wobei natürlich keine 16-stufige Abstufung des Betonungsgrades erwartet werden konnte). Bild 3.1 zeigt die Summe der Beurteilung durch die Testpersonen für eine Passage, Bild 3.2 zeigt die prozentuale Verteilung der Summenbewertung der Silben in allen Dialogen.

---

[2] Die Anzahl der gesprochenen Silben läßt sich nur ungefähr angeben, da sich insbesondere in Wörtern mit der End-Lautfolge "Liquid / Schwa-Laut / Nasal" durch Elision des Schwa-Lautes und an Vokal / Wortgrenze / Vokal durch Lautverschmelzungserscheinungen die Anzahl der gesprochenen Silben verändert. Allerdings spielt z.B. bei der Frage, ob ein Hörer eine spezielle Realisierung des Wortes "vielen" als ein- oder zweisilbig perzipiert, nicht nur der Grad der Reduziertheit der zweiten Silbe eine Rolle (/L.ERN./ → /L.N./), sondern auch der Gedächtniseindruck der eigenen Aussprache. Auf diese Problematik wird in Kap.6 noch genauer eingegangen, siehe auch z.B. [LINDNER 81].

```
Dialog 21:

 10  1        10   1    2         13 1       10
Bichler,Grüß Gott. Ich möchte am Montag von Nürnberg nach

 14   4              1       12            12    2
Köln fahr'n, so daß ich am Nachmittag in Köln  bin.
```

Bild 3.1: Summe der Beurteilungen durch 15 Testpersonen.

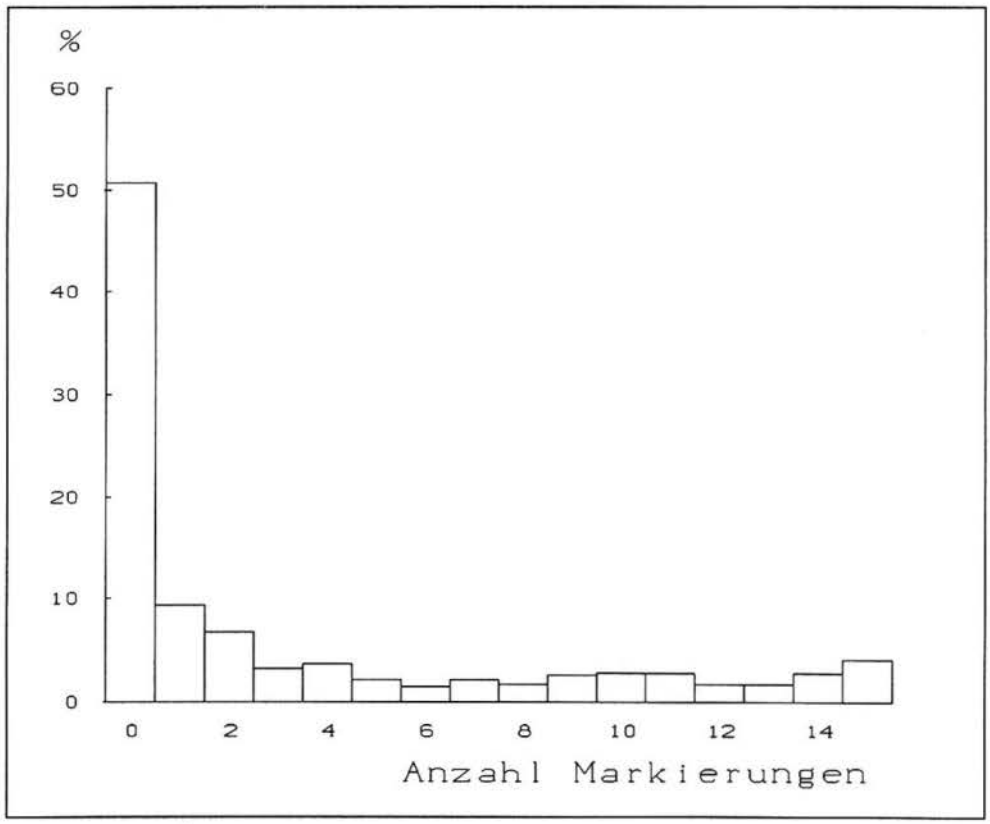


Bild 3.2 Verteilung der Summenbewertung für die betrachteten Silben bei der Beurteilung durch 15 Testpersonen. Nach rechts sind die möglichen Bewertungsstufen aufgetragen, nach oben der jeweilige prozentuale Anteil.

Bei einer Beurteilung des Ergebnisses ist zu bedenken, daß verschiedene Fehlerquellen das Ergebnis verfälschen können:

- Unkonzentriertheit der Testperson
- Fehler beim Auszählen
- Gedächtniseindruck der eigenen Aussprache
- kleine Stichprobe.

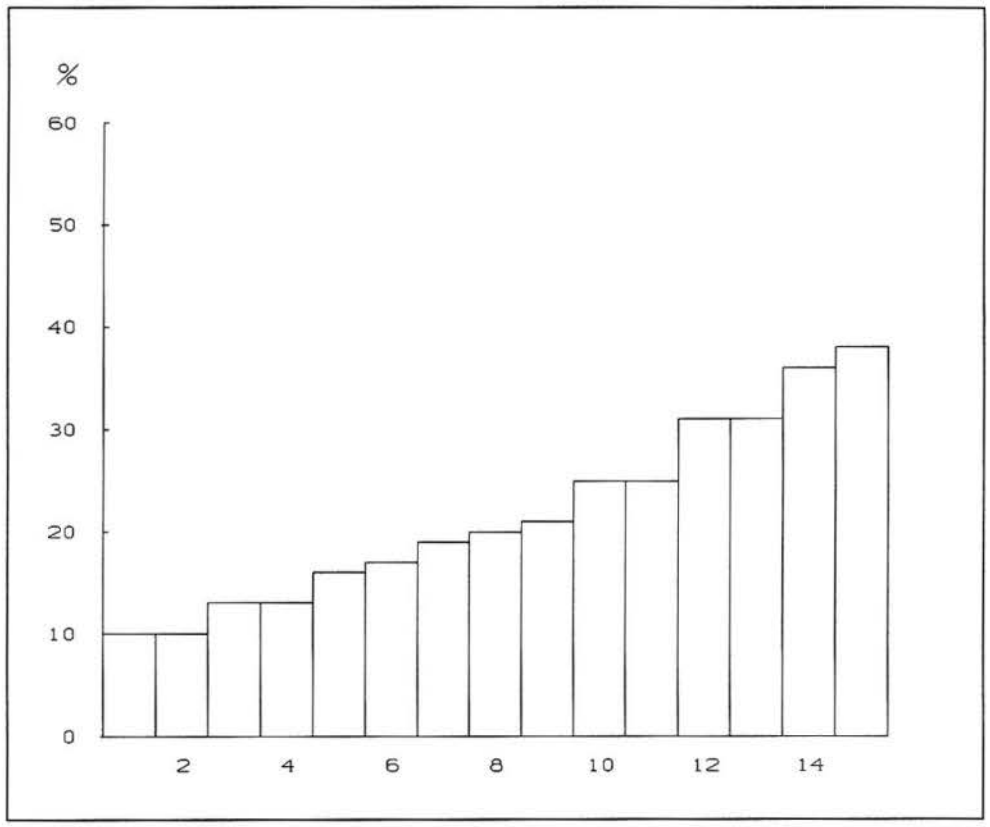


Bild 3.3: Anteil der markierten Silben. Nach rechts sind die Testpersonen aufgetragen, nach oben der jeweilige prozentuale Anteil markierter Silben.

Es zeigt sich, daß die Anzahl der Silben, die von den einzelnen Personen als betont perzipiert wird, relativ stark schwankt. Bild 3.3 zeigt für jede Testperson die Anzahl der als betont empfundenen Silben in Prozent. Wie man sieht, liegt der Anteil zwischen 10 und 38 Prozent, im Mittel werden 22 Prozent markiert.

Die Stichprobe wurde zum Vergleich auch einem ausgebildeten Phonetiker vorgelegt. Bild 3.4 zeigt für jede der 16 Betonungsstufen, wieviel Prozent der Silben jeder Stufe auch von dem Phonetiker als betont markiert werden. Zusätzlich ist die Regressionsgerade zwischen der Bewertungsstufe und der Übereinstimmung in Prozent als gestrichelte Linie aufgetragen. Der Korrelationskoeffizient zwischen den beiden Zahlenreihen beträgt 0.87 und belegt den engen Zusammenhang zwischen den beiden Urteilen. Auch die Zahl der von dem Phonetiker markierten Silben (23 Prozent) unterscheidet sich von der Zahl der im Mittel von den 15 Testpersonen markierten Silben (22 Prozent) nur geringfügig.

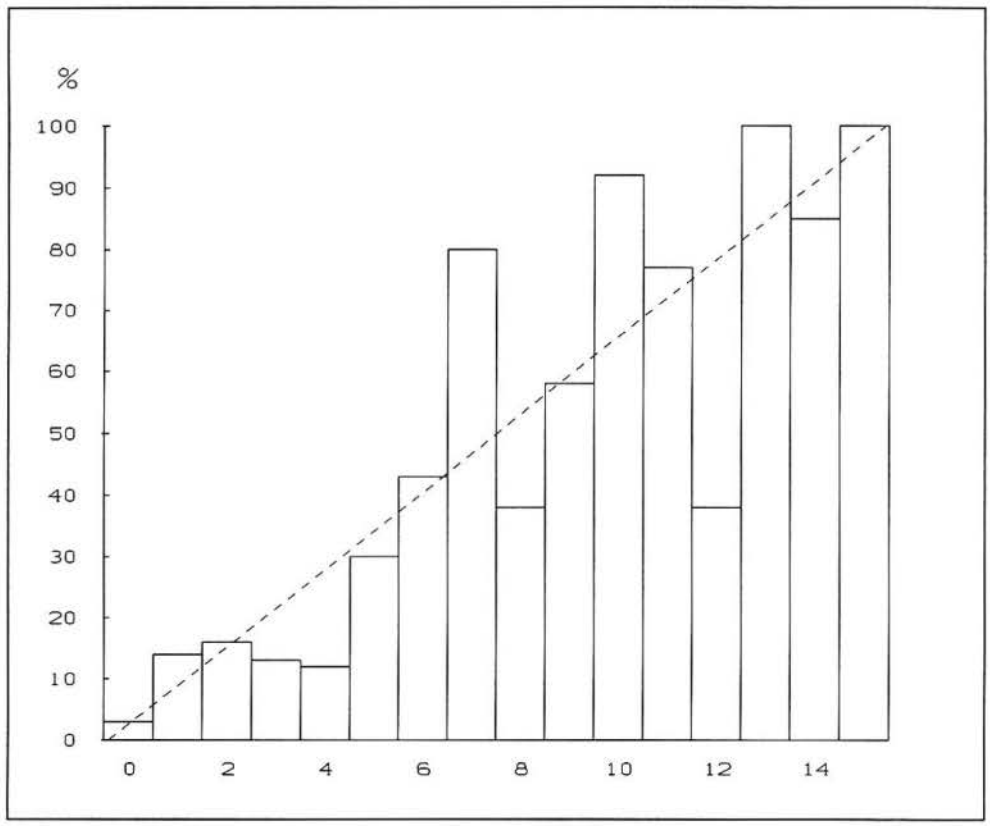


Bild 3.4: Übereinstimmung der Beurteilung durch die 15 Testpersonen und durch einen ausgebildeten Phonetiker. Nach rechts sind die Summenbewertungen aufgetragen, nach oben der jeweilige prozentuale Anteil der Silben, die auch von dem Phonetiker als betont markiert wurden. Die gestrichelte Linie ist die Regressionsgerade zwischen der Bewertungsstufe und der Übereinstimmung in Prozent.

Es ist, wie oben bereits erwähnt, sicherlich nicht angebracht, die Zahlen dahingehend als Betonungsmaß anzusehen, daß man zwei Silben mit derselben Zahl als gleich hervorgehoben betrachten kann und etwa eine Bewertung sieben von einer Bewertung acht unterscheiden kann. Trotzdem läßt sich sagen, daß die Summe der Beurteilungen den Verlauf der Hervorhebung innerhalb der Äußerung gut wiedergibt, und eine Einteilung der Silben z.B. in zwei oder vier Stufen ({unbetont, betont} oder {unbetont, neutral, betont, stark betont}) auf der Basis dieses Betonungsmaßes durchaus möglich ist.

Pragmatische Relevanz der Silben mit einer hohen Betonungszahl:

Um mögliche Auswirkungen einer automatisch erstellten Betonungsbeschreibung für die pragmatische Analyse zu untersuchen, werden zunächst einmal alle Wörter betrachtet, die den Nukleus der Phrasen darstellen, die für die pragmatische Analyse ([EHRLICH 89]) besonders wichtig sind. In der Dialog-Stichprobe sind dies 21 Prozent aller Wörter. Die durchschnittliche Betonungszahl der Wortakzent-Silbe dieser Wörter ist zwölf, während die durchschnittliche Betonungszahl *aller* Silben einen Wert von zwei hat.

Betrachtet man die Silben näher, die von mehr als 50 Prozent der Testpersonen als betont markiert werden, also eine Bewertung von mindestens acht haben, so erhält man eineinhalb mal soviele Wörter (31 Prozent aller Wörter der Stichprobe). Es stellt sich jedoch heraus, daß sich in allen Fällen das dazugehörige Wort in ein für die Anwendung *Zugauskunft* relevantes Pragmatik-Konzept[3] einordnen läßt, bzw. daß es sich um einen metakommunikativen (d.h. dialogsteuernden) Dialogschritt handelt. Tabelle 3.2 zeigt den prozentualen Anteil der verschiedenen Konzepte. In den angegebenen Beispielen sind die Silben mit einer Bewertung von mindestens acht unterstrichen. Außer bei den metakommunikativen Dialogschritten handelt es sich ausnahmslos um relevante Pragmatik-Konzepte.

Bei den metakommunikativen Dialogschritten, insbesondere am Anfang eines Auskunftsdialoges (Name, Begrüßungsformel, Einleitungsfloskeln wie "ich hätte eine Frage"), ist es noch nicht geklärt, in welchem Maß sie in einem Mensch-Maschine-Dialog überhaupt auftreten würden. Hier spielen auch Faktoren wie Dialogsteuerung durch die Maschine, Kompetenz des Systems und Akzeptanz durch den Benutzer eine Rolle. Falls sie auftreten, sollten sie sicherlich mit einer anderen Verarbeitungsstrategie behandelt werden als der restliche Dialog. Zum einen wären dann Wörter zu erwarten, die mit Sicherheit nicht im Lexikon sind (Eigennamen), zum anderen bietet die Position (Anfang des Dialoges) die Möglichkeit der Sprecheradaption.

---

[3] In EVAR wird das Wissen über den Anwendungsbereich als *Semantisches Netz* repräsentiert. Begriffe des Problemkreises wie z.B. *Fahrplanauskunft* werden als Konzepte modelliert. Einzelheiten hierzu finden sich z.B. in [EHRLICH 89] und [SAGERER 89].

Einteilung der Bewertungen ≥ 8	%
metakommunikativer Dialogschritt: z.B. Bichler, Grüß Gott	31
Pragmatik-Konzept Fahrplanauskunft: z.B. zurückfahren	18
Pragmatik-Konzept Abfahrts-/Zielort: z.B. nach Hamburg	24
Pragmatik-Konzept Abfahrts-/Ankunftszeit: z.B. morgen früh	22
sonstige Pragmatik-Konzepte: z.B. mit demselben Zug	4

Tab. 3.2: Zuordnung der Wörter, die eine Silbe mit einer Bewertung ≥ 8 enthalten, zu Konzepten, die im Pragmatik-Modul des Erlanger Spracherkennungssystems Verwendung finden.

3.3 Das Royé-Korpus

In dieser Arbeit wurde schon mehrfach auf die in [ROYÉ 83] durchgeführte Analyse einer Fernsehdiskussion eingegangen. Über das Institut für deutsche Sprache in Mannheim ist ein Tonband lieferbar, das den Mitschnitt dieser Fernsehdiskussion enthält. Zusätzlich stellte Herr Royé seine Tonhöhenkurven zur Verfügung[4].

Aus dem Analogband wurden Beispiele für verschiedene Hervorhebungsarten ([ROYÉ 83, Kap.4.1]) mit den identischen Filter- und Abtastratenwerten wie bei der Dialog-Stichprobe digitalisiert. Insgesamt handelt es sich um 22 Passagen mit einer Gesamtdauer von ca. 120 Sekunden. Aus den von Royé per Hand korrigierten F_0-Verläufen auf den Tonhöhenkurven (siehe die Beispiele in Kap.2.9 und [ROYÉ 83, Kap.1.4.2]) wurden mit Hilfe eines Digitalisierbretts Grundfrequenz-Referenzwerte erstellt.

---

[4] An dieser Stelle sei Frau Knetschke vom Institut für deutsche Sprache und Herrn Royé noch einmal ausdrücklich für ihre kollegiale Hilfe gedankt.

3.4 Die Modus-Fokus-Korpora

Im Rahmen des DFG-Projekts "Modus-Fokus-Intonation" an der Ludwig-Maximilians-Universität in München wurde Testmaterial auf dem Hintergrund des in [ALTMANN 84, ALTMANN 87] entwickelten Satzmodussystems erstellt (siehe Kap.2.5).

Das Korpus ist in vier Teilkorpora unterteilt, die in [BATLINER 89g] ausführlich beschrieben sind. Sprecher waren wiederum "naive" Versuchspersonen (Studenten der Germanistik). Die Testsätze waren in modus- und fokussteuernde Kontexte eingebettet, d.h. die Versuchspersonen hatten die Aufgabe, sich aufgrund des Kontextsatzes und/oder einer Situationsbeschreibung in eine Situation hineinzudenken und die Testsätze entsprechend zu intonieren. Das Material ist auf Verwechslungen ausgelegt, da bei identischer segmentaler Struktur der Satzmodus sowie die im Fokus stehende Phrase allein durch intonatorische Merkmale indiziert wird. Es handelt sich also um *intonatorische Minimalpaare*. Für alle Äußerungen wurden mit einem mechanischen Tonhöhenschreiber Grundfrequenzverläufe erzeugt (Mingogramme). Von den Mingogrammen wurden intonatorische Parameter, wie z.B. der Offset der Grundfrequenz, extrahiert. Zwei Sprechergruppen, bestehend aus je drei weiblichen und drei männlichen Sprechern, produzierten die insgesamt 155 Testsätze zwei- bis viermal (2074 Äußerungen). Da in 75 Fällen die Sprecher laryngalisierten, blieben 1999 Äußerungen übrig.

Die Untersuchungen im Rahmen der Arbeiten am EVAR-Prosodie-Modul wurden in sehr enger Kooperation mit Mitarbeitern der Münchner Intonationsgruppe durchgeführt. Daher liegen alle Daten zu diesen Korpora vor, bzw. sie wurden teilweise in Erlangen erstellt.

Tabelle 3.3 zeigt einer kurze Charakterisierung der vier Teilkorpora K1, K2, Fokus und Leo.

Korpus	Sprecher-gruppe	Test-sätze	Anzahl	syntaktische Strukturen	Minimal-paare
K1	1	71	896	Verb-Erst- Verb-Zweit-	Modus
K2	2	45	573	Verb-Letzt-	Modus
Fokus	2	26	355	Verb-Erst- Verb-Zweit-	Modus+ Fokus
Leo	2	13	175	Verb-Erst- Verb-Zweit-	Modus+ Fokus

Tab. 3.3: Kurze Charakterisierung der vier Modus-Fokus-Korpora.

Zwei Teilkorpora, das *Leo-Korpus* und das *Fokus-Korpus*, stehen in digitalisierter Form zur Verfügung (16 kHz; 0.1-6.4 kHz Bandpaßfilter). Obwohl die Analogaufnahmen im schallarmen Raum des Instituts für Phonetik in München erstellt wurden, sind einige der digitalisierten Aufnahmen durch mehrfaches Schneiden und Umkopieren des Analogbandes qualitativ nicht hochwertig.

3.4.1 Das Fokus-Korpus

Bei diesem Korpus handelt es sich um intonatorische Modus- und Fokus-Minimalpaare mit drei unterschiedlichen, aber ähnlich aufgebauten segmentalen Strukturen. Durch verschiedene Kontextsätze ergeben sich 26 verschiedene Konstellationen, von denen insgesamt 357 Realisierungen mit einer Gesamtdauer von ca. zwölf Minuten zur Verfügung stehen. Tabelle 3.4 zeigt die drei segmental unterschiedlichen Testsätze und die intendierten Satzmodi.

Testsatz				indizierter Satzmodus	
Matrixsatz	1.Phrase	2. Phrase	3. Phrase	Fragen	Nicht-Fragen
Sie läßt	die Nina	das Leinen	weben	Assertive Frage	Aussage
Lassen Sie	den Manni	die Bohnen	schneiden	Verb-Erst-Frage	Imperativ
Lassen wir	den Leo	die Blumen	düngen	Verb-Erst-Frage	Adhortativ

Tab. 3.4: Die segmental unterschiedlichen Testsätze des Fokus-Korpus sowie die intendierten Satzmodi.

Beispiel:

Modus-Fokus-Konstellation:

Aussagesatz, Fokus auf "Leinen"

Situationsbeschreibung:

In einem Textilbetrieb; eine Mutter erkundigt sich bei einer Angestellten nach den handwerklichen Fortschritten ihrer Tochter.

Kontextsatz:

Mutter: *"Was läßt die Meisterin meine Nina denn gerade weben?"*

Testsatz:

Angestellte: *"Sie läßt die Nina das Leinen weben."*

3.4.2 Das Leo-Korpus

Bei diesem Korpus handelt es sich um zwei Testsätze, einen mit Verb-Erst- und einen mit Verb-Zweit-Stellung (*"Säuft der Leo"* und *"Der Leo säuft"*), mit denen viele unterschiedliche Modus- und Fokus-Strukturen erzeugt werden können ([BATLINER 88b]). Das Korpus ist also sehr konsistent. Durch die verschiedenen Kontextsätze ergeben sich 13 Konstellationen, von denen insgesamt 180 Realisierungen von Kontext- und Testsatz mit einer Gesamtdauer von ca. 15 Minuten zur Verfügung stehen.

Beispiele:

Modus-Fokus-Konstellation: Aussagesatz, Fokus auf Leo
Kontextsatz: *"Ihr fragt mich, wer säuft?"* Testsatz: *"Der Leo säuft."*

Modus-Fokus-Konstellation: Aussagesatz, Fokus auf säuft
Kontextsatz: *"Was soll schon mit dem Leo sein?"* Testsatz: *"Der Leo säuft."*

Modus-Fokus-Konstellation: Verb-Erst-Fragesatz, Fokus auf säuft
Kontextsatz: *"Stimmt das mit dem Leo?"* Testsatz: *"Säuft der Leo?"*

4 Akzentuierungsmittel und Merkmalextraktion

In Kapitel 2 wurden die wichtigsten prosodischen Eigenschaften, die bei der Perzeption von Betonung eine Rolle spielen, sowie ihre physikalischen Korrelate eingeführt. In diesem Kapitel wird noch einmal - unter dem Aspekt der automatischen Merkmalextraktion - auf die physikalischen Korrelate eingegangen. Ziel der Merkmalextraktion ist es, das in digitalisierter Form vorliegende Sprachsignal in Untereinheiten zu zerlegen, die für die Erstellung einer automatischen Betonungsbeschreibung geeignet sind. Für diese Untereinheiten sind Merkmale zu berechnen, die den für die Betonungsempfindung relevanten Teil des Verlaufs der physikalischen Korrelate gut wiedergeben. Aufgrund dieser Merkmale ist dann eine Betonungsbeschreibung für das gegebene Sprachsignal zu erstellen (siehe Kap.5).

In diesem Kapitel werden nur Algorithmen zur Merkmalextraktion besprochen, die im Rahmen dieser Arbeit entwickelt bzw. implementiert wurden. In Kap.4.1 wird auf die Segmentierung des Sprachsignals in Pausen, Silbenkerne und Silbenrandbereiche eingegangen. Verfahren zur Berechnung von Merkmalen für die prosodischen Eigenschaften Tonhöhe, Lautheit und zeitliche Strukturierung werden in Kap.4.2-4.4 vorgestellt.

4.1 Silbenkerndetektion

Als kleinste betonbare Einheit gilt die Silbe, Träger der Betonung ist insbesondere der Silbenkern. Daher ist der erste Schritt bei der Betonungsberechnung die Lokalisation von Silbenkernen. Dabei werden in der Regel Energiemaße im Zeit- ([MILLER 75]) oder Frequenzbereich ([MERMELSTEIN 75], [WILLIAMS 88]) oder die Hypothesen eines groben ([AULL 84]) oder feinen Akustik-Phonetik-Klassifikators ([MERCIER 88]) verwendet. Die Grenzen zwischen dem, was als *akustisch-phonetische* und was als *prosodische* Information bezeichnet wird, sind hier sehr verschwommen: Es gibt mehrere ASE-Systeme, bei denen *silbenorientierte Einheiten* die unterste Segmentierungsebene des Akustik-Phonetik-Moduls darstellen, so daß die Extraktion der Silbenstruktur einer Äußerung als das Analyse-Ergebnis des Akustik-Phonetik-Moduls gilt (siehe z.B. [RUSKE 88, S.106ff] sowie die dort angegebenen Literaturstellen).

Im Rahmen dieser Arbeit wird ein zweistufiger Ansatz zur Silbenkerndetektion verfolgt:

1) Die Silbenkerne werden aufgrund der logarithmierten spektralen Energie in drei verschiedenen Frequenzbändern detektiert. Das Modul ist in Kap.4.1.2 beschrieben (weitere Einzelheiten in [SCHMÖLZ 85, 87]). Analyse-Ergebnis ist eine Zerlegung des Sprachsignals in Pausenbereiche, Silbenkernbereiche und Silbenrandbereiche.
2) Für die weitere Analyse kann es sinnvoll sein, die Silbenkerngrenzen mit den Segmentgrenzen (Lautgrenzen) des Akustik-Phonetik-Moduls abzugleichen: z.B. werden die Worthypothesen in EVAR über dem Segmentraster erzeugt, nicht über dem Frameraster. Möchte man jetzt an den betonten Stellen das Lexikon einschränken (siehe Kap.6.5), dann sollten die Grenzen der "verbotenen" Gebiete mit Segmentgrenzen zusammenfallen. In Kap.4.1.3 (weitere

Einzelheiten in [FISCHER 89] und [KUNZMANN 90]) wird alternativ zu dem in [REGEL 88, Kap.4.3] beschriebenen syntaktischen Segmentierer ein Segmentierungs-/Klassifikationsverfahren für Wortuntereinheiten beschrieben. Das Verfahren benutzt als Eingangsdaten die Lautkomponenten-Klassifikation und optional das Segmentierungsergebnis aus Punkt 1. Das Analyse-Ergebnis ist eine Zerlegung des Sprachsignals in klassifizierte, homogene Bereiche. Bei Benutzung der prosodischen Segmentierungs-Daten werden diese mit einem regelbasierten Ansatz korrigiert (Verschiebung der Silbenkerngrenzen und Auftrennung verschmolzener Silbenkerne).

4.1.1 Berechnung der Energie in verschiedenen Frequenzbereichen

In diesem Kapitel wird die Berechnung der Energie eines Sprachsignals in einem beliebigen, aber festen Frequenzbereich besprochen. (Das Verfahren ist natürlich auf beliebige eindimensionale Signale anwendbar.) Hierzu wird zunächst das Kurzzeitspektrum eines Sprachsignals definiert. (Eine ausgezeichnete Einführung in die Theorie der Fouriertransformation unter besonderer Berücksichtigung der Verarbeitung von Sprachsignalen findet sich z.B. in [RABINER 78], siehe auch [OPPENHEIM 75], [HESS 83].)

Seien $f_{k,k=0,\ldots,M-1}$ die mit einem Fenster gewichteten, mittelwertfreien Abtastwerte eines Sprachsignals S (es gilt $-2^{B-1} \leq f_k \leq 2^{B-1}-1$, wobei B die Auflösung des A/D-Wandlers ist); dann sind die Fourierkoeffizienten F_μ definiert als

$$F_\mu = \sum_{k=0}^{M-1} f_k \cdot e^{(-i2\pi\mu k/M)}$$

Es werden nur die Fourier-Koeffizienten $F_{\mu,\mu=0,\ldots,(M-1)/2}$ berechnet, da es sich um ein symmetrisches Spektrum handelt. Die Koeffizienten von 0 bis (M-1)/2 sind Stützpunkte des Spektrums im Bereich von 0 bis π bzw. von 0 bis Abtastrate/2 Hz. Somit entspricht der Koeffizient F_μ der Frequenz (μ·Abtastrate)/M Hz. Man erkennt sofort, daß die Frequenzauflösung des Kurzzeitspektrums umgekehrt proportional zur Zeitauflösung[1] ist.

Analog zur Definition des Schallpegels L ([ZWICKER 67])

$$L = 20 \cdot \lg(p_{eff}/p_0) \text{ dB}$$

soll nun für den betrachteten Zeitbereich die logarithmierte Energie $E_{ugf,ogf}$ in einem Frequenzband von ugf Hz bis ogf Hz berechnet werden. Hierzu werden die Koeffizienten F_μ mit ugf $\leq$ (μ·Abtastrate)/M $\leq$ ogf aufaddiert. Der Wertebereich für diese Summe hängt ab von der Breite des Energiebandes, der Breite des Analysefensters, der Abtastrate und der Auflösung des A/D-Wandlers. Da nicht die absoluten Werte, sondern die Energieunterschiede zwischen verschiedenen Bereichen des Sprachsignals interessieren, kann man den Bezugsenergiewert in

---

[1] Die Zeitauflösung ist bestimmt durch die Abtastrate und die Breite des Analysefensters.

Abhängigkeit von diesen Werten so wählen, daß $E_{ugf,ogf}$ in einem gut darstellbaren Zahlenbereich liegt. Sei norm der Bezugsenergiewert, u der Koeffizient, der der unteren Grenzfrequenz ugf entspricht und o der für die obere Grenzfrequenz ogf, so gilt

$$E_{ugf,ogf} = 2000 \cdot \lg\left(\left(\sum_{K=u}^{K=o} F_K\right) / \text{norm}\right) \quad \text{rmB} \quad \text{(relative Millibel)}$$

norm wurde so gewählt, daß $E_{ugf,ogf}$ für die benutzten Framelängen und Energiebänder ungefähr mittelwertfrei ist und im Bereich ±5000 liegt (so bewegen sich z.B. für das in Kap.4.1.2 benutzte Energieband 300-2300 Hz die Werte für die EVAR-Stichprobe zwischen -4800 und 3100 bei einem Mittelwert von 104). Die so berechneten Energiewerte haben einerseits eine gute Auflösung und können andererseits als 16 Bit-Integer-Werte speichergünstig dargestellt werden. (Die Benutzung eines anderen Bezugsenergiewertes norm' entspricht der Addition einer Konstanten.)

4.1.2 Lokalisierung von Silbenkernen aufgrund der spektralen Energie

Das im folgenden vorgestellte Verfahren zur Bestimmung der Silbenkerne basiert auf der *Schallfülletheorie* nach [SIEVERS 85], die z.B. in [KOHLER 77] beschrieben ist. Eine Lautfolge ist demnach nur dann eine Silbe, wenn ihre *natürliche Schallfülle* vom Anfang zur Mitte (Silbengipfel) hin ansteigt und zum Ende hin wieder abfällt. Unter *natürlicher Schallfülle* oder *Sonorität* ist (weitestgehend) die Schallintensität zu verstehen, wie sie das menschliche Ohr empfindet. Bild 4.1 zeigt für die einzelnen Lautoberklassen den Grad ihrer Schallfülle.

Bild 4.1 Schallfülleskala der Lautklassen (nach [KOHLER 77, S.80]).

Während die Bestimmung des Silbengipfels als Schallfülle-Gipfel vergleichsweise einfach durch Abbildung auf das lokale Maximum einer Energie-Kontur geschehen kann und auch die hier interessierenden Silbenkerngrenzen im akustischen Signal zu finden sind (allerdings nicht notwendigerweise in Energie-Konturen, siehe 4.1.3), ist die Bestimmung der Silbengrenzen allein aufgrund des akustischen Signals (also ohne Kenntnis der gesprochenen Äußerung) nicht immer möglich. (Man denke an Beispiele wie *tiefliegende Wolken* vs. *tieffliegende Flugzeuge.*)

Bei dem in [SCHMÖLZ 85] beschriebenen Verfahren wird die logarithmierte spektrale Energie im Frequenzbereich zwischen 300 und 2300 Hz zur Bestimmung des Silbengipfels und der Silbenkerngrenzen benutzt. In diesem Frequenzband liegt mit den ersten beiden Formanten der Vokale der Hauptteil der *sonoranten Energie,* so daß die Schallfülleabstufung aus Bild 4.1 vergleichsweise gut nachgebildet wird.

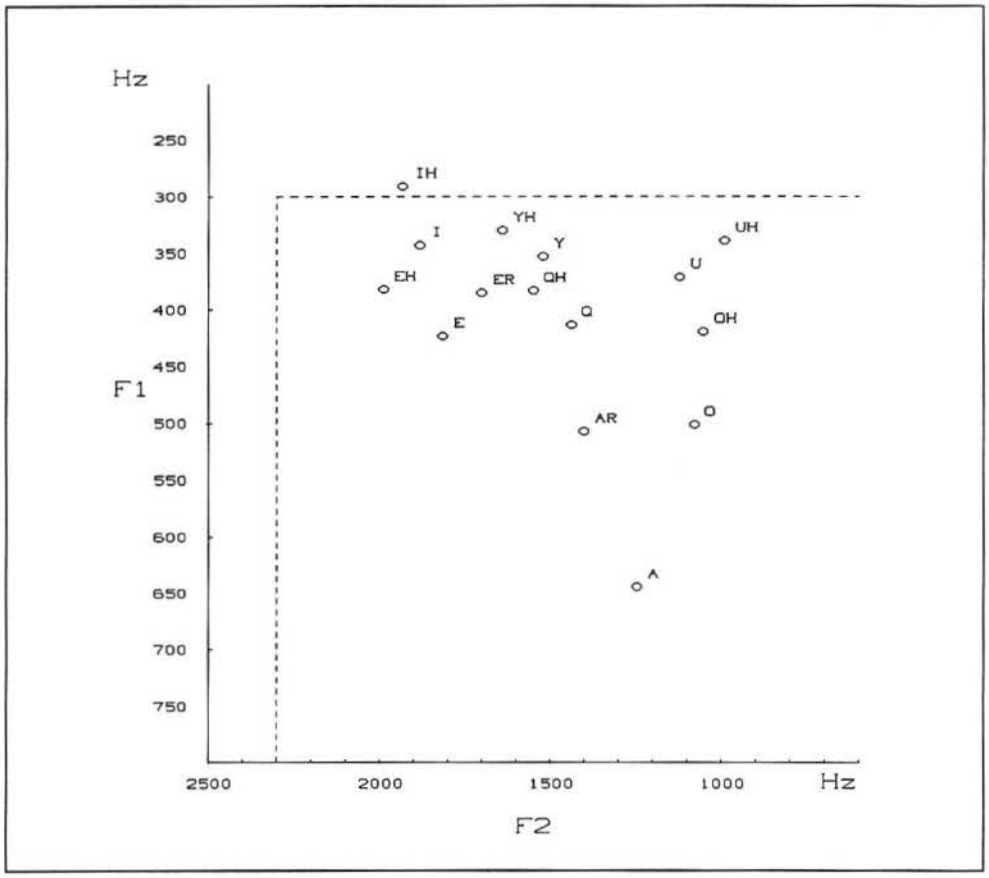


Bild 4.2 Mittlere Lage der ersten beiden Vokalformanten der EVAR-Stichprobe (nach [NÖTH 85]). Die gestrichelten Linien deuten die Grenzen des "sonoranten" Energiebandes an.

Bild 4.2 zeigt die mittlere Lage der Vokalformanten im Material des Erlanger Spracherkennungssystems. Die obere und untere Grenzfrequenz des Bandpasses sind gestrichelt eingezeichnet.

Während die obere Grenzfrequenz problemlos ist, stellt die untere einen Kompromiß dar. Bei den hohen Vorderzungenvokalen (/IH/, /I./, /YH/) liegt der erste Formant in der Nähe der unteren Grenzfrequenz (beim /IH/ sogar darunter). Auf der anderen Seite zeichnen sich Nasale und Liquide durch einen starken ersten Formanten im Bereich zwischen 250 und 300 Hz aus. Legt man die untere Grenzfrequenz des Bandpasses zu niedrig, so werden diese Lautklassen nicht genügend gedämpft.

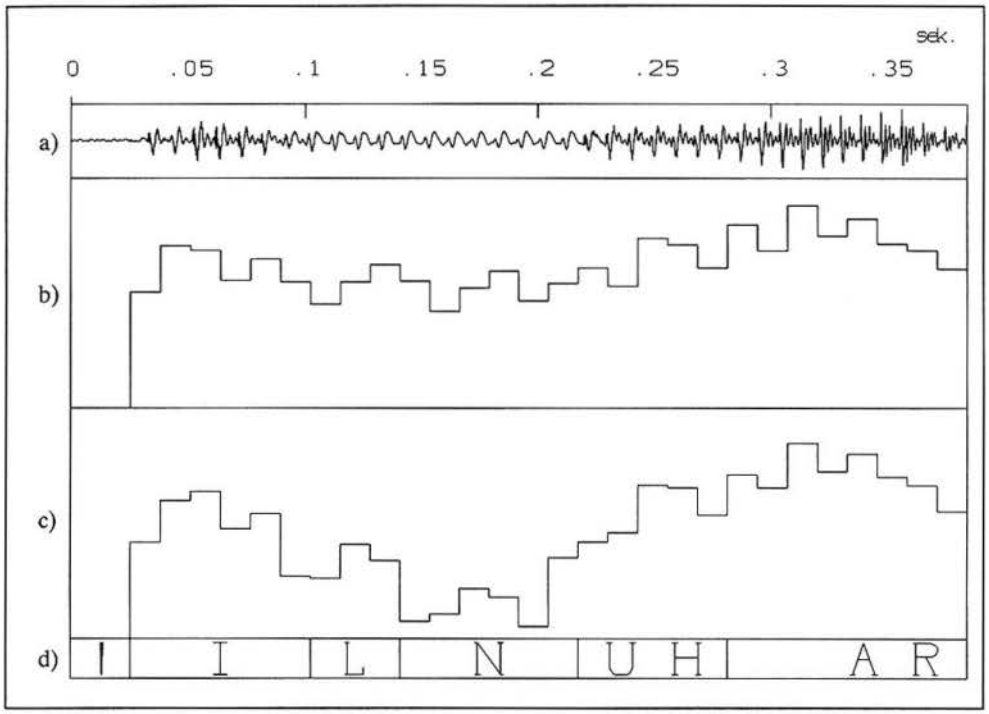


Bild 4.3 Einfluß der unteren Grenzfrequenz auf das Dämpfungsverhalten des Bandpasses bei Nasalen und Liquiden am Beispiel der Wortfolge *"will nur"*:

a) Sprachsignal
b) ungeglätteter Bandpaßausgang (100 Hz - 2300 Hz)
c) ungeglätteter Bandpaßausgang (300 Hz - 2300 Hz)
d) Handklassifikation nach [REGEL 88]

Bild 4.3 verdeutlicht dies für die Wortfolge *"will nur"*. Bild 4.3a zeigt das Zeitsignal, 4.3b den Energieverlauf für den Frequenzbereich zwischen 100 und 2300 Hz, 4.3c den Energieverlauf für den Frequenzbereich zwischen 300 und 2300 Hz, Bild 4.3d zeigt die Handklassifikation. Wie man sieht, wird beim 300-2300-Hz-Band sowohl das /L./ als auch das /N./ durch die höhere untere Grenzfrequenz besser gedämpft als beim 100-2300-Hz-Band.

Zum Finden der Silbenkerne werden in dem mit einem Medianfilter geglätteten Ausgangssignal des Bandpaßfilters folgende Punkte gesucht:

- lokale Maxima MAX_i
- benachbarte lokale Minima MIN_i^L und MIN_i^R
- die Positionen $SKGR_i^L$ und $SKGR_i^R$, an denen vom Maximum MAX_i aus gesehen die Werte $SKS_i^L = (MAX_i+MIN_i^L)/2$ bzw. $SKS_i^R = (MAX_i+MIN_i^R)/2$ erstmals unterschritten werden.

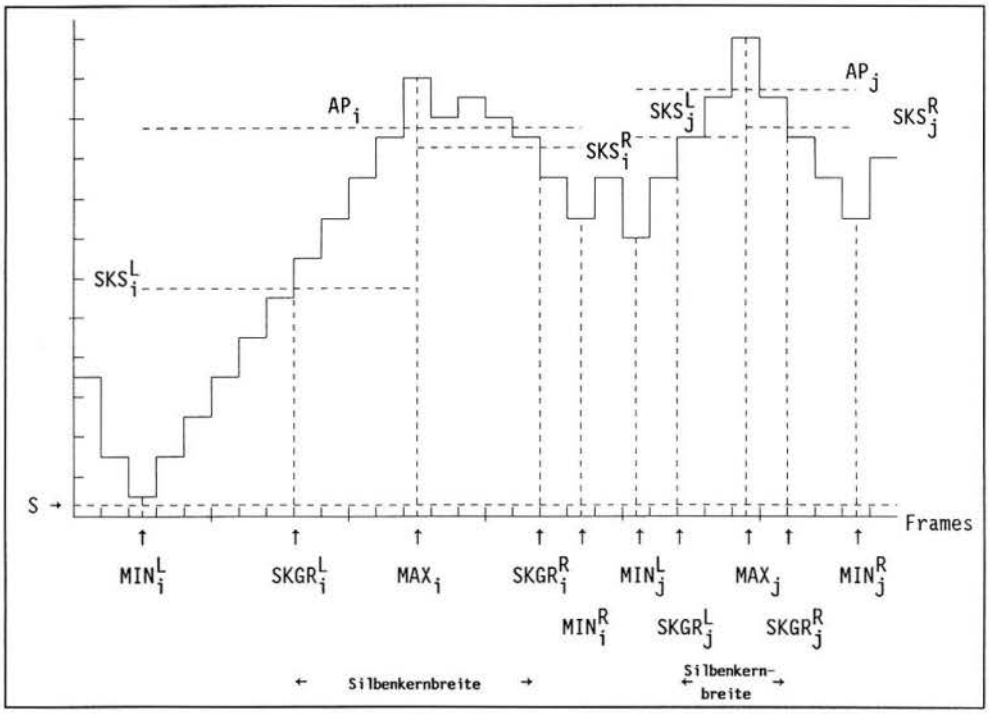


Bild 4.4: Für die Segmentierung des Sprachsignals in Silbenkern- und Silbenrandgebiete notwendige Positionen und Bedingungen.

Es werden nur Punkte berücksichtigt, die folgenden Bedingungen genügen:

- Signifikanz: Die Maxima müssen eine globale Schwelle S überschreiten. Die Schwelle entspricht ungefähr dem Mittelwert des sonoranten Energiebandes zwischen 300 und 2300 Hz über alle Äußerungen aus der EVAR-Stichprobe (siehe Kap.4.1.1).
- Mindestausprägung: Für jedes Maximum MAX_i wird eine lokale Schwelle $AP_i = (1-x)*MAX_i$, $0<x<1$ festgelegt, die von den beiden Minima unterschritten werden muß.
- Mindestbreite: Die Anzahl Frames zwischen $SKGR_i^L$ und $SKGR_i^R$ muß über einer vorgegebenen Schwelle MBR liegen.
- Mindestabstand: Die Anzahl Frames zwischen der rechten Grenze $SKGR_i^R$ und der linken Grenze $SKGR_j^L$ des nachfolgenden Silbenkernbereichs muß über einer vorgegebenen Schwelle MAB liegen.

Von allen so gefundenen Maxima MAX_i wird angenommen, daß es sich um Silbengipfel handelt, die dazugehörigen Silbenkerngrenzen sind durch $SKGR_i^L$ und $SKGR_i^R$ markiert. In Bild 4.4 sind die markierten Punkte sowie die Bedingungen, denen diese Punkte genügen müssen, für zwei Silbenkerne beispielhaft angedeutet. (Das Bild zeigt keine echte Kontur.)

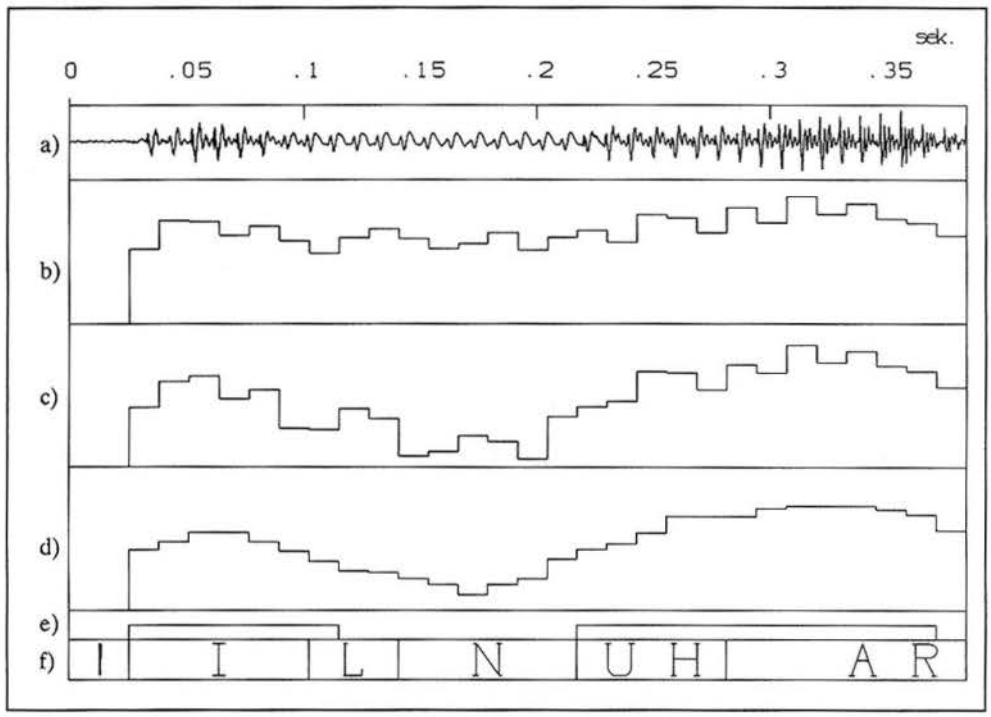


Bild 4.5 Die einzelnen Stadien bei der Silbenkernbestimmung am Beispiel der Wortfolge *"will nur"*:

a) Sprachsignal
b) ungeglättete Gesamtenergie (0 Hz - 5000 Hz)
c) ungeglätteter Bandpaßausgang (300 Hz - 2300 Hz)
d) mit einem Medianfilter der Länge 5 geglätteter Bandpaßausgang (300 Hz - 2300 Hz)
e) markierte Silbenkernbereiche
f) Handklassifikation nach [REGEL 88]

Bereiche, in denen die sonorante Energie länger als eine Schwelle `P` unter der Signifikanzschwelle `S` liegt, werden als Sprechpause markiert.

Bild 4.5 verdeutlicht noch einmal das Vorgehen beim Lokalisieren der Silbenkerne für denselben Signalausschnitt wie in Bild 4.3. Für das Zeitsignal (4.5a) wird nicht-überlappend für jeden Frame (12.8 Millisekunden) aus der Gesamtenergie (4.5b) das Frequenzband zwischen 300 und 2300 Hz (4.5c) betrachtet. Der Energieverlauf wird mit einem Medianfilter der Länge 5 geglättet (4.5d). In dem geglätteten Verlauf werden lokale Maxima als Silbenkerne markiert. Als Grenze eines Silbenkerns gilt dabei der Frame, bei dem der Bandpaßausgang den Wert `SKS = (Maximum+Minimum)/2` unterschreitet (4.5e). Bild 4.5f zeigt die Handklassifikation.

Eine Weiterentwicklung des soeben beschriebenen Verfahrens in drei Punkten ist in [SCHMÖLZ 87] beschrieben:

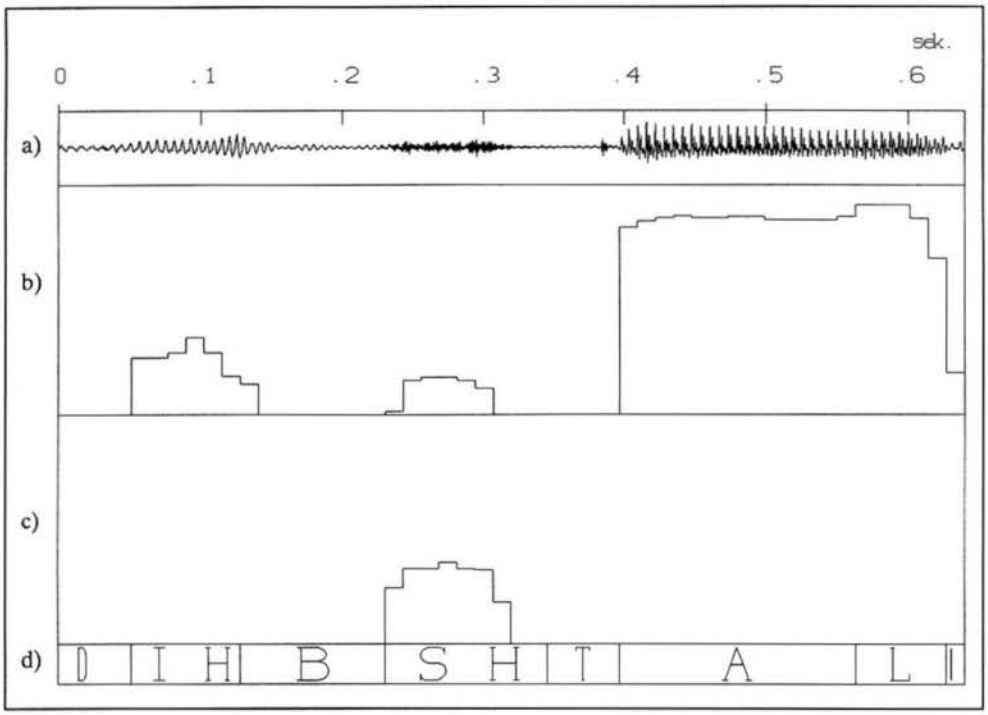


Bild 4.6: Identifikation eines stark ausgeprägten Frikativs durch Vergleich des *"sonoranten"* (300-2300 Hz) und *"frikativen"* (2300-5000 Hz) Energiebandes am Beispiel des Wortes *"Diebstahl"*:
a) Sprachsignal
b) geglätteter Bandpaßausgang (300 Hz - 2300 Hz)
c) geglätteter Bandpaßausgang (2300 Hz - 5000 Hz)
d) Handklassifikation nach [REGEL 88]

1) Ein vielzitiertes Gegenbeispiel zu der Schallfülletheorie ist das Wort "Obst". Obwohl es einsilbig perzipiert wird, weist es bei dem Frikativ /S./ ein weiteres Maximum auf. In [SCHMÖLZ 85] wird ein Drittel der fälschlicherweise eingefügten Silbenkerne auf stark ausgeprägte Frikative zurückgeführt. Problematisch ist allerdings nicht das /S./, sondern das /SH/. Das /S./ hat den Hauptteil seiner Energie im Bereich über 3000 Hz, während der Schwerpunkt der Energie des /SH/ ca. 1000 Hz niedriger liegt ([NÖTH 85]). Somit kommt es beim /SH/ vor, daß im sonoranten Band noch genügend Energie vorhanden ist. Um fälschlicherweise als Silbenkern identifizierte Frikative zu eliminieren, werden die *"sonorante"* Energie im Bereich zwischen 300 und 2300 Hz und die *"frikative"* Energie zwischen 2300 und 5000 Hz verglichen: Silbenkerne, die im *"frikativen"* Energieband mehr Energie aufweisen als im *"sonoranten"* Band, werden aus der Silbenkernliste entfernt. Bild 4.6 veranschaulicht dieses Vorgehen für das Wort *"Diebstahl"*.

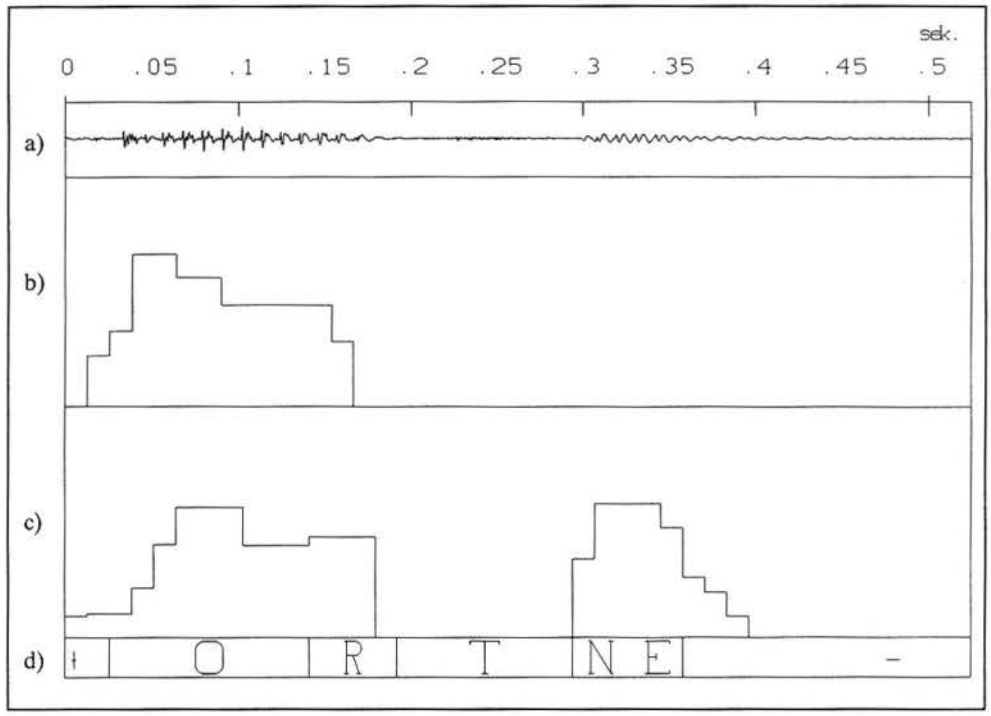


Bild 4.7: Identifikation eines konsonantischen Silbenkerns durch Vergleich des *"sonoranten"* (300-2300 Hz) und *"nasalen"* (100-300 Hz) Energiebandes am Beispiel des Wortes *"Orten"*:

a) Sprachsignal
b) geglätteter Bandpaßausgang (300 Hz - 2300 Hz)
c) geglätteter Bandpaßausgang (100 Hz - 300 Hz)
d) Handklassifikation nach [REGEL 88]

2) Die untere Grenzfrequenz von 300 Hz führt nicht nur dazu, daß die an- und auslautenden Nasale, sondern auch die silbischen Nasale und das silbische L (/LE/) gut gedämpft werden. In Silben, in denen der Reduktionsvokal /ER/ Silbenträger ist, kann durch Elision des Vokals ein direkt danach folgender Nasal die Silbenträgerfunktion übernehmen, wie z.B. in *"reden"* (/D.ERN./ → /D.NE/) oder *"Schnabel"* (/B.ERL./ → /B.LE/). Nasale und /L./, /LE/ haben, wie oben bereits erwähnt, einen hohen ersten Formanten mit schmaler Bandbreite im Bereich unter 300 Hz. Durch den Vergleich der *"sonoranten"* Energie im Bereich zwischen 300 und 2300 Hz mit der *"nasalen"* Energie zwischen 100 und 300 Hz sollen die konsonantischen Silbenkerne besser gefunden werden: Gipfel im unteren Energieband, die keinen im mittleren Energieband identifizierten Silbenkern berühren, werden als konsonantische Silben in die Silbenkernliste eingereiht. Da silbische Konsonanten immer unbetont sind, sind auch Konsequenzen für die zu erstellende Akzentbewertung aus dieser Information zu erwarten. Bild 4.7 veranschaulicht dieses Vorgehen für das Wort *"Orten"*.

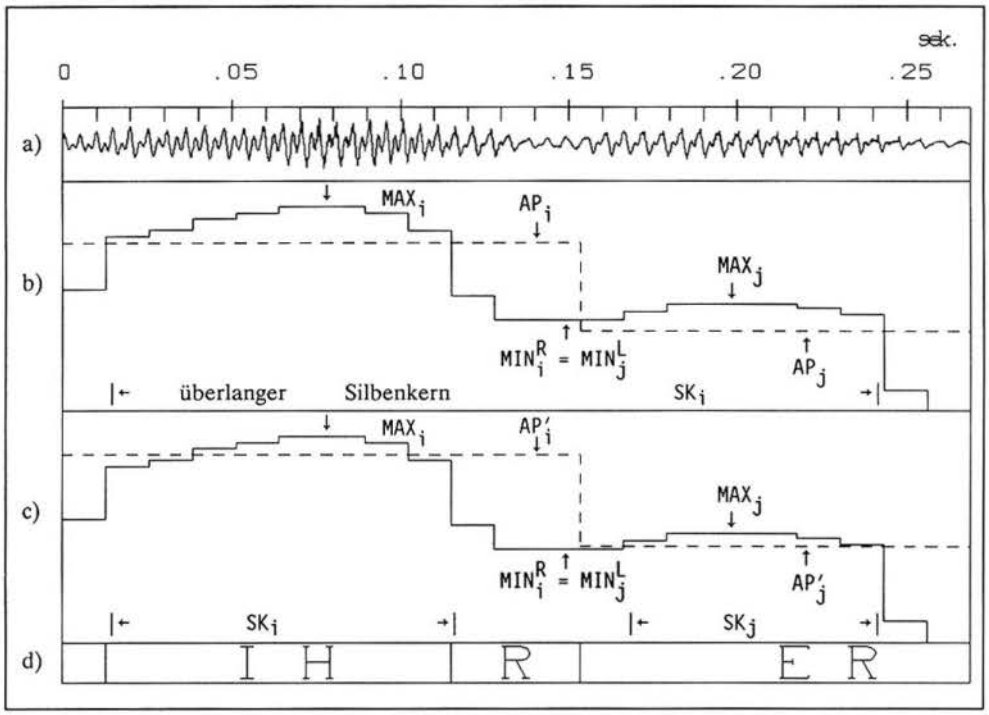


Bild 4.8: Bearbeitung eines überlangen Silbenkerns mit einer weicheren Schwelle AP' am Beispiel des Wortes *"ihre"*:

a) Sprachsignal
b) geglätteter Bandpaßausgang (300 Hz - 2300 Hz) mit der normalen Schwelleneinstellung AP=0.9*MAX
c) geglätteter Bandpaßausgang (300 Hz - 2300 Hz) mit der weicheren Schwelle AP'=0.95*MAX
d) Handklassifikation nach [REGEL 88]

3) Bei Vokal-Liquid-Vokal-Folgen, wie z.B. in dem Wort Bestellung, wird der Liquid oft nicht genügend gedämpft. Damit wird das Minimum zwischen den beiden Vokal-Maxima als nicht signifikant behandelt. Die beiden Silbenkerne werden zu einem Silbenkern verschmolzen. Daher werden überlange Silbenkerne - unter Benutzung eines weicheren Schwellwertes $AP'_i = (1-x')*MAX_i$, $0<x'<x$ - noch einmal getrennt betrachtet. Bild 4.8 veranschaulicht dieses Vorgehen für das Wort *"ihre"*.

Das lokale Maximum MAX_j wird zunächst nicht als signifikant eingestuft, da das linke lokale Minimum MIN_j^L größer als die Schwelle für die Mindestausprägung AP_j ist. Dadurch ergibt sich ein überlanger Silbenkern SK_i, der mit einer weicheren Schwelle AP'_j noch einmal untersucht wird. Da das Minimum MIN_j^L unter der Schwelle AP'_j liegt, wird der überlange Silbenkern in zwei Silbenkerne SK_i und SK_j aufgeteilt.

4.1.3 Korrektur der Silbenkerngrenzen durch Vergleich mit dem Analyseergebnis des Akustik-Phonetik-Moduls

Die Lokalisierung von Silbenkernen aufgrund der spektralen Energie hat den Vorteil der Robustheit, wie die folgende episodenhafte Beobachtung vermuten läßt: Aufgrund eines inzwischen behobenen Fehlers im Digitalisiergerät des IMMD5 wurde eine Reihe von Testsätzen digitalisiert, die von einem relativ starken hochfrequenten Rauschen überlagert waren (diese Äußerungen sind nicht Teil der in Kap.3 beschriebenen Stichproben). Der Fehler fiel auf, da das Akustik-Phonetik-Modul für diese Äußerungen sehr schlechte Ergebnisse brachte. Durch die unverhältnismäßig hohe und plausible Vokal/Frikativ-Verwechslung (z.B. /A/ → /XA/) auf Lautkomponentenebene wurden Silbenkerne vom Segmentierer häufig aufgetrennt, so daß bei der Generierung der Worthypothesen die sonst sehr sicher gefundenen Vokale auf drei Lauthypothesen abgebildet werden müssen. Bei den für die Lokalisierung der Silbenkerne verwendeten sehr groben Energie-Bändern konnte dagegen keine Störung beobachtet werden. Für die betreffenden Äußerungen wurden sogar alle Silbenkerne gefunden. (Dies ist selbstverständlich kein Argument gegen die im Akustik-Phonetik-Modul verwendeten Merkmale, denn für das vom Akustik-Phonetik-Modul zu lösende Klassifikationsproblem von 49 Klassen sind die im Prosodie-Modul verwendeten drei Bandpässe sicher ungeeignet.)

Wegen der höheren Empfindlichkeit gegenüber Veränderungen des akustischen Signals kann andererseits unter normalen Bedingungen ein mit sehr feiner Unterscheidung arbeitender Akustik-Phonetik-Analysator die akustische Oberfläche des Signals wesentlich besser wiedergeben als das sehr grobe *Sonorantenband.* Dies ist insbesondere bei zunehmender Sprechgeschwindigkeit und bei starken Verschleifungserscheinungen wichtig. In solchen Fällen sind zwischen Silbenkernen häufig keine relevanten Minima im Bandpaß-Ausgang zu beobachten, während eine feine akustisch-phonetische Analyse noch Veränderungen anzeigt. Bild 4.9 zeigt für einen Ausschnitt aus dem Wort *"verschiedenen"* die Handklassifikation nach Lautkomponenten, das Sprachsignal, die geglättete Energie des mittleren Energiebandes, die Silbenkernbestimmung nach Kap.4.1.2 sowie die Lautkomponenten-Hypothesen. Während das intervokalische /N/ auf der Ebene der Lautkomponentenhypothesen klar zu erkennen ist, sind in dem Energie-Verlauf keine markanten Veränderungen zu beobachten.

In [FISCHER 89] wird eine Segmentierung in lautähnliche Einheiten beschrieben, bei der eine Korrektur des Segmentierungsergebnisses mit Hilfe der Silbenkern-Detektion sowie auch umgekehrt eine Korrektur der Silbenkerndetektion mit Hilfe der akustisch-phonetischen Segmentierung möglich ist. Im folgenden wird zunächst das Prinzip der Segmentierung kurz beschrieben und danach die Vorgehensweise bei der Verknüpfung der akustisch-phonetischen und der prosodischen Segmentierung erläutert.

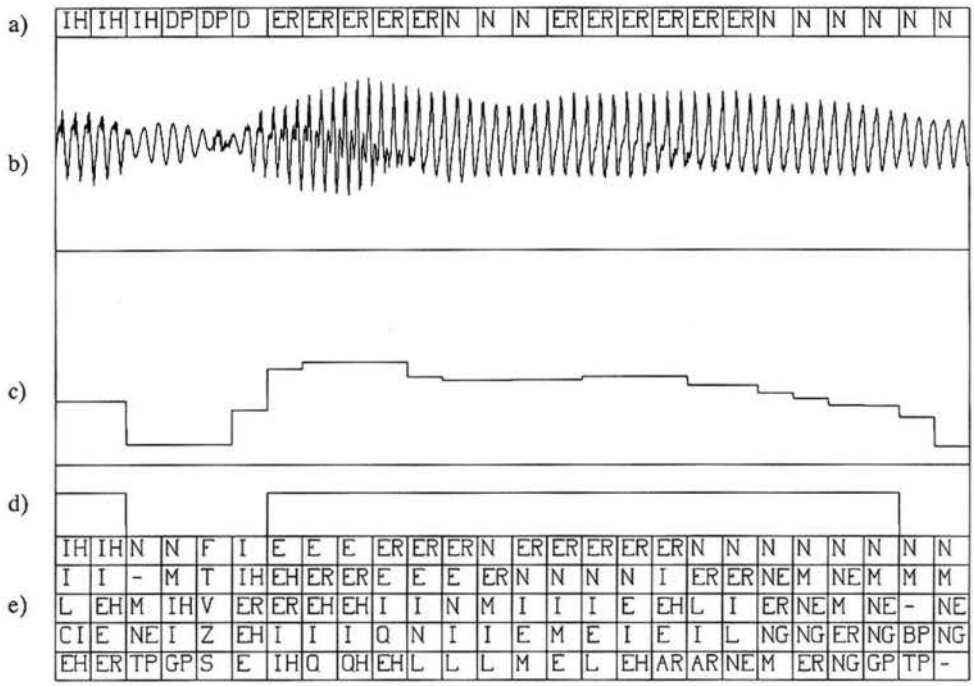


Bild 4.9: Mögliche Auftrennung verschmolzener Silbenkerne mit Hilfe der akustisch-phonetischen Segmentierung am Beispiel des Wortes *"(versch)iedenen"*

a) Handklassifikation auf Lautkomponenten-Ebene nach [REGEL 88]
b) Sprachsignal
c) geglätteter Bandpaßausgang (300 Hz - 2300 Hz)
d) aufgrund der Energie-Bänder berechnete Silbenkerngrenzen (Kap.4.1.2)
e) Lautkomponenten-Hypothesen

Ziel des neuen Segmentierers ist eine Zerlegung des Sprachsignals in lautähnliche Einheiten, wobei die Zahl der nicht gefundenen Laute so gering wie möglich zu halten ist. Dies ist sinnvoll, da Untersuchungen zur Worterkennung gezeigt haben, daß die Worthypothesen-Generierung auf Auslassungsfehler wesentlich empfindlicher reagiert als auf Einfügungsfehler (beim elementaren Hidden-Markov-Modell (HMM) mit zwei Zuständen entsprechen diese Fehler den Übergängen *Delete* und *Insert*, siehe [KUNZMANN 90]). Die Zahl der Auslassungsfehler kann aber i.allg. nur auf Kosten einer Erhöhung der Einfügungsfehler minimiert werden.

Ein grobes Maß für das Verhalten des Segmentierers ist der Grad der *Übersegmentierung*, d.h. die Anzahl der im Schnitt pro Laut hypothetisierten Segmente. Das für die Worthypothesen-Generierung benutzte Positionierungsverfahren (vertikale Summation) arbeitet erfahrungsgemäß dann am besten, wenn die Zahl der hypothetisierten Segmente ungefähr gleich der Zahl der zu vergleichenden elementaren HMM ist. Somit kann die Zahl der Auslassungsfehler nicht beliebig stark minimiert werden. Dies würde eine zu starke Übersegmentierung zur Folge haben.

Der Segmentierer beruht auf der Annahme, daß sich die Lautklasse, für die sich der Lautkomponentenklassifikator entscheidet (die beste der fünf Lautkomponentenhypothesen), oberflächengetreu verhält: In "konstanten" Bereichen, d.h. in der Mitte eines Lautes, ändert sich die Klassifikatorentscheidung nicht; an Lautübergängen schlägt sie mindestens einmal um. Ein Bereich, in dem sich die erste Alternative des Klassifikators nicht ändert, wird im folgenden als *Lauflänge* bezeichnet. Unter dieser Annahme reduzieren sich die potentiellen Lautanfänge (Segmentgrenzen) von jedem Frame auf jeden Beginn einer Lauflänge. In der EVAR-Stichprobe führt dies zu einer dreieinhalbfachen Übersegmentierung. Diese Übersegmentierung hat zwei Gründe:

1) Durch Klassifikationsfehler kann auch in der Mitte eines Lautes die Lauflänge umschlagen.
2) An den Lautgrenzen findet aufgrund kontextueller (koartikulatorischer) Einflüsse ein häufigerer Wechsel der Lauflängen statt.

Da zudem die Plosivlaute aus unterschiedlichen Lautkomponenten aufgebaut werden müssen, werden auf den Lauflängen *Segmentierungsalternativen* in Form eines *Segmentgraphen* gebildet. Bild 4.10 verdeutlicht das Vorgehen anhand zweier potentieller Lautanfänge:

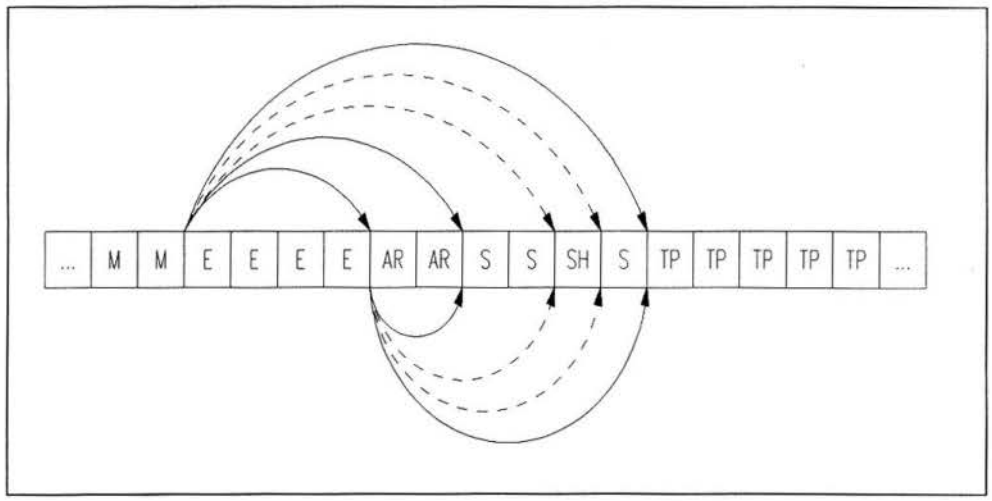


Bild 4.10: Mögliche Endpunkte für zwei Lautanfänge bei einer Beschränkung der Suchumgebung auf zehn Frames

Die Pfeile deuten Segmentkandidaten an, die innerhalb einer bestimmten Umgebung (hier: zehn Frames) angenommen werden. Sie stellen Alternativen dar, über denen alle Laute aus einem Lautinventar verifiziert werden ([FISCHER 89, Kap.3]). Die gestrichelt eingezeichneten Pfeile entfallen nach der Anwendung einer Glättungsoperation, bei der Lauflängen, die genau ein Frame lang sind und den gleichen linken und rechten Nachbarn besitzen, als Störungen aufgrund des Klassifikatorverhaltens interpretiert und eliminiert werden. Durch verschiedene Glättungsoperationen und den festzulegenden maximalen Suchbereich kann somit Einfluß auf die Größe des Segmentgraphen und den Grad der Übersegmentierung genommen werden.

Bei der Lautverifikation wird jeder Segmentkandidat mit dem Lautinventar verglichen. Die Bewertung der besten Lautalternative wird als Gewichtung in den Segmentgraphen eingetragen, d.h. nach der Verifikation liegt ein bewerteter Graph vor. Mit dem A*-Algorithmus wird in dem Graphen der optimale Pfad gesucht, wodurch eine optimale Zuordnung zwischen den verifizierten Lautalternativen und dem Sprachsignal erzeugt wird. Der A*-Algorithmus wurde verwendet, da in [KUHN 89] mit diesem Graphsuchverfahren Syntaxhypothesen erzeugt werden (siehe auch [KUNZMANN 90, Kap.10]). Somit ist eine durchgängiger Ansatz für die gesamte Erkennungsphase[2] in EVAR möglich.

Die Grenzen der Silbenkerne und Pausen, die mit dem in Kap.4.1.2 beschriebenen Verfahren erzeugt werden, können auf zwei Arten zur Aufwandsreduktion und Beeinflussung des Segmentierungsergebnisses eingesetzt werden:

1) Lauflängen, die Bereiche überdecken, die das Prosodie-Modul als Silbenkern oder Pause markiert hat, werden zu einer Lauflänge zusammengefaßt, wobei die Ränder dieser Gebiete, wenn nötig, angeglichen werden (siehe Bild 4.11).
2) Silbenkerne und Pausen werden als Begrenzungen des Segmentnetzes verwendet, d.h. über diesen Bereichen wird keine alternative Segmentierung erstellt. Dadurch reduziert sich nicht nur die Anzahl der zu verifizierenden Lauthypothesen, sondern es wird vor allem auch der Suchraum für den A* -Algorithmus eingeschränkt. Zusätzlich kann man dafür sorgen, daß über den Silbenkernen nur solche Lauthypothesen erzeugt werden, die potentiell Silbenkernträger sein können.

Bild 4.11 zeigt die Aufwandsreduktion exemplarisch anhand der Baumentwicklung eines Beispiel-Suchgraphen (es handelt sich um einen Ausschnitt aus einer Äußerung der EVAR-Stichprobe): Über dem Frameraster (Bild 4.11a) sind in Bild 4.11b die Lauflängen und in Bild 4.11c die Silbenkerngrenzen eingezeichnet. Die gestrichelt eingezeichneten Pfeile in Bild 4.11e zeigen die Segmentkandidaten aufgrund der Lauflängen, die durchgezogenen Pfeile zeigen die Segmentkandidaten, die nach Einbezug der Silbenkernsegmentierung übrigbleiben. Die Bilder 4.11d und 4.11f zeigen die Baumentwicklung der beiden Suchgraphen.

---

[2] Nach [NIEMANN 88a] kann die Verarbeitung einer Benutzeranfrage in einem sprachverstehenden Dialog-System in die drei Verarbeitungsphasen *Erkennung, Verstehen* und *Dialog* eingeteilt werden. Die Erkennungsphase kann in die Aufgaben *Segmentierung des Sprachsignals, Generieren und Verifizieren von Worthypothesen* sowie *Verkettung von Worthypothesen* unterteilt werden.

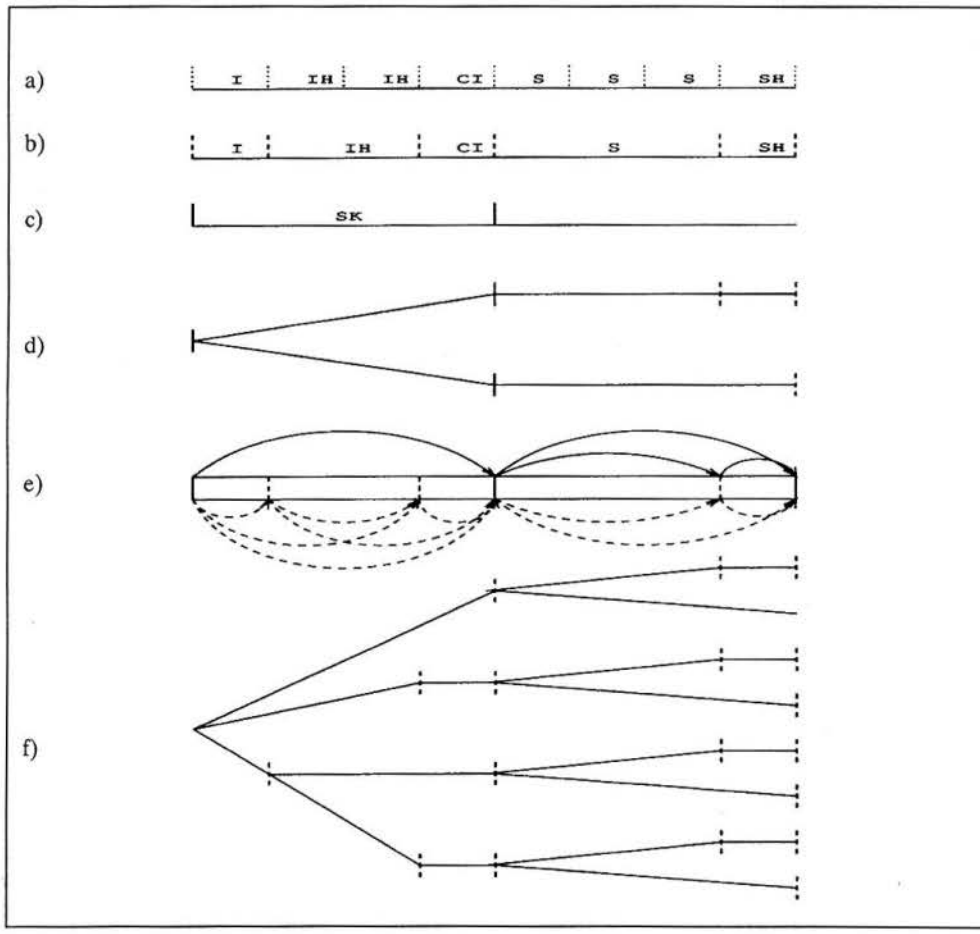


Bild 4.11: Suchraumreduktion für die Graphsuche durch den Einsatz prosodischer Information.

Bei der beschriebenen Vorgehensweise muß daran gedacht werden, daß sich insbesondere die Verschmelzung von Silbenkernen in Hinblick auf die intendierte Übersegmentierung und die Worterkennungsleistung negativ auswirken, da mindestens zwei Silbenkerne und eventuell vorhandene Laute im Silbenaus- und Silbenanlaut zu einer Lauthypothese verschmolzen werden. Man kann dieses Problem umgehen, indem man Silbenkernen ab einer gewissen Länge nicht mehr "glaubt", d.h. nach Angleichung der Silbenkerngrenzen an die Lauflängen auch innerhalb eines Silbenkernbereichs ein Segmentnetz aufbaut.

Aufgrund des mit dem A*-Algorithmus gefundenen optimalen Pfades innerhalb des Silbenkerns läßt sich die Segmentierung nach Kap.4.1.2 korrigieren, indem linksrandige und rechtrandige sonorante Konsonanten abgetrennt werden und Vokal-Vokal- bzw. Vokal-Konsonant-Vokal-Folgen in zwei Silbenkerne aufgetrennt werden.

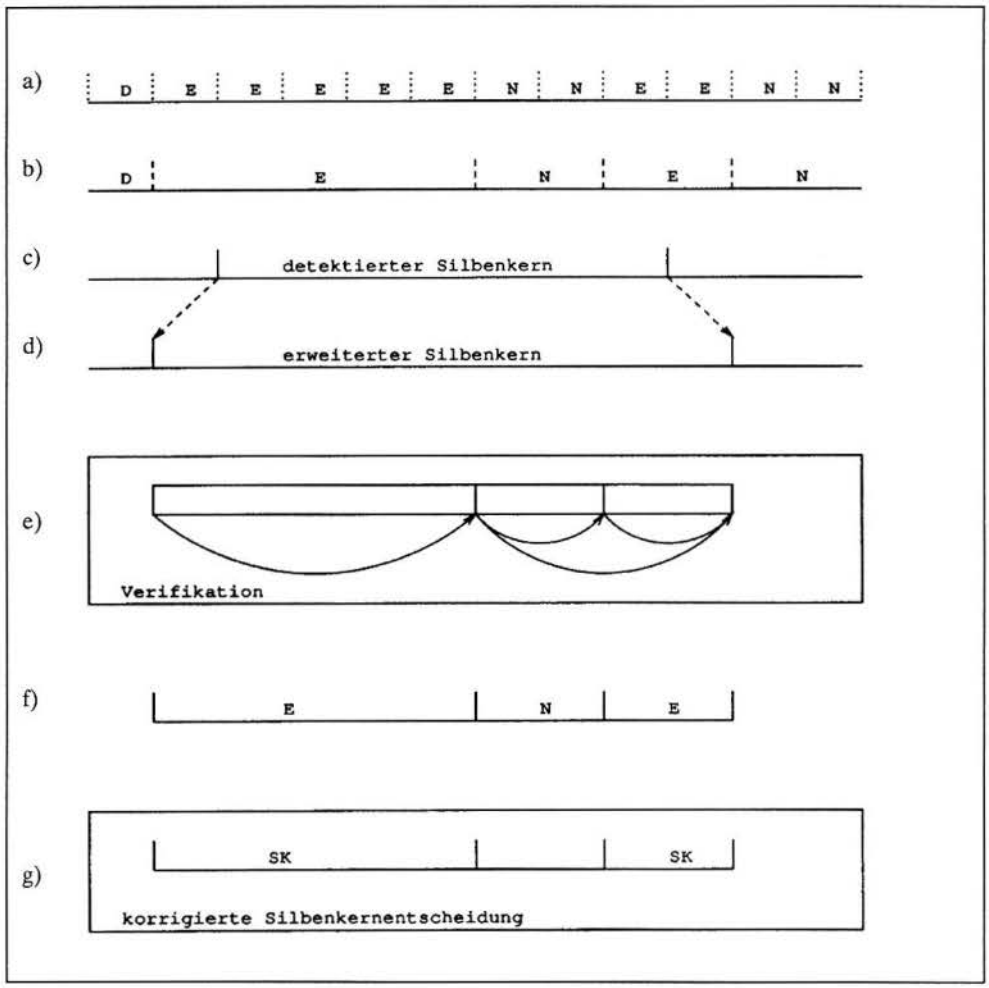


Bild 4.12: Auftrennung eines Silbenkerns aufgrund des optimalen Pfades der Lautsegmentierung innerhalb des Silbenkerns.

Bild 4.12 zeigt die Korrektur der Silbenkernsegmentierung exemplarisch für die Auftrennung von zwei verschmolzenen Silbenkernen. Über dem Frameraster (Bild 4.12a) sind in Bild 4.12b die Lauflängen und in Bild 4.12c die Silbenkerngrenzen eingezeichnet. In Bild 4.12d ist die Angleichung der Silbenkerngrenzen an die Lauflängen gezeigt. Innerhalb des Silbenkerns wird ein Segmentgraph aufgebaut (Bild 4.12e) und der optimale Pfad gesucht (Bild 4.12f). Aufgrund der Lautklassifikation wird die Silbenkern-Segmentierung korrigiert (Bild 4.12g).

4.2 Tonhöhenmerkmale

Im Rahmen der Arbeiten am Prosodie-Modul von EVAR wurden umfangreiche Untersuchungen zur Berechnung des Tonhöhenkorrelats *Grundfrequenz* und zur Berechnung signifikanter Merkmale aus dem F_0-Verlauf durchgeführt: In [HEUNISCH 86] werden verschiedene aus der Literatur bekannte Grundfrequenzalgorithmen auf ihre Brauchbarkeit für das Prosodie-Modul hin untersucht. In [KOMPE 89b] wird auf der Basis der in [HEUNISCH 86] implementierten Algorithmen ein eigener Grundfrequenzalgorithmus entwickelt. In [KIESSLING 89] wird eine interaktive Arbeitsumgebung zum periodenweisen Segmentieren des Sprachsignals beschrieben. Methoden zur Stilisierung der F_0-Kontur mit Linienelementen finden sich in [LANG 87] und [STALLWITZ 89]. In [NÖTH 87], [BATLINER 89a, 89b] werden verschiedene aus der F_0-Kontur abgeleitete Merkmale und Transformationen auf Brauchbarkeit für die Modus- und Fokusklassifikation hin untersucht.

Im folgenden wird die Grundfrequenz vom Standpunkt der Produktion her definiert (Kap.4.2.2). Hierfür ist vorher eine kurze Darstellung des Vorgangs der Spracherzeugung und seiner Modellierung durch ein lineares System notwendig (Kap.4.2.1). Danach wird anhand der beiden in [HEUNISCH 86] implementierten Algorithmen zur F_0-Bestimmung das prinzipielle Vorgehen bei der F_0-Berechnung dargestellt (Kap.4.2.3). In Kap.4.2.4 wird auf das Prinzip des in [KOMPE 89b] implementierten Mehrkanalverfahrens eingegangen. Normierungsoperationen und Extraktion von Merkmalen aus der F_0-Kontur werden in Kap.4.2.5 behandelt.

Eine genaue Darstellung des Vorgangs der Spracherzeugung und ein sehr ausführlicher Überblick über die verschiedenen Geräte und Algorithmen zur Berechnung der Grundfrequenz finden sich in [HESS 83] (einschließlich einer nach den verschiedenen Aspekten und Methoden der Grundfrequenzbestimmung klassifizierten Bibliographie).

4.2.1 Ein Modell der Spracherzeugung

Bild 4.13 zeigt eine schematische Darstellung des Atmungs- und Artikulationssystems. An der Lauterzeugung sind im wesentlichen die Atmungsorgane und Stimmbänder (zur Erzeugung des Anregungssignals), Mund- und Nasenraum (als Resonanzraum) sowie Lippen, Zunge, Zähne, Gaumen und Zäpfchen (als Artikulatoren) beteiligt. Die Stimmbänder können mindestens zwei Stellungen einnehmen: Weit geöffnet (bei der Artikulation von stimmlosen Lauten) oder quasi-periodisch geöffnet und geschlossen (bei der Artikulation der stimmhaften Laute). Ist bei stimmhafter Artikulation die Glottis geschlossen, so wird durch die Atmungstätigkeit ein Druck aufgebaut. Durch den ansteigenden Druck wird der Verschluß gesprengt. Die Luft entweicht vom *subglottalen* in den *supraglottalen* Bereich, der dadurch entstehende Druckabfall führt aufgrund des Bernoulli-Effekts zu einem erneuten Verschluß.

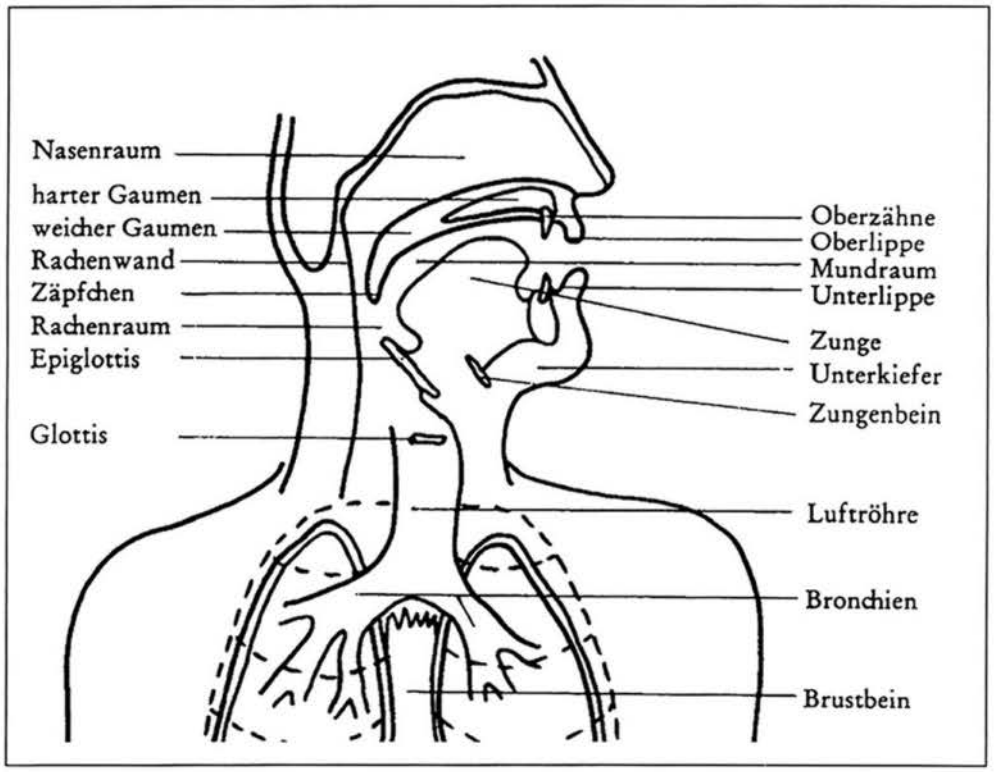


Bild 4.13: Atmungs- und Artikulationssystem (aus [Kohler 77]).

Dieser Vorgang wiederholt sich quasi-periodisch und erzeugt ein oberwellenreiches Signal. Als *glottaler* Zyklus wird der Zeitraum von einem vollständigen Verschluß der Glottis bis zum nächsten vollständigen Verschluß bezeichnet.

Der Mund- und Rachenraum bilden eine zeitlich variable akustische Röhre, die als *Ansatzrohr* oder *Vokaltrakt* bezeichnet wird. Durch Heben oder Senken des weichen Gaumens kann der Nasenraum "dazugeschaltet" werden. Unter Vernachlässigung des Nasenraumes kann der Vorgang der Spracherzeugung in guter Näherung durch ein zeitvariantes lineares Filter `H` modelliert werden (Bild 4.14). Das Filter, das die Resonanzeigenschaften des Vokaltrakts nachbildet, wird für stimmhafte Laute durch die Impulsfolge und für stimmlose Laute durch weißes Rauschen angeregt. Die Maxima im Modellspektrum `H(z)` entsprechen den für jede Vokaltraktstellung charakteristischen Resonanzfrequenzen (*Formanten*). Bild 4.15 zeigt das Schema der Lauterzeugung nach diesem Modell:

Das Spektrum (4.15d) des Anregungssignals (4.15a) wird mit dem Spektrum (4.15e) des Vokaltrakts multipliziert und ergibt das Spektrum (4.15f) des Sprachsignals (4.15c).

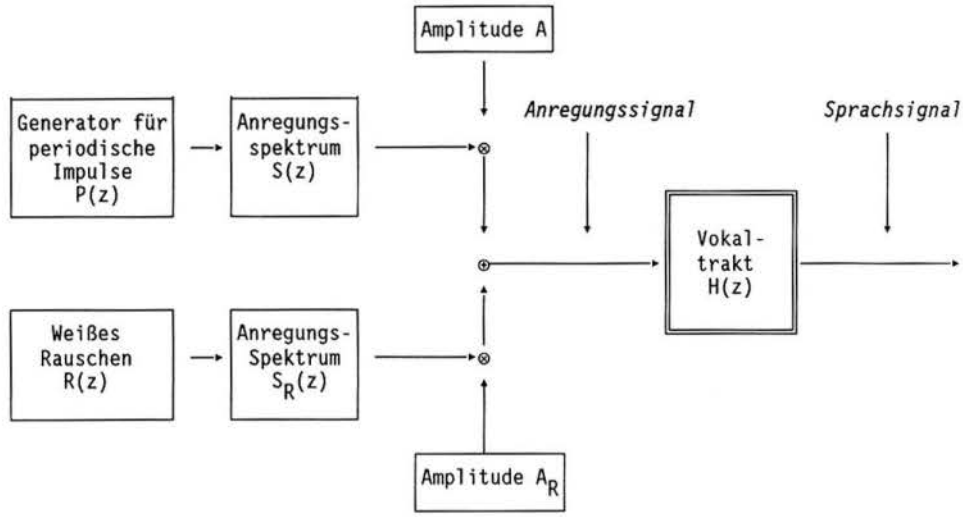


Bild 4.14: Lineares Modell der Spracherzeugung (nach [HESS 83]).

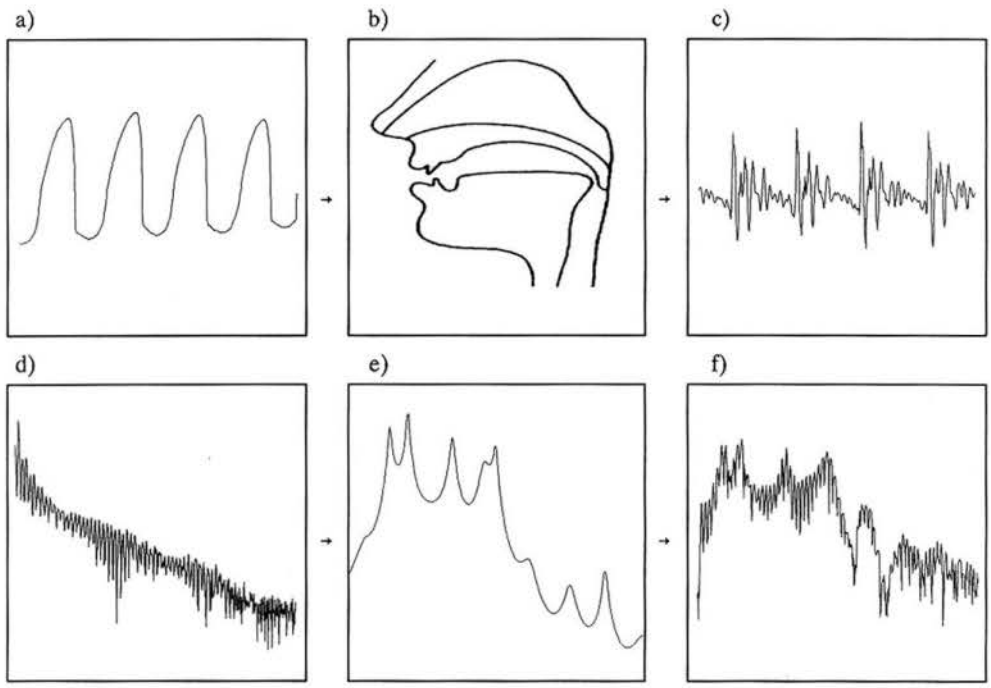


Bild 4.15: Schema zur Lauterzeugung:
a) Anregungssignal
b) Vokaltrakt
c) Sprachsignal
d) Spektrum des Anregungssignals
e) Spektrum des Vokaltrakts
f) Spektrum des Sprachsignals

4.2.2 Grundperiode und Grundfrequenz

Von der Produktion aus gesehen ist die *Grundperiode* T_o eines stimmhaften Lautsignals als die *Dauer eines glottalen Zyklus* eindeutig definiert. Die *Grundfrequenz* F_o des Lautsignals ergibt sich dann zu jedem Zeitpunkt t zu $1/T_0$. Die Stimmbandschwingungen (und damit die Grundfrequenz) lassen sich mit Hilfe eines Laryngographen eindeutig bestimmen ([HESS 83, S.116ff], [INDEFREY 88]). Bild 4.16 zeigt für einen 37.5 Millisekunden langen Ausschnitt aus dem ersten /A/ des Logatoms *"aba"* (4.16a) das Laryngogramm (4.16b) und in 4.16c das differenzierte Laryngogramm. Zum Ausgleich der Laufzeit des Sprachsignals vom Mund des Sprechers zum Mikrophon wurde das Sprachsignal um eine Millisekunde verschoben[3].

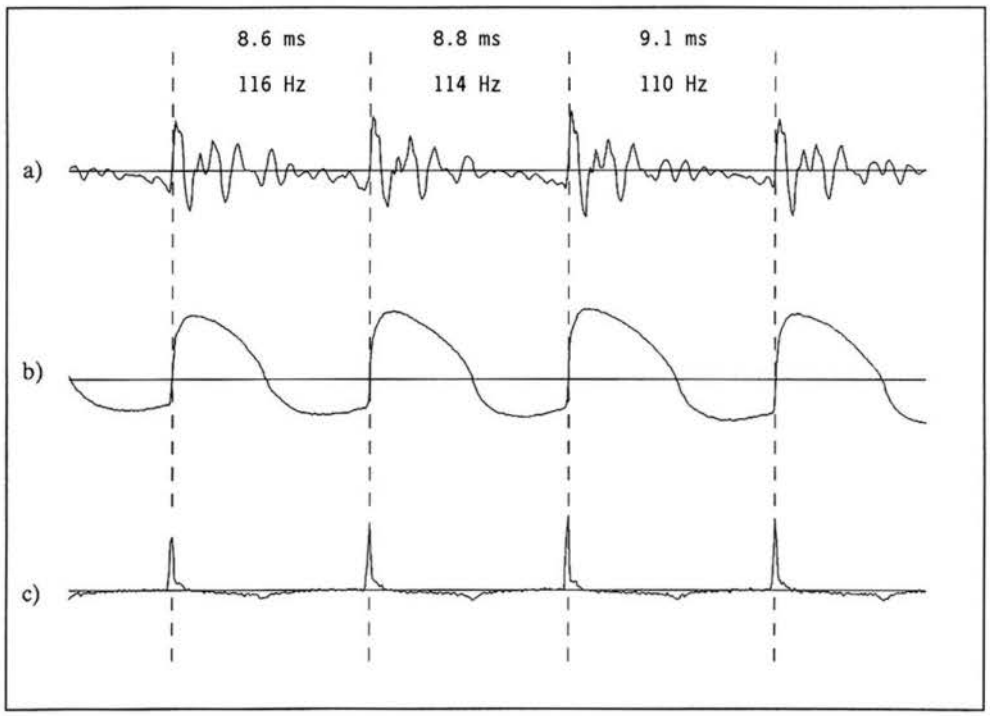


Bild 4.16: Ausschnitt aus dem ersten /A/ des Logatoms *"aba"*. Der Beginn eines glottalen Zyklus ist jeweils durch die gestrichelten Linien markiert.

a) Sprachsignal
b) Laryngogramm
c) Differenziertes Laryngogramm

---

[3] An dieser Stelle sei Herrn Ruske von der Technischen Universität in München noch einmal ausdrücklich für die Bereitstellung des Sprach- und Laryngosignals gedankt.

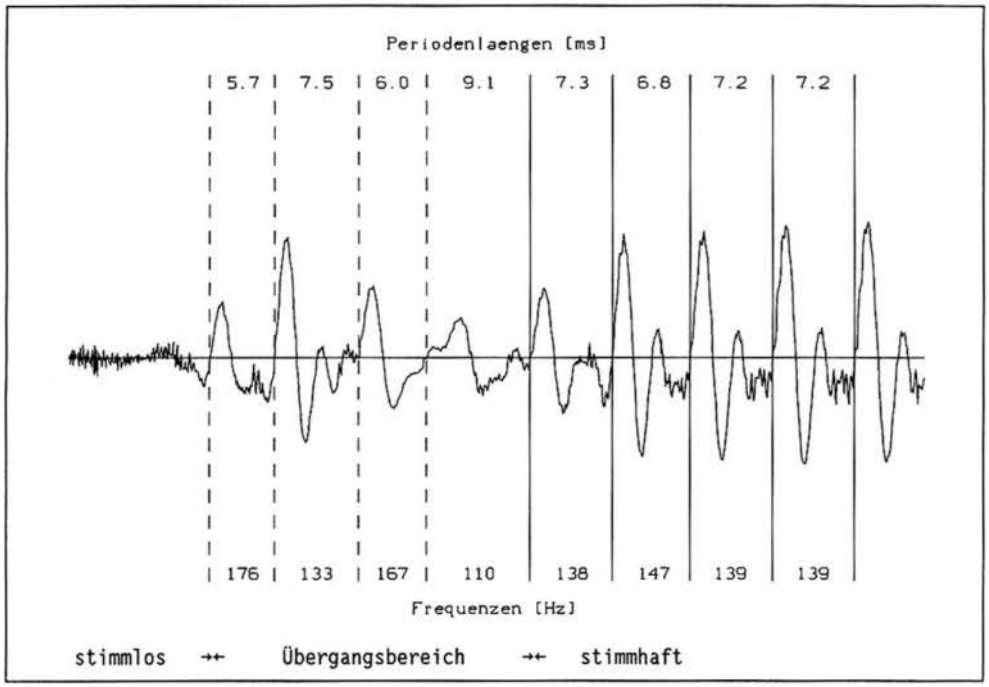


Bild 4.17: Typische Unregelmäßigkeit beim Stimmlos/stimmhaft-Übergang.

Der Zeitpunkt des vollständigen Verschlusses der Glottis ist im differenzierten Laryngogramm durch das Maximum deutlich gekennzeichnet ([INDEFREY 88, S.28]). Der größte positive Ausschlag im Laryngogramm zeigt den größten Kontakt der Stimmbänder (das Laryngogramm verhält sich also gerade umgekehrt zum supraglottalen Luftstrom; siehe [HESS 83, S.43] und die dort zitierte Untersuchung von [LECLUSE 77]). Die Dauer der drei vollständig dargestellten Grundperioden in Bild 4.16 sowie die entsprechende Grundfrequenz sind über dem Sprachsignal eingezeichnet. Man sieht sofort, daß es sich hier nur um ein quasi-periodisches Signal und nicht um ein periodisches Signal im mathematischen Sinne handelt. Den Maxima im differenzierten Laryngosignal (das i.allg. nicht zur Verfügung steht) entsprechen periodisch wiederkehrende Strukturmerkmale. Häufig werden die Grundperioden am positiven Nulldurchgang vor der *Leitamplitude* (die maximale Amplitude einer Grundperiode) markiert ([KIESSLING 89]). Da i.allg. nicht der exakte Beginn des glottalen Zyklus interessiert, werden auch andere Strukturmerkmale, insbesondere die Leitamplitude selbst, verwendet.

Das in Kap.4.2.1 vorgestellte Modell ist in mehreren Punkten zu vereinfachend, um den Vorgang der Spracherzeugung korrekt zu beschreiben. Die Transition von stimmhaften zu stimmlosen Lauten und umgekehrt geschieht nicht verzögerungsfrei, so daß es zu Einschwingvorgängen kommen kann. Bild 4.17 zeigt einen Ausschnitt aus einem Satz der Fokus-Stichprobe, gesprochen wurde *"(S) ie (läßt die Nina ...)"*.

Man erkennt deutlich (zumindest im Zeitsignal; das Laryngosignal liegt zu den Modus-Fokus-Korpora leider nicht vor) das Einschwingen beim Stimmlos/stimmhaft-Übergang von /S/ zu /I/. Die Striche deuten die "Grundperioden" an, wenn man am positiven Nulldurchgang vor der Leitamplitude schneidet.

Auch innerhalb von stimmhaften Bereichen kann es durch Lautübergänge und Formant-Transitionen vorkommen, daß ein Strukturmerkmal (Nulldurchgang vor der Leitamplitude, Leitamplitude, ...) nicht konsistent zum Markieren der Grundperiode verwendet werden kann, wie das Bild 4.18 verdeutlicht.

Obwohl die periodische Struktur des Signals deutlich zu erkennen ist, können die Grundperioden nicht konsistent an einem Strukturmerkmal (Nulldurchgang, Maximum der Periode, Minimum der Periode) markiert werden. Die durchgezogenen Linien zeigen die Grundperioden an, wenn man am Nulldurchgang vor der Leitamplitude schneidet, die gestrichelten zeigen eine mögliche Fortführung der periodenweisen Markierung im Transitionsbereich an.

Weiterhin können die Stimmbänder i.allg. mehr als die zwei beschriebenen Stellungen *stimmlos* und *stimmhaft* einnehmen. Neben *Flüstern* und *behauchter Phonation* sind dies der *Glottisverschluß (harter Vokaleinsatz)* und die *Laryngalisierung*. Die letzten beiden Phonationsarten sind hier von starkem Interesse, da sie bei der automatischen Grundfrequenzberechnung starke Probleme aufwerfen (siehe Kap.6.1). Der *Glottisverschluß* ist ein phonetisches Grenzsignal, das vor vokalischen Wort- oder Morphemanfängen auftreten kann (z.B. bei *"da ich"* vs. *"Deich"*). In [KOHLER 77] wird die Produktion des Glottisverschlusses folgendermaßen beschrieben:

> "Ein Sonderfall ist das einfache Unterbrechen des Luftstroms durch ein Verschließen der Glottis, die dann anschließend wieder geöffnet wird oder übergeht in eine periodische Öffnungs- und Schließbewegung, ...
>
> [KOHLER 77, S.60]

Bild 4.19 zeigt einen Ausschnitt aus der Äußerung *"(... ich möcht)e a(m nächsten Sonntag ...)"*. Über dem Zeitsignal ist die Dauer der Verschluß- und Lösungsphase des Glottisverschlusses in Millisekunden aufgetragen, die benachbarten Vokale sind nicht vollständig dargestellt.

Die *Laryngalisierung* (andere Bezeichnungen sind *"creaky voice"* und *"vocal fry"*) kann alternativ zum *Glottisverschluß* als Grenzsignal eingesetzt werden oder alternativ zum *tiefen Offset* bei der Markierung des Satzmodus. In [LEHISTE 70] wird die Produktion einer *Laryngalisierung* folgendermaßen beschrieben:

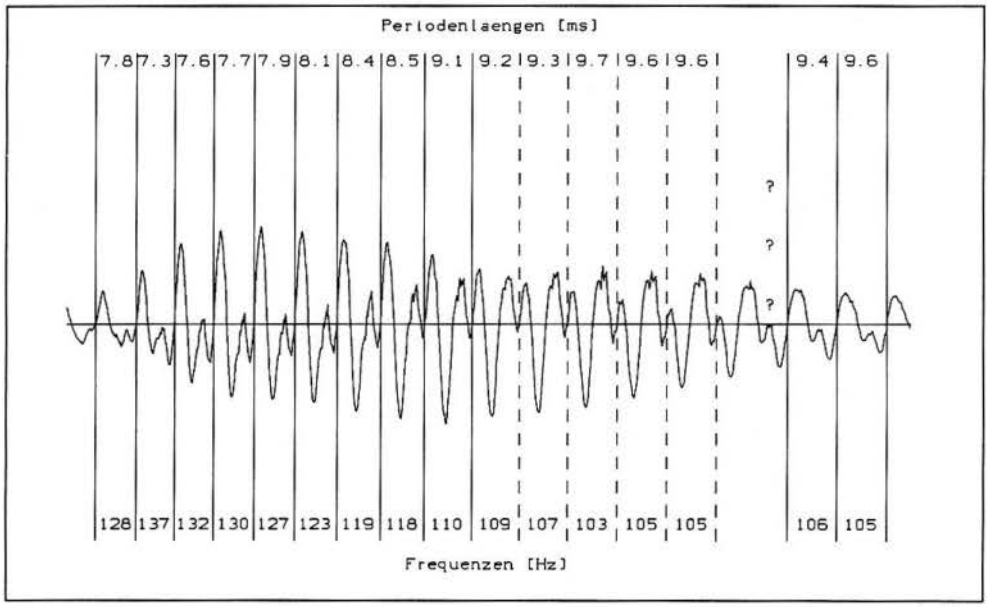


Bild 4.18: Stimmhaftes Sprachsignal, bei dem sich die Grundperioden nicht konsistent an einem Strukturmerkmal schneiden lassen.

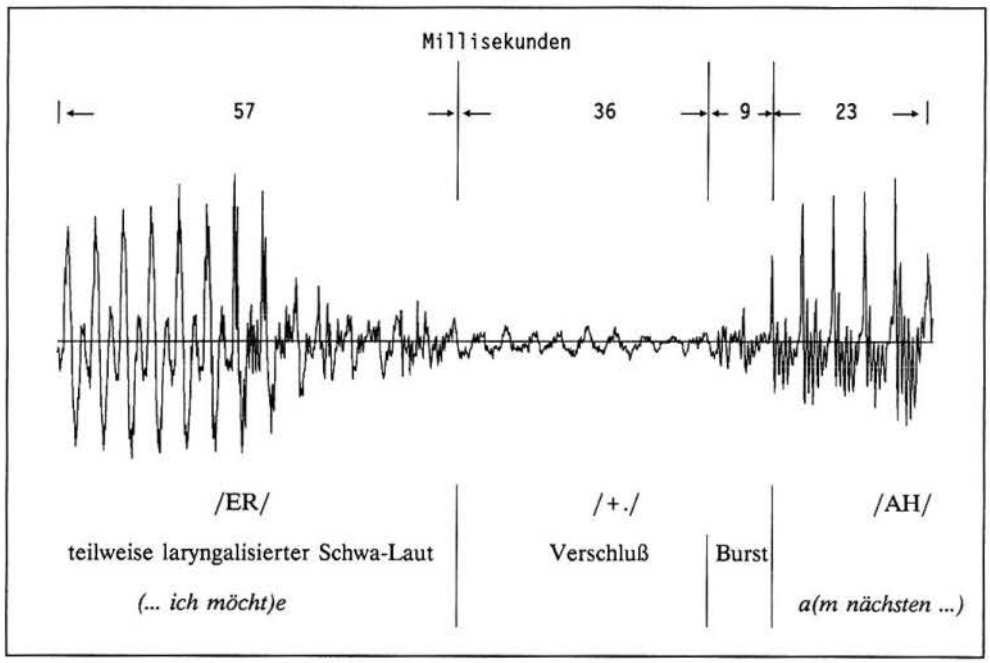


Bild 4.19: Beispiel für Glottisverschluß bei Vokal-Vokal-Übergang an einer Wortgrenze. Gesprochen wurde *"(... ich möcht)e a(m nächsten Sonntag ...)"*.

> "... It is the type called 'creak' by Catford, who describes the physiological process involved as a low-frequency periodic vibration of a small section of the vocal folds. Catford assumes that only a very small section of the ligamental glottis, near the thyroid end, is involved, and that the mean rates of airflow are very small. The vibrations have a frequency of about 40 cycles per second, ...
>
> When these modes of vibration are used in a linguistically significant way, I prefer to use the term *laryngealization,* referring to irregular, biphasic, or unusually slow vibration of the vocal folds.
>
> [LEHISTE 70, S.58-60]

Bild 4.20 zeigt einen Ausschnitt aus der Äußerung *"(Wer säuft de)nn ei(gentlich?)"*. Am oberen Bildrand sind die Periodenlängen eingezeichnet, am unteren die entsprechenden Grundfrequenzwerte (die Perioden wurden am Nulldurchgang vor der Leitamplitude geschnitten). Die ersten drei Perioden gehören zum /N/ und sind in normaler Phonation gesprochen, ebenso wie die letzten vier dargestellten Perioden des /AI/. In dem laryngal artikulierten Teil des /AI/ fällt die Grundfrequenz innerhalb weniger Perioden um eine Oktave.

Genaugenommen kann man nicht von einem alternativen Einsatz von Glottisverschluß und Laryngalisierung als Grenzsignal sprechen, denn der vollständige Glottisverschluß ist nur sehr selten zu beobachten, und der Übergang zwischen den beiden Phonationsarten ist fließend. Es handelt sich also nicht um zwei klar getrennte Kategorien, sondern um ein Kontinuum, wie Bild 4.21 verdeutlicht. Dargestellt ist ein Ausschnitt aus der Äußerung *"(... möchten S)ie a(breisen?)"*. Der "Glottisverschluß" ist im Zeitsignal deutlich erkennbar. Die relativ hohen Amplitudenwerte während der "Verschlußphase" und die teilweise laryngale Realisierung des nachfolgenden Vokals lassen allerdings vermuten, daß hier kein vollständiger Verschluß der Glottis vorlag.

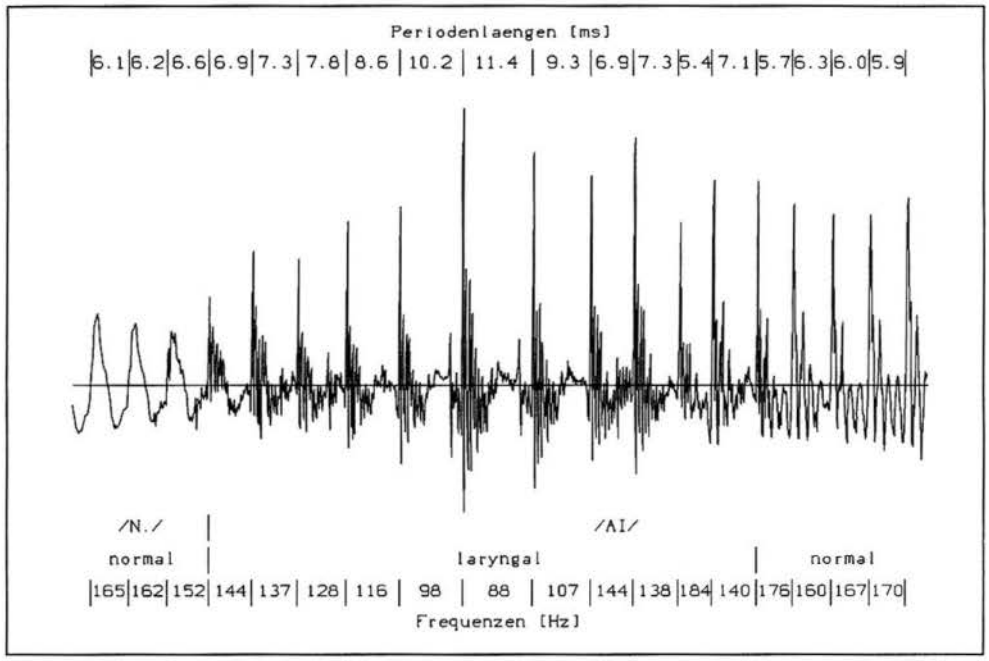


Bild 4.20: Wortgrenzenmarkierung durch laryngale Artikulation des Vokalanfangs an der Wortgrenze zwischen *"denn"* und *"eigentlich"*.

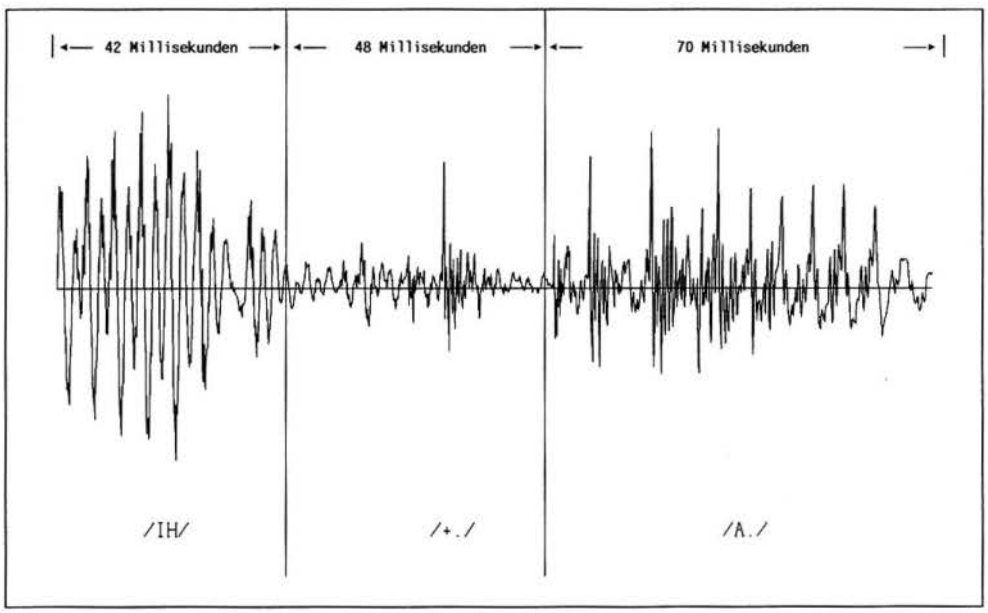


Bild 4.21: Beispiel für Glottisverschluß bei Vokal-Vokal-Übergang an einer Wortgrenze. Gesprochen wurde *"(Zu welcher Zeit möchten S)ie a(breisen?)"*.

4.2.3 Berechnung von Grundfrequenzschätzwerten im Zeit- und Frequenzbereich

Die Definition der Grundfrequenz als Kehrwert der aktuellen Grundperiode ist für diese Arbeit nicht brauchbar. Zum einen wird der F_0-Wert im festen Zeitraster (Frame) benötigt, zum anderen ist die Abbildung der Glottisbewegungen auf die Strukturmerkmale im Zeitsignal durch Lautübergänge (Einfluß energiereicher Formanten) zu stark gestört, wie in Bild 4.18 dargestellt ist. Für die automatische Grundfrequenzbestimmung werden deshalb meist *Kurzzeitanalyse-Verfahren* (*short-term analysis pitch determination algorithms* in [HESS 83]) verwendet.

Hierbei wird das Zeitsignal in Analysefenster zerlegt, die lang genug sein müssen, um "einige" Perioden zu enthalten und kurz genug, um die interessierenden F_0-Veränderungen nicht wegzuglätten. Abgesehen von Laryngalisierungen, bei denen die Grundfrequenz auf bis zu 28 Hz abfallen kann ([LEHISTE 70, S.59]), liegt die Grundfrequenz größenordnungsmäßig zwischen 50 (sehr tiefe Männerstimmen) und 550 Hz (sehr hohe Frauen- und Kinderstimmen). Möchte man diesen Bereich abdecken, so benötigt man ein Analysefenster von ca. 40 Millisekunden, um sicherzustellen, daß mindestens zwei volle Perioden im Analysefenster enthalten sind (50 Hz $\triangleq$ einer Grundperiodendauer von 20 Millisekunden). Bei sehr hohen Stimmen liegen ca. 10 Perioden in einem 40 Millisekunden-Fenster, so daß auch hier eine ausreichende Genauigkeit gewährleistet ist. Bei den Experimenten, über die in Kap.6 berichtet wird, wurde eine Fensterlänge von 38.4 Millisekunden bei den mit 10 kHz digitalisierten Signalen verwendet und eine Fensterlänge von 37.5 Millisekunden für die 16-kHz-Signale. Die Fortschaltzeit betrug 12.8 bzw. 12.5 Millisekunden. Für jedes Analysefenster wird ein Schätzwert für die im Fenster enthaltenen Grundperioden berechnet.

In [HEUNISCH 86] wurden verschiedene Grundfrequenzalgorithmen auf ihre Brauchbarkeit im Rahmen von EVAR untersucht. Dabei wurde insbesondere auf die Ergebnisse in [HESS 83, Kap.9.3], [RABINER 76], [HETTWER 85], und [MCGONEGAL 77] zurückgegriffen. Es zeigte sich, daß keines der dort untersuchten Verfahren für alle Stimmlagen (die durchschnittliche Grundfrequenz), Sprecher, Fehlerarten und Aufnahmebedingungen am besten abschnitt. Jedes Verfahren hatte seine Schwachpunkte. Daher wurden im Hinblick auf eine Mehrkanallösung (siehe Kap.4.2.4) das AMDF-Verfahren ([ROSS 74]) und das Seneff-Verfahren ([SENEFF 78]) implementiert. Für das SIFT-Verfahren ([MARKEL 72]) und das CEPSTRUM-Verfahren standen Implementierungen im ILS-Software-Paket[4] zur Verfügung.

Unter der für ein Auskunftssystem wichtigen Randbedingung von Beschränkung des Zeitsignals auf Telefonqualität (simuliert über einen Hochpaß-Filter mit Grenzfrequenz 300 Hz) zeigte das Seneff-Verfahren in [HEUNISCH 86] das beste Grobfehlerverhalten (siehe Kap.6.1 für eine Definition der verschiedenen Fehlerarten). Das SIFT-Verfahren zeigte in Vorversuchen ein sehr schlechtes Feinfehler-Verhalten und wurde nicht weiter untersucht. Das CEPSTRUM-Verfahren war bei Simulation von Telefonqualität nicht ganz so gut wie das Seneff-Verfahren. In weiteren

---

[4] ILS ist eingetragenes Warenzeichen der Firma Signal Technology Inc.

Versuchen für die Implementierung des in Kap.4.2.4 beschriebenen eigenen Grundfrequenzverfahrens (unveröffentlicht) zeigte das CEPSTRUM-Verfahren bei Veränderung der Stimmhaft/stimmlos-Entscheidung ein schlechteres Fehlerverhalten (siehe hierzu auch [MCGONEGAL 77]). Das AMDF-Verfahren schnitt schlechter ab als das Seneff-Verfahren, wurde aber weiterentwickelt, da es nach [RABINER 76] für hohe Stimmlagen gute Ergebnisse zeigte (das Fehlerverhalten in Abhängigkeit von der Stimmlage wurde im Rahmen dieser Arbeit nicht weiter untersucht).

Da das AMDF- und das Seneff-Verfahren aufgrund der Ergebnisse in [HEUNISCH 86] für den eigenen Mehrkanal-Grundfrequenz-Algorithmus verwendet werden, und da es sich bei den beiden Verfahren um typische Vertreter von Zeit- und Frequenzbereichsverfahren handelt, werden im folgenden die notwendigen Verarbeitungsschritte bei der Berechnung der Grundfrequenzschätzwerte anhand dieser beiden Verfahren erläutert. Das im Zeitbereich arbeitende *AMDF-Verfahren* (Kap.4.2.3.5) ist eine Korrelationstechnik. Die Korrelationsverfahren beruhen auf der Tatsache, daß die Autokorrelationsfunktion einer periodischen Funktion für ganzzahlige Vielfache der Grundperiode ein Maximum annimmt. Das im Frequenzbereich arbeitende *Seneff-Verfahren* (Kap.4.2.3.6) führt eine harmonische Analyse durch. Es beruht auf der Tatsache, daß ein periodisches Signal nur Frequenzanteile für ganzzahlige Vielfache der Grundfrequenz besitzt (Oberwellen).

Die zur Erläuterung verwendeten Abbildungen beziehen sich auf den 37.5 Millisekunden langen Ausschnitt aus dem ersten /A/ des Logatoms *"aba"*, der in Bild 4.16 dargestellt ist.

Eine genauere Beschreibung der Algorithmen wird hier verzichtet, da es sich um bekannte Verfahren handelt, die außer in den Original-Veröffentlichungen auch in [HESS 83] ausführlich beschrieben sind.

4.2.3.1 Bestimmung der stimmhaften Bereiche eines Zeitsignals

Die Bestimmung der Grundfrequenz ist nur in stimmhaft angeregten Teilen des Sprachsignals sinnvoll möglich. Daher muß das Sprachsignal zunächst in stimmhafte Bereiche (*SH*) sowie in stimmlose Bereiche und Sprechpausen (*SL*) zerlegt werden (*SH/SL-Entscheidung*). Hierzu eignen sich Parameter wie Energie, Nulldurchgangsrate, maximale Amplitude oder das Ergebnis des Akustik-Phonetik-Moduls, d.h eine Klassifikation nach Lauten oder Lautkomponenten.

Da in SL-Bereichen die Grundfrequenzanalyse i.allg. zufällige Werte erzeugt, hängt die Zahl der fehlerhaften Schätzwerte stark von der SH/SL-Entscheidung bzw. von den verwendeten Schwellwerten ab. Sind die SH-Schwellen zu weich eingestellt, so steigt die Zahl der fehlerhaften (konturverfälschenden) Werte; ist sie zu hart eingestellt, so wird die Kontur durch fehlende Werte verfälscht (siehe hierzu jedoch die Ergebnisse in Kap.6.1).

Im Rahmen der Arbeiten am EVAR-Prosodie-Modul wurden vier unterschiedliche SH/SL-Entscheidungen mit verschiedenen Schwellen untersucht (das Fehlerverhalten der beiden besten SH/SL-Entscheidungen - der dritten und der vierten - ist in Kap.6.1.1 beschrieben):

1) F_0-Werte nur über den Silbenkern-Bereichen
2) F_0-Werte in Bereichen, in denen die Energie im Bereich zwischen 300 und 2300 Hz eine Schwelle S überschreitet
3) F_0-Werte in Bereichen, in denen die Summe der abgeschätzten *a posteriori-Wahrscheinlichkeiten* der sonoranten Lautkomponentenklassen eine Schwelle S übersteigt.
4) F_0-Werte in Bereichen, in denen die Strukturmerkmale *Anzahl Nulldurchgänge, Amplitudenquadrat* und *Amplitudenmaximum* eines Frames gewisse Schwellwert-Bedingungen erfüllen. Diese SH/SL-Entscheidung ist in [KIESSLING 89] für einen interaktiven periodensynchronen Grundfrequenzalgorithmus entwickelt worden und brachte bei dem in Kap.4.2.4 beschriebenen Verfahren die besten Ergebnisse. Es wird daher kurz erläutert.

Sei $a_i(k)$, k=1,...,N der k-te Abtastwert des i-ten Frames und W der Wertebereich des Signals, also die Differenz zwischen größter positiver Amplitude und kleinster negativer Amplitude der Äußerung (das Signal wird vor der SH/SL-Klassifikation global mittelwertfrei gemacht). Bei der SH/SL-Entscheidung nach [KIESSLING 89] wird das Frame i als *stimmhaft* klassifiziert, wenn die drei folgenden Bedingungen erfüllt sind:

1) Relative Anzahl Nulldurchgänge	ZeroX/N	<	VUV_1
2) Amplitudenquadratmittel	$(\sum a_i^2(k))/N/W$	>	VUV_2
3) Amplitudenbetragsmaximum	$\max\|a_i(k)\|/W$	>	VUV_3

Sehr kurze stimmhafte Bereiche werden häufig mit einem Median-Filter weggeglättet. Hier werden zunächst stimmlose Bereiche der Länge eins mit den benachbarten stimmhaften Bereichen verschmolzen und danach stimmhafte Bereiche der Länge eins mit den benachbarten stimmlosen.

4.2.3.2 Tiefpaßfilterung

Da die maximale Grundfrequenz bei ca. 500 Hz liegt, wird das Sprachsignal meistens mit einer Grenzfrequenz von ca. ein kHz tiefpaßgefiltert und mit ca. zwei kHz neu abgetastet (downsampling). Durch diese Datenreduktion ist eine erhebliche Aufwandsreduktion gewährleistet (z.B. bei der für die Frequenzbereichsverfahren notwendigen *schnellen Fouriertransformation*). Zusätzlich werden der Einfluß der höheren Formanten und Rauschanteile im Signal unterdrückt.

Bei einer Grenzfrequenz von ca. 1 kHz ist gewährleistet, daß i.allg. mindestens zwei Harmonische (ganzzahlige Vielfache der Grundfrequenz) im Spektrum des reduzierten Signals vorhanden sind, selbst wenn durch die Beschränkung auf Telefonbandbreite (0.3-3.4 kHz) die Grundfrequenz nicht im Spektrum enthalten ist.

Bild 4.22 zeigt den oben erwähnten Ausschnitt aus dem Logatom *"aba"* vor (4.22a) und nach der Tiefpaßfilterung (4.22b). Es wurde ein digitales Bandpaßfilter (100-1100 Hz) verwendet.

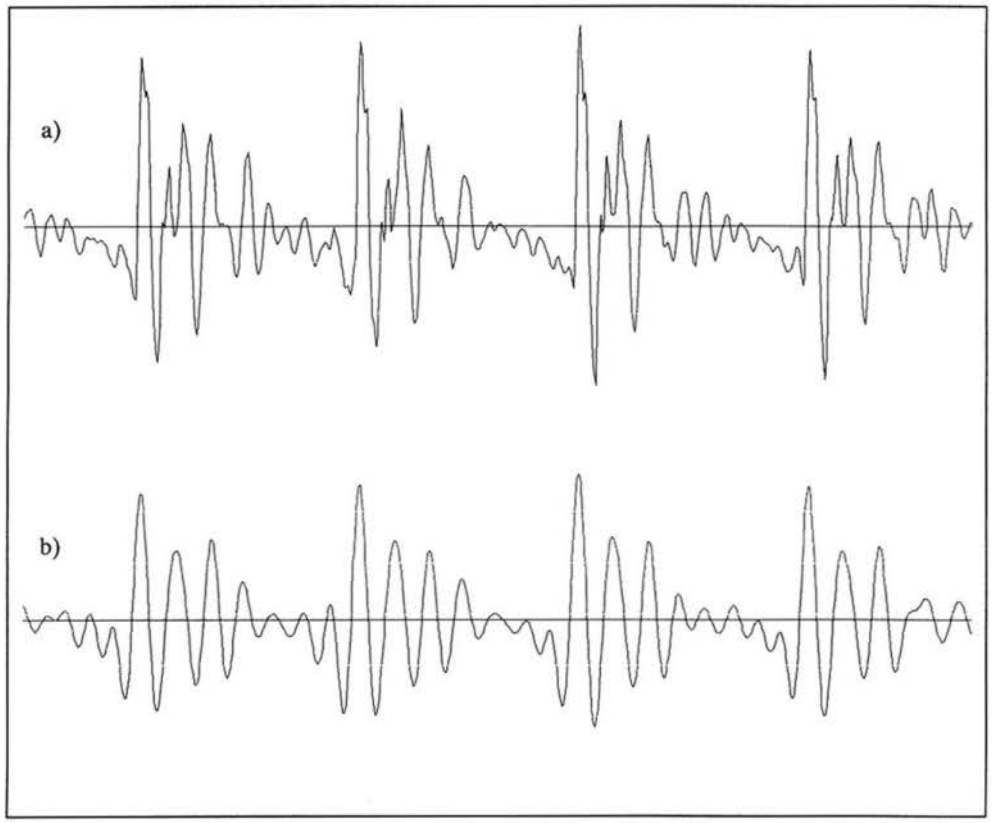


Bild 4.22: Analysefenster (37.5 Millisekunden) vor (a) und nach (b) der Tiefpaßfilterung (0.1-1.1 kHz).

4.2.3.3 Center Clipping und dreistufige Quantisierung

Center Clipping ([SONDHI 68]) ist eine nichtlineare Verzerrungstechnik mit dem Ziel, die Formantstruktur zu zerstören, aber die Periodizitäten im Sprachsignal unverändert zu lassen. Durch das Entfernen der Resonanzfrequenzen sollen die Harmonischen der Grundfrequenz auf gleiche Amplitudenhöhe gebracht werden (*spectrum flattening*). Hierzu werden die lokalen Extremwerte, die nicht auf den glottalen Verschluß, sondern auf die Impulsantwort des Vokaltrakts zurückzuführen sind, aus dem Spektrum entfernt:

In Abhängigkeit vom Maximalwert des Analysefensters wird ein lokaler Schwellwert bestimmt. Alle Signalwerte, deren Betrag unter dieser Schwelle liegt, werden auf Null gesetzt. Eine weitere Vereinfachung erhält man durch eine dreistufige Quantisierung, d.h. alle Signalwerte, die nach dem Center Clipping noch ungleich Null sind, werden auf +1 bzw. -1 gesetzt (Peak Clipping, [DUBNOWSKI 76]). Bild 4.23 zeigt das Zeitsignal vor (4.23a) und nach (4.23b) dem Center Clipping mit dreistufiger Quantisierung.

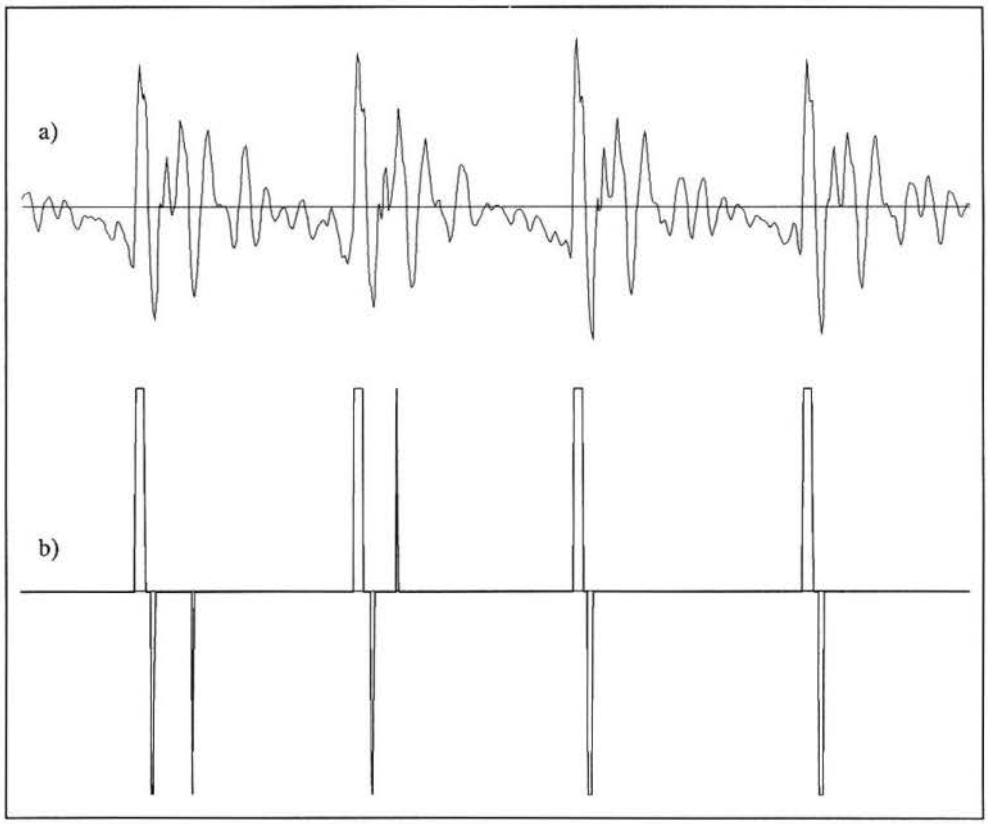


Bild 4.23: Analysefenster (37.5 Millisekunden) vor (a) und nach (b) dem Center Clipping mit dreistufiger Quantisierung.

Bild 4.24 zeigt das Spektrum desselben Signalausschnitts, wiederum vor (4.24a) und nach (4.24b) dem Center Clipping mit dreistufiger Quantisierung. Man erkennt deutlich die Dämpfung der mit "↓" markierten ersten vier Formanten.

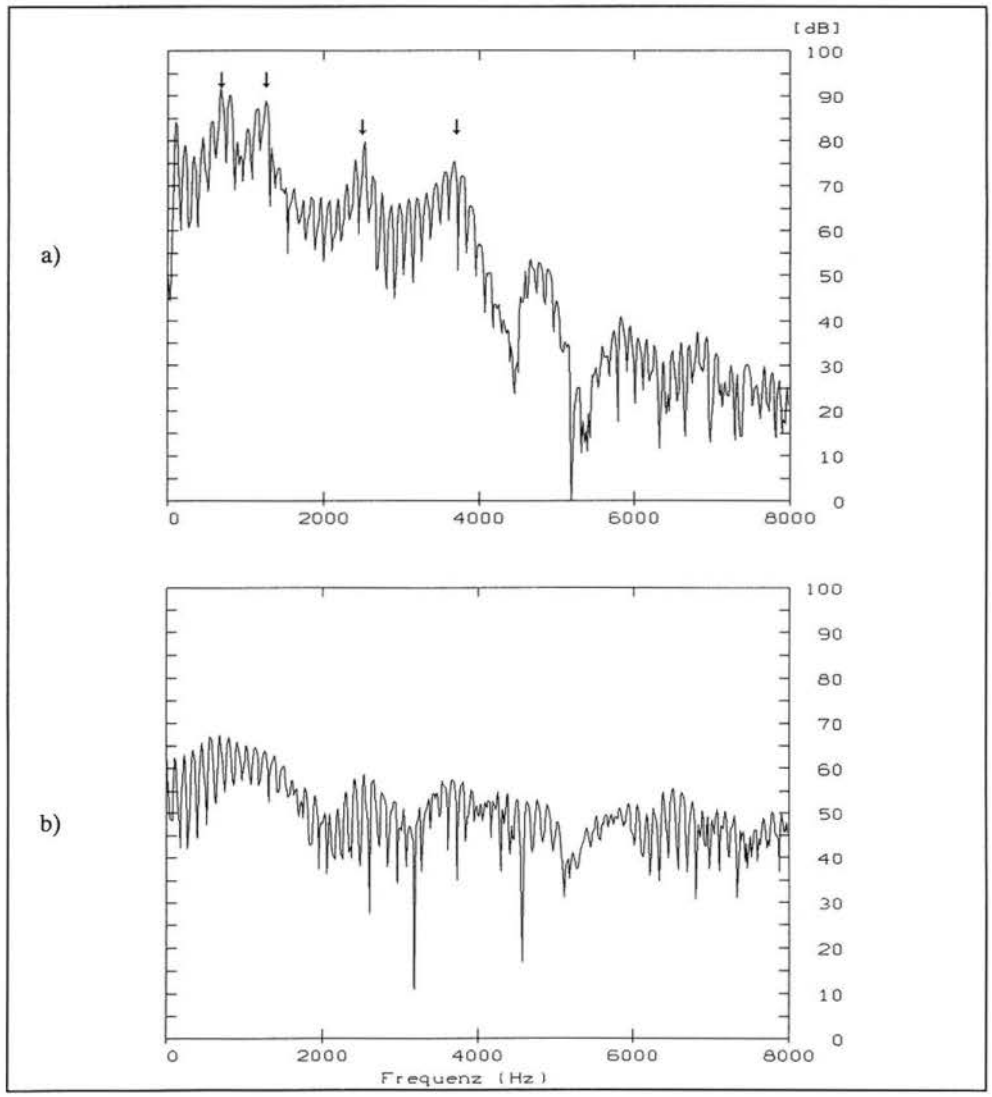


Bild 4.24: Spektrum des Analysefensters aus Bild 4.22 vor (a) und nach (b) dem Center Clipping mit dreistufiger Quantisierung. Die ersten vier Formanten sind mit einem "↓" gekennzeichnet.

4.2.3.4 Fensterfunktionen

Bei den auf dem Zeitsignal arbeitenden Korrelationsverfahren werden die Werte des Analysefensters in der Regel unverändert aus dem Signalstrom genommen (Rechteckfenster). Bei den auf dem Spektrum arbeitenden Methoden werden dagegen die Signalwerte am Rand des Analysefensters gedämpft, wobei meistens das Hamming-Fenster benutzt wird. Bild 4.25 zeigt für den Signalausschnitt des /A./ das Spektrum im Frequenzbereich zwischen 100 und 1100 Hz. In Bild 4.25a wurde das Rechteckfenster verwendet, in Bild 4.25b wurde das Hamming-Fenster verwendet. Durch das Hamming-Fenster werden die Harmonischen zwar breiter, aber die beim Rechteckfenster beobachtbaren Nebenmaxima zwischen den Harmonischen werden gedämpft.

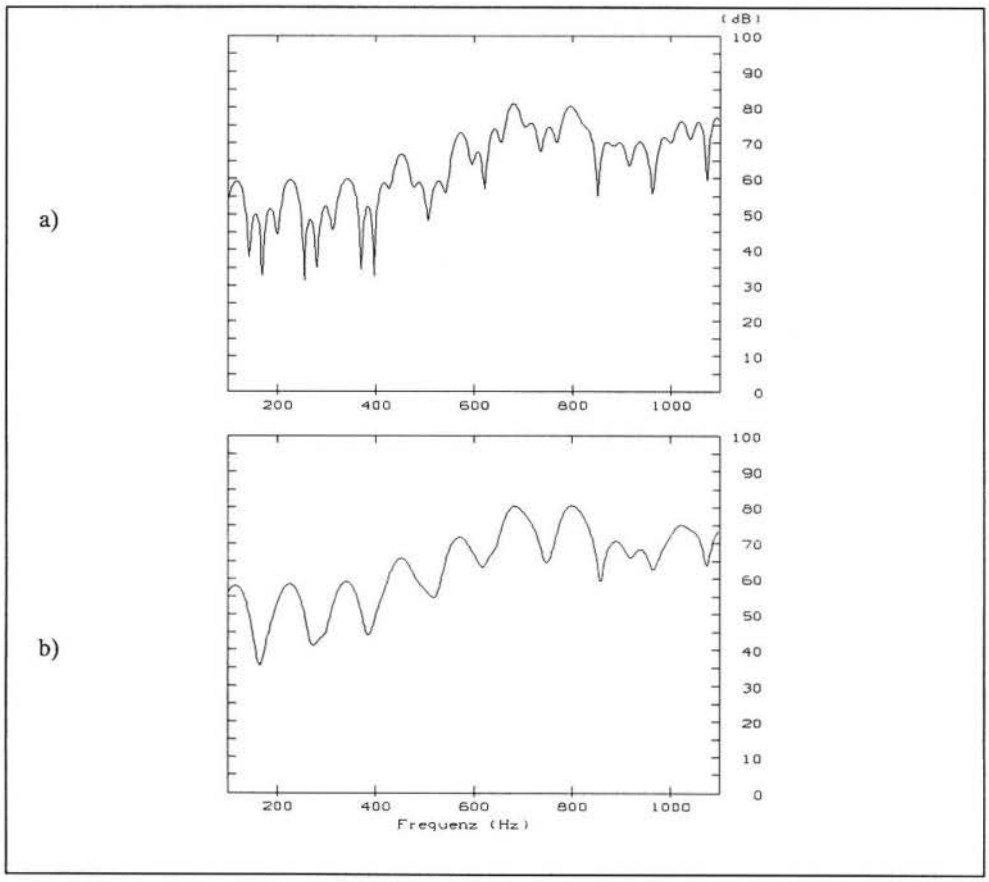


Bild 4.25: Spektrum des Analysefensters aus Bild 4.15 im Bereich von 100 bis 1100 Hz, wenn das Analysefenster mit dem Rechteckfenster (a) bzw. dem Hamming-Fenster (b) gewichtet wird.

Anschaulich läßt sich die Notwendigkeit der Dämpfung der Randbereiche damit begründen, daß i.allg. Beginn und Ende des Fensters nicht mit dem Beginn bzw. dem Ende einer Grundperiode zusammenfallen. Somit treten an den Rändern Unstetigkeiten auf, wenn man sich den Signalausschnitt unendlich periodisch fortgesetzt denkt (eine ausführliche Diskussion der Eigenschaften verschiedener Fensterfunktionen sowie eine theoretische Begründung für die Dämpfung der Randbereiche findet sich z.B. in [OPPENHEIM 75, Kap.5]).

4.2.3.5 Das AMDF-Verfahren

Berechnet man von einer periodischen Funktion f(t) mit Periode P die Autokorrelationsfunktion $\phi(k)$, so nimmt $\phi(k)$ für $k = \pm nP$, $(n = 0,1,2,\ldots)$ maximale Werte an. Somit kann die Periode einer Funktion aus ihrer Autokorrelationsfunktion berechnet werden.

Die sehr aufwendige Berechnung der Autokorrelation (hohe Zahl von Multiplikationen) kann man vermeiden, indem man die Abtastwerte nicht zeitversetzt multipliziert, sondern subtrahiert. Die sich daraus ergebende Funktion nimmt für periodische Funktionen für ganzzahlige Vielfache der Grundperiode einen minimalen Wert an, weshalb man von einer *"Antikorrelationsfunktion"* ([HESS 83, S.372]) spricht. Die bekannteste Antikorrelationsfunktion ist die AMDF-Funktion (Average Magnitude Difference Function, [ROSS 74]). Die für die Kurzzeitanalyse angepaßte Version der AMDF-Funktion ist folgendermaßen definiert (nach [HEUNISCH 86, S.71], für die F_0-Berechnung leicht modifiziert):

Seien f(1),...,f(N) die Funktionswerte des Analysefensters, GP_u die kürzeste zu erwartende Grundperiode, GP_o die längste. Dann ergibt sich die AMDF-Funktion zu

$$AMDF(d) = \sum_{n=1}^{N-GP_o} |f(n) - f(n+d)|$$

Die Funktion AMDF(d) muß nur für $GP_u \leq d \leq GP_o$ berechnet werden. Handelt es sich bei dem Analysefenster um einen Ausschnitt aus einem periodischen Signal mit Periode P, so gilt

$$AMDF(d) = 0 \Leftrightarrow d = nP$$

Für quasiperiodische stimmhafte Sprachsignale nimmt die Funktion bei Vielfachen der Grundperiode i.allg. ein Minimum an. Vor der Berechnung der AMDF-Funktion sollte das Center Clipping mit dreistufiger Quantisierung durchgeführt werden. Bild 4.26 zeigt für das Analysefenster des /A./ die AMDF-Funktion nach Anwendung des Center Clipping. Aus Darstellungsgründen wurde die AMDF-Funktion soweit verschoben, daß der (nicht berechnete) Funktionswert für d=0 zeitsynchron zum positiven Nulldurchgang vor der ersten dargestellten Leitamplitude liegt. Aufgrund der Tatsache, daß es sich hier nur um ein quasiperiodisches Signal handelt, erhält man häufig das absolute Minimum bei einem Vielfachen der Grundperiode, wie dies auch für das dargestellte Analysefenster in Bild 4.26 der Fall ist.

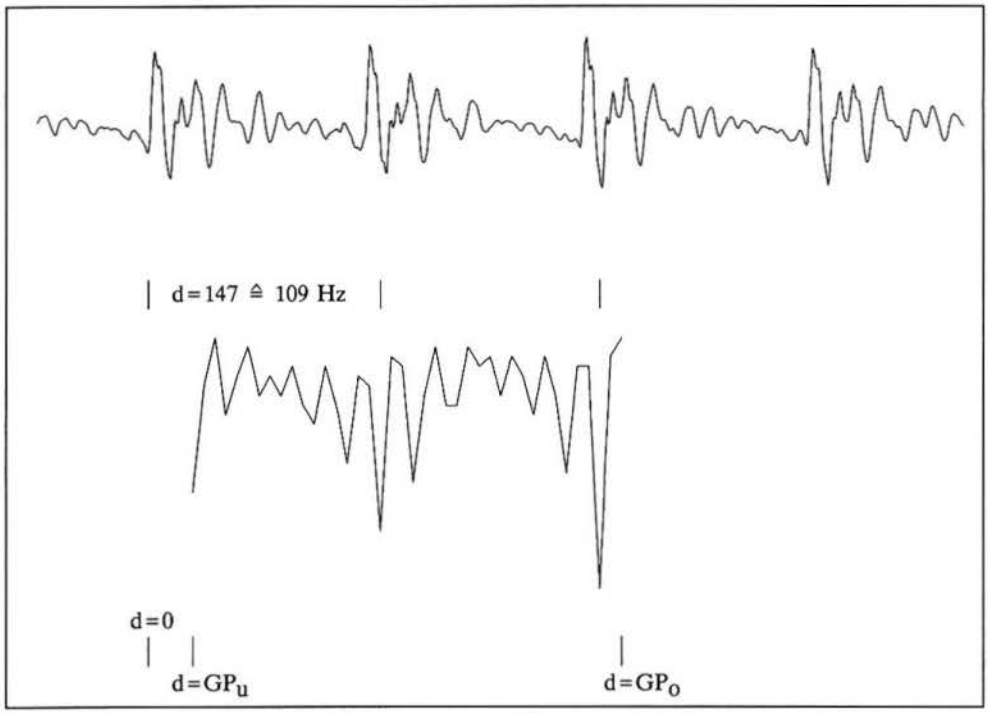


Bild 4.26: Analysefenster (600 Abtastpunkte = 37.5 Millisekunden) und AMDF-Funktion für die potentiellen Grundperioden zwischen 28 Abtastpunkten (≙ 571 Hz) und 300 Abtastpunkten (≙ 53 Hz).
Die Zahlen beziehen sich auf 16 kHz. Die AMDF-Funktion wurde mit dem Nulldurchgang vor der ersten Leitamplitude synchronisiert.

4.2.3.6 Das Seneff-Verfahren

Im folgenden wird das Seneff-Verfahren ([SENEFF 78]) als Beispiel für die Vorgehensweise bei der harmonischen Analyse vorgestellt. Die harmonische Analyse basiert auf der Tatsache, daß die meisten der Harmonischen (ganzzahlige Vielfache der Grundfrequenz) im Spektrum enthalten sind, da es sich beim Erregungssignal um ein oberwellenreiches Signal handelt und die überlagerte Formantstruktur des Vokaltrakts zwar die Amplituden der Harmonischen stark verändert, aber die feine wellige Struktur beibehält. Für die auf die harmonische Struktur zurückzuführenden Maxima gilt, daß der Abstand zwischen zwei benachbarten Maxima gleich der Grundfrequenz ist. Bild 4.27 zeigt noch einmal das Spektrum des /A./ für den Frequenzbereich von 100 bis 1100 Hz. Die lokalen Maxima sind mit einem Strich gekennzeichnet und bezüglich der Größe des Amplitudenwertes durchnumeriert. Bis auf das 7. und 9. Maximum entsprechen alle lokalen Maxima einer Harmonischen. Die 8. Harmonische würde zwischen dem 7. und 9. Maximum liegen.

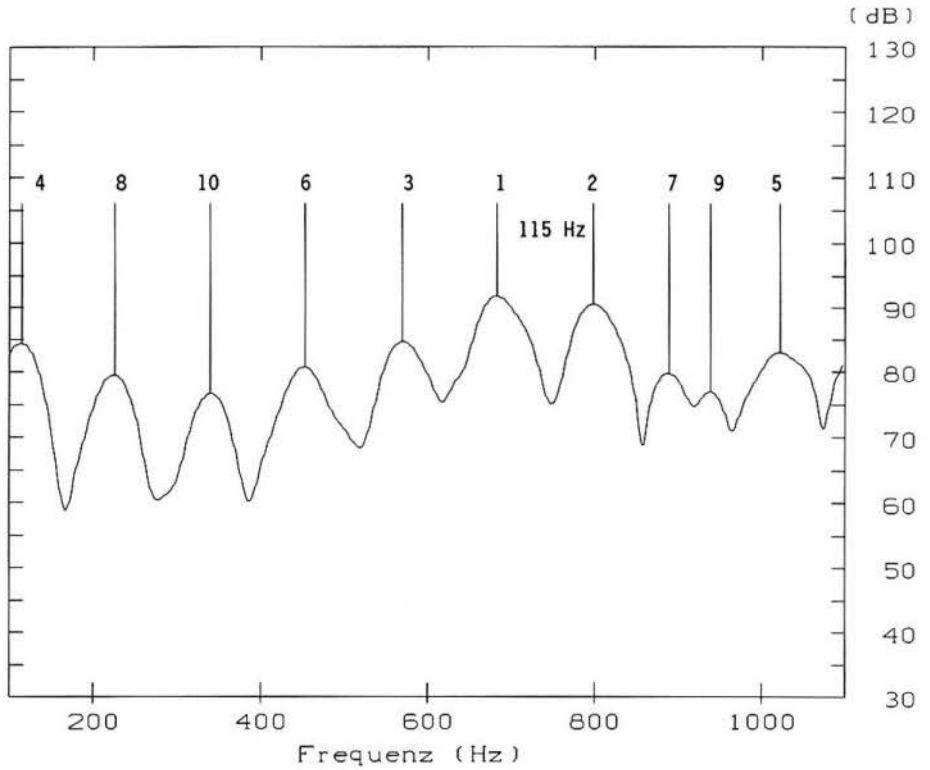


Bild 4.27: Ausschnitt aus dem Spektrum des Lautes /A./ im Bereich von 100 bis 1100 Hz. Die lokalen Maxima sind der Größe nach numeriert. Bis auf das 7. und 9. Maximum handelt es sich bei den Maximas um Harmonische.

Der Abstand zwischen zwei Harmonischen entspricht der Grundfrequenz. Beispielhaft wurde der Abstand zwischen den beiden größten Maxima angegeben. Den drei vollständigen im Analysefenster enthaltenen Grundperioden würden die Grundfrequenzen 116, 114 und 110 Hz entsprechen (siehe Bild 4.16).

Bei dem Seneff-Verfahren wird die Grundfrequenz folgendermaßen iterativ ermittelt: Im Spektrum des tiefpaßgefilterten und mit dem Hamming-Fenster multiplizierten Analysefensters werden für `k=2,3,4...` die `k` größten signifikanten Amplitudenmaxima geordnet. Als Maß für die Größe eines Maximums wird, anders als in Bild 4.27 dargestellt, das Integral der Amplitudenwerte zwischen den links und rechts von einem lokalen Maximum liegenden lokalen Minima gewählt. Ob ein lokales Maximum signifikant ist, wird über Schwellen für Kriterien wie *"Abstand zum nächsten lokalen Maximum"* gesteuert. In jedem Iterationsschritt werden die Abstände zwischen den benachbarten signifikanten Maxima in ein Histogramm eingetragen. Für das Analysefenster in Bild 4.27 wären dies die Differenzen

k=2 M_1-M_2

k=3 M_3-M_1, M_1-M_2

...

Das Verfahren bricht spätestens ab, wenn alle signifikanten Maxima abgearbeitet sind, meist jedoch früher, wenn sich ein Histogrammeintrag stabilisiert hat.

Problematisch erweisen sich bei der harmonischen Analyse irrelevante "falsche" Maxima und fehlende Harmonische. Da allerdings i.allg. nur einzelne Harmonische fehlen und durch den Einfluß des ersten Formanten die größten Harmonischen nebeneinander liegen, werden bei dem Seneff-Algorithmus normalerweise pro Iterationsschritt mehr richtige als falsche Abstände in das Histogramm eingefügt.

4.2.4 Berechnung der Grundfrequenzkontur mit der Dynamischen Programmierung

Im Rahmen der Arbeiten am Prosodie-Modul von EVAR wurde ein Grundfrequenzalgorithmus entwickelt, der sich durch eine hohe Zuverlässigkeit auszeichnet und trotzdem auch noch sehr feine Grundfrequenzveränderungen feststellen kann. Der Algorithmus wurde in [KOMPE 89b] implementiert und in bezug auf Grob- und Feinfehlerverhalten mit dem in Kap.4.2.3.5 beschriebenen AMDF- und dem in 4.2.3.6 beschriebenen Seneff-Verfahren verglichen. Die wichtigsten Ergebnisse dieser Fehlerauswertung sind in Kap.6.1 beschrieben.

Genaugenommen handelt es sich bei dem neuen Algorithmus um eine Klasse von Algorithmen bzw. um eine Vorgehensweise: Mit einem Standard-Grundfrequenzalgorithmus wird Information über die globale und lokale Stimmlage (die mittlere Grundfrequenz über einem gewissen zeitlichen Bereich) der Äußerung gewonnen. Danach werden für jedes Analysefenster mehrere mögliche Grundfrequenz-Schätzwerte bestimmt (hierfür wird die globale Stimmlageninformation benutzt). In jedem stimmhaften Bereich wird in der Folge der möglichen Grundfrequenzwerte nach einer optimalen Folge von Grundfrequenz-Schätzwerten gesucht (hierfür wird die lokale Stimmlageninformation benutzt). Für das Verfahren werden keine Annahmen darüber gemacht, wie die globale und wie die lokale Stimmlageninformation gewonnen werden. Daher wird der Algorithmus so weit wie möglich unabhängig von der in [KOMPE 89b] gewählten Realisierung beschrieben (Einzelheiten zur besten Version aus [KOMPE 89b] können dem Struktogramm in Bild 4.29b entnommen werden).

Der Algorithmus basiert auf der Annahme, daß das absolute Maximum im Kurzzeitspektrum eines stimmhaften Signalbereichs ein ganzzahliges Vielfaches der Grundfrequenz, also eine Harmonische, ist. Da das Sprachsignal aufgrund der Tiefpaßfilterung nur noch Frequenzanteile im Bereich von 100 bis 1100 Hz enthält, liegt das Maximum meist in der Nähe des ersten Formanten; so hat in Bild 4.27 die 6. Harmonische die maximale Amplitude.

Unter der Voraussetzung, daß das Maximum im Spektrum des Analysefensters eine Harmonische ist, besteht das Problem darin, den richtigen Teiler zu finden. Sei $GM_{i,j}$ die Frequenz des größten Amplitudenmaximums des i-ten Analysefensters $f_{i,j}$ im j-ten stimmhaften Bereich eines Sprachsignals und GR ein globaler Grundfrequenz-Schätzwert für dieses aktuelle Sprachsignal.

Bestimmt man nun den Teiler $n_{i,j} = \lfloor GM_{i,j}/GR + 0.5 \rfloor$, so kann man bei richtiger Wahl von m davon ausgehen, daß sich für $t_{i,j,k} = k+n_{i,j}$ $(k=-m,..,m)$ unter den sich ergebenden $2m+1$ Grundfrequenzkandidaten $K_{i,j,k} = GM_i/t_{i,j,k}$ der richtige Grundfrequenz-Schätzwert befindet.

Einen globalen Grundfrequenz-Schätzwert erhält man, indem man an einigen wenigen Stellen SB_j, von denen man annimmt, daß sich die Grundfrequenz an diesen Stellen gut berechnen läßt, mit einem sehr zuverlässigen Standard-Grundfrequenzalgorithmus bzw. einem aus mehreren Algorithmen abgeleiteten Mehrkanalverfahren die Grundfrequenz berechnet. Aus diesen Grundfrequenz-Zielwerten Z_j (lokale Stimmlageninformation) läßt sich über eine Mittelungsoperation (Median oder Mittelwert) ein globaler Stimmlagen-Schätzwert GR berechnen. In [KOMPE 89b] wird eine Stelle SB_j pro stimmhaftem Bereich für die lokale Zielwert-Bestimmung benutzt.

Die Grundfrequenzkandidaten $K_{i,j,k}$ eines stimmhaften Bereichs SH_j bilden eine Matrix, in der der Anteil dieses stimmhaften Bereichs an der Grundfrequenzkontur der Äußerung als Pfad gesucht werden muß. Die i-te Spalte der Matrix besteht aus den $2m+1$ Grundfrequenzkandidaten $K_{i,j,k}$ des i-ten Frames. Unter der Annahme, daß sich die Grundfrequenz innerhalb eines SH-Bereichs zwar sehr stark ändern kann, die Veränderung von Analysefenster zu Analysefenster aber dazu tendiert, geringe Werte anzunehmen, kann man den Abstand zwischen zwei Grundfrequenzkandidaten $K_{i,j,k}$ und $K_{i-1,j,k'}$ als Kosten ansehen und den Pfad mit minimalen Kosten durch die Matrix suchen. Zusätzlich geht der Abstand zwischen dem Grundfrequenzkandidaten $K_{i,j,k}$ und dem lokalen Grundfrequenz-Zielwert Z_j in die Kosten jedes Pfades durch diesen Grundfrequenzkandidaten mit ein. Zur Normierung der Abstände empfiehlt es sich, den Logarithmus der Grundfrequenzkandidaten bzw. Zielwerte für die Abstandsberechnung zu benutzen, da sonst Pfade auf niedrigem Frequenzniveau zu stark bevorzugt werden (siehe auch Kap.4.2.5):

$$|\ln(K_{i,j,k}/K_{i-1,j,k'})| \quad \text{bzw.} \quad |\ln(K_{i,j,k}/Z_j)|$$

Da der Abstand zum Zielwert für jedes Analysefenster berechnet wird und der Zielwert somit mehrfach in die Kosten des Gesamtpfades eingeht, wird eine gewichtete Summe der beiden Abstände als Kostenfunktion gebildet. Mit der **D**ynamischen **P**rogrammierung (DP, siehe z.B. [BELLMAN 72]) läßt sich unter Vorgabe der eben beschriebenen lokalen Kosten der optimale Pfad durch die Matrix der Grundfrequenzkandidaten effizient berechnen.

Bild 4.28 erläutert das Vorgehen bei der Suche nach dem optimalen Pfad anhand des Beginns der Matrix für einen stimmhaften Bereich SB_j. Für jeden Frame i sind die Frequenz des größten Amplitudenmaximums $GM_{i,j}$ und der Teiler $n_{i,j}$ sowie die sich daraus ergebenden Grundfrequenzkandidaten $K_{i,j,k}$ dargestellt. Der globale Richtwert GR der Äußerung, aus der dieser stimmhafte Bereich entnommen wurde, beträgt 180 Hz, der lokale Zielwert Z_j dieses stimmhaften Bereichs SB_j beträgt 170 Hz.

Sucht man für jeden Knoten den Vorgänger mit dem geringsten Abstand (Logarithmus des Quotienten der beiden Hz-Werte), so ergeben sich fünf Pfade (gestrichelt). Werden in Bild 4.28 entlang der gestrichelten Pfade diese Abstände aufsummiert, so ergeben sich an den Endpunkten der unteren vier Pfade die gleichen Kosten. Die richtige Lösung (den durchgezogenen Pfad) erhält man, wenn in jedem Knoten der Abstand zum Vorgänger und der Abstand zum Zielwert Z_j gewichtet aufaddiert werden.

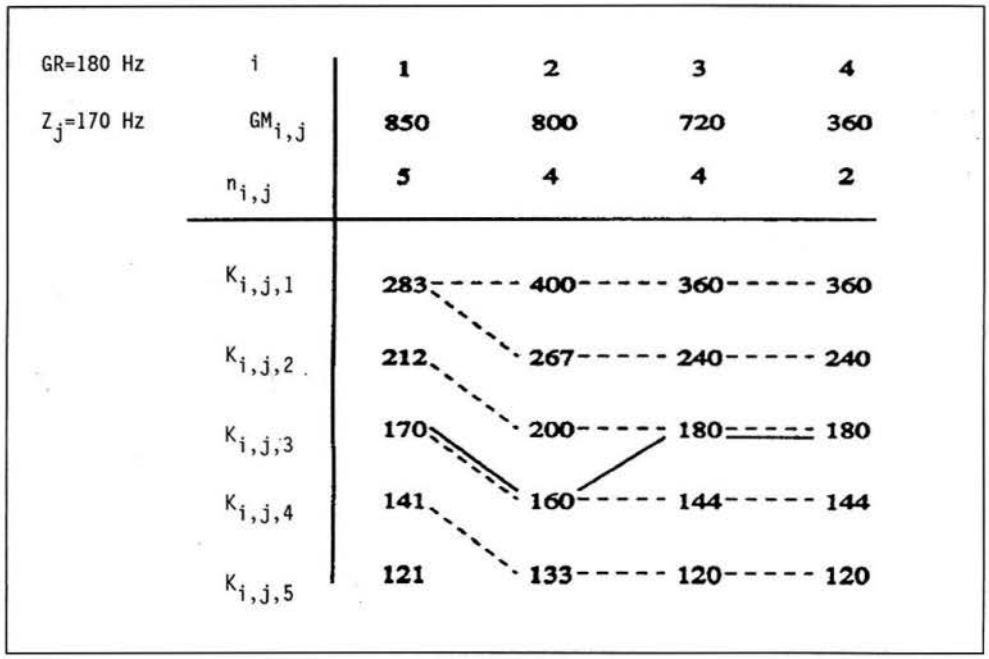


Bild 4.28: Matrix der Grundfrequenzkandidaten für den Beginn eines stimmhaften Bereichs. Der optimale Pfad ist durchgezogen eingezeichnet.

Bei der in [KOMPE 89b] beschriebenen Implementierung wird für jeden stimmhaften Bereich SH_j als "stabiler" Bereich SB_j die Stelle mit der maximalen sonoranten Energie (300-2300-Hz-Band) bestimmt. Da die Energie bereits für die Silbenkernbestimmung berechnet werden muß, fällt hier kein zusätzlicher Berechnungsaufwand an. Ein anderes brauchbares Kriterium ist z.B. die maximale Summe der abgeschätzten *a posteriori-Wahrscheinlichkeiten* der sonoranten Lautkomponentenklassen, falls die dritte SH/SL-Entscheidung aus Kap.4.2.3.1 verwendet wird.

Die Grundfrequenz-Zielwerte werden mit einem Mehrkanalverfahren berechnet, welches das AMDF- und das Seneff-Verfahren kombiniert.

Bei dem für die Implementierung benutzten Trainings- und Testmaterial, das mit einem Wertebereich von 85-500 Hz (ohne Laryngalisierungen) und bis zu 315 Hz Stimmumfang pro Äußerung einen sehr großen Bereich abdeckt, erwiesen sich fünf Grundfrequenzkandidaten pro Frame als ausreichend.

Durch die Verknüpfung von lokaler und globaler Stimmlagen-Information (Richtwert und Zielwert) und einen sehr großen Grundfrequenzbereich können einerseits starke Grundfrequenzschwankungen innerhalb eines stimmhaften Bereichs noch erfaßt werden (bei der Verwendung von fünf Teilern wird in dem Beispiel im Bild 4.28 für das zweite Frame der Bereich von 133 bis 400 Hz abgedeckt). Andererseits ist es bei kurzen stimmlosen Bereichen, bei denen das Frequenzmaximum eher zufällige Werte annimmt, durch die Kostenfunktion bei der Pfadsuche meistens

möglich, einen Grundfrequenzwert zu finden, der die Gesamtkontur nicht verfälscht. Im Gegensatz zu vielen anderen Verfahren (vergleiche [MCGONEGAL 77], wo für das Cepstrum-Verfahren ein starkes Ansteigen der Fehlerrate bei Veränderung der SH/SL-Entscheidung berichtet wird) produziert der Algorithmus auch für sehr "weiche" Stimmhaft-Entscheidungen meistens noch korrekte F_0-Konturen.

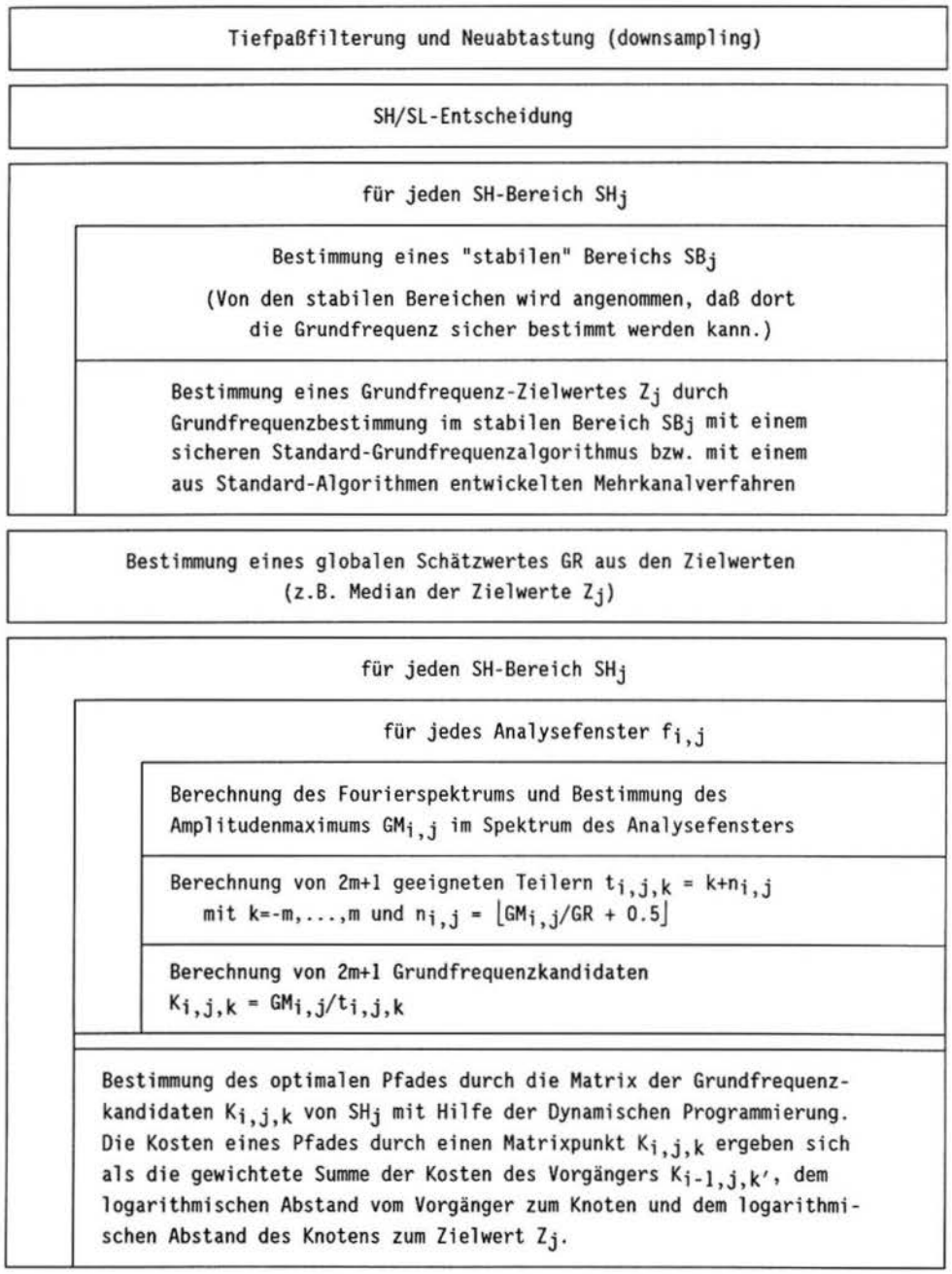
Während bei den Standardverfahren (AMDF- und Seneff-Verfahren), die zu jedem Zeitpunkt nur Information über ein Analysefenster zur Verfügung haben, eine Glättung der harmonischen Fehler z.B. mit einem Medianfilter i.allg. unbedingt notwendig ist, ist die Glättung bei dem hier beschriebenen Verfahren nicht so wichtig. Ein harmonischer Fehler liegt vor, wenn das Doppelte der Grundfrequenz (Oktav-Fehler) oder die Hälfte der Grundfrequenz (subharmonischer Fehler) als Grundfrequenzschätzwert berechnet wird. Dies tritt z.B. bei Lautübergängen aufgrund des Einflusses der Formanten auf. In einem solchen Fall wird durch die Kostenfunktion schon eine Art Glättung durchgeführt. Da auch bei dem DP-Verfahren durch eine Glättung (Medianfilter der Länge 3, gefolgt von Medianfilter der Länge 5) eine leichte Senkung der harmonischen Fehler erzielt wurde (im Gegensatz zu einer drastischen Senkung beim AMDF- und Seneff-Verfahren), wurde für die meisten der in Kapitel 6 beschriebenen Versuche die F_0-Kontur mediangefiltert.

Als problematisch erwiesen sich bei den Untersuchungen die stimmhaften Bereiche, bei denen der Zielwert in einem laryngalisierten Bereich berechnet wurde (siehe Kap.6.1.3). In solchen Fällen wurde für den ganzen SH-Bereich eine zu niedrige Grundfrequenz-Kontur berechnet. Lag dagegen ein korrekter Zielwert vor, so wurde durch die Pfadkosten des DP häufig über die Laryngalisierung "hinweggeglättet". Für die Betrachtung der F_0-Kontur ist ein solches Verhalten sehr wünschenswert, auch wenn im laryngalen Teil die Grundfrequenzschätzwerte genaugenommen falsch sind, zumindest wenn man die Grundfrequenz vom Standpunkt der Produktion her definiert.

Der Algorithmus setzt nicht voraus, daß das gesamte Sprachsignal vorliegt, da sich der globale Richtwert GR bereits nach wenigen stimmhaften Bereichen stabilisiert und sich pro Sprecher von Äußerung zu Äußerung nicht stark ändert.

Es sollte noch einmal betont werden, daß der Grundfrequenzalgorithmus weitestgehend methodenunabhängig ist, d.h. keine Anforderungen an die Module zur SH/SL-Entscheidung, zur Bestimmung der stabilen Bereiche und zur Bestimmung der Zielwerte stellt.

Bild 4.29 zeigt das Flußdiagramm des eben beschriebenen Algorithmus. In Bild 4.29a ist das prinzipielle Vorgehen gezeigt, Bild 4.29b zeigt Details zur besten Version des Algorithmus aus [KOMPE 89b].



a)

Bild 4.29: Flußdiagramm zur Grundfrequenzberechnung mit der Dynamischen Programmierung
a) prinzipelle Vorgehensweise b) (nächste Seite) Details zur Implementierung in [KOMPE 89b]

Bandpaßfilterung (100-1100 Hz Bandbegrenzung) und
Neuabtastung im Verhältnis 1:4 (10 kHz) bzw. 1:7 (16 kHz)

SH/SL nach [KIESSLING 89] (Kap.4.2.3.1): $VUV_1=0.3$ $VUV_2=0.0004 \cdot M^2$ $VUV_3=0.017 \cdot M$

für jeden SH-Bereich SH_j

Bestimmung des Frames M mit maximaler Energie im 300-2300-Hz-Band

Bestimmung eines Grundfrequenz-Zielwertes Z_j
(Median der 9 folgendermaßen bestimmten Grundfrequenz-Schätzwerte)
Grundfrequenzbestimmung mit dem AMDF-Verfahren für den Frame M
Grundfrequenzbestimmung mit dem Seneff-Verfahren für die Frames
M-2 bis M+2 (Breite des Analysefensters jeweils 3 Frames) und
M-1 bis M+1 (Breite des Analysefensters jeweils 5 Frames)

Bestimmung eines globalen Schätzwertes GR (Median der Zielwerte Zj)

für jeden SH-Bereich SH_j

für jedes Analysefenster $f_{i,j}$

Berechnung des Fourierspektrums und Bestimmung des Amplitudenmaximums $GM_{i,j}$ im Spektrum des Analysefensters (Länge 3 Frames)

Berechnung von 5 geeigneten Teilern $t_{i,j,k} = k+n_{i,j}$
mit $k=-2,\ldots,2$ und $n_{i,j} = \lfloor GM_{i,j}/GR + 0.5 \rfloor$

Berechnung von 5 Grundfrequenzkandidaten $K_{i,j,k} = GM_{i,j}/t_{i,j,k}$

Bestimmung des optimalen Pfades durch die Matrix der Grundfrequenzkandidaten $K_{i,j,k}$ von SH_j mit Hilfe der Dynamischen Programmierung. Die Kosten eines Pfades durch einen Matrixpunkt $K_{i,j,k}$ ergeben sich als die gewichtete Summe der Kosten des Vorgängers $K_{i-1,j,k'}$, dem logarithmischen Abstand vom Vorgänger zum Knoten und dem logarithmischen Abstand des Knotens zum Zielwert Z_j. Der Abstand zum Zielwert geht mit der Gewichtung $\gamma_3=0.1$ für $SH_j \leq 200$ Millisekunden und $\gamma_3=0.01$ für $SH_j>200$ Millisekunden in die Kosten ein. Die anderen beiden Summanden gehen mit der Gewichtung $\gamma_1=\gamma_2=1$ in die Kosten ein.

b)

4.2.5 Normierungsoperationen und Extraktion der Tonhöhenmerkmale

Nach der frameweisen Berechnung der Grundfrequenzschätzwerte müssen aus diesen Werten für jeden Silbenkern Merkmale abgeleitet werden, die Veränderungen der Tonhöhe wiedergeben. Vorher können verschiedene Normierungsmaßnamen durchgeführt werden, z.B. um einige der intrinsischen Eigenschaften auszugleichen. Da das Prosodie-Modul in EVAR als eigenständiges Modul konzipiert ist, sollte eine datengetriebene (bottom up) Erstellung der Betonungsbeschreibung ermöglicht werden. Daher sind lautspezifische Normierungsmaßnahmen an dieser Stelle sehr schwierig, selbst unter Verwendung der Ergebnisse des Akustik-Phonetik-Moduls. Sie können aber z.B. in einer Verifikationsphase (z.B. prosodische Verifikation und Interpretation einer Satzhypothese) berücksichtigt werden.

Hier soll nur auf eine globale Normierungsoperation eingegangen werden: die **Halbtontransformation**. Ansätze zur Entfernung mikroprosodischer Effekte an SL/SH-Grenzen sowie zur Normierung der F_0-Kontur in Bezug auf die Deklination (Kap.2.7.1) werden zur Zeit implementiert und in [STALLWITZ 89] beschrieben.

Für die weitere Verarbeitung werden die Hz-Werte nach der Formel

$$HT(i) = \frac{12}{\ln(2)} \cdot \ln(F_0)$$

in Halbtöne bezüglich der Frequenz 1 Hz umgewandelt. Man erkennt sofort, daß die Differenz zwischen einem Hz-Wert hz und dem um eine Oktave höheren doppelten Hz-Wert $hz' = 2 \cdot hz$ zwölf Halbtöne beträgt.

Für diese Normierung spricht, daß in Wahrnehmungsexperimenten für den Bereich unter 500 Hz (also dem Bereich der Grundfrequenz der menschlichen Stimme) ein Ton mit doppelter Frequenz als doppelt so hoch empfunden wurde ([ZWICKER 67, S.78ff]). Weiterhin ist der Stimmumfang (range) für Sprecher mit hoher Stimmlage höher als für Sprecher mit niedriger. Bild 4.30 zeigt für 1999 Äußerungen der vier Modus-Fokus-Korpora (Kap.3.4) den Stimmumfang der Äußerung in Abhängigkeit von der Stimmlage der Äußerung. Die restlichen 75 der insgesamt 2074 Äußerungen wurden wegen Laryngalisierung ausgesondert. Der Stimmumfang wurde mit der Differenz aus maximalem und minimalem Grundfrequenzwert abgeschätzt, die Stimmlage mit dem Mittelwert aus Onset (F_0 am Äußerungsanfang), Offset (F_0 am Äußerungsende), F_0-Maximum und F_0-Minimum. In Bild 4.30a sind Stimmlage und Stimmumfang in Hz aufgetragen, in Bild 4.30b in Halbtönen. Die durchgezogene Linie ist jeweils die Regressionsgerade. Durch die Halbtontransformation werden Stimmumfang und Stimmlage weitestgehend dekorreliert: Der Korrelationskoeffizient beträgt für die Hz-Werte 0.7 und für die Halbton-Werte -0.07.

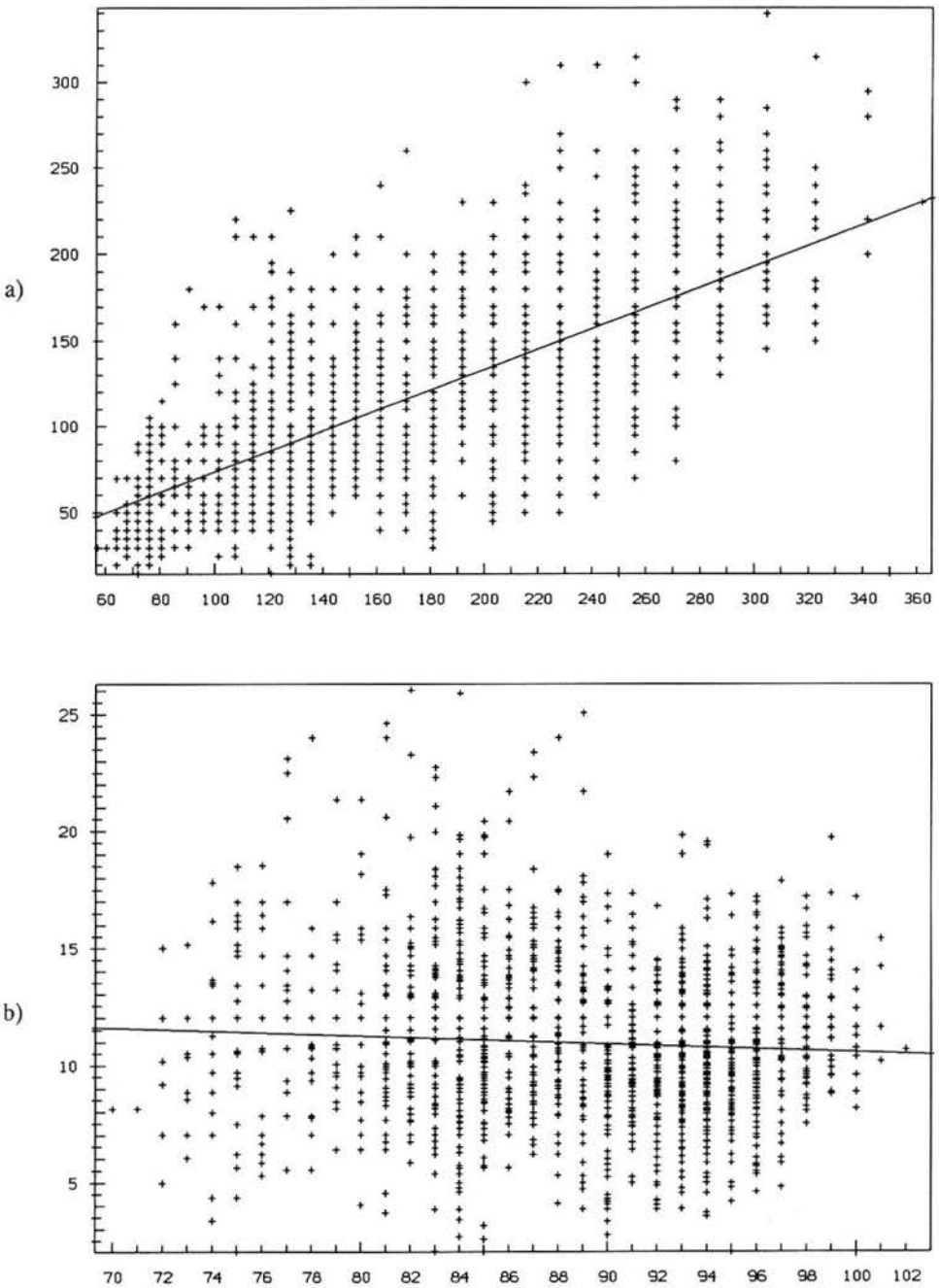


Bild 4.30 Darstellung des Stimmumfangs (Y-Achse) in Abhängigkeit von der Stimmlage (X-Achse) für 1999 Äußerungen von 12 Sprechern. Bild 4.30a zeigt die Werte in Hz, Bild 4.30b in Halbtönen.

Ob die "gehöradäquate" Halbtontransformation tatsächlich eine bessere Repräsentation des Tonhöhenkorrelats Grundfrequenz gewährleistet, ist allerdings noch Gegenstand der Forschung. Betrachtet man für dasselbe Korpus von 1999 Sätzen den Stimmumfang getrennt für Frauen und Männer, so beträgt er bei den Frauen 146 Hz bzw. 10.3 Halbtöne, bei den Männern 91 Hz bzw. 11.7 Halbtöne, was ein Hinweis auf eine "Übernormierung" sein könnte (siehe Kap.6.2 und [BATLINER 89a] zur Klassifikation von Fragen und Nicht-Fragen mit Hz- und Halbton-Werten sowie [RIETVIELD 85] zur Frage, ob die Betonungswahrnehmung eher einer Hz- oder Halbton-Skala folgt). Ob die Grundfrequenz-Werte in Hz oder Halbtönen weiterverarbeitet werden, ist daher in dem hier beschriebenen Modul über eine Option steuerbar.

Es wurde schon mehrfach erwähnt, daß für die Wahrnehmung der Betonung nicht so sehr die absoluten Werte, sondern die relativen Werte im Vergleich zur direkten Umgebung entscheidend sind. Daher werden für jeden Silbenkern Merkmale berechnet, die dann in Relation zu den Merkmalen der beiden benachbarten Silbenkerne betrachtet werden.

Zur Beurteilung der prosodischen Eigenschaft Tonhöhe werden für die Silbenkerne zwei Merkmale berechnet, mit denen die Stimmführungsfiguren *Tonbruch* und *Schleifton* modelliert werden sollen:

1) Eine Tonhöhenveränderung, die innerhalb weniger Perioden sehr abrupt verläuft, wird in [ROYÉ 83] als Tonbruch bezeichnet. Ein Tonbruch im eigentlichen Sinn, d.h. ein Sprung in der Grundfrequenzkontur von einer Periode zur nächsten, kommt fast nie vor. Die meisten Tonbrüche lassen sich so charakterisieren, daß zu Beginn/Ende eines SH-Bereiches in einem sehr kurzen Bereich von ca. 30 Millisekunden sich die Grundfrequenz auf ein deutlich höheres/tieferes Niveau bewegt und im restlichen Teil des SH-Bereich vergleichsweise konstant ist. Somit läßt sich der Tonbruch dadurch modellieren, daß die durchschnittliche Grundfrequenz für jeden Silbenkern berechnet wird und mit den Durchschnittswerten der benachbarten Silbenkerne verglichen wird.
   Sei SK_i ein Silbenkern, der von Frame $fr_{i,1}$ bis Frame $fr_{i,n}$ geht. Sei $f0_{i,j}$ der Grundfrequenzschätzwert in Hz oder Halbtönen für Frame $fr_{i,j}$. Dann ist

$$AVF0_i = \frac{1}{n} \sum_{j=1}^{n} f0_{i,j}$$

   der Mittelwert der Grundfrequenzschätzwerte für den Silbenkern SK_i, und

$$NIVF0_i = MAX((AVF0_i - AVF0_{i-1}),(AVF0_i - AVF0_{i+1}))$$

   nimmt positive Werte an, falls die durchschnittliche Grundfrequenz des Silbenkerns SK_i höher ist als die von mindestens einem der beiden direkt benachbarten Silbenkerne. Mit diesem Maß können die beiden mit Abstand am häufigsten auftretenden relevanten Tonbrüche, der steigende Vorakzent-Stimmbruch und der fallende Nachakzent-Tonbruch, erfaßt werden.
2) Weit häufiger als der Tonbruch ist der Schleifton zu beobachten. Dabei ändert sich die Grundfrequenz innerhalb des Silbenkerns. Die Veränderung kann *steigend, fallend, steigend-*

fallend und *fallend-steigend* verlaufen. Um diese Tonhöhenveränderung zu beschreiben, wird der Verlauf der Grundfrequenzschätzwerte innerhalb einer Silbe durch Liniensegmente approximiert. Dabei wird das in [LANG 87] beschriebene Verfahren eingesetzt. Bei diesem Ansatz wird zunächst die Ausgleichsgerade durch die Grundfrequenzschätzwerte eines Silbenkerns berechnet. Mit dem in [NIEMANN 74, S.63ff] beschriebenen Verfahren werden die optimalen Koeffizienten a,b und c der allgemeinen Geradengleichung

$$ax + by = c, \quad a^2 + b^2 > 0$$

berechnet. Nimmt man als X-Koordinate den Index eines Frames und als Y-Koordinate den Grundfrequenzschätzwert des Frames (Nomenklatur wie unter Punkt 1), so ergibt sich der mittlere quadratische Ordinatenabstand d_i der Grundfrequenzschätzwerte des Silbenkerns SK_i von der Ausgleichsgeraden als

$$d_i = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{a_i}{b_i} fr_{i,j} + f0_{i,j} - \frac{c_i}{b_i}\right)^2$$

Überschreitet dieser Abstand eine vorgegebene Schwelle, so wird angenommen, daß in dem Silbenkernbereich eine doppelte Stimmführungsfigur vorliegt. In diesem Fall werden zwei Geraden iterativ berechnet, indem für $1 < j < n$ zwei Linien durch die Punkte $(fr_{i,1}, f0_{i,1})$, $(fr_{i,j}, f0_{i,j})$, $(fr_{i,n}, f0_{i,n})$ gelegt werden. Ausgewählt wird das Geradenpaar mit den Koeffizienten $a_{i,1}$, $b_{i,1}$, $c_{i,1}$ und $a_{i,2}$, $b_{i,2}$, $c_{i,2}$, für das der Ordinatenabstand minimal wird. Als Maß STG_i für die Grundfrequenzveränderung innerhalb eines Silbenkerns wird der Absolutwert der größten Steigung der für die Approximation benutzten Geraden verwendet, also

$$STG_i = \begin{cases} |a_i/b_i| & \text{falls eine Gerade verwendet wird} \\ MAX(|a_{i,1}/b_{i,1}|, |a_{i,2}/b_{i,2}|) & \text{falls zwei Geraden verwendet werden} \end{cases}$$

Durch die Betragsbildung wird der Tatsache Rechnung getragen, daß die Akzentuierung sowohl durch Anheben als auch durch Absenken der Stimme markiert werden kann.

Im Gegensatz zu den anderen Merkmalen, die für die Akzentbewertung berechnet werden, wird das Merkmal STG nicht in Relation zum Verhalten des Merkmals in den beiden benachbarten Silbenkernen gesetzt.

In [LANG 87] wird aufgrund dieser Approximation noch eine Gesamtkontur erstellt. Dabei wird mit Hilfe von Schwellwerten entschieden, ob der Bereich zwischen dem Endpunkt $(fr_{i,n}, f0_{i,n})$ eines Silbenkerns und dem Anfangspunkt $(fr_{i+1,1}, f0_{i+1,1})$ des nachfolgenden Silbenkerns "schleifend" durch eine Linie verbunden wird oder "tonbruchartig" durch eine Linie von $(fr_{i,n}, f0_{i,n})$ nach $(fr_{i+1,1}, f0_{i,n})$ sowie durch eine Linie von $(fr_{i+1,1}, f0_{i,n})$ nach $(fr_{i+1,1}, f0_{i+1,1})$. Bereiche, in denen ein Schleifton angenommen wird, werden zusätzlich geglättet. Auf diesen Teil der Kontur-Stilisierung soll hier nicht weiter eingegangen werden.

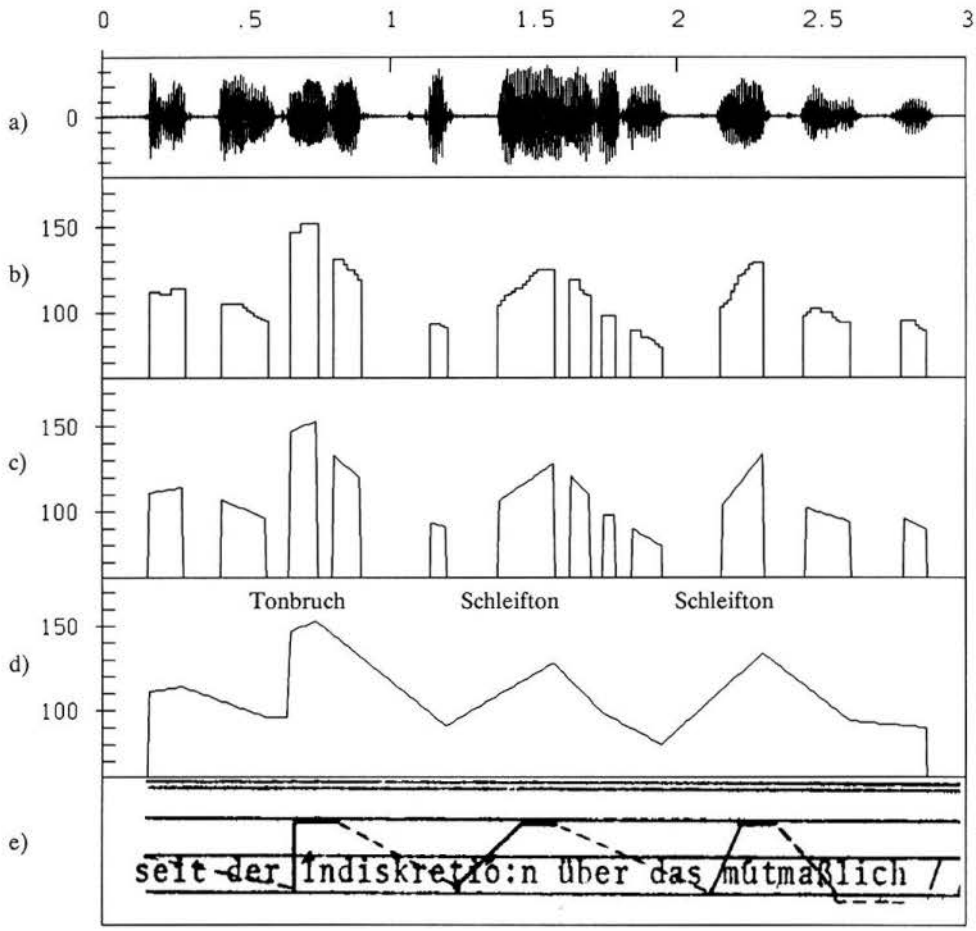


Bild 4.31: Stilisierung der Grundfrequenzkontur mit stückweise linearen Funktionen am Beispiel der Äußerung *"seit der Indiskretion über das mutmaßlich ..."*

a) Sprachsignal
b) geglättete Grundfrequenzschätzwerte in Hz (die Grundfrequenzwerte sind nur für die Silbenkernbereiche dargestellt)
c) mit maximal 2 Linien approximierter Grundfrequenzverlauf innerhalb der Silbenkerne
d) daraus abgeleitete Stilisierung der Grundfrequenzkontur mit Linienelementen
e) Transkription des Tonhöhenverlaufs aus [ROYÉ 83, S.183]

Bild 4.31 zeigt einen Ausschnitt aus dem Royé-Korpus, in dem ein Tonbruch und zwei Schleiftöne zu sehen sind. Unter dem Sprachsignal (4.31a) sind die Grundfrequenz-Schätzwerte für die Analysefenster innerhalb der Silbenkerne aufgetragen (4.31b). Die Approximation der F_0-Kontur innerhalb der Silbenkerne mit maximal zwei Linien ist in 4.31c zu sehen (aus der Steigung der Linien ergibt sich das Merkmal STG). Die Transkription des Tonhöhenverlaufs nach [ROYÉ 83] ist in 4.31e dargestellt. Bild 4.31d zeigt zusätzlich die nach [LANG 87] berechnete Stilisierung der Grundfrequenzkontur mit Linienelementen.

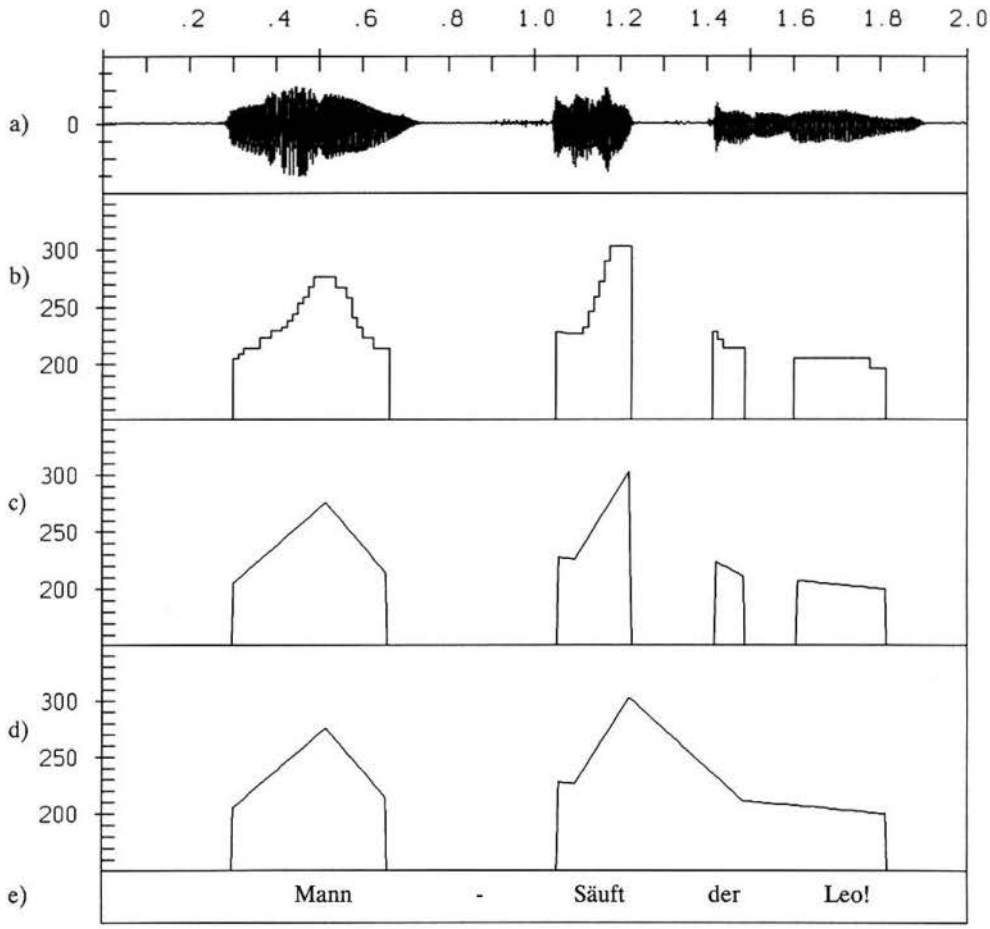


Bild 4.32: Stilisierung der Grundfrequenzkontur mit stückweise linearen Funktionen am Beispiel der Äußerung *"(Gestern war ich mit dem Leo im Wirtshaus.) Mann - säuft der Leo!"*

a) Sprachsignal
b) geglättete Grundfrequenzschätzwerte in Hz (die Grundfrequenzwerte sind nur für die Silbenkernbereiche dargestellt)
c) mit maximal 2 Linien approximierter Grundfrequenzverlauf innerhalb der Silbenkerne
d) daraus abgeleitete Stilisierung der Grundfrequenzkontur mit Linienelementen
e) orthographische Transkription der Äußerung

Bild 4.32 zeigt mit derselben Bildaufteilung eine Äußerung aus dem Leo-Korpus als Beispiel für einen steigend-fallenden Schleifton, der innerhalb eines Silbenkerns verläuft und mit zwei Linien approximiert werden muß. Dargestellt ist der Verb-Erst-Exklamativsatz *"Säuft der Leo!"* sowie das Wort *"Mann"* des zugehörigen Kontextsatzes *"Gestern war ich mit dem Leo im Wirtshaus. Mann - "*. (In Bild 4.32e ist lediglich die orthographische Transkription dargestellt.)

4.3 Dauermerkmale

Im Vergleich zur Besprechung der Merkmalberechnung für das Akzentuierungsmittel Tonhöhe sind die Unterkapitel für die Merkmale zur zeitlichen Strukturierung und zur Lautheit sehr kurz. Damit soll keine Wertung der Wichtigkeit ausgedrückt werden. Vielmehr werden durch die Silbenkerndetektion (Kap.4.1) schon mehrere sehr wichtige Schritte bei der Berechnung von Dauer- und Lautheitsmerkmalen durchgeführt.So sind die Silbenkernlänge und die sonorante Energie schon besprochen worden.

Grundlage der Merkmalberechnung für die prosodische Eigenschaft *zeitliche Strukturierung* ist die Länge des Silbenkerns. Ausgehend von den Tatsachen, daß

- betonte Silben i.allg. gedehnt werden
- die meisten der kurzen geschlossenen Vokale nie in betonter Stellung auftreten ([DUDEN 73, S.36])
- im Deutschen bei mehrsilbigen Wörtern die Silbenstruktur *Silbe mit Langvokal gefolgt von Wortakzentsilbe* so gut wie nie auftritt (Kap.2.7.3)

wird als Dauermerkmal gemessen, ob der aktuelle Silbenkern länger ist als der Durchschnitt der Silbenkernlängen der beiden benachbarten Silbenkerne.

Sei n_i die Länge des aktuellen Silbenkerns, dann ist das Dauermerkmal DAU_i definiert als

$$DAU_i = 2n_i - n_{i-1} - n_{i+1}$$

Eine getrennte Betrachtung von Lang- und Kurzvokalen sowie eine Normierung der Einflüsse der Lautumgebung wird nicht durchgeführt. Da das Prosodie-Modul als eigenständiges Modul eine datengetriebene Analyse durchführen soll, liegt die Information zur Korrektur der Kontexteinflüsse nicht vor. Auch nach dem Vergleich der Silbenkerne mit der Segmentierung durch das Akustik-Phonetik-Modul (Kap.4.1.3) ändert sich an dieser Tatsache nichts. Eine bessere Betrachtung von Kontextinformation ergibt sich in einer Verifikationsphase, da hier (unter der Annahme, daß eine zu verifizierende Hypothese zutrifft) nicht nur Information über die Lautumgebung, sondern auch syntaktische, semantische und pragmatische Information über die Hypothese vorliegt. Somit kann die Dauer eines Silbenkerns in einer Verifikationsphase z.B. unter der Annahme betrachtet werden, daß es sich um ein *langes offenes e* in einem *Verb* handelt, die Nachbarsilbe zu einem *Funktionswort* gehört und die Äußerung mit der *Sprechgeschwindigkeit* "Anzahl Silben in der Standardaussprache der Hypothese / Dauer der Hypothese" gesprochen wurde.

4.4 Lautheitsmerkmale

Für die Lautheitsmerkmale gilt ähnliches wie für das Dauermerkmal. Auch hier ist mit dem sonoranten Bandpaß von 300 bis 2300 Hz bereits ein Energiemaß pro Frame beschrieben worden, das für die Berechnung von Lautheitsmerkmalen verwendet werden kann.

Die Berechnung der Kurzzeit-Energie ist bei weitem nicht so kompliziert wie die Berechnung der Grundfrequenz und die Segmentierung des Sprachsignals. Daher ist der Parameter *sonorante Energie* auch mit wesentlich weniger Fehlern behaftet. Dem steht allerdings gegenüber, daß die Unterschiede in den intrinsischen Werten der einzelnen Vokale sich für die Energie am stärksten auswirken (siehe Kap.2.7.4).

Problematisch sind insbesondere der systematisch unbetonte Zentralvokal (/ER/ wie z.B. in *"bitte"*) und das vokalisierte R-Allophon (/AJ/ wie z.B. in *"nur"*). Der Zentralvokal hat eine hohe intrinsische Energie, ist aber im Deutschen systematisch unbetont. Das vokalisierte R-Allophon ist datengetrieben so gut wie nicht vom Silbenkern trennbar. Dies führt häufig zu einem zu hohen Wert für die Lautheitsmerkmale (und für das Dauermerkmal).

Als Merkmale zur Beschreibung der Lautheit werden zwei Parameter pro Silbenkern berechnet und mit derselben Verarbeitungsvorschrift in Relation zu den Nachbarsilbenkernen gesetzt: die maximale Energie und das Energie-Integral. Sei $E_{i,j}$ die sonorante Energie des j-ten Frames im i-ten Silbenkern (Nomenklatur wie oben). Dann sind ME_i und SE_i definiert als

$$ME_i = Max(E_{i,1}, \ldots, E_{i,n})$$

$$SE_i = \sum_{j=1}^{n} E_{i,j}$$

Mit der Verknüpfungsvorschrift

$$MAXE_i = ME_i/ME_{i-1} + ME_i/ME_{i+1} - 2$$

bzw.

$$SUME_i = SE_i/SE_{i-1} + SE_i/SE_{i+1} - 2$$

werden zwei Lautheitsmerkmale berechnet, die angeben, ob die maximale Energie bzw. das Energie-Integral der aktuellen Silbe größer ist als der Durchschnitt der beiden benachbarten Silbenkerne.

5 Erstellung einer Betonungsbeschreibung

Im letzten Kapitel wurden Algorithmen zur Merkmalberechnung vorgestellt. In diesem Kapitel wird die Erstellung einer Betonungsbeschreibung besprochen, die auf den beschriebenen fünf Merkmalen aufbaut. Die Aufgabe, für jeden Silbenkern i eine Abbildung

$$Merkmal1_i \times Merkmal2_i \times Merkmal3_i \times Merkmal4_i \times Merkmal5_i \rightarrow \{ betont , unbetont \}$$

zu erstellen, wird stufenweise gelöst. Zunächst werden die Merkmalausprägungen mit Hilfe der *Theorie der vagen Mengen (fuzzy set theory)* bewertet (zum Einsatz von vagen Mengen in der Sprachverarbeitung siehe [DE MORI 83]). Die Bewertungsfunktionen $U1_{BR}$, ... ,$U5_{BR}$ sind Trapezfunktionen und können als *charakteristische Funktionen (fuzzy membership functions)* verstanden werden: Die Wertebereiche der Merkmale

- NIVFO — die durchschnittliche Grundfrequenz im Vergleich mit dem Durchschnitt der benachbarten Silben
- STG — die Veränderung der Grundfrequenz innerhalb des Silbenkerns
- DAU — die Länge des Silbenkerns im Vergleich mit der Länge der benachbarten Silbenkerne
- MAXE — die maximale Energie im 300-2300-Hz-Band im Vergleich zum entsprechenden Maximum der benachbarten Silbenkerne
- SUME — das Energie-Integral im 300-2300-Hz-Band im Vergleich mit dem entspechenden Integral der benachbarten Silbenkerne

sind mit Hilfe je einer charakteristischen Funktion Uj_{BR} auf die Menge "**B**etonungs**R**elevanz von Merkmalwerten" abzubilden.

Nimmt z.B. die Funktion $U2_{BR}(STG_i)$ einen Wert von 0 an, so steht dies für die Aussage:

> Die Veränderung der Grundfrequenz innerhalb der Silbe i, repräsentiert durch das Merkmal STG_i, liefert keinen Beitrag zur Betonungsmarkierung mit der prosodischen Eigenschaft Tonhöhe.

Nimmt $U2_{BR}(STG_i)$ einen Wert von 1 an, so steht dies für die Aussage:

> Die Veränderung der Grundfrequenz innerhalb der Silbe i, repräsentiert durch das Merkmal STG_i, liefert einen sehr starken Beitrag zur Betonungsmarkierung mit der prosodischen Eigenschaft Tonhöhe.

Die Bewertungen der Parameter, die eine prosodische Eigenschaft repräsentieren, werden mit der *Vereinigungsfunktion für vage Mengen* zu einer weiteren charakteristischen Funktion zusammengefaßt. Diese gibt an, welchen Beitrag die prosodische Eigenschaft zur Gesamtbetonung leistet. Aus den drei Einzelbewertungen für die prosodischen Eigenschaften Tonhöhe, zeitliche Strukturierung und Intensität wird, wiederum über eine charakteristische Funktion, eine Bewertung des Grades der Akzentuierung der jeweiligen Silbe erstellt.

Bei dieser Vorgehensweise kann über eine einfache Intervallbildung eine Einteilung in n Betonungsstufen erstellt werden, also z.B.

für n=2 [0,1] → { *betont , unbetont* }

oder für n=4 [0,1] → { *stark betont , betont , neutral , unbetont* }

Im Gegensatz zu einer direkten Abbildung

Parameterausprägungen —→ Bewertungsstufen

bleibt die Information über den Beitrag der einzelnen prosodischen Eigenschaften erhalten, da neben der Gesamtbewertung auch die Bewertung der drei prosodischen Eigenschaften weitergegeben wird. Die stufenweise Ermittlung der Betonungsbeschreibung bietet außerdem den Vorteil der einfachen Erweiterung und Änderung. So ist z.B. bei der Verwendung von Hz- statt Halbton-Werten bei den Tonhöhen-Merkmalen lediglich die entsprechende Funktion zur Merkmalbewertung zu ändern, die beiden anderen Bewertungsebenen bleiben unberührt.

Bei der Verwendung von *vagen Mengen* werden die charakterischen Funktionen üblicherweise heuristisch festgelegt. Es sind auch Parametertraining und -adaption möglich ([DE MORI 83]).

Bild 5.1 zeigt das Vorgehen bei der Erstellung der Betonungsbeschreibung für jeden Silbenkern aufgrund der Merkmalausprägungen.

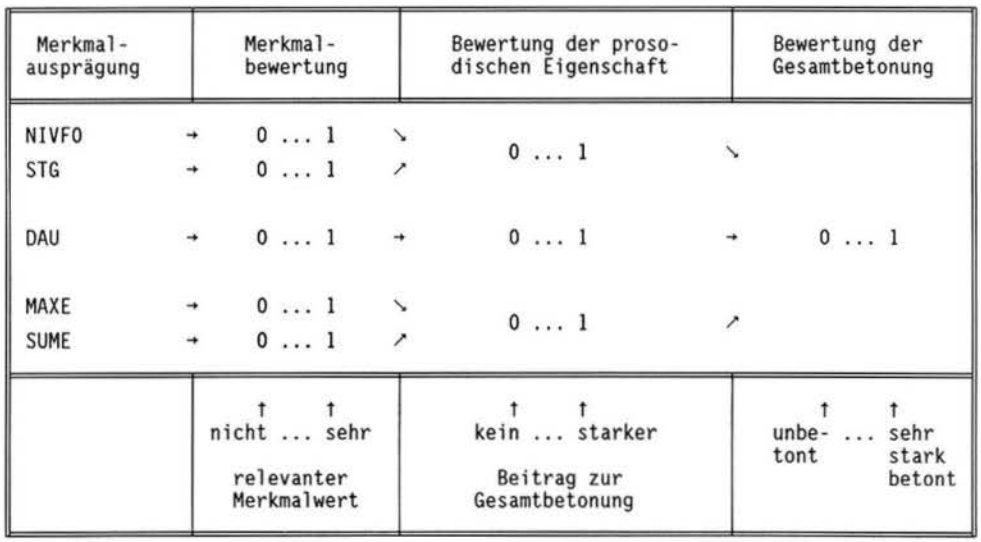


Bild 5.1: Stufenweise Berechnung der Betonungsbeschreibung mit der Theorie der vagen Mengen.

5.1 Bewertung der Merkmale

Da die Vorgehensweise bei der Abbildung der Merkmalausprägungen auf die charakteristischen Funktionen Uj_{BR} für alle Merkmale identisch ist, wird im folgenden lediglich die Abbildung des Merkmals MAXE besprochen, d.h. die maximale Energie eines Silbenkerns im sonoranten Energieband im Vergleich zu der maximalen Energie der beiden benachbarten Silbenkerne. Wie in Kap.4.4 bereits ausgeführt, nimmt $MAXE_i$ positive Werte an, wenn die maximale Energie eines Silbenkerns i größer ist als der Durchschnitt der maximalen Energie der benachbarten Silbenkerne. Daher wird die charakteristische Funktion $U4_{BR}(MAXE_i)$ für $MAXE_i \leq 0$ auf 0 gesetzt. Ist die maximale Energie eines Silbenkerns um den Faktor 1.3 größer als der Durchschnitt der maximalen Energie-Werte der benachbarten Silbenkerne ($MAXE_i \geq 0.3$), so wird angenommen, daß es sich um einen relevanten Merkmalwert handelt. Ist $MAXE_i \geq 4$, so sinkt die charakteristische Funktion wieder, da bei einem solch hohen Unterschied zum linken und zum rechten Nachbarn vermutet wird, daß es sich um einen Fehler in der Silbenkerndetektion handelt. Für $MAXE_i \geq 5$ wird angenommen, daß es sich sicher um einen Berechnungsfehler handelt. Somit ergibt sich die charakteristische Funktion zu

$$U4_{BR}(MAXE_i) = \begin{cases} 0 & \text{für } MAXE_i \leq 0 \\ MAXE_i \cdot 10/3 & \text{für } 0 < MAXE_i < 0.3 \\ 1 & \text{für } 0.3 \leq MAXE_i \leq 4 \\ -MAXE_i + 5 & \text{für } 4 < MAXE_i < 5 \\ 0 & \text{für } 5 \leq MAXE_i \end{cases}$$

Bild 5.2 zeigt die charakteristische Funktion für das Merkmal MAXE. Die Berechnung der charakteristischen Funktionen für die anderen Merkmale geschieht analog.

5.2 Bewertung der prosodischen Eigenschaften

Aus den Bewertungsfunktionen für die Merkmale sind Bewertungsfunktionen für die Akzentuierungsmittel Tonhöhe, zeitliche Strukturierung und Lautheit zu erstellen. Das Prosodie-Modul soll auch als eigenständiges Modul arbeiten können, d.h. ohne auf das Wissen des Akustik-Phonetik-Moduls über die Vokalklasse der zu bewertenden Silbenkerne zuzugreifen. Daher wird das Akzentuierungsmittel Klangfarbe bei der datengetriebenen Betonungsbeschreibung nur soweit berücksichtigt, als konsonantische Silbenkerne automatisch als unbetont eingestuft werden. Konsonantische Silbenkerne sind solche, die aufgrund der Energie im Frequenzband 100-300 Hz detektiert werden (siehe Kap.4.1.2).

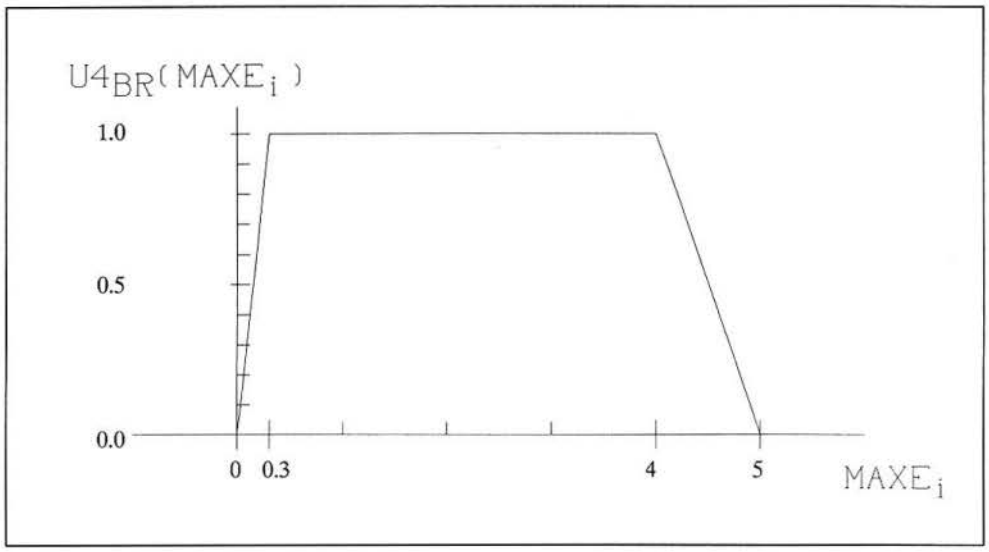


Bild 5.2: Charakteristische Funktion für die Bewertung des Energiemaximums eines Silbenkerns im Vergleich mit den Energiemaxima der beiden benachbarten Silbenkerne.

Die charakteristischen Funktionen $U_{MA}(i)$ (≙ der prosodischen Eigenschaft *Tonhöhe*), $U_{TA}(i)$ (≙ der prosodischen Eigenschaft *zeitliche Strukturierung*) und $U_{DA}(i)$ (≙ der prosodischen Eigenschaft *Lautheit*) für die Mengen der **M**elodisch, **T**emporal und **D**ynamisch **A**kzentuierten Silbenkerne werden ebenfalls mit der Algebra der vagen Mengen gebildet. Sie ergeben sich aus der Vereinigung der charakteristischen Funktionen für die entsprechenden Merkmale (im Fall der zeitlichen Strukturierung, bei der nur ein Merkmal berechnet wird, handelt es sich um die identische Abbildung):

$$
\begin{aligned}
U_{MA}(i) &= MAX(\, U1_{BR}(NIVFO_i) \,,\, U2_{BR}(STG_i) \,) \\
U_{TA}(i) &= U3_{BR}(DAU_i) \\
U_{DA}(i) &= MAX(\, U4_{BR}(MAXE_i) \,,\, U5_{BR}(SUME_i) \,)
\end{aligned}
$$

5.3 Bewertung der Gesamtbetonung

Aus den Bewertungen der einzelnen Akzentuierungsmittel ist eine Bewertung der Gesamtbetonung der Silbe zu erstellen. Da i.allg. nicht alle Akzentuierungsmittel eingesetzt werden, wird die charakteristische Funktion $U_{BET}(i)$ der Menge der **BET**onten Silben folgendermaßen berechnet:

- Wurde bei der Silbenkerndetektion ein konsonantischer Silbenkern hypothetisiert, so wird die Betonungsbewertung 0 zugewiesen.
- Falls mindestens zwei der drei Einzelbewertungen eines vokalischen Silbenkerns über einer Signifikanzschwelle S (S wurde heuristisch auf 0.85 festgelegt) liegen, wird die Gesamtbewertung auf 1 gesetzt. Ansonsten wird das arithmetische Mittel der drei Einzelbewertungen gebildet.

Somit ergibt sich $U_{BET}(i)$ zu

$$U_{BET}(i) = \begin{cases} 0 & \text{falls ein konsonantischer Silbenkern hypothetisiert wurde} \\ 1 & \text{falls } \mathrm{MEDIAN}(U_{MA}(i), U_{TA}(i), U_{DA}(i)) \geq S \\ (U_{MA}(i)+U_{TA}(i)+U_{DA}(i)) \,/\, 3 & \text{sonst} \end{cases}$$

Bild 5.3 verdeutlicht noch einmal das Vorgehen bei der Silbenkerndetektion (ohne Abgleich mit den Ergebnissen des Akustik-Phonetik-Moduls) und bei der Berechnung der Betonungsbeschreibung.

Sprachsignal

↓

Berechnung der logarithmischen spektralen Energie in drei Frequenzbändern:
300-2300 Hz 2300-5000 Hz 100-300 Hz

↓

Glättung mit einem Medianfilter der Länge 5

↓

Detektion lokaler Maxima im 300-2300-Hz-Band und Markierung als Silbenkerne

↓

Elimination von Kernen mit höherer Energie im 2300-5000-Hz-Band

↓

Ermittlung lokaler Maxima im 100-300-Hz-Band und Markierung als reduzierte Silbenkerne
(falls sie keinen vokalischen Silbenkern berühren)

↓

für jeden Silbenkern:
Berechnung einer Betonungszahl zwischen
0 ≙ unbetont und
1 ≙ stark betont
aufgrund der folgenden Merkmale des Silbenkerns:

- Durchschnittliche Grundfrequenz im Vergleich mit dem Durchschnitt der benachbarten Silbenkerne
- Veränderung der Grundfrequenz innerhalb des Silbenkerns

--> prosodische Eigenschaft **Tonhöhe**

- Länge des Silbenkerns im Vergleich mit der Länge der benachbarten Silbenkerne

--> prosodische Eigenschaft **zeitliche Strukturierung**

- Maximale Energie im 300-2300-Hz-Band im Vergleich zum Maximum der benachbarten Silbenkerne
- Energie-Integral im 300-2300-Hz-Band im Vergleich mit dem Integral der benachbarten Silbenkerne

--> prosodische Eigenschaft **Lautheit**

(konsonantischen Silbenkernen wird die Betonungszahl 0 zugewiesen)

Bild 5.3: Detektion der Silbenkerne und Berechnung einer Betonungszahl für jeden Silbenkern.

6 Experimentelle Untersuchungen zum Einsatz der prosodischen Information in einem sprachverstehenden System

In diesem Kapitel werden einige Untersuchungen vorgestellt, die im Rahmen der Arbeit am Prosodie-Modul des sprachverstehenden Systems EVAR durchgeführt wurden. In Kap.6.1 werden Ergebnisse zu den in Kap.4.2.3 und Kap.4.2.4 beschriebenen F_0-Algorithmen beschrieben. Die Klassifikation von Fragen und Nicht-Fragen anhand intonatorischer Mittel, genauer gesagt anhand von aus dem Grundfrequenzverlauf extrahierten Merkmalen, ist in Kap.6.2 beschrieben. Kap.6.3 befaßt sich mit den Ergebnissen zur Silbenkerndetektion aufgrund der spektralen Energie (Kap.4.1.2) sowie den Ergebnissen zur automatischen Betonungsbeschreibung. Untersuchungen zur prosodischen Wortakzent-Verifikation sowie zur prosodischen Lexikoneinschränkung an Satzakzent-Stellen werden in Kap.6.4 und Kap.6.5 diskutiert. Kap.6.6 geht auf die Auswirkung des in Kap.4.1.3 beschriebenen Segmentierers auf die Worterkennung ein.

6.1 Fehleranalyse für die Grundfrequenzbestimmung

Im Rahmen von [HEUNISCH 86] wurden das AMDF- (Kap.4.2.3.5) und das Seneff-Verfahren (Kap.4.2.3.6) implementiert. Aufbauend auf diesen Verfahren wurde in [KOMPE 89b] ein eigenes Mehrkanalverfahren entwickelt (Kap.4.2.4). Eine Fehleranalyse für die drei Algorithmen wurde in [KOMPE 89b] durchgeführt. Hier soll nur auf die wichtigsten Ergebnisse eingegangen werden.

Bei der Betrachtung von Grundfrequenzalgorithmen sind drei Fehlerklassen zu unterscheiden:

1) Von SH/SL-Fehlern spricht man, wenn stimmhaft erzeugte Teile einer Äußerung als stimmlos klassifiziert werden bzw. wenn stimmlose als stimmhaft klassifiziert werden. Beispielhaft wird das Verhalten der dritten (die *Sonoranten-Entscheidung* ergibt sich aus der Summe der abgeschätzten *a posteriori-Wahrscheinlichkeiten* der sonoranten Lautkomponenten) und vierten (das in [KIESSLING 89] implementierte Schwellwertverfahren arbeitet auf Strukturmerkmalen des Zeitsignals) SH/SL-Entscheidung aus Kap.4.2.3.1 für die EVAR-Stichprobe (Kap.3.1) vorgestellt.
2) Von Feinfehlern spricht man, wenn der Grundfrequenzschätzwert zwar von dem tatsächlichen Grundfrequenzwert abweicht, aber die Grundfrequenzkontur nicht verfälscht wird. Feinfehler sind in erster Linie auf Auflösungsungenauigkeiten zurückzuführen. So beträgt z.B. der Abstand zwischen zwei Fourier-Koeffizienten bei den für die Untersuchungen verwendeten Parametereinstellungen (FFT-Länge, etc.) 8.9 Hz. Mit dem in [KIESSLING 89] beschriebenen interaktiven Verfahren wurden bei 24 Äußerungen des Fokus-Korpus die Grundperioden markiert. Die Werte wurden von zwei Phonetikern korrigiert. Anhand dieser Referenzkonturen wird das Feinfehlerverhalten der drei Grundfrequenzalgorithmen diskutiert.
3) Als Grobfehler werden die Fehler bezeichnet, bei denen durch den Fehler die Grundfrequenzkontur verfälscht wird. Die Grobfehler der drei Algorithmen werden für die Dialog-

(Kap.3.2), die Fokus- (Kap.3.4.1) und die Leo-Stichprobe (Kap.3.4.2) diskutiert. Als Referenzkonturen wurden die per Hand korrigierten Mingogramme der Äußerungen verwendet.

6.1.1 SH/SL-Fehler

Zur EVAR-Stichprobe existiert für jeden Frame eine Handklassifikation nach Lautkomponenten, die in [REGEL 88] ausführlich beschrieben ist. Zusätzlich zu der Lautkomponentenklasse ist für jeden Frame vermerkt, ob es sich bei dem Frame um einen Signalausschnitt aus einem Übergangsbereich zwischen zwei Lauten handelt oder nicht (Rand- und Mitte-Frames). Die Lautkomponenten wurden in drei Klassen eingeteilt:

1) Sonorante Lautkomponentenklassen (Vokale, Nasale und /L./) wurden als stimmhafte Lautkomponenten klassifiziert (SH).
2) Stimmlose Plosive und stimmlose Frikative sowie die Sprechpause wurden als stimmlose Lautkomponenten klassifiziert (SL).
3) Stimmhafte Plosive und stimmhafte Frikative wurden als gemischt angeregt klassifiziert (MIX).

Tabelle 6.1 zeigt die Größe der Stichprobe in Anzahl Frames und in Sekunden, sowie den jeweiligen Anteil der drei Klassen SH, SL und MIX.

	alle Frames				Mitte-Frames			
	SH	MIX	SL	gesamt	SH	MIX	SL	gesamt
Anzahl Frames	24589	6326	14133	45048	15136	3934	10230	29300
Zeit in Sekunden	315	81	181	577	194	50	131	375

Tab. 6.1: Größe der für die Beurteilung der SH/SL-Entscheidungen verwendeten EVAR-Stichprobe in Anzahl Frames und in Sekunden, sowie die Größe der drei Lautoberklassen stimmhaft (SH), stimmlos (SL) und gemischt (MIX) angeregt.

Tabelle 6.2 zeigt für die beiden SH/SL-Entscheidungen, die Sonoranten-Entscheidung und die Schwellwert-Entscheidung nach [KIESSLING 89], den jeweiligen Anteil der drei Lautoberklassen, der als stimmhaft klassifiziert wurde. Dabei wurden folgende Schwellen verwendet (Kap.4.2.3.1):

- SON = 0.1, d.h. die Summe der a posteriori Wahrscheinlichkeiten der sonoranten Lautkomponenten-Klassen muß größer als 0.1 sein
- VUV_1 = 0.3, d.h. ein Frame wird dann als stimmlos klassifiziert, wenn bei mehr als 30 Prozent der Abtastpunkte ein Nulldurchgang vorliegt.
  VUV_2 = $0.0004 \cdot M^2$, d.h. ein Frame wird dann als stimmlos klassifiziert, wenn das arithmetische Mittel der Amplitudenquadrate weniger als 0.04 Prozent des (theoretisch) maximal möglichen Wertes beträgt.

VUV_3 = 0.017•M, d.h. wenn die größte Amplitude im Frame kleiner als 1.7 Prozent der größtmöglichen ist.

Beide SH/SL-Entscheidungen wurden mit einem Medianfilter der Länge 5 geglättet.

Man erkennt in Tabelle 6.2, daß die SH/SL-Entscheidung nach [KIESSLING 89] bei den verwendeten Schwellwerten wesentlich weicher ist als das Sonoranten-Kriterium. Die Tatsache, daß Sonoranten so gut wie sicher als stimmhaft erkannt werden, erkauft man mit einer höheren Fehlerrate SL nach SH.

Es wird sich zeigen, daß sich diese Tatsache bei dem Mehrkanalverfahren eher positiv auf die Anzahl der Grobfehler auswirkt: Da vor allem sehr kurze stimmlose Bereiche als stimmhaft klassifiziert werden, wird aufgrund der Pfadbeschränkung die F_0-Kontur nicht verfälscht, während die Zahl der Zielwertfehler gesenkt werden kann.

SH/SL Entscheidung		alle Frames			Mitte-Frames		
		SH	MIX	SL	SH	MIX	SL
Sonoranten	SH	94.6	15.3	8.3	98.2	9.0	2.4
Kießling	SH	99.2	49.2	16.6	99.7	44.9	11.1

Tab. 6.2: Anteil der Frames (in Prozent) aus der EVAR-Stichprobe, die gemäß Handklassifikation einer der drei Lautoberklassen stimmhaft (SH), stimmlos (SL) und gemischt (MIX) angehören und vom Sonoranten-Kriterium einerseits und von der SH/SL-Entscheidung nach [KIESSLING 89] andererseits als stimmhaft klassifiziert werden. Alle Frames, die nicht als stimmhaft klassifiziert werden, gehören der Klasse stimmlos an.

6.1.2 Feinfehler

Für die Betrachtung der Feinfehler standen 24 Äußerungen aus der Fokus-Stichprobe als Referenzkonturen zur Verfügung (je zwölf von einem männlichen und einem weiblichen Sprecher). Mit dem in [KIESSLING 89] beschriebenen interaktiven Verfahren wurden automatisch die Grundperioden am positiven Nulldurchgang vor der Leitamplitude markiert. Das Ergebnis wurde von zwei Phonetikern korrigiert. Die Stichprobe umfaßt insgesamt 2474 stimmhafte Frames mit einer Länge von je 12.5 Millisekunden und 5116 markierten Grundperioden, was einer durchschnittlichen Grundfrequenz von 165 Hz entspricht. Da die Grundfrequenz-Schätzwerte über einem Bereich von drei Frames ermittelt werden (37.5 Millisekunden für 16 kHz Signale, 38.4 Millisekunden für 10 kHz Signale), wird der Schätzwert über durchschnittlich neun Grundperioden gebildet. Als Grundfrequenz-Referenzwert R für ein Analyse-Fenster wurde das arithmetische Mittel der Grundfrequenzwerte genommen, die sich aus den vollständigen im Analysefenster enthaltenen Grundperioden ergeben. Als Feinfehler wurden alle Abweichungen angenommen, bei denen der Grundfrequenz-Schätzwert A und der Referenzwert R (in Hz) um weniger als 10 Prozent voneinander abweichen, bei denen also gilt: 0.1•R > |R-A| .

Frames mit höheren Abweichungen wurden als Grobfehler klassifiziert. Tabelle 6.3 zeigt für jeden der drei Grundfrequenz-Algorithmen das Feinfehler-Verhalten. Das eigene Mehrkanalverfahren ist mit DPGF (**D**ynamisch **P**rogrammierte **G**rund**F**requenz-Kontur) bezeichnet. Die Feinfehlerauswertung wurde für die ungefilterte Kontur (MED 0) sowie für die mit einem Medianfilter der Länge 3, gefolgt von einem Medianfilter der Länge 5 gefilterte Kontur (MED 3&5) durchgeführt. Als Maß für das Feinfehlerverhalten wurde der Absolutbetrag der Differenz von Referenzwert und Schätzwert in Hz und in Halbtönen verwendet. Die 24 Äußerungen wurden bezüglich der maximalen Abweichung (Max), des Mittelwerts (μ), der Streuung (σ) und der 95%-Schwelle untersucht. Sortiert man die Abstände der Größe nach, so gibt die 95%-Schwelle den größten der 95 Prozent kleinsten Abstände an. Die Spalte "Problem" gibt den prozentualen Anteil der Frames an, bei denen Schätzwert und Referenzwert um mehr als 10 Prozent voneinander abweichen, bei denen also nach dieser vorläufigen Definition ein Grobfehler vorliegt. Diese Fälle gingen in die Spalten Max, μ, σ und 95% nicht ein.

Die Zeile "Periodenschwankg." gibt die "Feinfehler" der Referenzwerte an: Hier wurden für jedes Analysefenster die Abweichungen der Grundfrequenzwerte, die sich aus den einzelnen Grundperioden ergeben, vom jeweiligen Referenzwert R ausgewertet.

Das Feinfehlerverhalten aller drei Algorithmen kann als sehr gut bezeichnet werden; die Abweichungen liegen fast immer unterhalb der Abweichungen der Grundperioden eines Analysefensters, wobei das DPGF-Verfahren die besten Ergebnisse brachte.

		Problem	|R-A|				|HT(R)-HT(A)|			
		%	Max	μ	σ	95%	Max	μ	σ	95%
Med 0	DPGF	1.7	22	3.1	2.7	8.5	1.6	0.31	0.27	0.88
	AMDF	6.7	27	3.7	3.7	11.1	1.7	0.36	0.31	1.01
	Seneff	9.5	25	4.7	3.9	11.7	1.8	0.46	0.37	1.24
Med 3&5	DPGF	1.1	22	3.3	3.0	9.3	1.7	0.33	0.30	0.98
	AMDF	2.9	27	3.7	3.7	11.1	1.8	0.36	0.31	1.01
	Seneff	4.6	32	4.5	4.1	12.2	1.8	0.43	0.36	1.23
Periodenschwankg.		./.	81	5.0	6.2	16.2	4.6	0.43	0.47	1.34

Tab. 6.3: Über alle stimmhaften Analysefenster gemittelte Feinfehler der Grundfrequenzwerte von drei Verfahren bzgl. verschiedener Abstandsmaße für ungeglättete Konturen (*Med 0*). Die mit *Med 3&5* bezeichneten Grundfrequenzkonturen wurden zunächst mit einem Medianfilter der Breite 3 Frames und anschließend mit einem der Breite 5 Frames gefiltert. In der Spalte *Problem* ist der Anteil stimmhafter Frames angegeben, die um mehr als 10 Prozent vom Referenzwert abweichen. Alle anderen Frames gehen in die Werte der übrigen Spalten ein. Die Zeile *Periodenschwankg.* gibt die Abweichungen jeder einzelnen Grundperiode vom Mittelwert des betreffenden Analysefensters an. Mit *R* ist der Referenzwert und mit *A* der automatisch berechnete Grundfrequenzwert, jeweils in Hz bezeichnet. *HT(R), HT(A)* sind die entsprechenden Frequenzen in Halbtönen.

Durch die Medianglättung wird die Zahl der problematischen Fälle um 35 Prozent (DPGF), 57 Prozent (AMDF) und 52 Prozent (Seneff) verringert, ohne daß sich die Feinauflösung der Verfahren nennenswert ändert. In bezug auf die Anzahl der Problemfälle zeigt das DPGF-Verfahren ein deutlich besseres Verhalten als die anderen beiden Verfahren.

In [KOMPE 89b] wurden für die ungeglätteten DPGF-Werte die 1.7 Prozent Problemfälle (42 Frames) genauer untersucht. Aufgrund dieser Analyse kommt Kompe zu dem Ergebnis, daß

- nur 12 der 42 Fälle zu einer echten Verfälschung der F_0-Kontur führen
- zumindest für die (sehr kleine) Stichprobe von 24 Äußerungen sich eine absolute Schwelle von 30 Hz Abweichung zwischen Schätzwert und Referenzwert als geeigneter herausstellte als der relative Wert von 10 Prozent .

6.1.3 Grobfehler

Die Grobfehlerrate ist für den Einsatz der Grundfrequenz-Algorithmen in einem ASE-System sicherlich das wichtigste Fehlerkriterium. In [KOMPE 89b, Kap. 5.3.2] wurde das Grobfehlerverhalten der drei Verfahren für die Leo-, die Fokus- und die Dialog-Stichprobe untersucht. Der stimmhafte Anteil an diesen drei Stichproben beträgt insgesamt ca. 14 Minuten. Die Leo-Stichprobe wurde in [KOMPE 89b] als Lernstichprobe zum Einstellen von Schwellwerten bei der Entwicklung des DPGF-Algorithmus verwendet. Zusätzlich zu den 360 Äußerungen der Leo-Stichprobe (180 Leo- und 180 Kontext-Sätze) der 6 Testsprecher standen 30 Realisierungen des männlichen Sprechers BA zur Verfügung (je eine Realisierung aller Kontext- und Leo-Sätze, siehe die Zeile BA in Tabelle 6.4 und 6.5). Bei diesem Sprecher traten ungewöhnlich viele Oktav-Sprünge auf.

Als Referenz-Konturen lagen von allen Äußerungen Mingogramm-Aufzeichnungen vor. Diese waren im Rahmen der Arbeiten am DFG-Projekt Modus-Fokus-Intonation [ALTMANN 89a] erstellt worden. Zweifelsfälle wurden nach Anhören der Analog-Aufnahmen verifiziert bzw. korrigiert. Da keine frameweisen Referenzwerte vorlagen, mußte die Auswertung durch den Vergleich der berechneten F_0-Konturen mit dem Mingogramm-Verlauf erfolgen. Als Fehlerkriterium wurde die Eigenschaft *grob konturverfälschend* für die Klassifikation von Grobfehlern benutzt. Dieses Kriterium ist sicher sehr subjektiv. Da allerdings die Auswertung von einer Person vorgenommen wurde, ist zumindest die Vergleichbarkeit über die Stichproben und Algorithmen hinweg gewährleistet. Weiterhin wurden zweifelhafte Fälle mit einem Phonetiker des DFG-Projekts besprochen.

Laryngalisierungen, über die eines der Verfahren hinwegglättet (es liefert also während der Laryngalisierung genaugenommen einen falschen, zu hohen Wert), wurden nicht als Grobfehler gezählt. In Bild 4.20 würde eine solche Glättung über der dargestellten Laryngalisierung F_0-Schätzwerten von ca. 160 Hz, also einer Art Interpolation zwischen den normal artikulierten benachbarten Bereichen entsprechen. Diese Vorgehensweise läßt sich damit begründen, daß der Mensch zwar eine Laryngalisierung als Grenzsignal hört, aber der perzipierte Tonhöhenverlauf nicht von der Laryngalisierung beeinflußt wird.

Es stellt sich hier die Frage, wie die Grobfehler zu zählen sind. Da bei der SH/SL-Entscheidung i.allg. nicht zwischen *stimmlos* und *Sprechpause* unterschieden wird, erscheint der prozentuale Anteil der fehlerhaften Analyse-Frames an allen Frames nicht geeignet. Die so gut wie sicher als stimmlos klassifizierbaren Sprechpausen können - je nach Anteil an der Gesamtstichprobe - das Ergebnis verfälschen. Als besser vergleichbares Maß ist der prozentuale Anteil der grob fehlerhaften Schätzwerte an der Anzahl aller *stimmhaften Frames* anzusehen. Für eine automatische Analyse ist es sicherlich sehr wichtig, in wieviel Prozent der Äußerungen mindestens ein Grobfehler zu erwarten ist.

Die Auswertung der drei Stichproben nach diesen beiden Fehlermaßen ist in den Tabellen 6.4 (Prozent fehlerhafter Äußerungen) und 6.5 (Prozent fehlerhafter stimmhafter Frames) zu sehen.

Stich-probe		Anzahl Sätze	SH/SL-Kießling Med 0 DPGF	SH/SL-Kießling Med 3&5 DPGF	SH/SL-Kießling Med 3&5 AMDF	SH/SL-Kießling Med 3&5 Seneff	SH/SL-Sonoranten Med 0 DPGF	SH/SL-Sonoranten Med 3&5 AMDF	SH/SL-Sonoranten Med 3&5 Seneff
Leo	BA	30	33.3	26.6	50.0	76.6	30.0	53.3	56.6
	m	178	6.2	5.1	32.6	5.6	15.2	39.9	6.7
	w	182	14.8	9.8	24.7	20.3	7.6	22.0	17.0
	g	390	12.3	8.9	30.3	17.9	12.8	32.6	15.4
Fokus	m	175	7.9	6.9	45.1	17.7	14.9	62.9	14.9
	w	183	8.2	6.0	38.8	37.7	6.6	28.4	30.6
	g	358	8.1	6.4	41.9	27.9	10.6	45.3	22.6
Dialog	g	16	43.8	43.8	100.0	81.3	31.3	100.0	87.5
Gesamt		764	11.0	8.5	37.2	24.0	12.2	40.0	20.3

Tab. 6.4: Anteil der Sätze in Prozent mit mindestens einem Grobfehler bei einem der drei angegebenen Verfahren. Die Zeilen *m*, *w* bzw. *g* beziehen sich ausschließlich auf männliche, weibliche bzw. alle Sprecher (gesamte Stichprobe). Die Zeile *BA* bei der Leo-Stichprobe bezeichnet die 30 Sätze des männlichen Sprechers BA. Diese sind in den Werten der Zeile *m* nicht enthalten. Somit wurden die Äußerungen der Zeilen *m* bei Leo- und Fokus-Stichprobe von denselben Sprechern produziert. In der Zeile *Gesamt* ist jeweils die Summe aus allen drei Stichproben eingetragen. Für DPGF, Seneff und AMDF wurden nichtlinear geglättete Konturen verwendet (*Med 3&5*: zuerst Median der Länge 3 Frames, dann der Länge 5 Frames). Für DPGF sind auch die Fehlerraten für ungeglättete Konturen angegeben (*Med 0*).

Stich-probe		SH/SL-Kießling Anzahl Frames	Med 0 DPGF	Med 3&5 DPGF	Med 3&5 AMDF	Med 3&5 Seneff	SH/SL-Sonoranten Anzahl Frames	Med 0 DPGF	Med 3&5 AMDF	Med 3&5 Seneff
	BA	2030	5.7	5.6	2.5	7.8	1901	4.9	3.8	7.3
	m	12141	1.5	1.5	1.8	0.3	10862	2.0	2.3	0.2
Leo	w	12439	1.2	1.0	2.1	2.0	11183	1.0	1.9	1.8
	g	26610	1.7	1.6	1.9	1.7	23946	1.8	2.2	1.5
	m	19023	0.8	0.7	1.9	0.6	16568	1.7	2.6	0.3
Fokus	w	18513	0.5	0.5	1.9	1.9	15879	0.3	1.4	1.3
	g	37536	0.6	0.6	1.9	1.3	32447	1.0	2.0	0.8
Dialog	g	4235	5.5	5.5	5.7	5.0	4038	2.5	5.8	4.8
Gesamt		68381	1.3	1.3	2.2	1.6	60431	1.4	2.3	1.3

Tab. 6.5: Anteil der Frames mit Grobfehler an der Gesamtzahl der stimmhaften Frames der betreffenden Stichprobe (Bezeichnungen wie Tab. 6.4).

Tabelle 6.6 zeigt beispielhaft für drei SH/SL- und Glättungs-Einstellungen die sprecherabhängigen äußerungsweisen Grobfehlerraten für die 6 Sprecher der Leo- und der Fokus-Stichprobe. Zum Vergleich sind auch die Fehlerraten für den männlichen Sprecher BA aufgeführt.

	weiblich			männlich			
	LO	ST	SC	PA	AU	EB	BA
DPGF Med 0 Sonoranten	5.9	6.3	11.7	8.6	13.3	13.7	30.0
DPGF Med 0 SH/SL-Kießling	5.9	11.0	13.3	7.8	9.2	6.8	33.3
DPGF Med 3&5 SH/SL-Kießling	4.2	4.7	11.7	6.7	7.5	6.8	26.6

Tab. 6.6: Sprecherabhängige Grobfehlerraten (in Prozent der Sätze) für Leo- und Fokus-Stichproben gemeinsam (ca. 120 Äußerungen pro Sprecher) für das DPGF-Verfahren. Zum Vergleich sind die Ergebnisse für die 30 Äußerungen des männlichen Sprechers BA aufgeführt.

Die Grobfehleranalyse für die drei Grundfrequenz-Algorithmen läßt sich wie folgt zusammenfassen:

1) Das DPGF-Verfahren zeigte das mit Abstand beste Fehlerverhalten unter den drei F_0-Algorithmen. Das Seneff-Verfahren war deutlich besser als das AMDF-Verfahren (das letzte Ergebnis steht in Einklang mit den Ergebnissen in [HEUNISCH 86], siehe auch Kap.4.2.3).
2) Das verwendete Medianfilter brachte in allen Fällen eine Verbesserung der Ergebnisse. Während bei AMDF und Seneff die Filterung unbedingt durchgeführt werden sollte, ist die Verbesserung bei DPGF zwar immer noch deutlich, aber nicht so stark (in den Tabellen ist für AMDF und Seneff nur das Ergebnis für die mediangefilterten Konturen angegeben).
3) Die SH/SL-Entscheidung nach [KIESSLING 89] brachte für das DPGF-Verfahren fast durchweg bessere Ergebnisse als das Sonoranten-Kriterium, obwohl im Schnitt 13 Prozent mehr Frames als stimmhaft klassifiziert wurden. Durch die weichere SH/SL-Entscheidung wurden häufig zwei benachbarte stimmhafte Bereiche zu einem Bereich verschmolzen. Im Vergleich zur Sonoranten-Entscheidung wurden daher weniger Zielwertfehler erzeugt.
4) Bei der Auswertung nach Frames war das DPGF-Verfahren den anderen Verfahren nicht so stark überlegen wie bei einer äußerungsweisen Auswertung. Dies liegt daran, daß sich ein falscher F_0-Wert durch die Pfadeinschränkungen der Dynamischen Programmierung über einen längeren Bereich auswirkt. Somit werden mit dem DPGF-Verfahren im Mittel wesentlich weniger fehlerhafte Konturen berechnet, aber ein Fehler wirkt sich stärker aus.
5) Der Sprecher BA schnitt bei allen Verfahren deutlich schlechter ab. Eine mögliche Erklärung könnte die im Vergleich zu den anderen Sprechern ausgesprochen tiefe Stimmlage des Sprechers sein. Diese Vermutung müßte an einem größeren Korpus verifiziert werden. Für die anderen Sprecher konnten keine sprecher- oder stimmlagenspezifischen Fehler beobachtet werden.
6) Die Dialog-Stichprobe zeigte wesentlich schlechtere Ergebnisse als die anderen Stichproben. Eine Erklärung für dieses Verhalten ist die hohe Anzahl der Laryngalisierungen in diesem Material. Im Gegensatz zu den anderen Korpora handelt es sich bei der Dialog-Stichprobe um frei gesprochenes Material mit sehr starken Reduktionsformen und einer vergleichsweise hohen Sprechgeschwindigkeit. 13 Prozent der Vokale wurden zumindest teilweise laryngal artikuliert. Dies erklärt auch das bessere Fehlerverhalten des DPGF-Verfahrens mit der Sonoranten-Entscheidung, während ansonsten das Kießling-Verfahren bessere Ergebnisse brachte: Der Lautkomponenten-Klassifikator klassifizierte einige laryngale Stellen als /XI/ oder /XA/ (ICH- und ACH-Laut) und somit als stimmlos. In den benachbarten normal artikulierten Lauten wurde die korrekte F_0-Kontur gefunden. Bei der SH/SL-Entscheidung nach Kießling wurden diese laryngalen Bereiche als stimmhaft klassifiziert. Wurden diese Stellen für die Zielwertbestimmung benutzt, so sorgten sie dafür, daß dem gesamten SH-Bereich ein deutlich zu niedriger Konturverlauf zugeordnet wurde.

Die Anzahl der Laryngalisierungen ist sprecher- und sprechtempospezifisch. Laryngalisierungen wurden bisher noch nicht sehr gründlich untersucht und häufig als Störfaktor von der Analyse sogar ausgeschlossen. Eine Erweiterung des Mehrkanalverfahrens in Bezug auf Erkennung und Korrektur von Laryngalisierungen ist unbedingt notwendig und dürfte zu einer starken Verbesserung des Grobfehler-Verhaltens führen. Das Wissen über laryngale Stellen kann im übrigen auch für die häufig schwierige Unterscheidung von Vokal-Vokal-Folge vs. Diphthong sehr wichtig sein.

6.2 Prosodische Satzmodus-Bestimmung

Die Untersuchungen zur prosodischen Satzmodus-Bestimmung wurden in enger Zusammenarbeit mit den Mitarbeitern des DFG-Projekts *Modus-Fokus-Intonation* durchgeführt. Weitere Einzelheiten zu den im folgenden präsentierten Ergebnissen finden sich in [BATLINER 89a, 89b, 89c, 89h], [STALLWITZ 89], [NÖTH 87] und [LANG 87]. Um eine möglichst große Datenmenge zur Verfügung zu haben, wurden auch per Hand aus Mingogrammen extrahierte intonatorische Merkmale benutzt. Somit stand das gesamte Modus-Fokus-Korpus (Kap.3.4) zur Verfügung. Für die beiden Teilkorpora, die in digitalisierter Form vorliegen, die Leo- und die Fokus-Stichprobe, wurden zusätzliche Experimente mit automatisch extrahierten Merkmalen durchgeführt. Bei den per Hand extrahierten Merkmalen wurden irreguläre Werte nicht in die Analyse mit einbezogen. Bei diesen Problemfällen handelt es sich um Laryngalisierungen am Äußerungsende, wodurch kein sinnvoller Wert für den satzmodus-relevanten *Offset* (F_0-Wert am Äußerungsende) ermittelt werden konnte. Der Anteil der aussortierten Äußerungen betrug knapp vier Prozent. Somit verblieben 1999 von ursprünglich 2074 Äußerungen in der Analyse. Bei den automatisch extrahierten Merkmalen wurden fehlerhafte Merkmalwerte nicht ausgesondert. Auf diese Weise konnten an einem großen und segmental uneinheitlichen Korpus verschiedene intonatorische Merkmale unter verschiedenen Transformationen auf ihre Relevanz für die intonatorische Markierung des Satzmodus überprüft werden. Gleichzeitig waren Aussagen über die Übertragbarkeit dieser Ergebnisse auf ein vollständig automatisches Verfahren zu erwarten.

6.2.1 Ein einfaches Modell der intonatorischen Markierung des Satzmodus

Gegenstand der im folgenden beschriebenen Untersuchungen ist die für sprachverstehende Dialogsysteme äußerst wichtige intonatorische Unterscheidung von Fragen vs. Nicht-Fragen (siehe das Beispiel in Kap.1.2). Nicht-intonatorische Merkmale (syntaktische, semantische, etc.) werden für die Unterscheidung hier ebensowenig betrachtet wie die im Satzmodus-System nach [ALTMANN 84] mögliche weitere Differenzierung der fünf Satzmodus-Grundtypen *Frage, Aussage, Imperativ, Wunsch und Exklamativ*. Es werden nur Tonhöhen-Merkmale untersucht, obwohl z.B. für die Markierung des *Exklamativ* auch die prosodische Eigenschaft Lautheit eine Rolle spielt.

Es ist bekannt, daß der F_0-Wert am Äußerungsende für die Frage/Nicht-Frage-Unterscheidung eine große Rolle spielt. Viele Intonationsbeschreibungen, insbesondere diejenigen der meisten deskriptiven Grammatiken des Deutschen, stützen sich auf [VON ESSEN 56], der neben der weiterweisenden *progredienten Intonation* die *terminale* und die *interrogative Intonation* unterscheidet. Die terminale Intonation (tiefer F_0-Wert am Äußerungsende bzw. fallender Tonverlauf) charakterisiert Aussagen und W-Fragen, die interrogative (hoher F_0-Wert am Äußerungsende bzw. steigender Tonverlauf) wird bei Entscheidungsfragen eingesetzt. (Der Begriff *Intonation* wird, anders als in dieser Arbeit, bei [VON ESSEN 56] mit Tonhöhenverlauf gleichgesetzt.) Der Parameter *F_0-Wert am Äußerungsende* wird dort als alleiniger Entscheidungsfaktor genannt. [KLEIN 82] belegt die Tatsache, daß die "von Essensche Intonationsanalyse" Eingang in die meisten deskriptiven Grammatiken gefunden hat, anhand eines Zitats aus [HELBIG 74]:

> Im nicht zusammengesetzten Aussagesatz steht das finite Verb gewöhnlich an zweiter Stelle. Die Intonation ist terminal. ...
>
> Das finite Verb tritt in der Entscheidungsfrage an die Satzspitze. Die Intonation ist interrogativ.
> [HELBIG 74, S.541-542]

Kleins Kommentar zu diesem Zitat ist genauso kurz wie die zitierte Intonationsbeschreibung:

> Das ist sehr knapp, sehr klar, vielleicht ein bißchen undifferenziert, aber es gibt im Kern das wieder, was sich auch sonst in den Grammatiken findet. Es ist völlig falsch.
> [KLEIN 82, S.291]

Das Urteil von Klein, das später von ihm differenziert und relativiert wird, ist sicherlich, zumindest quantitativ gesehen, zu hart. Zwar zeigt Klein, daß, qualitativ gesehen, auch ganz andere intonatorische Realisierungsformen für Fragen möglich sind, aber da die terminale und interrogative Form häufig benutzt wird, ist das Modell eben nicht *völlig* falsch. Es soll daher die Frage untersucht werden, *wie* falsch das Modell ist, d.h.

- inwieweit reicht der Grundfrequenzwert am Äußerungsende bzw. die Charakterisierung steigende/fallende Grundfrequenzkontur als Frage-/Nicht-Frageindikator aus?
- wie soll dieser Wert aussehen (Rohwert in Hz oder umgerechnet zu Bezugsgrößen)?
- gibt es sprecherspezifische Strategien beim Einsatz der Parameter?

6.2.2 Quantitative Gültigkeit des Modells für die Modus-Fokus-Korpora

Mit der Diskriminanzanalyse wurden die folgenden Konstellationen für die in der Analyse verbliebenen 1999 Äußerungen benutzt, um den durch den Kontextsatz und die Situationsbeschreibung vorgegebenen Satzmodus zu bestimmen (eine ausführliche Beschreibung der Korpora findet sich in [BATLINER 89g], siehe auch Kap.3.4):

Prädiktorvariablen

- der F_0-Wert am Äußerungsende (*Off*)
- der F_0-Wert am Äußerungsanfang (*Ons*)
- der größte F_0-Wert (*Max*)
- der kleinste F_0-Wert (*Min*)

Transformationen

- F_0-Rohwerte (*Hz*)
- F_0-Werte umgerechnet in Halbtonwerte (*Ht*)
- *Hz*- und *Ht*-Werte minus sprecherspezifischem Basiswert (*Hzbas* bzw. *Htbas*)
- *Hz*- und *Ht*-Werte minus einem Äußerungsmittelwert aus den vier Werten *Ons*, *Off*, *Max* und *Min* (*Hzmit* bzw. *Htmit*)

Lernstichproben

- alle Sprecher (*ln*); hiermit können die interessanten Fälle gefunden werden, die selbst im günstigsten Fall aus dem Modell herausfallen.
- n-1 Sprecher (*ln-1*); mit der "leave-one-out"-Methode kann Sprecherunabhängigkeit simuliert werden.
- ein Sprecher (*l1*); Generalisierung von einem auf andere Sprecher.

Die im weiteren besprochenen Analysen werden wie folgt gekennzeichnet: '*ln/Off/Htbas*' steht für eine Diskriminanzanalyse, bei der Lern- gleich Teststichprobe ist, und als Prädiktorvariable der Offset in Halbtönen umgerechnet zum sprecherspezifischen Basiswert benutzt wurde.

Die Ergebnisse der Diskriminanzanalyse lassen sich wie folgt zusammenfassen:

- Prädiktorvariablen:

  Bei *Hz* und *Ht* verbesserte zwar die Hinzunahme weiterer Prädiktoren die Klassifikation, da damit automatisch die Stimmlage (s.u.) mit in die Berechnung einging, bei *Htbas* und *Htmit* war die Verbesserung aber minimal und immer unter 1 Prozent. Zum einen liegt das an der Korrelation der Variablen untereinander, zum anderen daran, daß diese anderen Variablen für die Unterscheidung von Frage/Nicht-Frage kaum relevant sind. (Das gilt nicht für die Unterscheidung anderer Satzmodi, siehe [NÖTH 87]).

- Transformationen:

  Die im allgemeinen als "gehörsadäquat" angesehene Transformation der *Hz* in *Ht*-Werte zur Normierung des unterschiedlichen *Stimmumfangs (Range)* von Männern und Frauen ergab eher schlechtere Ergebnisse. Bei *Hzmit* und *Htmit* sind die Unterschiede zu vernachlässigen (siehe Tabelle 6.8). Der Range betrug bei den weiblichen Sprechern im Mittel 146 *Hz* bzw. 10.3 *Ht*, bei den Männern 91 *Hz* bzw. 11.7 *Ht*. In *Hz* ist der Range der Frauen im Mittel also größer als der der Männer, in *Ht* ist es umgekehrt. Möglicherweise liegt eine adäquatere Transformation *zwischen* den *Hz*- und *Ht*-Werten.

  Eine Normierung der *Stimmlage* zu einem Bezugswert, egal ob zum Basis- oder Mittelwert, ergab dagegen immer bessere Ergebnisse als bei den *Hz*- oder *Ht*-Werten (siehe

Tab. 6.8). Der Grund wird klar, wenn man z.B. *Ht* und *Htbas* bei *ln-1/Off* miteinander vergleicht und das Ergebnis nach Fragen (F) und Nicht-Fragen (N) sowie nach männlichen (M) und weiblichen Sprechern (W) aufschlüsselt (Tab. 6.7): Wegen der unterschiedlichen Stimmlagen werden ohne die Normierung bei den Frauen viele Nicht-Fragen fälschlich als Fragen klassifiziert und umgekehrt bei den Männern viele Fragen als Nicht-Fragen. (Die beiden Einträge sind in Tabelle 6.7 markiert.)

			klassifiziert als			
			F		N	
			Ht	Htbas	Ht	Htbas
Testsatz	F	W	92	78	8	23
		M	41	83	59	17
	N	W	46	10	55	90
		M	1	8	99	92

Tab. 6.7: Erkennungsraten für Fragen und Nicht-Fragen bei den Konstellationen *ln-1/Off/Ht* und *ln-1/Off/Htbas*, getrennt nach dem Geschlecht der Sprecher.

- Sprecherspezifische Strategien:

  Tabelle 6.8 zeigt, daß der Unterschied bei *Off* zwischen *ln*, *ln-1* und *l1* bei *Hz* jeweils ca. fünf Prozent beträgt, bei *Hzmit*, *Htmit* und *Htbas* aber vernachlässigbar ist; die einzelnen Sprecher wenden also grundsätzlich die gleiche Strategie an. Ein Blick auf die einzelnen Sprecher bei *ln-1/Off/Htbas* zeigt, daß bei der Prädiktion der Fragen eine Sprecherin deutlich schlechter abschneidet (48 Prozent) als alle anderen Sprecher, die zwischen 75 Prozent und 93 Prozent liegen. Wenn der Fokus nicht auf der letzten Phrase liegt, sondern auf der vorletzten, werden Frage und Fokus normalerweise durch einen F_0-Abfall und einen ausgeprägten F_0-Anstieg auf dieser Phrase indiziert; ein hoher finaler Offset zur Markierung des Satzmodus *Frage* ist dann fakultativ und 'normal', wird aber von dieser einen Sprecherin nicht realisiert (20 Fälle aus dem Fokus-Korpus, siehe Bild 2.6c).

Prädiktor-variable	Lernstichprobe		
	ln	ln-1	11
Off/Hz	75	71	65
Off/Ht	68	67	65
Off/Hzmit	87	87	86
Off/Htmit	87	87	86
Off/Htbas	87	87	86

Tab. 6.8: Erkennungsraten für unterschiedliche Lernstichproben.

6.2.3 Automatische Merkmalextraktion - Reproduzierbarkeit der Ergebnisse und weitere Merkmale

Die Erkennungsraten für die automatischen Paramter (z.B. *ln-1/off/htmit:* 91 Prozent für das Fokus-Korpus und 88 Prozent für das Leo-Korpus) sind zwar etwas niedriger, aber mit denen für handextrahierte Merkmale (93 Prozent für Leo- und Fokus-Korpus) durchaus vergleichbar. Bei diesen Experimenten wurden lediglich die Äußerungen der Teilkorpora zum Trainieren und Testen verwendet. Die Einzel- und Gesamtergebnisse für die handextrahierten Merkmale konnten bis auf die leicht schlechteren Erkennungsraten reproduziert werden.

Zusätzlich wurde die Steigung der Ausgleichsgerade der F_0-Werte (*Steig*) als Prädiktorvariable untersucht. Die Annahme, daß sich der fallende Tonverlauf in Aussagesätzen und der steigende Tonverlauf in Fragesätzen in der Steigung der Ausgleichsgeraden über alle stimmhaften F_0-Werte niederschlägt, wurde bestätigt. Für das Fokus-Korpus und die *ln-1/Steig/Hz*-Konstellation z.B. betrug die Erkennungsrate 86 Prozent. Auch bei der Ausgleichsgeraden brachte die Halbton-Transformation keine Verbesserung der Ergebnisse, ebensowenig wie verschiedene Methoden zur Entfernung von mikroprosodischen Einflüssen aus der F_0-Kontur ([STALLWITZ 89]).

6.2.4 Fälle, in denen das einfache Modell nicht zutrifft

Bild 6.1 zeigt die Verteilung in Prozent der *Off/Htbas*-Werte für Fragen und Nicht-Fragen. Für die Fragen zeigt sich eine klare bimodale Verteilung, wobei der kleinere Gipfel im Zentrum der Nicht-Fragen liegt. Im Gegensatz dazu ist die Verteilung der Nicht-Fragen unimodal, aber rechtsschief und reicht nicht ins Zentrum der Fragen. Dies schlägt sich auch in den Erkennungsraten für Fragen und Nicht-Fragen nieder (z.B. für *ln-1/Off/Htbas* 80 Prozent für Fragen vs. 91 Prozent für Nicht-Fragen).

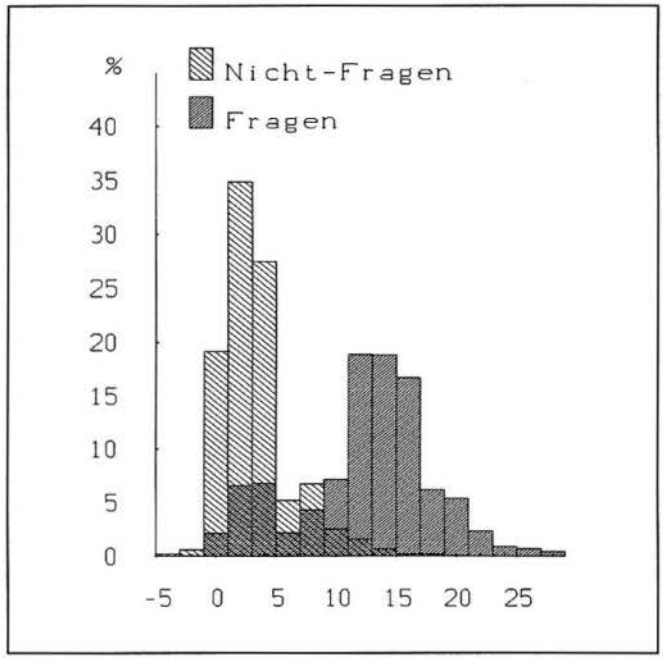


Bild 6.1: *Off/Htbas*-Werte für Fragen und Nicht-Fragen.

		Diskriminanzanalyse (*ln/Off/Htbas*)					
		richtig			falsch		
Hörer-urteil	richtig	I	1625	(81%)	II	221	(11%)
	falsch	III	105	(5%)	IV	47	(2%)

Tab. 6.9: Klassifikation durch Hörer und Diskriminanzanalyse.

Um zu überprüfen, wie oft das einfache Modell der intonatorischen Satzmodus-Markierung zutrifft, wurde das Ergebnis einer Diskriminanzanalyse mit dem Ergebnis eines Hörtests verglichen. Durchschnittlich 12 phonetisch untrainierte Hörer klassifizierten alle Äußerungen ohne Kontext nach den Satzmodus-Grundtypen *Frage, Aussage, Imperativ, Wunsch und Exklamativ* (siehe [OPPENRIEDER 88a]). Diese Klassifizierung wurde umkodiert in einen Wert zwischen 0.0 und 1.0 für Frage/Nicht-Frage. Wie bei der Wahrscheinlichkeit der Gruppenzugehörigkeit, die sich aus der Diskriminanzanalyse ergibt, wurden nun die Werte unter 0.5 als 'falsch' und die über 0.5 als 'richtig' bezeichnet. Damit ergibt sich die Aufteilung der 1999 Fälle in Tabelle 6.9.

Geht man davon aus, daß in Fällen, in denen Hörerurteil und automatische Klassifikation übereinstimmen, der Offset die dominierende Rolle bei der Satzmodus-Markierung spielt, so trifft diese Annahme bei dem untersuchten Datenmaterial in gut 80 Prozent der Fälle zu (Gruppe I). Eine genaue Diskussion der Gruppen II, III und IV findet sich in [BATLINER 89h]. Interessant sind insbesondere die Fälle aus Gruppe II, da hier anzunehmen ist, daß andere Merkmale als der Offset die Klassifikation bestimmen. Die meisten dieser Fälle sind für den linken Gipfel bei der bimodalen Verteilung der Fragen verantwortlich, weisen also einen tiefen Offset auf.

47 Fälle (2.4 Prozent) sind Alternativfragen wie *Möchten Sie Mohn oder Streusel?*, die bei expliziter Angabe aller Wahlmöglichkeiten regulär mit tiefem Offset produziert werden. In diesen wie in ähnlichen Fällen disambiguieren also *nicht-intonatorische Merkmale* wie Verbstellung und Verbsemantik.

20 Fälle (1 Prozent) sind auf die sprecherspezifische Strategie des tiefen Offset bei Fokus in präfinaler Position zurückzuführen (siehe Kap.6.2.2).

In einigen Fällen, besonders aus Gruppe IV, muß der *Kontext* allein disambiguieren, so z.B. bei *Schlafen Sie ?/!* (Frage oder Imperativ). Hier ist auch bei Fragen ein tiefer Offset regulär.

Zusammenfassend läßt sich sagen, daß das sehr einfache Modell der intonatorischen Satzmodus-Markierung mit hohem F_0-Wert am Äußerungsende für Fragen und tiefem F_0-Wert für Nicht-Fragen bei dem untersuchten Datenmaterial quantitativ gut zutraf und daß sich die abweichenden Fälle gut erklären ließen.

6.3 Datengetriebene Silbenkernbestimmung und Betonungszuweisung

Die Ergebnisse zur Silbenkerndetektion beziehen sich auf die Äußerungen von vier Sprechern aus der EVAR-Stichprobe (Kap.3.1) und auf die Dialog-Stichprobe (Kap.3.2). Die Teilstichprobe der EVAR-Stichprobe umfaßt 90 Äußerungen mit insgesamt ca. 1100 Silben. Weitere Einzelheiten zu den im folgenden präsentierten Ergebnissen finden sich in [SCHMÖLZ 85, 87]. Die Äußerungen von drei der vier Sprecher aus der EVAR-Stichprobe dienten als Lernstichprobe für das Einstellen der Schwellwerte bei der Silbenkerndetektion sowie für die Erstellung der Bewertungsfunktionen bei der Betonungszuweisung.

6.3.1 Ergebnisse zur Silbenkerndetektion

Die folgenden Ergebnisse beziehen sich auf die Lokalisierung der Silbenkerne aufgrund der spektralen Energie. Mit dem in [SCHMÖLZ 85] implementierten Verfahren (Kap. 4.1.2) konnten 92 Prozent der Silbenkerne in der EVAR-Stichprobe detektiert werden. Die Zahl der zusätzlich gefundenen Silbenkerne betrug zwei Prozent der gesprochenen Silbenkerne. Durch die in [SCHMÖLZ 87] implementierten Erweiterungen (Kap. 4.1.2) des Algorithmus ergaben sich folgende Veränderungen:

1) Die Zahl der gefundenen konsonantischen Silbenkerne konnte durch die Betrachtung des nasalen Energiebandes von 71 auf 87 Prozent gesteigert werden. Gleichzeitig verdoppelte sich die Zahl der Einfügungen auf fünf Prozent. Eine genaue Analyse dieser zusätzlichen Silbenkerne ergab, daß es sich in erster Linie um Nachhall bei den isoliert gesprochenen Wörtern (die Namen der Städte mit IC-Bahnhof) handelte.
2) Durch die Betrachtung des frikativen Energiebandes konnte ein halbes Prozent der ursprünglichen Einfügungen eliminiert werden, ohne daß ein tatsächlicher Silbenkern gelöscht wurde.
3) Die Korrektur der überlangen Silbenkerne führte ebenfalls zu einer Verbesserung der Ergebnisse um ein halbes Prozent.
4) Die Gesamterkennungsrate stieg auf 94 Prozent gefundene Silbenkerne. Der starke Anstieg der fehlerhaften Einfügungen auf fünf Prozent wurde in Kauf genommen, da es sich in erster Linie um ein Artefakt bei den isoliert gesprochenen Einzelwörtern handelte. Den größten Anteil an den Veränderungen hatte die Bestimmung der konsonantischen Silbenkerne.

Die Erkennungsraten für die Dialog-Stichprobe waren deutlich schlechter als die für die EVAR-Stichprobe. Insgesamt wurden nur 87 Prozent der Silbenkerne detektiert und drei Prozent zusätzliche Silbenkerne eingefügt.

Bild 6.2 zeigt das Erkennungsverhalten der Silbenkerndetektion (V2) bei der EVAR-Stichprobe und der Dialog-Stichprobe in Bezug auf vokalische Silbenkerne, konsonantische Silbenkerne, alle Silbenkerne und Einfügungen. Für die EVAR-Stichprobe sind die Erkennungsraten auch für den Grundalgorithmus aus [SCHMÖLZ 85] mit angegeben (V1).

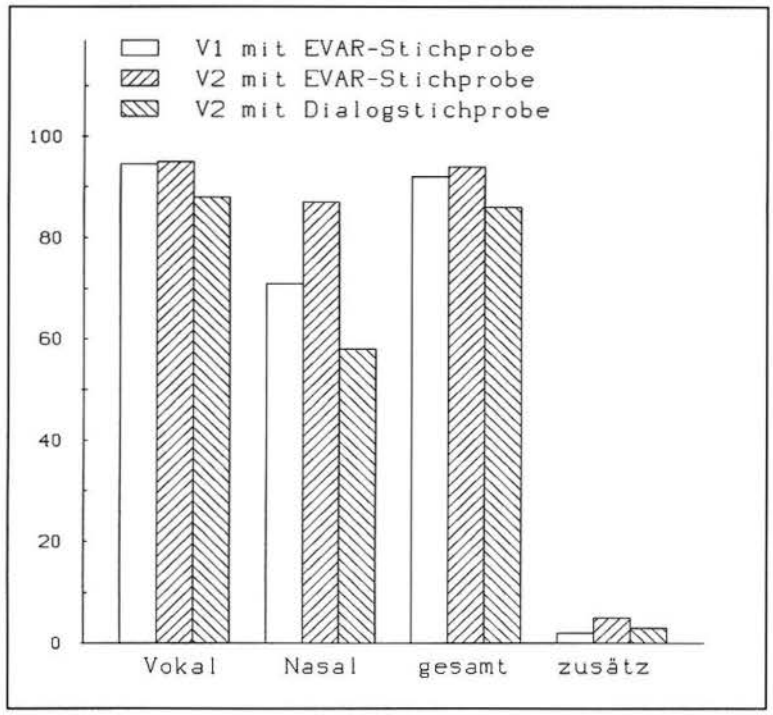


Bild 6.2: Ergebnis der Silbenkernbestimmung für die vokalischen Silbenkerne, für die konsonantischen Silbenkerne, Gesamtergebnis sowie Anzahl der zusätzlich eingefügten Silbenkerne. Nach oben ist der Anteil der identifizierten Silbenkerne in Prozent angegeben, bzw. welcher Prozentsatz der tatsächlich vorhandenen Silbenkerne zusätzlich eingefügt wurde. Bei V1 handelt es sich um den in [SCHMÖLZ 85] beschriebenen Grundalgorithmus, bei V2 um die in [SCHMÖLZ 87] erweiterte Version.

Während die Ergebnisse für die EVAR-Stichprobe vergleichbar mit Ergebnissen aus der Literatur sind, trifft dies für die Dialog-Stichprobe nicht zu. [MERMELSTEIN 75] berichtet von 93 Prozent erkannter Silben und drei Prozent Einfügungen, [LEA 80b] von 91 Prozent erkannter Silben und 1 Prozent Einfügungen. Eine genaue Analyse der nicht gefundenen Silbenkerne in der Dialog-Stichprobe zeigte, daß der Unterschied in den Erkennungsraten fast ausschließlich auf die unterschiedliche Aufnahmesituation für die beiden Stichproben zurückzuführen ist. Während die Sätze der EVAR-Stichprobe nach einer schriftlichen Vorlage gesprochen wurden, handelt es sich bei der Dialog-Stichprobe um freie Rede. Um eine möglichst natürliche Betonung unter Dialogsituation zu erhalten, wurde bewußt auf eine Instruktion der Sprecher verzichtet. Dies hatte allerdings u.a. auch zur Folge, daß sich der Sprecher gelegentlich innerhalb einer Äußerung vom Mikrophon wegdrehte, und daß sehr starke Verschleifungserscheinungen zu beobachten waren. Insbesondere bei Silbengrenzen mit Vokal-Vokal-Übergang, aber auch bei intervokalischen Sonoranten trifft in schneller Sprechweise und bei Verwendung starker Reduktionsformen das Modell der Schallfülle-Theorie mit einem ausgeprägten Gipfel pro Sprechsilbe nicht mehr zu. Dies führt zu Silbenkernverschmelzungen (siehe Bild 2.3):

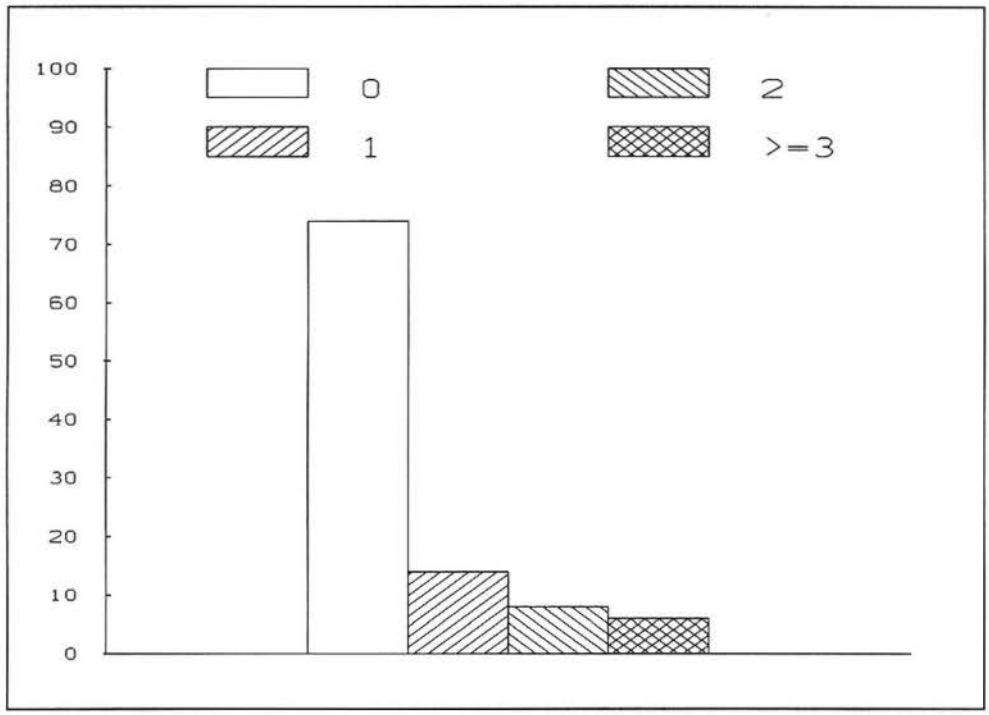


Bild 6.3: Prozentualer Anteil der nicht gefundenen Silbenkerne in Abhängigkeit von ihrer Summenbewertung.

Bild 6.3 zeigt für die nicht gefundenen Silben die prozentuale Verteilung in Abhängigkeit von der in Kap.3.2.3 beschriebenen Summenbewertung. Man erkennt, daß ca. 95 Prozent der nicht gefundenen Silbenkerne eine Bewertung ≤ 2 haben. Es besteht somit ein starker Zusammenhang zwischen dem Akzentuierungsgrad bzw. der Reduktionsstufe und der Erkennungsrate.

6.3.2 Ergebnisse zur Betonungsbeschreibung

Zur Beurteilung des automatisch erzeugten Betonungsmaßes wurde das Ergebnis mit dem Resultat des in Kap.3.2.3 beschriebenen Perzeptionstests verglichen. Dabei wurden die beiden Intervalle in jeweils drei Bereiche (unbetont, betont, stark betont) unterteilt. Für die einzelnen Stufen wurden die in Tabelle 6.10 angegebenen Grenzen heuristisch festgelegt.

Bild 6.4 zeigt den prozentualen Anteil der von der automatischen Akzentbestimmung als "unbetont", "betont" und "stark betont" bewerteten Silben in Abhängigkeit von der auditiven Bewertung der Hörer.

Tabelle 6.11 zeigt das Ergebnis eines weiteren Vergleichs der Hörer- und der automatischen Bewertung: Der Wertebereich des automatischen Betonungsmaßes wurde so in vier Intervalle unterteilt, daß ungefähr gleich große Silbenkernmengen entstanden. Für die Silbenkernmengen wurde jeweils die durchschnittliche Hörerbewertung ermittelt.

	Perzeptions-test	automatische Akzentbestimmung
unbetont	0 - 2	0.0 - 0.33
betont	3 - 8	0.34 - 0.66
stark betont	9 - 15	0.67 - 1.0

Tab. 6.10: Einteilung der bei der auditiven Beurteilung und bei der automatischen Akzentbestimmung möglichen Ergebnisse in die drei Klassen "unbetont", "betont" und "stark betont".

automatische Betonungsbewertung	0.0-0.2	0.2-0.4	0.4-0.8	0.8-1.0
durchschnittliche Hörerbewertung	1.9	2.9	4.2	5.2
prozentualer Anteil an der Silbenkernmenge	22	24	25	30

Tab. 6.11: Durchschnittliche Hörerbewertung und prozentualer Anteil an der Gesamtmenge der Silbenkerne für vier nach der automatischen Bewertung geordnete Teilmengen der Silbenkerne.

Bild 6.4 und Tabelle 6.11 zeigen einen deutlichen Zusammenhang zwischen der automatischen Akzentbewertung und der Hörer-Bewertung. Man erkennt auch, daß das momentane Ergebnis noch verbessert werden muß. Die Fehlentscheidungen des Algorithmus sind in [SCHMÖLZ 87] anhand der Fälle mit hohem Hörerurteil und niedriger automatischer Bewertung bzw. mit niedrigem Hörerurteil und hoher automatischer Bewertung ausführlich untersucht worden. Die Fehlerursachen lassen sich in drei Gruppen einteilen:

1) *Fehler in der Hörerbewertung*

   Stark ausgeprägte Pausenfüller (*"äh"*) wurden von den Hörern aufgrund der semantischen Bedeutung als *unbetont* , von der automatischen Bewertung aufgrund der starken Signalveränderungen jedoch als *stark betont* eingestuft.

2) *Fehlerhafte Eingangsdaten*

   Fehlerhafte Grundfrequenzwerte führten zu starken Veränderungen der Parameterwerte zur Bewertung der prosodischen Eigenschaft Tonhöhe (siehe Kap.6.1).

   Durch die Verschmelzung von zwei unbetonten Silben wurde diesen eine zu hoher Dauer-Bewertung und (über das Energie-Integral) eine zu hohe Lautheits-Bewertung zugewiesen. Somit wurden diese Silbenkerne als *betont* eingestuft, während ein benachbarter, betonter Silbenkern u.U. gleichzeitig als *unbetont* bewertet wurde.

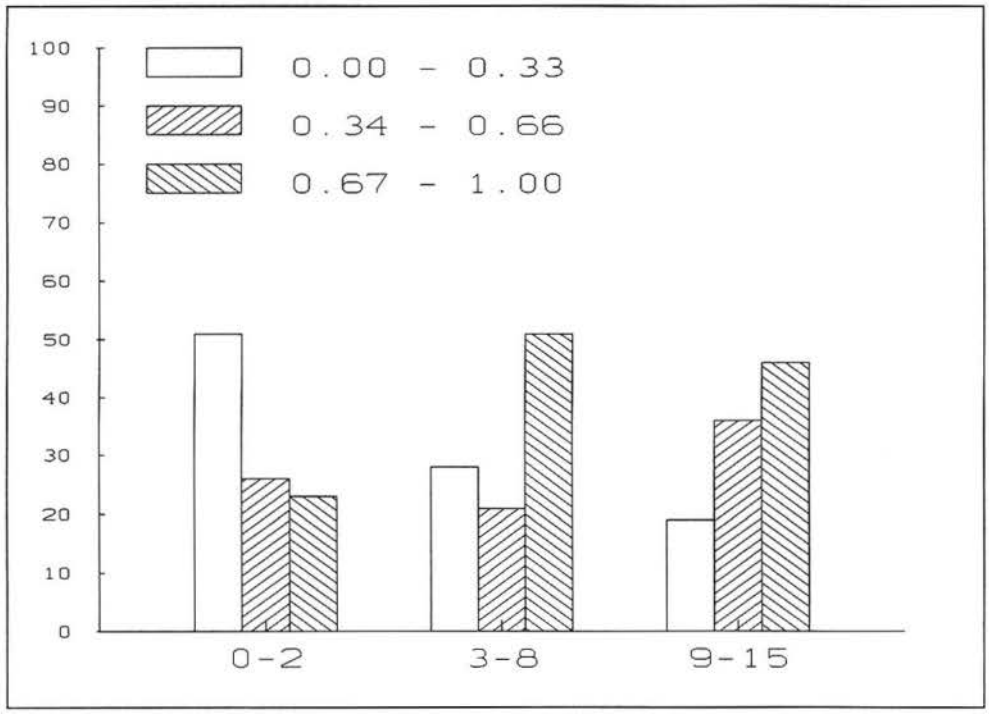


Bild 6.4: Prozentualer Anteil der Silbenkerne sk_i mit einer Akzentbewertung $AKZB_i$ im Intervall [0.0,0.33], [0.34,0.66] oder [0.67,1.0] in Abhängigkeit von einer Summenbewertung durch die Hörer in den Intervallen [0,2], [3,8] und [9,15].

3) *Intrinsische Eigenschaften*

Zu hohe intrinsische Energiewerte für die tiefen Vokale, insbesondere für /A/ und /AR/ führten zu einer Überbewertung der tiefen Vokale in unbetonter Stellung und zu einer Unterbewertung der hohen Vokale in betonter Stellung.

Bei kurzen betonten Vokalen fanden die Tonhöhenbewegungen oft schon im davor gesprochenen als unbetont eingestuften Silbenkern statt. Genaugenommen könnte man diese Fehlerursache auch unter Punkt 1) einordnen, denn signalphonetisch und perzeptiv ist der vom automatischen Verfahren als betont eingestufte Silbenkern auffällig, die Betonungszuweisung an die benachbarte Silbe durch den Hörer geschieht jedoch teilweise auch aufgrund der semantischen Interpretation.

6.4 Prosodische Verifikation der Silbenstruktur und des lexikalischen Wortakzents

Der folgende Abschnitt baut auf den Ergebnissen aus [NÖTH 88b] und [KOMPE 89a] auf. Die Erkennungsraten, über die dort berichtet wird, beziehen sich auf den in [REGEL 88] beschriebenen Segmentierer und die damit generierten Worthypothesen ([KUNZMANN 90]). Die hier präsentierten zusätzlichen Ergebnisse wurden mit dem in Kap.4.1.3 beschriebenen Segmentierer erzielt. Die Verbesserung der Erkennungsleistung durch die Verwendung des neuen Segmentieres wird in Kap.6.6 besprochen. Zum Zeitpunkt der Experimente lag allerdings erst eine Vorversion des Segmentierers vor, die ohne den Abgleich der Lauflängensegmentierung mit der Silbenkernsegmentierung arbeitete. Die Worthypothesen, die für die Untersuchungen in diesem Kapitel und in Kap.6.5 verwendet wurden, sind unter folgenden Bedingungen erzeugt worden:

- Es wurde das GLEXS-Lexikon (ca. 4200 Vollformen) verwendet.
- Es wurden 50 Worthypothesen/Lauthypothese erzeugt.
- Als Teststichprobe wurde die Pragmatik-Stichprobe verwendet, als Trainingsstichprobe der HMM auf Laut-und Wortebene die restlichen Äußerungen der EVAR-Stichprobe.
- Auf Laut- und auf Wortebene wurde jeweils das elementare Markov-Modell **S**ubstitute**I**nsert-**D**elete (SID-Variante, siehe [KUNZMANN 90, Kap.7-8]) benutzt.
- Bei der Erstellung der Lautsegmentierung wurden Lauflängen der Länge 1 mit gleichen linken und rechten Nachbarn weggeglättet.
- Die maximale zeitlich folgende Umgebung, in der für einen Lautanfang nach potentiellen Endpunkten gesucht wurde, war auf fünf Frames gesetzt.
- Die Silbensegmentierung des Prosodie-Moduls wurde für die Lautsegmentierung nicht verwendet.

6.4.1 Vorbemerkungen zur Leistungsbeurteilung bei der Worthypothesengenerierung und zur Verifikation

Bevor auf die Experimente im einzelnen eingegangen wird, werden einige Erkennungsmaße für die Worterkennung erläutert, weitere Einzelheiten finden sich in [KUNZMANN 90]. Geht man davon aus, daß die Worthypothesen nach einem Bewertungsmaß geordnet sind, so kann man bei einem **Einzelwort-Erkennungssystem** die Leistungsfähigkeit des Systems am *absoluten Rang* einer Worthypothese messen. Dieser ergibt sich für das richtige Wort aus seiner Position plus der Hälfte der gleich bewerteten Hypothesen. Da man für jedes Wort aus dem Lexikon eine Hypothese erzeugen kann, kann man davon ausgehen, daß das richtige Wort immer in der Hypothesenmenge vorliegt. Aussagekräftiger als der mittlere absolute Rang (über alle Wörter der Teststichprobe gemittelt) ist die absolute Rangkurve. Diese gibt für jede Zahl n, $0<n\leq100$ an, wieviele Hypothesen man pro zu erkennendem Wort erzeugen muß, um über alle Äußerungen der Teststichprobe n Prozent der richtigen Wörter zu erhalten.

Bei Erkennung **kontinuierlicher Sprache** ist der mittlere absolute Rang unzureichend. Hier muß man damit rechnen, daß nicht alle gesprochenen Wörter richtig positioniert in der Hypothesenmenge enthalten sind. Um dies zu gewährleisten, müßte man für jeden Teilbereich, d.h. für je zwei Lauthypothesen, alle Wörter des Lexikons hypothetisieren. Weiterhin ist es sinnvoll, den Rang auf die Länge der Äußerung zu normieren (relativer Rang), da man sehr unterschiedlich lange Äußerungen verarbeiten muß. Die Normierungseinheit sollte so gewählt werden, daß sie auch während eines Analyselaufes verwendet werden kann. Ein Kontroll-Modul kann z.B. von der Worterkennungseinheit nicht x Hypothesen pro gesprochenem Wort anfordern, da nicht bekannt ist, wieviele Wörter gesprochen wurden. Im EVAR-System wurde daher als Normierungseinheit die Anzahl der *Lauthypothesen (Segmente)* verwendet. Die *relative Rangkurve* gibt somit für jede Zahl n an, wieviele Worthypothesen pro Lauthypothese erzeugt werden müssen, um n Prozent der gesprochenen Wörter zu hypothetisieren. Da der neue Segmentierer allerdings ca. 15 Prozent mehr Lauthypothesen erzeugt als der Segmentierer nach [REGEL 88], wurde bei den hier vorgestellten Untersuchungen aus Gründen der Vergleichbarkeit für die relativen Rangkurven die Anzahl der mit dem Segmentierer nach [REGEL 88] erzeugten Lauthypothesen verwendet. Somit konnte die Verbesserung durch den Einsatz des neuen Segmentieres im Vergleich zum Segmentierer nach [REGEL 88] exakt dargestellt werden.

Bei der prosodischen Verifikation von Worthypothesen wird im folgenden davon ausgegangen, daß die Hypothesen nach der Güte geordnet vorliegen. Aufgrund prosodischer Information können dann die Hypothesen eine modifizierte Bewertung erhalten oder verworfen werden. Eine Neubewertung entspricht einer Umsortierung der Hypothesenmenge, eine Verwerfung einer Reduktion. Bei der Umsortierung hat man das Problem, eine neue Bewertung erstellen zu müssen. Die Bewertung einer Hypothese ist bei dem im EVAR-System verwendeten HMM das Ergebnis eines statistischen Mustervergleichs, also ein Ähnlichkeitsmaß. Möchte man die Hypothesen neu bewerten, stellt sich die Frage, wie die prosodische Bewertung, in diesem Fall ein Abstandsmaß zwischen einer Fuzzy-Bewertung für die Betonung und der lexikalischen Wortbetonung, mit dem statistischen Ähnlichkeitsmaß verknüpft werden kann. Insbesondere geht die Möglichkeit einer sinnvollen Interpretation u.U. verloren. Im folgenden werden zunächst Experimente zur prosodischen Verifikation beschrieben, bei denen die zu verifizierenden Hypothesen nicht umbewertet, sondern verworfen werden. Es handelt sich also um Filter, die für jede Worthypothese aus einer Hypothesenmenge ein Kriterium überprüfen und die Hypothese entweder unverändert akzeptieren oder verwerfen.

Ein Problem bei diesem Vorgehen ist das Verwerfen richtiger Hypothesen. Die Auswirkung eines Filters soll anhand von drei konstruierten Beispielen erläutert werden, bei denen die Teststichprobe aus einer einzigen Äußerung besteht: Angenommen, es liegt eine Äußerung mit fünf gesprochenen Wörtern und 50 Lauthypothesen vor und es werden 50 Hypothesen/Segment erzeugt. Somit liegen 2500 Worthypothesen vor, die prosodisch zu verifizieren sind. Die richtigen Worthypothesen seien an den Positionen 100, 200, 300, 400 und 500 in der nach der Bewertung geordneten

Hypothesenmenge. Um 60 Prozent der gesprochenen Wörter zu erhalten, müssen also sechs Hypothesen/Segment generiert werden.

Hat man ein Filter F1, das aufgrund einer fehlerhaften Betonungsbestimmung die richtige Hypothese an Position 300 verwirft und im Bereich unterhalb der Hypothese mit Position 500 keine weiteren Hypothesen mehr, dann müssen acht Hypothesen/Segment erzeugt werden, um 60 Prozent der Worthypothesen zu erhalten. Die relative Rangkurve verschlechtert sich also ab sechs Hypothesen/Segment, darunter bleibt sie gleich.

Hat man ein Filter F2, das alle Hypothesen ab Position 501 verwirft, so wird zwar die Hypothesenmenge um 80 Prozent reduziert, aber die relative Rangkurve verbessert sich dadurch nicht. Das Ergebnis ließe sich auch durch Erzeugung von weniger Hypothesen erreichen, d.h., wenn man von vorneherein nur 10 Hypothesen/Segment erzeugt.

Ein drittes Filter F3 verwirft die folgenden Hypothesen: 50-99, 150-199, 251-350 und 450-499; durch dieses Filter wird zwar eine richtige Worthypothese verworfen, aber im Bereich unter acht Hypothesen wird die Erkennungsleistung stark verbessert. Berechnet man die durchschnittliche Verbesserung für die ganzzahligen Stützpunkte 1, 2, ..., 8 Hypothesen/Segment, ergibt sich eine durchschnittliche Verbesserung um 20 Prozentpunkte, d.h. bei gleicher Hypothesenmenge werden im Schnitt 20 Prozent mehr gesprochene Wörter erkannt. Um beispielsweise 60 Prozent der gesprochenen Wörter in der Hypothesenmenge zu haben, benötigt man nur noch vier statt sechs Hypothesen/Segment, so daß das Filter die notwendige Hypothesenmenge um 33 Prozent reduziert. Bild 6.5 zeigt die relative Rangkurve für die gefilterte und ungefilterte Hypothesenmenge, die sich für das Demonstrationsbeispiel mit dem Filter F3 ergeben würde. In den folgenden Auswertungen werden die Stützpunkte für die relativen Rangkurven anhand der 62 Äußerungen der Pragmatik-Stichprobe nach derselben Vorgehensweise erstellt.

6.4.2 Das DEL-Filter

In [NÖTH 88b] wurde die Verifikation der Worthypothesen folgendermaßen durchgeführt: Für alle Hypothesen von mehrsilbigen Wörtern (84 Prozent der Worthypothesenmenge) wurden aus der Standardaussprache und dem Zuordnungspfad der Worthypothese die Silbenkernbereiche bestimmt. Für die so bestimmten Silbenkerne wurde eine Betonungsbeschreibung mit dem in Kap.5 beschriebenen Verfahren erstellt. Es wurden verschiedene Ansätze zur prosodischen Verifikation untersucht. Die besten Ergebnisse wurden mit dem DEL-Filter erzielt, das nicht das Betonungsmuster sondern das Silbenmuster verifiziert: Das DEL-Filter verwirft alle Worthypothesen, bei denen ein Silbenkern im Zuordnungspfad der Worthypothesengenerierung durch einen Delete-Übergang realisiert wird, d.h. bei denen die Zahl der Silben in der Standardumschrift größer ist als die Zahl der hypothetisierten Silbenkerne.

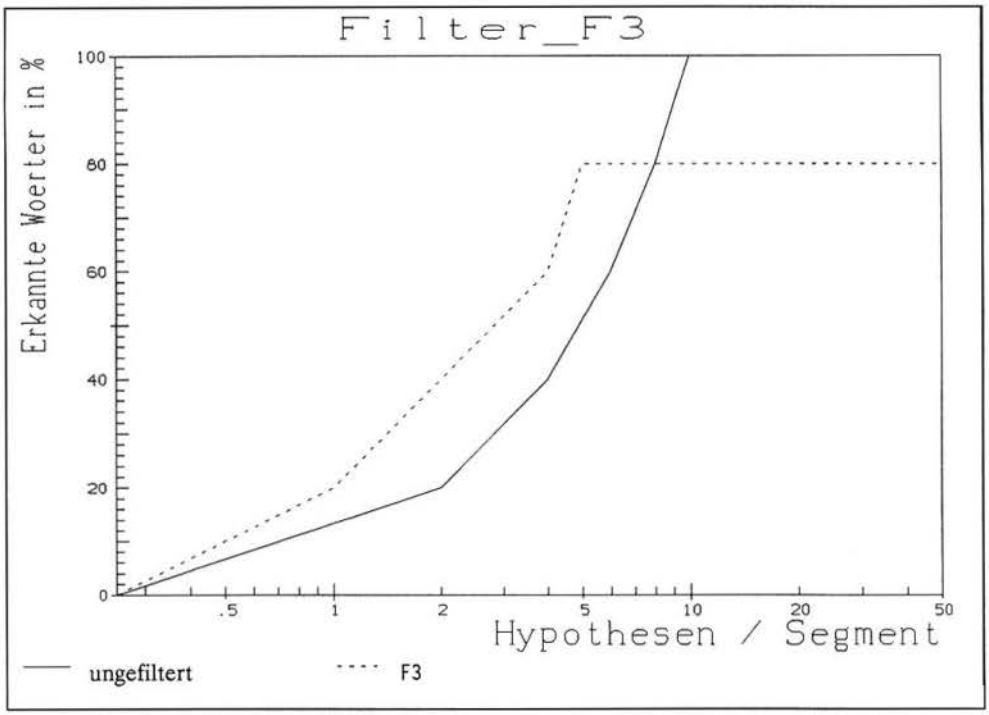


Bild 6.5: Relative Rangkurven für die ungefilterte und die mit dem F3-Filter bearbeitete Worthypothesenmenge des Demonstrationsbeispiels.

Eine Bestimmung der Silbenkerne aufgrund der spektralen Energie (Kap.4.1.2) wurde nicht durchgeführt, da die Fehler der beiden Segmentierungsverfahren Probleme aufwarfen. Trennt z.B. das Akustik-Phonetik-Modul einen Vokal auf, die Silbenkernsegmentierung aber nicht, so liegt für die Worthypothesen, die mit dem zweiten Vokal beginnen, der zur ersten Silbe der Worthypothese gehörende Silbenkern teilweise außerhalb des Worthypothesenbereichs. Neben der Verschlechterung der Silbenkerndetektion bei der Dialogstichprobe war dies der zweite Grund, eine Synchronisierung der beiden Segmentierungen anzustreben (Kap 4.1.3).

Mit dem DEL-Filter wurden in [NÖTH 88b] 31 Prozent der ursprünglichen Hypothesenmenge verworfen, so daß von den ursprünglich 50 Hypothesen/Segment noch 34 Hypothesen/Segment übrigblieben. Bis zu diesem Schnittpunkt lag die relative Rangkurve immer über der ungefilterten Kurve. Betrachtet man die ganzzahligen Stützpunkte der relativen Rangkurven, so wurde im Bereich unter 20 Hypothesen/Segment die Erkennungsrate um durchschnittlich zwei Prozentpunkte verbessert.

Die erfolgversprechendsten Versuche aus [NÖTH 88b] und [KOMPE 89a] wurden noch einmal für die neue Segmentierung durchgeführt. Bei der neuen Segmentierung findet eine stärkere Übersegmentierung statt (s.o.). Wie erwartet, wurden daher weniger Worthypothesen verworfen

(23 Prozent). Allerdings verbesserte sich im Bereich unter 20 Hypothesen/Segment das Verhältnis von richtigen und falschen verworfenen Hypothesen, so daß sich die Wirksamkeit des Filters trotz der geringeren Gesamtzahl von verworfenen Hypothesen verbesserte. Die um acht Prozentpunkte höhere "Wirksamkeit" des DEL-Filters in Bezug auf die Verwerfung von Hypothesen bei der alten Segmentierung ist so zu interpretieren, daß ein Großteil der verworfenen Hypothesen von vorneherein eine schlechte Bewertung hatte.

Bild 6.6 zeigt die relative Rangkurven für die ungefilterte Hypothesenmenge und für die mit dem DEL-Filter bearbeitete Hypothesenmenge bei Verwendung der neuen Segmentierung. Tabelle 6.12 stellt das Erkennungsverhalten des DEL-Filters für die alte und die neue Segmentierung gegenüber.

Es sollte noch einmal darauf hingewiesen werden, daß die Anzahl der verworfenen Hypothesen kein Leistungskriterium darstellt. In [KOMPE 89a] wurden 500 Hypothesen/Segment erzeugt und mit dem DEL-Filter bearbeitet. Obwohl das Ergebnis im Bereich bis zu 34 Hypothesen/Segment identisch ist, da ja die ersten 50 Hypothesen/Segment der ungefilterten Hypothesenmenge gleich sind, stieg bei der großen Hypothesenmenge die Anzahl der verworfenen Hypothesen von 31 auf 47 Prozent. Dieses Ergebnis ist insofern interessant, als das nachträgliche Verwerfen der Worthypothesen mit von der Standardaussprache abweichender Silbenstruktur (DEL-Filter) bei Einzelworterkennung mit großem Wortschatz der Lexikon-Einschränkung während der Generierungsphase entspricht (siehe z.B. [AULL 84]).

Für die Verwerfung der richtigen Worthypothesen sind zwei Hauptgründe zu nennen:

1) Fehler in der Lautsegmentierung
2) Silbenkern-Elisionen durch den Testsprecher (z.B. bei dem Wort *haben, /H.A.B.ERN./* → */H.A.M./,* siehe Kap.2.7.5), bei denen bereits die Silbenstruktur des gesprochenen Wortes nicht mit der lexikalischen Silbenstruktur übereinstimmt.

Mit zunehmender Sprechgeschwindigkeit (siehe z.B. die Ergebnisse zur Dialogstichprobe) ist zu erwarten, daß sich ohne eine Modellierung der Aussprachevarianten die Ergebnisse für das DEL-Filter stark verschlechtern, da dann zuviele richtige Hypothesen verworfen werden.

In [KOMPE 89a] wurden mehrere Experimente zur Neubewertung statt zur Verwerfung der Hypothesen durchgeführt, die das DEL-Kriterium nicht erfüllen. In einem Koordinatenabstiegsverfahren wurden verschiedene Konstanten auf die Bewertung der Hypothesen addiert, und die Hypothesenmenge wurde neu sortiert. Bis zum Schnittpunkt von 34 Hypothesen/Segment, ab dem das DEL-Filter durch die Verwerfung der richtigen Hypothesen zwangsläufig schlechter werden muß als die Neubewertung der Hypothesen, sind die Kurven für das DEL-Filter mit Verwerfung und das beste DEL-Filter mit Neubewertung praktisch identisch. Aus diesem Grund, und da die Hypothesenbewertungen nach der Neubewertung nicht mehr sinnvoll interpretiert werden können (s.o.), wurde diese Vorgehensweise (Umbewertung statt Verwerfung von Hypothesen) nicht mehr weiter verfolgt.

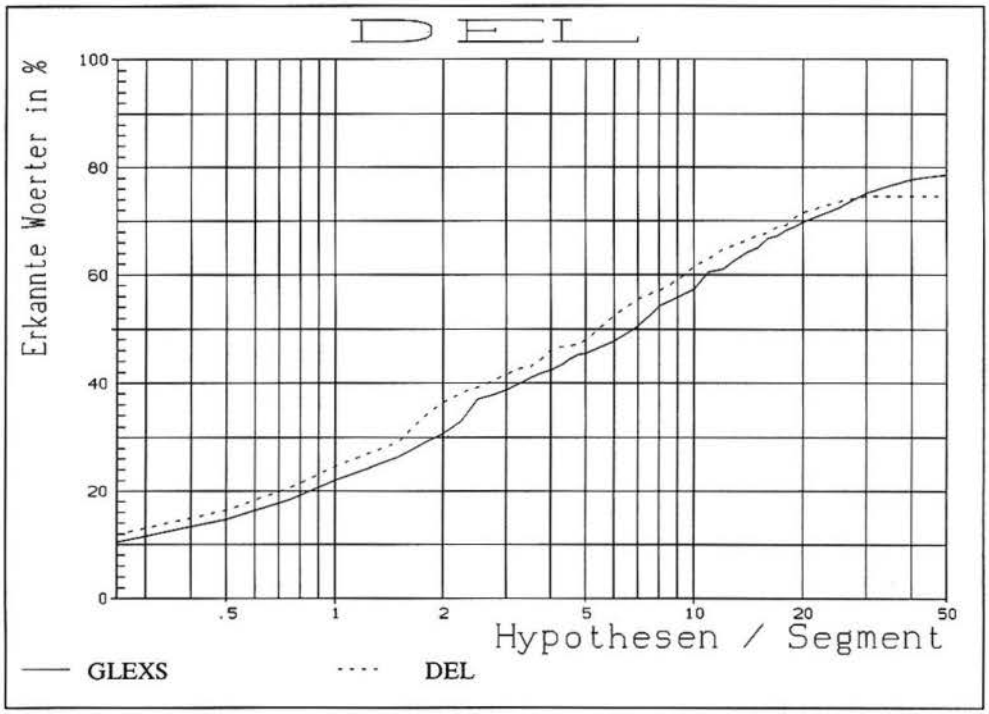


Bild 6.6: Relative Rangkurven für die ungefilterte (GLEXS) und die mit dem DEL-Filter (DEL) bearbeitete Worthypothesenmenge für die Pragmatikstichprobe bei Verwendung der neuen Lautsegmentierung.

DEL-Filter	10H/S	20H/S	50%	Schnitt-punkt	richtige verworf.	insgesamt verworfen
REGEL-Segmentierer	2	2	11	34	4	31
neuer Segmentierer	4	3	20	30	4	23

Tab. 6.12: Erkennungsverhalten des DEL-Filters für Worthypothesen, die mit dem Segmentierer nach [REGEL 88] und dem in Kap.4.1.3 beschriebenen Segmentierer erzeugt wurden. Die Spalten haben folgende Bedeutung:

10H/S	Durchschnittliche Verbesserung der Erkennungsraten unter 10 Hypothesen/Segment
20H/S	Durchschnittliche Verbesserung der Erkennungsraten unter 20 Hypothesen/Segment
50%	Reduktion der ursprünglichen Worthypothesenmenge in Prozent, um 50 Prozent der gesprochenen Wörter zu erhalten
Schnittpunkt	Kleinste Hypothesenmenge (in Hypothesen/Segment), bei der die Erkennungsrate nach der Bearbeitung durch das DEL-Filter schlechter ist als die der ungefilterten Hypothesenmenge
richtige verworf.	Anzahl der insgesamt verworfenen richtigen Hypothesen in Prozent, gemessen an der Zahl aller gesprochenen Wörter (354)
insgesamt verworfen	Anzahl verworfener Hypothesen in Prozent, gemessen an der Ausgangsmenge von 50 Hypothesen/Segment (ca. 85000 Hypothesen für die 62 Äußerungen).

Eine Verifikation des lexikalischen Betonungsmusters mit der automatisch erzeugten Betonungsbeschreibung erbrachte für die Hypothesenmenge, die nach der DEL-Filterung übrigblieb, keine Verbesserung des Ergebnisses. Folgende Gründe waren für die Verschlechterung verantwortlich:

1) Fehler bei der Segmentierung (Kap.6.3) und bei der Grundfrequenzbestimmung (Kap.6.1), die sich auf die Bewertung der prosodischen Eigenschaften *Lautdauer, Lautstärke (es wird das Energie-Integral berechnet)* und *Tonhöhe* auswirken
2) Intrinsisch hohe Energie-Werte für den systematisch unbetonten Zentralvokal (siehe Kap.2.7.4)
3) Nicht realisierter Wortakzent (s. Kap.2.3)
4) Falsch realisierter Wortakzent durch den Testsprecher.

6.5 Einschränkung des Lexikons an betonten Stellen

Im letzten Abschnitt wurde mit dem DEL-Filter ein Hypothesen-Filter vorgestellt, das die Übereinstimmung einer Worthypothese mit dem Zeitsignal unter prosodischen Gesichtspunkten überprüft. Die im folgenden beschriebenen PRAG-Filter verwerfen Worthypothesen nicht aufgrund fehlender Übereinstimmung mit dem Sprachsignal, sondern aufgrund ihrer Zugehörigkeit zu gewissen Wortklassen. Voruntersuchungen zu den hier präsentierten Ergebnissen finden sich in [NÖTH 88b, 89a], [KOMPE 89a]. Ebenso wie für das DEL-Filter beziehen sich die Ergebnisse der Voruntersuchungen auf den Segmentierer nach [REGEL 88]. Die wichtigsten Experimente aus [NÖTH 89a] wurden mit dem neuen Segmentierer noch einmal durchgeführt. Die Ausgangshypothesenmenge stimmt mit der aus dem letzten Abschnitt überein.

Geht man davon aus, daß der Benutzer in einem Mensch-Maschine-Dialog ebenso wie in einem Mensch-Mensch-Dialog (siehe Kap.3.2.3) die für seine Anfrage wichtigen Wörter betont, so können an betonten Stellen Wortklassen ausgeschlossen werden. Dies gilt trotz guter Übereinstimmung mit dem Sprachsignal. So wird z.B. bei der Äußerung: *"Ich möchte am Wochenende nach **Mainz** fahren."* das Wort *ein* über dem betonten (und damit deutlich artikulierten) Wort *Mainz* sehr gut passen und damit als Worthypothese mit sehr hoher Bewertung vorkommen. Wird jedoch dieser Bereich als stark betont markiert, so kann diese falsche Worthypothese verworfen werden, da es sich um einen unbestimmten Artikel und damit um ein (in dem betrachteten Kontext und von Ausnahmen abgesehen) *erwartungsgemäß nicht betonbares* Wort handelt. Erwartungsgemäß nicht betonbar sind Wörter bestimmter syntaktischer Klassen (Funktionswörter wie z.B. Artikel) und solche Wörter, die keinen Pragmatik-Eintrag besitzen, d.h. alle Wörter, die im Anwendungsbereich des Systems keine wichtige Funktion einnehmen können ([EHRLICH 89a]). So kann z.B. davon ausgegangen werden, daß im Anwendungsbereich von EVAR in dem Satz *"Ich möchte mit meiner Schwester nach Würzburg fahren."* das Wort *"Schwester"* nicht betont wird. *Erwartungsgemäß nicht betonbar* soll aber nicht heißen, daß in Ausnahmefällen oder in anderen Dialogsituationen diese Wörter nicht doch

betont sein können. Ein Beispiel wäre das Funktionswort *"auch"* in *"Dürfen wir in München auch den nächsten Zug nehmen?"*. Allerdings ist das Verwerfen dieser Wörter trotzdem akzeptabel, da, wie im letzten Abschnitt bereits erläutert, von einem Hypothesen-Filter nicht gefordert wird, *keine richtigen Worthypothesen* zu verwerfen, sondern *möglichst viele falsche* und *möglichst wenige richtige.* Die eine Aufgabe erfordert ein möglichst rigoroses Verwerfen und die andere ein möglichst vorsichtiges, weshalb ein Verifikations-Filter zwangsläufig einen Kompromiß darstellt. Für die PRAG-Filter gilt zusätzlich, daß die pragmatische Analyse u.U. trotz eines verworfenen richtigen Wortes durchgeführt werden kann, da keine pragmatisch relevanten Wörter verworfen werden.

Die Informationen über syntaktische Klassen und Pragmatik-Einträge sind für jedes Wort im Lexikon kodiert (siehe [EHRLICH 89a] und Kap.1.3). Tabelle 6.13 zeigt die Wortarten, die im Rahmen der Anwendung als *potentiell betonbar* und *erwartungsgemäß nicht betonbar* eingestuft wurden.

potentiell betonbar		erwartungsgemäß nicht betonbar	
Adjektiv	mit Pragmatik-eintrag	Adjektiv	ohne Pragmatik-eintrag
Adverb		Adverb	
Nomen		Nomen	
Nomen_proprium		Nomen_proprium	
unflektiertes_Adjektiv		unflektiertes_Adjektiv	
Verb		Verb	
Floskel		Determinans	
Frageadverb		Hilfsverb_"haben"	
Fragedeterminans		Hilfsverb_"sein"	
Fragepronomen		Hilfsverb_"werden"	
Konjunktion		Infinitivpartikel	
Koordination		Infinitivsubjunktion	
Modalverb		Praeposition	
Negationspartikel		Praepraeposition	
Ordinalzahl		Pronomen	
Satzwort		Reflexivpronomen	
Subjunktion		Relativpronomen	
Verbpraefix		Vergleichspartikel	
Zahlwort			

Tab. 6.13: Wortarten nach [EHRLICH 86], die im Rahmen der Anwendung als *potentiell betonbar* und *erwartungsgemäß nicht betonbar* eingestuft wurden.

In [NÖTH 88b] wurde gezeigt, daß die pragmatisch wichtigen Wörter (Pragmatik-Wörter, siehe Kap.3.1 und Anhang D) wesentlich besser erkannt werden als die restlichen Wörter. Im Bereich unter 10 Hypothesen/Segment ist die Erkennungsrate für die Pragmatik-Wörter um 14 Prozentpunkte höher als die für alle Wörter, im Bereich unter 20 Hypothesen/Segment um elf Prozentpunkte. Geht man von einem Prosodie-Modul aus, das für die Pragmatik-Wörter (24 Prozent aller Wörter) die Position der Wortakzentsilbe fehlerfrei findet, und an diesen Stellen alle Worthypothesen von *erwartungsgemäß nicht betonbaren* Wörtern verwirft, so verbessert sich die Gesamterkennungsrate im Bereich unter 20 Hypothesen/Segment um vier Prozentpunkte (siehe die Zeile MPRAG in Tab. 6.14). Zwar handelt es sich bei den Pragmatik-Wörtern nicht unbedingt um die prominentesten Stellen des Sprachsignals, aber in Kap.3.2.3 wurde ein hinreichender Zusammenhang zwischen Pragmatik-Wörtern und betonten Stellen des Signals festgestellt.

Es wurde nun versucht, an den automatisch als betont eingestuften Stellen des Signals die *erwartungsgemäß nicht betonbaren* Wörter zu verwerfen: Mit dem in Kap.4.1.2 beschriebenen Verfahren (Silbenkern-Detektion aufgrund der spektralen Energie) wurden Silbenkerne detektiert. Für diese Silbenkern-Segmentierung wurde mit dem in Kap.5 beschriebenen Verfahren eine Betonungsbeschreibung erstellt. Heuristisch wurden zwei Kriterien festgelegt, nach denen eine Silbe als betont eingestuft wird: Nach A1 ist eine Silbe betont, falls die Gesamtbetonung U_{BET} gleich 1 ist. Nach A2 müssen zwei der Einzelbewertungen U_{MA}, U_{TA} und U_{DA} gleich 1 und die dritte größer als 0.5 sein.

Anhebung von Lautstärke und Tonhöhe werden zur Betonung von Silben eingesetzt. Aus diesem Grund kann man davon ausgehen, daß es sich bei den größten lokalen Maxima im Energie- und Grundfrequenzverlauf einer Äußerung um betonte Stellen handelt. Um eine möglicherweise fehlerhafte Silbenkerndetektion zu vermeiden, wurden pro Äußerung die zwei stärksten lokalen Energie- (E) und Grundfrequenzmaxima (GF) bestimmt.

An den so markierten Stellen wurde das Lexikon auf die potentiell betonbaren Wörter eingeschränkt. Dies wurde durch das nachträgliche Herausfiltern der Worthypothesen für die erwartungsgemäß nicht betonbaren Wörter realisiert. Das Filter wurde jeweils an den automatisch als prominent markierten Stellen A1, A2, E und GF angewandt. Die Ergebnisse aus [NÖTH 89a] sind in Tabelle 6.14 zusammengestellt. Die Zeile DEL+A1 zeigt das beste Ergebnis, das in [NÖTH 89a] für die Kombination von DEL-Filter und PRAG-Filter erzielt wurde. Die Zeile MPRAG zeigt das Ergebnis bei fehlerfreier Bestimmung der Pragmatik-Wörter.

Bezüglich der Erkennungsraten brachte A1 unter den verschiedenen PRAG-Filtern die größte Verbesserung. Durch das Filtern an den Stellen A2 wurden bedeutend weniger richtige Hypothesen verworfen als bei A1, und die Verbesserung der Erkennungsraten war ebenfalls hoch. Das Filtern an den Stellen E und GF brachte eine geringere Erhöhung der Erkennungsraten und einen wesentlich früheren Schnittpunkt. Der Anteil der verworfenen Hypothesen war, im Gegensatz zum DEL-Filter, von der Größe der Hypothesenmenge nahezu unabhängig.

PRAG-Filter	20H/S	Schnitt-punkt	richtige verworf.	insgesamt verworfen
MPRAG	4	50	0	19
A1	3	34	4	29
A2	2	40	1	16
E	1	17	3	16
GF	1	17	3	15
DEL+A1	4	30	7	49

Tab. 6.14: Erkennungsverhalten der PRAG-Filter A1, A2, E, GF für Worthypothesen, die mit dem Segmentierer nach [REGEL 88] erzeugt wurden (aus [NÖTH 89a]). In der Zeile MPRAG wurden die Pragmatik-Stellen von Hand bestimmt, in der Zeile DEL+A1 wurde das PRAG-Filter A1 auf die Worthypothesenmenge angewandt, die nach der DEL-Filterung übrigblieb. Die Spalten haben folgende Bedeutung:

20H/S	Durchschnittliche Verbesserung der Erkennungsraten unter 20 Hypothesen/Segment
Schnittpunkt	Kleinste Hypothesenmenge (in Hypothesen/Segment), bei der die Erkennungsrate nach der Bearbeitung durch das DEL-Filter schlechter ist als die der ungefilterten Hypothesenmenge
richtige verworf.	Anzahl der insgesamt verworfenen richtigen Hypothesen in Prozent, gemessen an der Zahl aller gesprochenen Wörter (354)
insgesamt verworfen	Anzahl verworfener Hypothesen in Prozent, gemessen an der Ausgangsmenge von 50 Hypothesen/Segment (ca. 85000 Hypothesen für die 62 Äußerungen).

Die besten PRAG-Filter A1, A2 und DEL+A1 wurden auch auf die Hypothesen für die neue Segmentierung angewendet (gleiche Hypothesenmenge wie im letzten Abschnitt). Tabelle 6.15 zeigt das Erkennungsverhalten der beiden PRAG-Filter A1 und A2 für die ungefilterte und die DEL-gefilterte Worthypothesenmenge.

Im Bereich unter 10 Hypothesen/Segment war die Kombination DEL+A1 etwas besser als A1 oder DEL alleine, aber der Abstand war deutlich geringer als bei den mit der Segmentierung nach [REGEL 88] erzeugten Worthypothesen. Tabelle 6.16 zeigt die Erkennungsraten in Prozent aller tatsächlich gesprochenen Wörter für die folgenden Worthypothesenmengen: Die ungefilterte Hypothesen für die Segmentierung nach [REGEL 88] und die neue Segmentierung sowie die PRAG- und DEL-gefilterten Hypothesen für die neue Segmentierung. Die Bilder 6.7 bis 6.9 zeigen die relativen Rangkurven für die Filter A1, A2 und DEL+A1 für die neue Segmentierung im Vergleich zu der ungefilterten Hypothesenmenge für die neue Segmentierung und die Segmentierung nach [REGEL 88].

Filter	10H/S	20H/S	50%	Schnitt-punkt	richtige verworf.	insgesamt verworfen
A1	4	3	20	30	4	21
A2	3	2	15	35	2	12
DEL+A1	5	3	27	20	8	39

Tab. 6.15: Erkennungsverhalten der PRAG-Filter A1 und A2 für Worthypothesen, die mit dem in Kap.4.1.3 beschriebenen Segmentierer erzeugt wurden. In der Zeile DEL+A1 wurde das PRAG-Filter A1 auf die Menge der Worthypothesen angewandt, die nach der Filterung mit dem DEL-Filter übrigblieb. Die Spalten haben folgende Bedeutung:

10H/S	Durchschnittliche Verbesserung der Erkennungsraten unter 10 Hypothesen/Segment
20H/S	Durchschnittliche Verbesserung der Erkennungsraten unter 20 Hypothesen/Segment
50%	Reduktion der ursprünglichen Worthypothesenmenge in Prozent, um 50 Prozent der gesprochenen Wörter zu erhalten
Schnittpunkt	Kleinste Hypothesenmenge (in Hypothesen/Segment), bei der die Erkennungsrate nach der Bearbeitung durch das PRAG-Filter schlechter ist als die der ungefilterten Hypothesenmenge
richtige verworf.	Anzahl der insgesamt verworfenen richtigen Hypothesen in Prozent, gemessen an der Zahl aller gesprochenen Wörter (354)
insgesamt verworfen	Anzahl verworfener Hypothesen in Prozent, gemessen an der Ausgangsmenge von 50 Hypothesen/Segment (ca. 85000 Hypothesen für die 62 Äußerungen).

	Relativer Rang Prozent erkannte Wörter bei Erzeugung von								
Hypothesen/Segment	1	2	3	4	5	7	10	20	50
REGEL-Segmentierer	17	26	32	36	41	46	54	64	77
neuer Segmentierer	22	31	39	42	45	51	57	70	78
PRAG-Filter A1	26	37	41	45	48	55	61	71	74
PRAG-Filter A2	25	36	41	45	47	54	62	71	76
DEL-Filter	25	36	42	46	48	56	62	71	75
DEL+A1	27	37	43	46	50	56	61	69	70

Tab. 6.16: Erkennungsraten für die ungefilterten und die PRAG- und DEL-gefilterten Hypothesen bei unterschiedlich großen Hypothesenmengen. Zum Vergleich ist auch die Erkennungsrate für die ungefilterten Hypothesen, die mit dem Segmentierer nach [REGEL 88] erzeugt wurden, mit angegeben.

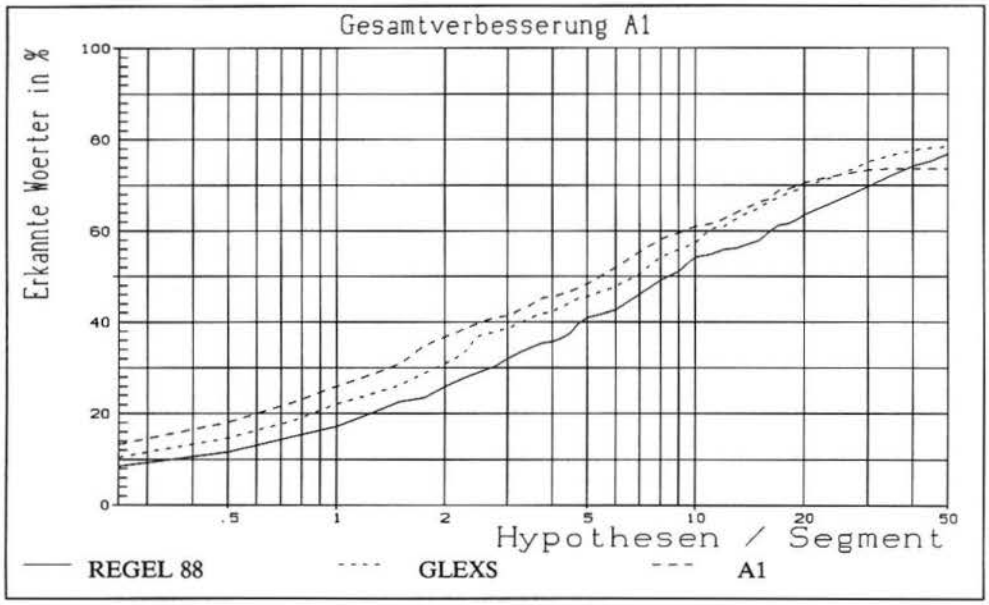


Bild 6.7: Relative Rangkurven für die ungefilterte (GLEXS) und die mit dem A1-Filter (A1) bearbeitete Worthypothesenmenge für die Pragmatik-Stichprobe bei Verwendung der neuen Lautsegmentierung sowie für die ungefilterte Hypothesenmenge bei Verwendung des Segmentierers nach [REGEL 88].

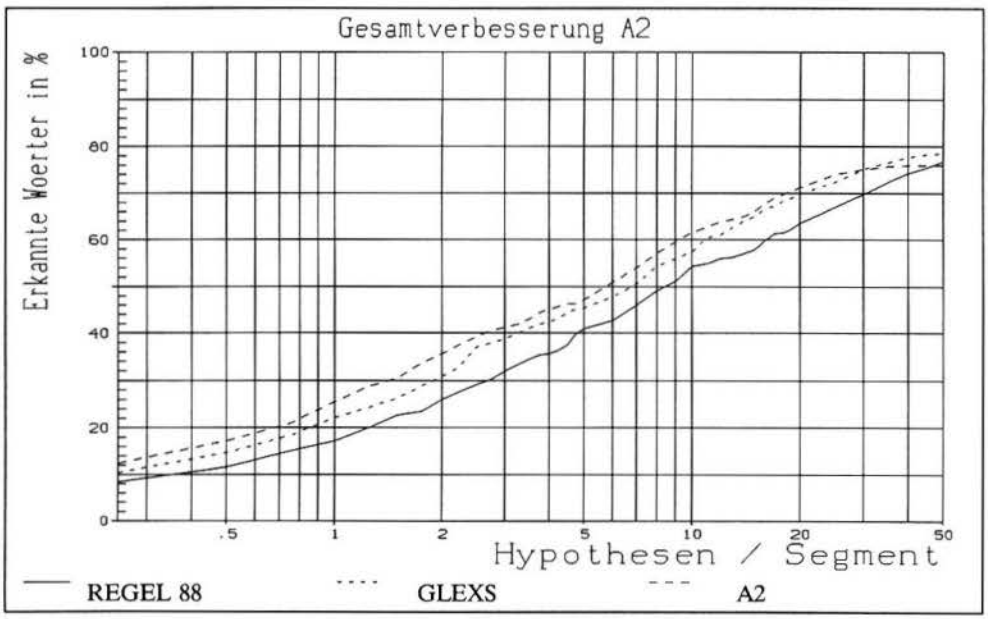


Bild 6.8: Relative Rangkurven für die ungefilterte (GLEXS) und die mit dem A2-Filter (A2) bearbeitete Worthypothesenmenge für die Pragmatik-Stichprobe bei Verwendung der neuen Lautsegmentierung sowie für die ungefilterte Hypothesenmenge bei Verwendung des Segmentierers nach [REGEL 88].

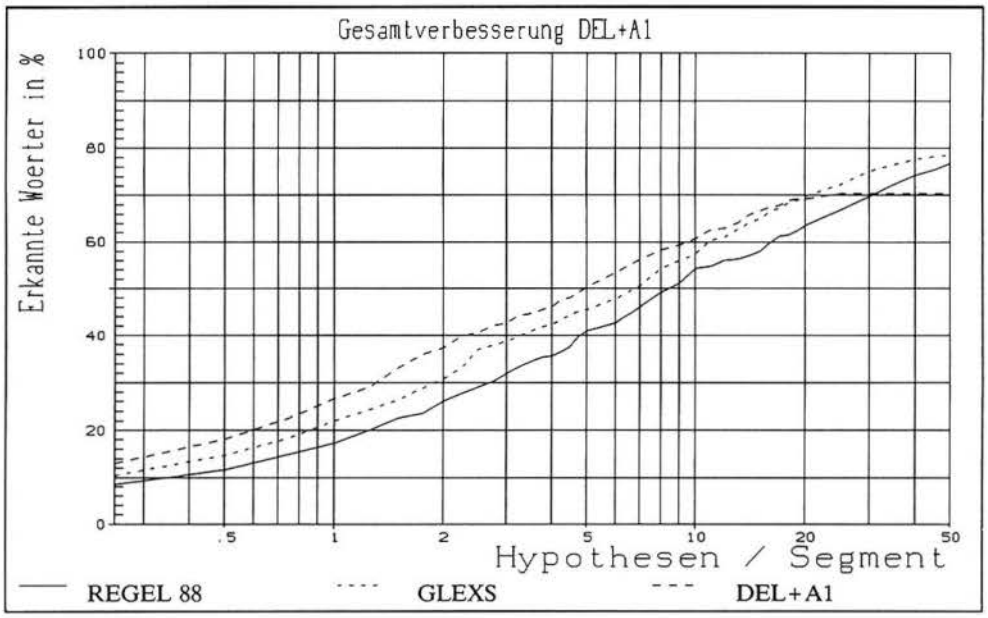


Bild 6.9: Relative Rangkurven für die ungefilterte (GLEXS) und die mit dem DEL+A1-Filter (DEL+A1) bearbeitete Worthypothesenmenge für die Pragmatik-Stichprobe bei Verwendung der neuen Lautsegmentierung sowie für die ungefilterte Hypothesenmenge bei Verwendung des Segmentierers nach [REGEL 88].

6.6 Verbesserung der Worterkennung durch die neue Lautsegmentierung - Ein Zwischenbericht

In diesem Abschnitt wird auf das Verfahren zur Lautsegmentierung eingegangen (Kap.4.1.3), das im Rahmen der Arbeiten am Prosodie-Modul und am Worterkennungs-Modul zur Zeit entwickelt wird. Für die Verbesserung der automatischen Betonungsbeschreibung waren die Arbeiten zur Lautsegmentierung aus den folgenden Gründen sehr wichtig:

1) Fehler bei der Silbenkerndetektion wirken sich auf die Betonungsbeschreibung aus, da fehlerhafte Dauer- und Lautheitsmerkmale berechnet werden. Häufigster Segmentierungsfehler bei der Silbenkerndetektion ist die Verschmelzung von Silbenkernen bei den Lautfolgen *Vokal - Vokal* und *Vokal - sonoranter Konsonant - Vokal.* Insbesondere bei der Verwendung stark reduzierter Ausspracheformen (siehe die Ergebnisse zur Dialog-Stichprobe, Kap.6.3) war ein starker Anstieg der Fehlsegmentierungen zu beobachten. Durch den Vergleich der Silbenkernsegmentierung aufgrund spektraler Energie mit einer auf der Lautkomponenten-Klassifikation basierenden Segmentierung schien es möglich, einen Teil der fehlerhaften Silbenkernverschmelzungen zu korrigieren.
2) Für die prosodischen Wortfilter zur Verifikation der Silbenstruktur und des lexikalischen Akzents (Kap.6.4) sowie zur Einschränkung des Lexikons an betonten Stellen (Kap.6.5) sollten die Silbenkerngrenzen mit Segmentgrenzen zusammenfallen.

3) Die Experimente zur Verifikation der Silbenstruktur für die Worthypothesen, die mit der Lautsegmentierung nach [REGEL 88] erzeugt wurden ([NÖTH 88b] und Kap.6.3), zeigten, daß bei einer sehr großen Zahl von Hypothesen (31 Prozent bei 50 Hypothesen/Segment) ein Silbenkern der Standardaussprache durch einen Delete-Übergang modelliert wurde. Es erschien sinnvoll, das Wissen über die höhere Prominenz von Silbenkernen gegenüber Silbenrändern bereits während der Generierung der Worthypothesen zu verwenden.
4) Es war zu untersuchen, inwieweit die Silbenkerngrenzen für die Lautsegmentierung als Ankerpunkte dienen können.

Um das Verhalten des Segmentierers besser beurteilen zu können, wurde die Silbenkernsegmentierung zunächst nicht verwendet, bei den Entwurfs-Entscheidungen wurde allerdings immer darauf geachtet, daß die Verwendung und Korrektur der Silbenkernsegmentierung möglich sein muß. Als Beurteilungskriterium diente immer die relative Rangkurve für die Worthypothesengenerierung. Da die Experimente in [KUNZMANN 90] ausführlich beschrieben sind, wird hier nur auf die wichtigsten Ergebnisse eingegangen. Für die Experimente wurde das Lexikon KLEXS (549 Wörter) verwendet. Die Äußerungen von sechs der zwölf Sprecher aus der EVAR-Stichprobe wurden für das Training verwendet, die restlichen Äußerungen zum Testen.

In einem ersten Schritt wurden verschiedene Ansätze zur Lautklassifikation untersucht. Bei Verwendung der Segmentgrenzen des Segmentierers nach [REGEL 88] wurden verschieden große Lautinventare (14 bis 39 unterscheidbare Lautklassen) über jedem Segment mit HMM verifiziert. Bei vergleichbarer Erkennungsrate auf der Lautebene wurde eine deutlich bessere Worterkennungsrate erzielt. Die besten Ergebnisse wurden mit 36 Lautklassen erzielt. Die Erkennungsrate auf Lautebene war mit der des Segmentieres nach [REGEL 88] vergleichbar (bei diesem Segmentierer werden ebenfalls 36 Klassen unterschieden), aber die relative Rangkurve der Worterkennung für die neue Lautklassifikation war immer besser als die für die alte Segmentierung (im Schnitt ca. sechs Prozentpunkte). Die zwei notwendigen Schritte bei der Generierung von Lauthypothesen, die *Segmentierung* des Signals und die *Klassifikation* der Segmente, konnten somit getrennt verbessert werden. Die Vorgehensweise bei der Klassifikation erlaubt außerdem die Verwendung anderer Wortuntereinheiten (z.B. Diphone, Halbsilben oder Silben), ohne daß die Algorithmen geändert werden müssen.

Um die Zahl der nicht gefundenen Laute zu verringern, wurde eine möglichst starke Übersegmentierung angestrebt, also eine Zerlegung des Sprachsignals in Einheiten, die im Mittel deutlich kürzer als Laute sind. Diese Initialzerlegung sollte dann in einem zweiten Schritt wieder zu einer Segmentierung nach Lauten zusammengefaßt werden. Realisiert wurde dies durch die Erzeugung eines Segmentgraphen: Jede Grenze der Initialzerlegung wurde als potentieller Anfangspunkt eines Lautes betrachtet. In einem gewissen, zeitlich folgenden Bereich wurden potentielle Endpunkte gesucht. Die Segmentierung nach Lauten ergab sich somit als optimaler Pfad im Segmentgraphen. Als geeignet erwies es sich, die Frames für die Initialzerlegung zu verwenden, bei denen die Entscheidung des Lautkomponentenklassifikators umschlägt (im folgenden wird diese

Zerlegung als Lauflängenzerlegung bezeichnet). Über die Zahl der potentiellen Anfangspunkte in der Initialzerlegung und über die Länge des Suchbereichs für die zugehörigen Endpunkte kann Einfluß auf die Segmentierung genommen werden. Beide Parameter wurden systematisch variiert:

1) Eine feinere Initialzerlegung als die Lauflängenzerlegung erhält man, wenn man jeden Frame als Anfangspunkt einer Initialzerlegung verwendet; gröbere Zerlegungen erhält man, wenn man die Lauflängenzerlegung verschieden stark glättet.
2) Je kürzer der Bereich ist, in dem für jeden potentiellen Anfangspunkt nach zugehörigen Endpunkten gesucht wird, um so stärker ist die Übersegmentierung. Der Suchraum für jeden Anfangspunkt wurde zwischen drei und 25 Frames variiert. Als potentielle Endpunkte wurden alle Frames oder der jeweils letzte Frame einer Lauflänge betrachtet (falls die Lauflänge länger war als der zu untersuchende zeitlich folgende Bereich, wurde das Ende dieser Lauflänge als einziger potentieller Endpunkt betrachtet).

Die Ergebnisse lassen sich wie folgt zusammenfassen:

1) Das Wissen über stabile Bereiche (Betrachtung der Lauflängen) führte zu einer starken Verbesserung gegenüber der Version, bei der alle Frames als potentielle Anfangspunkte zugelassen waren.
2) Eine sehr vorsichtige Glättung (Lauflängen der Länge eins mit gleichem linken und rechten Nachbarn werden eliminiert) führte zu einer Verbesserung der Ergebnisse, eine weitere Glättung zu einer Verschlechterung.
3) Die besten Ergebnisse wurden mit einer Beschränkung des zeitlich folgenden Bereichs auf fünf Frames erzielt. Die Länge des Suchbereichs (64 Millisekunden) lag somit unter der durchschnittlichen Lautlänge in der EVAR-Stichprobe (83 Millisekunden). Dies führte zu der intendierten leichten Übersegmentierung: pro gesprochenem Laut wurden im Schnitt 1.3 Lauthypothesen erzeugt.

Zusammen mit der in [KUNZMANN 90, Kap.6.3] beschriebenen neuen Trainingsmethode führte die durch die Punkte 1) - 3) charakterisierte Lautsegmentierung zu der in Bild 6.10 dargestellten Verbesserung der Worterkennungsleistung.

In [FISCHER 89] wurden Vorversuche zur Einbeziehung der Silbenkernsegmentierung für eine Stichprobe von 26 Sätzen durchgeführt. Dabei wurden die Silbenkerne als Ankerpunkte im Segmentgraphen verwendet, d.h. Segmentierungsalternativen wurden nur zwischen den Silbenkernen zugelassen. Zur Modellierung von Silbenkernverschmelzungen wurden auch Experimente durchgeführt, bei denen nur Silbenkerne bis zu einer gewissen Länge als Ankerpunkte verwendet wurden. Ziel dieser Voruntersuchungen war es, Erfahrungen über den Zusammenhang zwischen Reduktion des Rechenaufwands und Verringerung der Erkennungsleistung bei verschieden starken Einschränkungen des Segmentgraphen zu gewinnen. Eine Verschlechterung der Erkennungsleistung war zu erwarten, da die Silbenkernverschmelzungen der intendierten Übersegmentierung entgegenwirken. Zwar konnte bei der Betrachtung aller Silbenkerne der Rechenzeitaufwand um rund 50 Prozent gesenkt werden, aber dies führte zu einer starken Verschlechterung der

Erkennungsleistung für die Worterkennung. Die besten Erkennungsraten wurden erzielt, wenn nur Silbenkerne bis zu einer maximalen Länge von fünf Frames betrachtet wurden. Die relative Rangkurve lag im Durchschnitt um ca. zwei Prozentpunkte unter der Kurve für die Lautsegmentierung ohne Einbezug der Silbenkernsegmentierung. Der Rechenzeitaufwand sank um ca. 15 Prozent.

Die Ergebnisse der Voruntersuchung lassen erwarten, daß die Synchronisation der Laut- und der Silbenkernsegmentierung nach dem in Kap.4.1.3 beschriebenen Verfahren ohne Verlust der Erkennungsleistung für die Worterkennung möglich ist. Das Verfahren zur Synchronisation ist implementiert; Experimente zur Worterkennung und zur Verbesserung der Betonungsbeschreibung stehen noch aus.

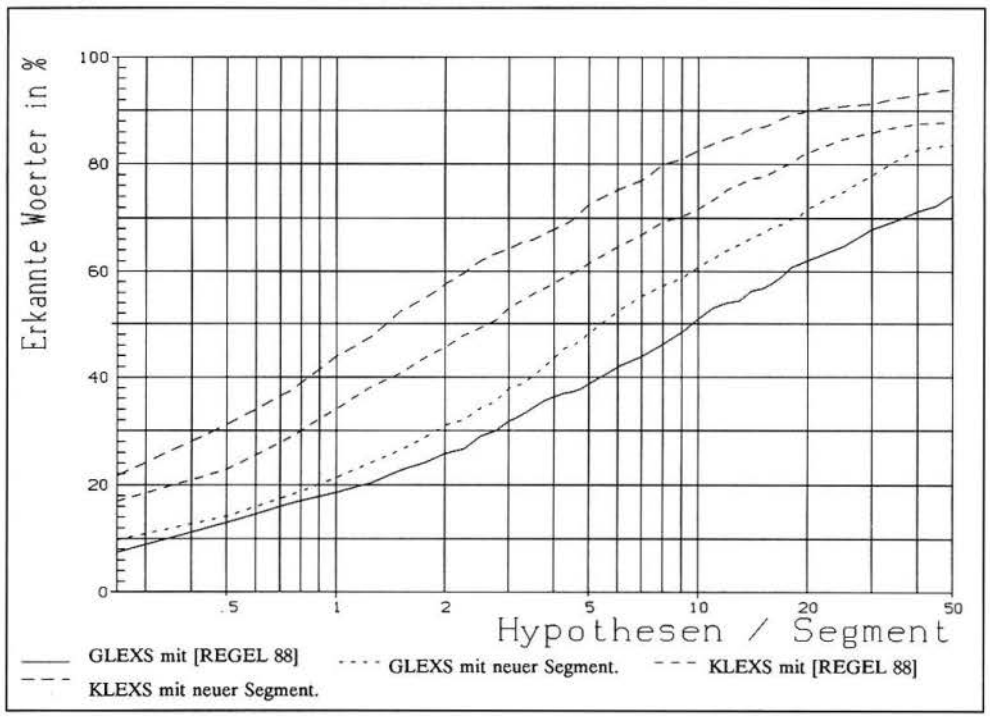


Bild 6.10: Momentan beste relative Rangkurven für die Lexika KLEXS und GLEXS bei der neuen Segmentierung im Vergleich zu den Rangkurven bei der Lautsegmentierung nach [REGEL 88]. Die Kurven beziehen sich auf die Äußerungen von sechs Sprechern der EVAR-Stichprobe.

7 Grundlegende Voruntersuchungen zur prosodischen Verifikation von Satzhypothesen - Ein Ausblick

In diesem Kapitel soll noch einmal - unter dem Aspekt der Verwendung in einem ASE-System - auf die Untersuchungen eingegangen werden, über die in Kap.2.10 berichtet wurde. Ebenso wie bei den Untersuchungen zur Bestimmung des Satzmodus mit intonatorischen Mitteln (Kap.6.2) wurden die Untersuchungen in enger Kooperation mit den Mitarbeitern des DFG-Projekts Modus-Fokus-Intonation durchgeführt. Es handelt sich dabei um Untersuchungen zu einem *Modell der intonatorischen Markierung von Modus und Fokus im Deutschen.* Bedingt durch die Komplexität des Untersuchungsgegenstandes *deutsche Sprache* und die Überlagerungen der verschiedenen Rollen der Intonation (siehe Kap. 2) bedingt, kann es sich nur um ein sehr *vorläufiges* und *bruchstückhaftes* Modell handeln.

Der Begriff *Fokus* bezeichnet die *wichtigste* Information einer Äußerung. Akustisch wird der Fokus durch die Position des Satzakzents realisiert (siehe Kap.2.4). Die Akzentuierung geschieht mit prosodischen Eigenschaften. Im folgenden wird dies mit dem Begriff *Markierung des Fokus* bezeichnet.

Eine exakte Begriffsbildung (Was ist ein *Modell*?) und eine Darstellung der verschiedenen Wissensrepräsentationsschemata würde den Rahmen dieser Arbeit bei weitem sprengen. Einführungen in dieses Gebiet unter dem Aspekt der Repräsentation von Wissen im Computer finden sich z.B. in [NIEMANN 81, Kap.7] und [SAGERER 89]. Im Rahmen dieses Abschnitts wird eine maschinelle Darstellung des Intonationsmodells mit *semantischen Netzen* konzipiert. Soweit nicht hier definiert, werden Definitionen und Begriffe im Sinne von [SAGERER 88, 89] verwendet.

7.1 Motivation

Fast alle Fehler bei der automatischen Betonungsbeschreibung (Kap.6.3), der prosodischen Wortakzentverifikation (Kap.6.4) und der prosodischen Lexikoneinschränkung an Satzakzentstellen (Kap.6.5) sind auf

1) fehlerhafte Segmentierung
2) fehlerhafte Grundfrequenzwerte
3) intrinsische Eigenschaften der Vokalklassen (insbesondere bei den Energie-Parametern)
4) Betrachtung eines zu geringen Kontexts

zurückzuführen. Nach der Entwicklung des Algorithmus zur datengetriebenen automatischen Betonungsbeschreibung wurde - neben Untersuchungen zur Verbesserung der Worterkennung aufgrund dieser Betonungsbeschreibung (Kap.6.4 und 6.5) - mit mehreren Ansätzen zur Reduktion der Fehlerpunkte 1) - 3) begonnen:

1) Für den in Kap.4.1.3 beschriebenen Ansatz zur Verbesserung der Silbenkern-Detektion ist eine Verringerung der fehlerhaften Silbenkernverschmelzungen zu erwarten.

2) Mit dem im Rahmen der Arbeiten am Prosodie-Modul entwickelten Grundfrequenz-Algorithmus konnte die Zahl der Äußerungen, bei denen mindestens ein Grobfehler zu beobachten war, um den Faktor 2-3 reduziert werden. Eine weitere Verbesserung ist von einer genaueren Suche nach laryngalisierten Stellen zu erwarten. Vorversuche mit dem in [REETZ 89] beschriebenen periodensynchronen Verfahren zeigten gute Ergebnisse und lassen eine Verbesserung der Zielwertbestimmung und somit eine Verringerung der Grobfehler möglich erscheinen[1].
3) Am Institut für Phonetik und Sprachliche Kommunikation der Ludwig-Maximilians-Universität München werden zur Zeit Perzeptionsexperimente mit Vokalpaaren durchgeführt. Aufgrund dieser Experimente sind Untersuchungen zur Normierung der Energie-Merkmale geplant.

Diese Ansätze zur Verbesserung der existierenden Betonungsbeschreibung können als Verfeinerungen des bestehenden Ansatzes betrachtet werden. Die unter Punkt 4) genannte Fehlerquelle erfordert ein prinzipiell anderes Vorgehen. Es genügt nicht, als größeren Kontext mehr benachbarte Silbenkerne zu betrachten, sondern man muß alles bereits vorhandene Wissen über das Gesprochene verwenden: Sollen während der linguistischen Analyse in einem sprachverstehenden System Satzhypothesen prosodisch verifiziert werden, so liegt i.allg. bereits sehr viel mehr Wissen über das Gesprochene vor als bei der Erstellung einer datengetriebenen Betonungsbeschreibung. Dieses Wissen über die akustisch-phonetische Realisierung und über grammatische Merkmale der gesprochenen Äußerung sollte für die prosodische Verifikation bereit stehen.

Anhand eines Beispiels soll verdeutlicht werden, welche Art von Wissen u.U. bereits zur Verfügung steht: Während eines Auskunftsdialogs aus dem Anwendungsgebiet von EVAR generiert das Dialog-System eine Auskunft, und der Benutzer reagiert mit einer Zusatzfrage. Aufgrund der akustisch-phonetischen, syntaktischen, semantischen und pragmatischen Analyse stehen zwei akustisch sehr ähnliche Satzhypothesen zur Auswahl, die prosodisch zu verifizieren sind:

1) *Da fährt noch einer?*

und

2) *Der fährt um ein Uhr?*

Für die Verifikationsaufgabe kann u.a. folgende Information eingesetzt werden:

- In beiden Fällen ist aufgrund des Fehlens anderer grammatischer Merkmale (Verbstellung oder w-Fragewort) zu erwarten, daß der Satzmodus *Frage* intonatorisch markiert wird.
- In dem betreffenden Kontext, charakterisiert durch die Kommunikationssituation *Informationsabfrage* und die Vorgeschichte des Dialogs, kann davon ausgegangen werden, daß im Fall 1) das Verb *"fahren"* im Fokus steht, im Fall 2) die Phrase *"um ein Uhr"*, wobei das Wort *"ein"* Träger des Satzakzents ist. Dies stellt eine sehr starke Einschränkung dar, denn es sind für *jedes* Wort der beiden Äußerungen andere Kontextsituationen denkbar, in denen

---

[1] An dieser Stelle sei Herrn Reetz noch einmal ausdrücklich für die Bereitstellung des Algorithmus gedankt.

es Träger des Satzakzents ist (mit Ausnahme des Wortes *"Uhr"*, für das es sehr schwierig sein dürfte, eine Situation zu konstruieren). Diejenige Phrase, die das akzentuierte Wort enthält, steht im Fokus. Somit kann prinzipiell jede Phrase der beiden Äußerungen im Fokus stehen.

- Die möglichen Vokalklassen sind für jeden Silbenkern stark eingeschränkt.

Soll dieses Wissen für die Verifikation benutzt werden, so benötigt man ein *Modell* zur intonatorischen Markierung von *Satzmodus* und *Fokus* im Deutschen.

7.2 Vorgehensweise und Untersuchungsmaterial

Untersuchungsmaterial war das Fokus-Korpus (Kap.3.4.1). Für die Konstruktion des Korpus war maßgebend, daß bei gleicher segmentaler Struktur möglichst viele unterschiedliche Fokuskonstellationen realisiert werden konnten (z.B. *enger Fokus und Fokusprojektion;* für Einzelheiten siehe [OPPENRIEDER 89a]). Aus Aufwandsgründen wurde dabei darauf geachtet, daß die Position des Fokus auf die letzte bzw. vorletzte Phrase beschränkt war. Für die Darstellung in diesem Kapitel wird die mögliche weitere Unterscheidung von Modus und Fokus nicht betrachtet, d.h. unter einer Modus/Fokus-Konstellation soll eine Kombination aus einer der in diesem Korpus möglichen Modi *Frage* und *Nicht-Frage* mit einer der möglichen Fokus-Positionen *letzte* (d.h. für die untersuchten Äußerungen die *dritte*) und *vorletzte* (d.h. für die untersuchten Äußerungen die *zweite*) Phrase verstanden werden. Weiterhin soll angenommen werden, daß durch die Konstruktion des Korpus alle im Anwendungsbereich interessierenden Konstellationen repräsentiert sind. 48 Prozent der Fälle waren Fragen, 52 Prozent Nicht-Fragen; in 76 Prozent der Fälle war der Fokus auf der zweiten, in 24 Prozent auf der dritten Phrase.

Das Korpus wurde mit zwei grundsätzlich verschiedenen Ansätzen untersucht.

1) Finden besonders *"schöner"* Fälle über Perzeptionsexperimente:
   Eine Gruppe phonetisch untrainierter Hörer beurteilte alle Äußerungen in Bezug auf Satzmodus, Position des Satzakzents und Natürlichkeit der Äußerung. Bei denjenigen Äußerungen, die in allen drei Kategorisierungstests besonders gut abschnitten, konnte angenommen werden, daß es sich sozusagen um *Paradebeispiele* von Realisierungen der jeweiligen Modus/Fokus-Konstellation handelte. Eine Zusammenfassung ähnlicher Realisierungen resultierte in intonatorisch unterschiedlichen Modus/Fokus-Markierungen (siehe Kap.7.4). Im folgenden werden diese unterschiedlichen Realisierungen als *Prototypen* bezeichnet. Dieser Ansatz führte zu einem *qualitativen Modell* (wie sehen die *gültigen* intonatorischen Markierungen der möglichen Modus/Fokus-Konstellationen aus?).
2) Statistische Auswertung von Merkmalwerten, welche die prosodischen Eigenschaften *Tonhöhe, zeitliche Strukturierung,* und *Lautheit* beschreiben:
   Es wurden verschiedene akustische Merkmale automatisch und per Hand berechnet. Ihre Relevanz in bezug auf die Markierung von Modus und Fokus wurde dadurch überprüft, daß sie als Merkmale für statistische Klassifikatoren benutzt wurden (Diskriminanzanalyse). Es handelte sich jeweils um Klassifikatoren für die Zwei-Klassen-Probleme *Frage / Nicht-Frage* und *Fokus auf der 2. Phrase / Fokus auf der 3. Phrase.* Für die relevanten Merkmale wurden

die Mittelwerte der vier Konstellationen bestimmt. Die Merkmale der Prototypen wurden mit den Mittelwerten verglichen. Für die Konstellationen mit mehr als einem Prototypen führte dies zu den Standardrealisierungen, die im folgenden als Kern-Prototypen bezeichnet werden, und zu selteneren, aber akzeptablen Realisierungen, die als Rand-Prototypen bezeichnet werden. Dieser Ansatz führte zu einem *quantitativen Modell* (wie sehen die *üblichen* intonatorischen Markierungen der möglichen Modus/Fokus-Konstellationen aus?).

In der Terminologie der *semantischen Netze* kann man die vier Konstellationen als *spezialisierte Konzepte* eines Konzepts *Modus/Fokus-Konstellation* modellieren. Die zu findenden verschiedenen intonatorischen Prototypen sind dann weitere *Spezialisierungen* der entsprechenden Konzepte. Das Wissen über die Häufigkeit der Prototypen kann man als Parameter in die Bewertungsprozedur des entsprechenden *spezialisierten Konzeptes* einfließen lassen.

7.3 Merkmalberechnung und Bewertung

Für die Äußerungen und die fokussierbaren Phrasen wurden Merkmale extrahiert und verschiedenen Normierungstransformationen unterworfen. Die Relevanz der Merkmale und der Transformationen wurde mit einem statistischen Klassifikator getestet. Auf die Merkmalextraktion und Normierungstransformationen soll hier nicht eingegangen werden, da sie für die Modellbildung von untergeordneter Bedeutung sind. Einzelheiten finden sich in [BATLINER 89a, 89b]. In diesem Zusammenhang ist lediglich wichtig, daß die Merkmale zwar zum großen Teil per Hand aus den Mingogrammen extrahiert wurden, daß aber eine automatische Extraktion möglich ist, sofern man die Phrasengrenzen zur Verfügung hat. Im Falle der Verifikation von Satzhypothesen wären die Grenzen durch den Zuordnungspfad gegeben. Auch im Fall der Normierungstransformationen sind vergleichbare automatisch berechenbare Transformationen möglich. Tabelle 7.1 zeigt die berechneten Merkmale sowie diejenige Transformation, mit der die besten Klassifikationsergebnisse erzielt wurden.

Tabelle 7.2 zeigt für die verschiedenen Merkmale die Erkennungsraten für die Klassifikatoren MODUS und FOKUS (siehe Tabelle 7.1). Die Merkmale wurden einzeln (Spalte *un*) bzw. zusammen mit dem entsprechenden Merkmal der anderen Phrase (Spalte *bi*) ausgewertet. Es wurden alle Fälle zum Lernen und zum Testen gewählt (mit Lernstichprobe=Teststichprobe). In den Spalten FokusF und FokusNF wurden nur Fragen bzw. Nicht-Fragen klassifiziert. Die letzten drei Zeilen zeigen die Erkennungsraten für die Klassifikation unter Verwendung aller Merkmale bei verschiedenen Lern- und Teststichproben. Folgende Schlüsse lassen sich aus den Erkennungsraten ziehen:

Tab. 7.1 (nächste Seite): Berechnete Merkmale, durchgeführte Perzeptionsexperimente und für die Einschätzung der Merkmalrelevanz benutzte Klassifkatoren. In der Spalte M/F steht für die Merkmale, ob sie nur für den Satzmodus-Klassifikator MODUS (M), nur für den Fokus-Klassifikator FOKUS (F) oder für beide Klassifkatoren (M/F) verwendet wurden.

Merkmale	Abk.	M/F	Berechnung/Transformationen
F_0-Offset	Off	M	Grundfrequenzwert am Äußerungsende in Halbtönen minus sprecherspezifischem Basiswert
Steigung	Steig	M	Steigung der Ausgleichsgeraden für die Grundfrequenzwerte in Halbtönen
F_0-Maximum der 2. und 3. Phrase	Max2 Max3	M/F M/F	Maximaler und minimaler Grundfrequenzwert der 2. und 3. Phrase jeweils in Halbtönen minus sprecherspezifischem Basiswert
F_0-Minimum der 2. und 3. Phrase	Min2 Min3	M/F M/F	
Relative Position von F_0-Max./Min. 2. und 3. Phrase	Pos2 Pos3	M/F M/F	Differenz der Werte von F_0-Maximum und F_0-Minimum auf der Zeitachse; positiver Wert, wenn Maximum vor Minimum, sonst negativer Wert.
Dauer der 2. und 3. Phrase	Dau2 Dau3	F F	jeweils [(aktuelle Dauer / mittl. Phrasendauer) · (Dauer / (Äußerungsdauer/Silbenzahl))]
Intensität der 2. und 3. Phrase	Int2 Int3	F F	Maximale Energie der Phrase im Frequenzbereich 0-5000 Hz, gemessen in relativen Millibelwerten
Hörerurteile			
Moduszuweisung	MOD		Prozent Fragekategorisierung, errechnet aus dem Kategorisierungstest nach *Frage*, *Aussage*, *Exklamativ*, *Imperativ* oder *Wunsch*
Fokuszuweisung	FOK		(SA2-SA3)/(SA1+SA2+SA3), wobei SAi für die Zahl der Hörer steht, welche die i-te Phrase als Träger des Satzakzents hörten. FOK nahm Werte zwischen +1.0 und -1.0 an und diente als Handklassifikation für den Klassifikator FOKUS
Natürlichkeit	NAT		Die Hörer mußten der Kombination Kontext- und Testsatz eine Zahl zwischen 1 (≙ *paßt gut zusammen*) und 5 (≙ *paßt nicht zusammen*) zuweisen
Klassifikatoren			
Satzmodus	MODUS		durch Kontextsatz und Situationsbeschreibung intendierter Satzmodus; es wurde nur zwischen Frage und Nicht-Frage unterschieden
Satzfokus	FOKUS		aus FOK errechnete Position des Satzakzents; nur die 2. oder 3. Phrase konnte akzentuiert sein

Merkmale	Modus un	Modus bi	Fokus un	Fokus bi	FokusF un	FokusF bi	FokusNF un	FokusNF bi
Off	93		-		-		-	
Steig	85		-		-		-	
Max2	73	> 94	60	> 78	80	> 81	54	> 92
Max3	94		60		63		88	
Min2	65	> 84	59	> 62	53	> 71	62	> 70
Min3	84		47		70		47	
Pos2	76	> 85	51	> 69	78	> 82	69	> 52
Pos3	78		51		62		48	
Dau2	-		66	> 74	60	> 71	70	> 82
Dau3	-		72		70		82	
Int2	-		58	> 66	52	> 56	62	> 70
Int3	-		55		51		53	
multiv-ln	97		93		95		96	
multiv-ln-1	92		84		86		94	
multiv-ll	92		78		76		82	

Tab. 7.2: Erkennungsraten für die Klassifikation nach Modus und Fokus für die einzelnen Merkmale (Spalte *un*) bzw. zusammen mit dem entsprechenden Merkmal der anderen Phrase (Spalte *bi*). Es wurden alle Äußerungen zum Lernen und zum Testen verwendet (mit Lernstichprobe=Teststichprobe). In den Spalten FokusF und FokusNF wurden nur Fragen bzw. Nicht-Fragen klassifiziert.
In den letzten drei Zeilen wurden alle Merkmale verwendet (acht für die Modus- und zehn für die Fokus-Klassifikation). Dabei steht *ln* für Lernstichprobe = Teststichprobe, *ln-1* für n-1 (fünf) Sprecher Lern- und einen Sprecher Teststichprobe und *ll* für einen Sprecher Lern- und n-1 Sprecher Teststichprobe. Die Erkennungsraten für die beiden Konfigurationen mit Lernstichprobe ≠ Teststichprobe wurden jeweils im "leave-one-out"-Verfahren erstellt.

1) Es handelt sich um adäquate Merkmale zur Beschreibung der vorhandenen intonatorischen Modus/Fokus-Markierungen. Die Erkennungsraten für die Klassifikation von Modus und Fokus lagen für die Konfiguration *multiv-ln-1* (n-1 Sprecher Lernstichprobe, 1 Sprecher Teststichprobe) jeweils über 90 Prozent. Für die Fokus-Klassifikation ergibt sich die Gesamterkennungsrate aus den Erkennungsraten der Spalten *FokusF* und *FokusNF* (ohne vorherige Trennung der Fragen und Nichtfragen beträgt die Erkennungsrate 84 Prozent). Zwar wurde für die Spalten *FokusF* und *FokusNF* die Trennung von Fragen und Nicht-Fragen per Hand durchgeführt, aber die Benutzung der automatischen Modus-Klassifikation führte sogar zu einer minimalen Verbesserung.
2) Die Tatsache, daß sich bei Unterscheidung der Satzmodi (die Spalten FokusF und FokusNF bzw. die Spalte Fokus) die Klassifikation des Fokus verbessert, läßt sich damit erklären, daß sich insbesondere bei der prosodischen Eigenschaft *Tonhöhe* starke Überlagerungen der beiden Markierungen ergeben. Die Überlagerung wirkt sich bei den Fragen i.allg. stärker aus

als bei den Nicht-Fragen. So ist z.B. bei dem Testsatz *"Sie läßt die Nina das Leinen weben"* der aufgrund der syntaktischen Struktur "übliche" Satzmodus *Aussage* intonatorisch nicht so stark markiert wie der Satzmodus *Frage*.

7.4 Bestimmung der Prototypen

Für die Ergebnisse der drei Perzeptionstests (siehe Tabelle 7.1) wurden folgende sehr hohe Schwellen angesetzt:

1) MOD $\geq$ 80 für Fragen und $\leq$ 20 für Nicht-Fragen, d.h. die Hörer waren sich bei der Moduszuweisung weitgehend einig.
2) |FOK| = 1, d.h. die Hörer waren sich bei der Fokuszuweisung absolut einig.
3) NAT $\leq$ 2, d.h. die Hörer beurteilten die Äußerungen als sehr natürlich.

Die Merkmale der 24 Fälle (sieben Prozent), die alle drei Schwellen passierten, wurden graphisch aufgetragen. Die Merkmale aller Fälle zu einer Modus/Fokus-Konstellation wurden miteinander verglichen. Auf diese Weise konnten sieben Prototypen identifiziert werden. Die Bilder 7.1 - 7.7 zeigen für je einen Repräsentanten der Prototypen die Grundfrequenzkontur für die zweite und dritte Phrase. Die Grundfrequenzkonturen sind in Halbtönen über dem sprecherspezifischen Basiswert[2] dargestellt. Die Zeitachse ist in Centisekunden ab Beginn der Äußerung aufgetragen. Bild 7.1 und 7.2 zeigen die Konstellationen (*Nicht-Frage, Fokus auf vorletzter Phrase* und *Frage, Fokus auf letzter Phrase*), bei denen nur ein intonatorischer Prototyp gefunden wurde. Bild 7.3 und 7.4 zeigen die Prototypen der Konstellation *Fokus auf letzter Phrase, Nicht-Frage*. Die Bilder 7.5 - 7.7 beziehen sich auf die Konstellation *Fokus auf vorletzter Phrase, Frage*.

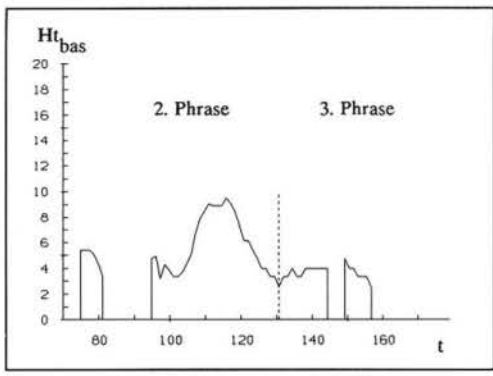


Bild 7.1: Nicht-Frage, Fokus auf der zweiten Phrase.

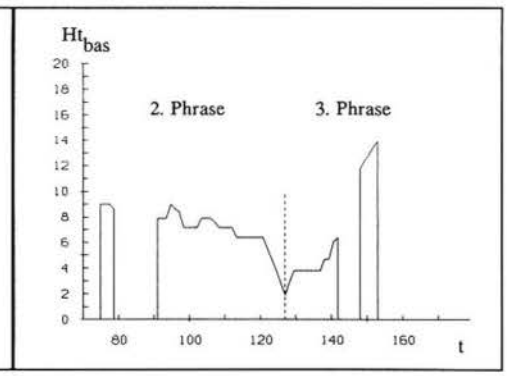


Bild 7.2: Frage, Fokus auf der dritten Phrase.

[2] Der tiefste von einem Sprecher erreichbare Grundfrequenzwert

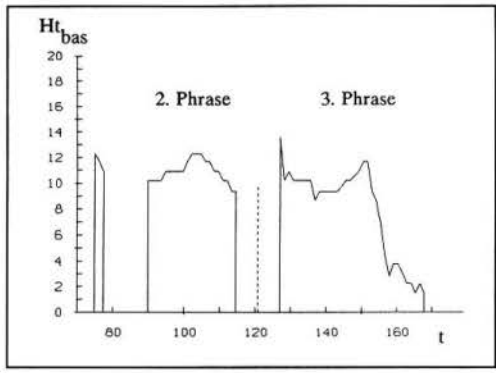


Bild 7.3: Nicht-Frage, Fokus auf der dritten Phrase.

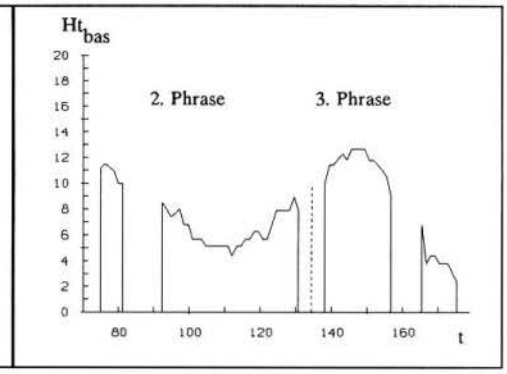


Bild 7.4: Nicht-Frage, Fokus auf der dritten Phrase.

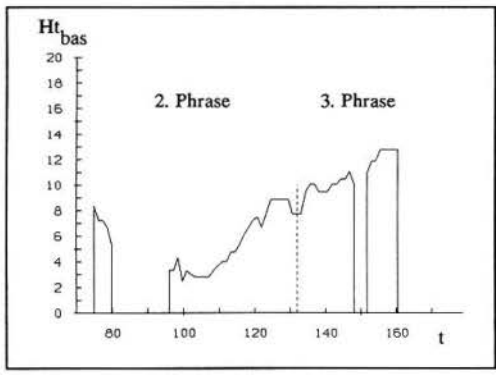


Bild 7.5: Frage, Fokus auf der zweiten Phrase.

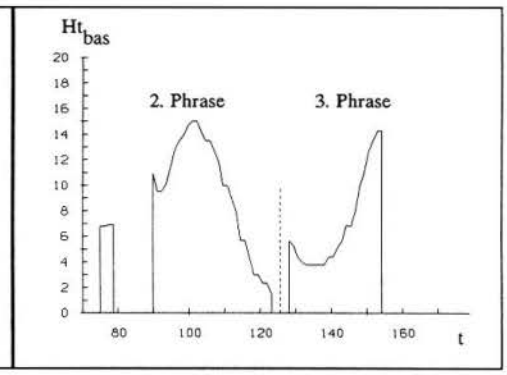


Bild 7.6: Frage, Fokus auf der zweiten Phrase.

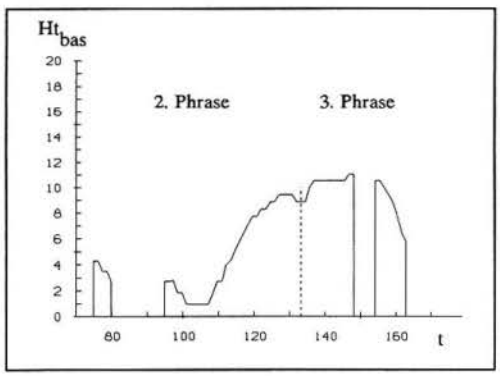


Bild 7.7: Frage, Fokus auf der zweiten Phrase.

7.5 Identifikation der Kern- und Rand-Prototypen

Für die vier Modus/Fokus-Konstellationen wurden die Mittelwerte der 10 Merkmale gebildet, die für die Fokus-Klassifikation verwendet wurden. Durch einen Vergleich der prototypischen Realisierungen mit diesen Mittelwerten wurden die Kern- und Rand-Prototypen gefunden. Diese Vorgehensweise ist nur gerechtfertigt, wenn man wie in der vorliegenden Situation *einen* Kern-Prototypen hat: Angenommen, es gäbe zwei Kern-Prototypen, die gleich häufig auftreten, und der Mittelwert eines Merkmals ist a für den einen Kern-Prototypen und -a für den anderen, so kann man aus dem Gesamtmittelwert 0 keinen Prototypen als Kern-Prototypen identifizieren, sondern man muß sein Modell zunächst verfeinern.

Bild 7.8 zeigt die Mittelwerte der für die Fokus-Klassifikation benutzten Merkmale (siehe Tab.7.1) für die vier Modus/Fokus-Konstellationen. In Bild 7.8a sind die Merkmale für die zweite Phrase beschriftet. Es handelt sich um dieselbe Konstellation wie in Bild 7.1). Die Y-Position der vier schwarzen Quadrate gibt den Mittelwert des minimalen und maximalen Grundfrequenzwertes einer Phrase (*Maxi, Mini*) an. Bezugsachse ist die linke Y-Achse in Halbtönen minus sprecherspezifischem Basiswert. Die Differenz der X-Position der beiden Extremwerte einer Phrase ist der Wert des Merkmals *Posi* in Centisekunden. Die Länge der beiden waagrechten Linien gibt die Mittelwerte der Dauermerkmale *Daui* in Centisekunden an. Von den Lautheitsmerkmalen *Inti* wurde für diese Darstellung der Wert *Int3-100 (rmB)* subtrahiert, und das Ergebnis wurde durch Division mit 100 in Dezibel umgewandelt. Somit gibt die Differenz der Y-Koordinaten der beiden Linien in bezug auf die rechte Y-Achse die Differenz der Mittelwerte in dB an.

Durch den Vergleich dieser Mittelwertdarstellungen mit den sieben Prototypen wurden die vier Prototypen der Bilder 7.1, 7.2, 7.3 und 7.5 als Kern-Prototypen identifiziert, und die der Bilder 7.4, 7.6, 7.7 als Rand-Prototypen. Eine Charakterisierung der Prototypen findet sich in [BATLINER 89b, 89c].

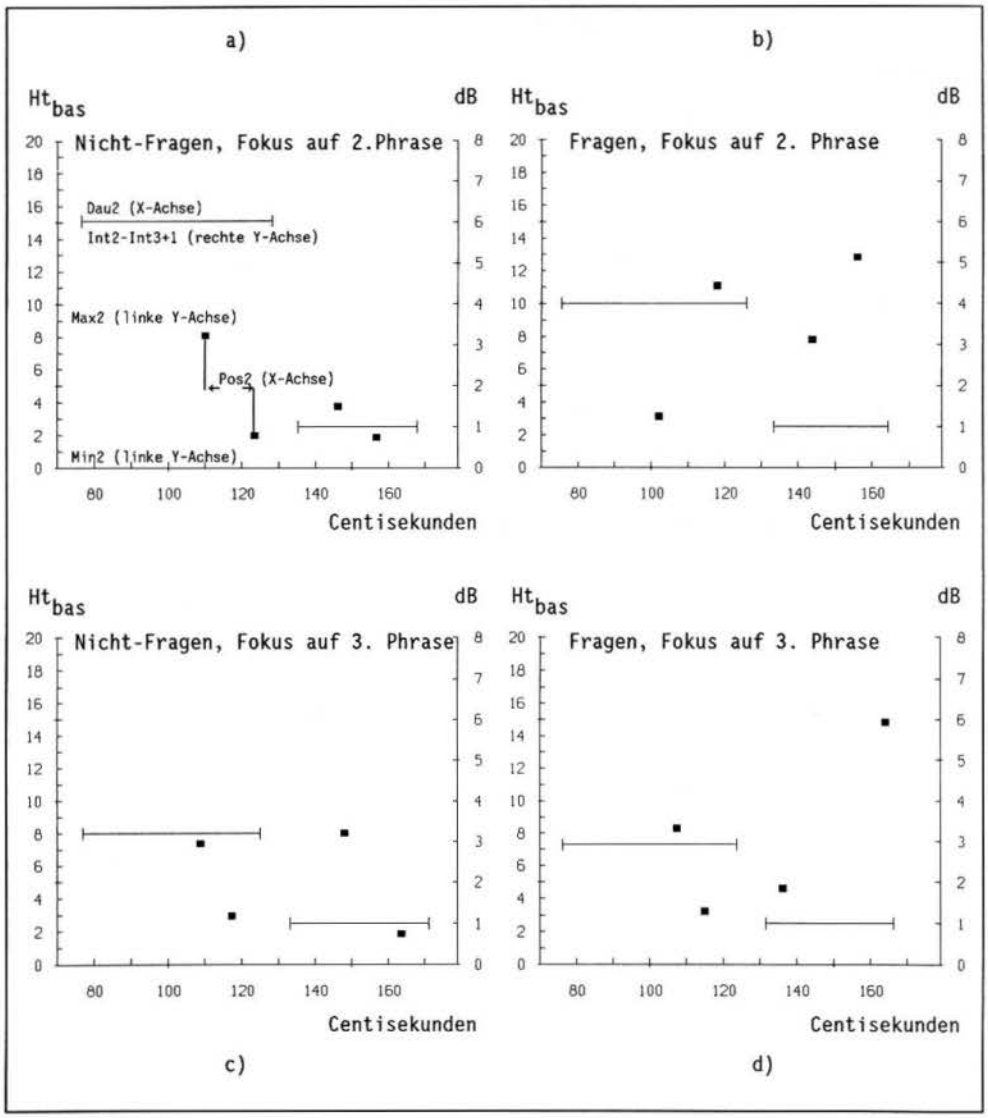


Bild 7.8: Mittelwerte der zehn für die Fokus-Klassifikation benutzten Merkmale. Die Y-Position der scharzen Quadrate gibt den Mittelwert der F_0-Extremwerte Maxi und Mini der i-ten Phrase an (in bezug auf die linke Y-Achse). Die Differenz der X-Koordinaten charakterisiert das Merkmal Posi. Die Länge der waagrechten Linien gibt die Größe der Dauermerkmale Daui an. Die Differenz der Y-Koordinaten gibt die Differenz der Lautheitsmerkmale Inti in dB an (in bezug auf die rechte Y-Achse).

7.6 Darstellung des erstellten Intonationsmodells in einem maschinellen Wissensrepräsentationsschema

Im folgenden soll zum Zweck der Darstellung in einem semantischen Netz eine eigene symbolische Beschreibung der gefundenen Prototypen vorgestellt werden, die auf den Tonhöhenmerkmalen *Maxi, Mini* und *Posi* aufbaut:

- Die Symbole A, a sollen eine F_0-Kontur innerhalb einer Phrase bezeichnen, bei der die dominante Bewegung fallend ist (*Posi* > 0).
- Die Symbole B, b sollen eine F_0-Kontur innerhalb einer Phrase bezeichnen, bei der die dominante Bewegung steigend ist (*Posi* < 0).
- Die Kontur einer Phrase wird mit einem kleinen Buchstaben bezeichnet, falls die F_0-Bewegung (*Maxi* - *Mini*) der benachbarten Phrase deutlich stärker ist. Ansonsten werden Großbuchstaben benutzt. Man beachte, daß hiermit keine Aussage über das absolute Ausmaß der F_0-Bewegung in der Phrase gemacht wird, sondern über das relative.

Tabelle 7.3 zeigt eine Charakterisierung der sieben Prototypen in dieser Nomenklatur.

Modus/Fokus-Konstellation (spezialisiertes Konzept)	Kern-Prototyp Inton_K_Typ	Rand-Prototyp Inton_R_Typ	Rand-Prototyp Inton_R_Typ
Nicht-Frage, Fokus auf 2. Phrase	Aa (Bild 7.1)	-	-
Nicht-Frage, Fokus auf 3. Phrase	aA (Bild 7.3)	bA (Bild 7.4)	-
Frage, Fokus auf 2. Phrase	Bb (Bild 7.5)	AB (Bild 7.6)	Ba (Bild 7.7)
Frage, Fokus auf 3. Phrase	aB (Bild 7.2)	-	-

Tab. 7.3: Beschreibung des Grundfrequenzverlaufs für die sieben Prototypen in der oben vorgestellten Nomenklatur.

Es wurde bereits in Kap.2 darauf hingewiesen, daß eine ganze Reihe von verschiedenen Schemata zur Beschreibung des Tonhöhenverlaufs existieren. Typische Beschreibungskategorien wie *Hochton, Tiefton* ([PIERREHUMBERT 80]), *Schleifton, Tonbruch* ([ROYÉ 83]) oder *konvex, konkav* ([OPPENRIEDER 88a]) beschreiben mehr oder weniger stilisiert einen Teilaspekt der Verlaufskurve. Für die Darstellung in diesem Kapitel wurde trotz der vielen bereits existierenden Beschreibungsformen eine eigene Symbolik benutzt, da sie mit wenigen, leicht nachvollziehbaren Grundsymbolen eine eindeutige Identifikation der gefundenen Prototypen erlaubt, ohne daß eine zugrundeliegende Theorie eingeführt werden muß.

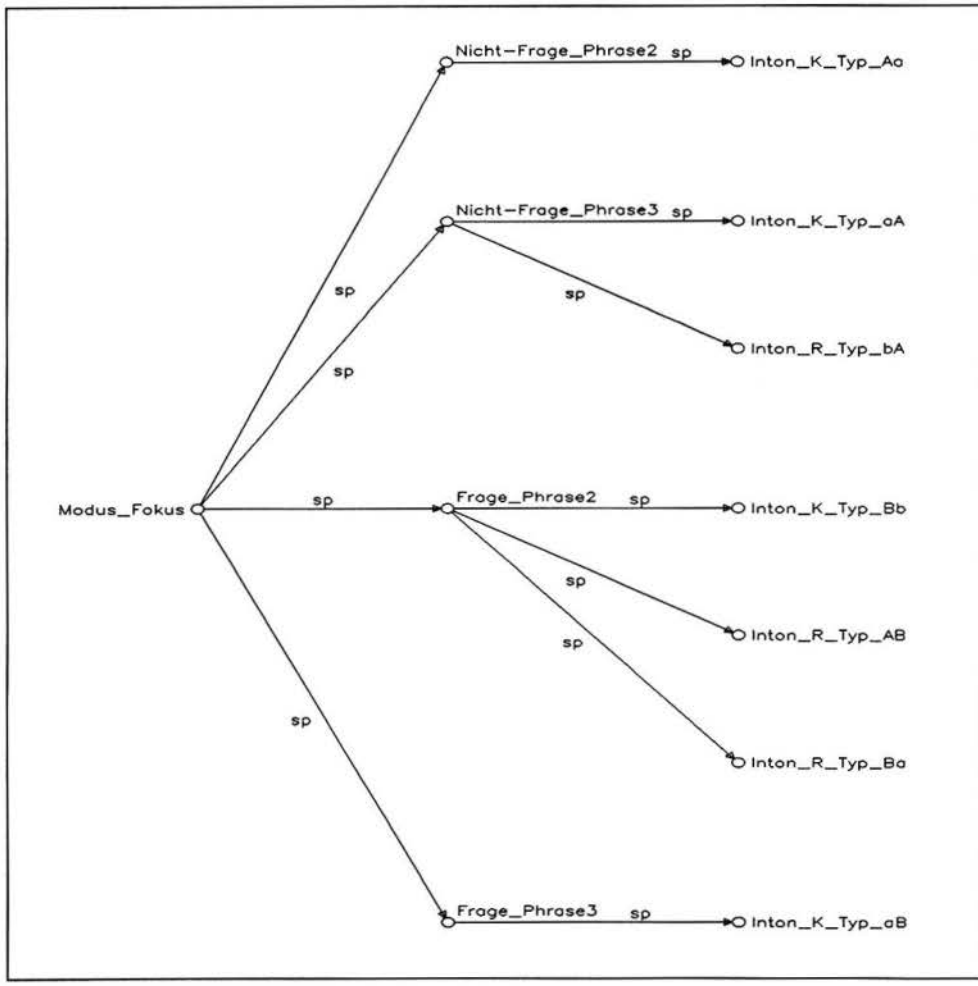


Bild 7.9: Darstellung des erstellten Intonationsmodells in einem semantischen Netz.

Bild 7.9 zeigt ein semantisches Netz für *Modus/Fokus-Konstellationen.* Die vier untersuchten Konstellationen sind spezialisierte Konzepte (*sp*-Kante) des allgemeinen Konzepts *Modus_Fokus.* Für jedes spezialisierte Konzept sind die verschiedenen intonatorischen Prototypen ebenfalls als spezialisierte Konzepte realisiert. Die Bezeichnung dieser Konzepte gibt an, ob es sich um einen Kern-Prototyp (Inton_K_Typ_XX) oder einen Rand-Prototyp (Inton_R_Typ_XX) handelt. Die letzten beiden Buchstaben im Konzeptnamen charakterisieren den Verlauf der Grundfrequenzkontur in der oben vorgestellten Nomenklatur (Tabelle 7.3).

Somit kann das entwickelte Modell der intonatorischen Markierung von Modus und Fokus zum Zwecke der prosodischen Verifikation von Satzhypothesen folgendermaßen in einem *semantischen Netz* dargestellt werden:

- Es wird die Strukur aus Bild 7.9 verwendet.
- In jedem spezialisierten Konzept wird das Wissen über die Form der intonatorischen Markierung in Form von Steueranweisungen für die prosodische Komponente eines Sprachsynthesegeräts abgelegt. Somit kann das Wissen, das aus den Merkmalausprägungen eines segmental eingeschränkten Korpus gewonnen wurde, in einer Darstellung repräsentiert werden, die es erlaubt, Instanzen zu segmental anderen Äußerungen aussagekräftig zu bewerten.
- Das Wissen über die Häufigkeit des Prototypen fließt in die Bewertung der Instanzen ein.

Eine prosodische Verifikation von Satzhypothesen ist dann folgendermaßen denkbar:

- Pro Satzhypothese wird für jeden in dem vorgegebenen Kontext möglichen Prototypen von fokussierter Phrase und Satzmodus mit einem Synthesegerät eine Referenzkontur erzeugt und mit der aktuellen Kontur verglichen. Das Abstandsmaß stellt eine Bewertung dar.
- Die Bewertung der Kern-Prototypen wird im Vergleich zu den Bewertungen der Rand-Prototypen um einen gewissen Faktor verbessert.
- Man entscheidet sich für die Modus/Fokus-Konstellation der Instanz mit der besten Bewertung.

Es muß noch einmal darauf hingewiesen werden, daß das hier vorgestellte *erwartungsgesteuerte Modell zur intonatorischen Markierung von Modus und Fokus im Deutschen* sehr vorläufig und bruchstückhaft ist. Es bietet allerdings den Vorteil, daß eine schrittweise Verfeinerung möglich ist. Weiterhin ermöglicht die Implementierung in einem maschinellen Wissensrepräsentationsschema ebenso wie die automatische Extrahierbarkeit der benötigten Parameter jederzeit eine Verifikation des erstellten Modells bzw. die Verifikation einer weiteren Verfeinerung des Modells.

Die im letzten Kapitel präsentierten Ergebnisse haben gezeigt, daß eine datengetriebene Betonungsbeschreibung zu einer Verbesserung der Analyseergebnisse während der Erkennungsphase (Erkennung bis zur Wortketten-Ebene, siehe [NIEMANN 88b]) eines sprachverstehenden Systems führen kann. Während der Verstehensphase wird allerdings mit jedem Analyseschritt mehr Wissen über das Gesprochene gewonnen. Deswegen ist ein schrittweises Einbeziehen der erreichten Analyseergebnisse unbedingt sinnvoll. Die hier vorgestellte Vorgehensweise ermöglicht eine solche schrittweise Einbeziehung und stellt somit eine sinnvolle Erweiterung der datengetriebenen Betonungsbeschreibung dar.

8 Zusammenfassung

Die enorme Verbesserung der Leistungsfähigkeit von Computer-Hardware, verbunden mit einem ebenso starken Preisverfall, lassen immer mehr ASE-Anwendungen realistisch erscheinen. Robuste Einzelworterkenner sind schon seit mehreren Jahren kommerziell verfügbar, während sich Systeme zum Erkennen und insbesondere zum Verstehen kontinuierlich gesprochener Sprache noch immer in einem experimentellen Stadium befinden. In den meisten sprachverstehenden Systemen konzentriert sich die Analyse des Zeitsignals auf die Extraktion von segmentaler Information, z.B. Information über die gesprochenen Laute. In letzter Zeit hat das Interesse an suprasegmentaler Information, die auch als *prosodische Information* bezeichnet wird, in Zusammenhang mit ASE-Systemen stark zugenommen, wofür sich im wesentlichen zwei Gründe anführen lassen:

- Prosodische Information wird in der zwischenmenschlichen Kommunikation sehr stark eingesetzt. In der Mensch-Maschine-Kommunikation sollte, zumindest langfristig gesehen, ein Anwender seine gewohnte Sprechweise benutzen können.
- Es ist zu erwarten, daß zwischen den mit prosodischen Mitteln als *betont markierten Stellen* einer Äußerung und den für den *Verstehensprozeß wichtigen Stellen* ein starker Zusammenhang besteht.

Die vorliegende Arbeit befaßte sich mit dem Einsatz prosodischer Information in dem an der Universität Erlangen entwickelten sprachverstehenden System EVAR. Nach einer Erörterung der potentiellen Einsatzgebiete prosodischer Information in ASE-Systemen wurden in Kapitel 1 das System EVAR und der Beitrag dieser Arbeit zum Einsatz prosodischer Information in ASE-Systemen vorgestellt: Schwerpunkt der Arbeit war die Erstellung einer automatischen Betonungsbeschreibung für das Deutsche. Die Einbettung in ein ASE-System bedingte dabei zunächst einen datengetriebenen Ansatz, d.h. die Betonungsbeschreibung wurde unter alleiniger Verwendung des Zeitsignals erstellt. Während der Verstehensphase im Analyseprozeß liegt bereits sehr viel Wissen über eine Äußerung in Form von Phrasen- und Satzhypothesen vor, das bei der Erstellung einer Betonungsbeschreibung verwendet werden kann. Daher wurden in enger Zusammenarbeit mit Mitarbeitern des DFG-Projekts "Modus-Fokus-Intonation" grundlegende Untersuchungen zur Erstellung eines *intonatorischen Modells des Deutschen* sowie zur Darstellung des Modells in einem maschinellen Wissensrepräsentationsschema durchgeführt.

In Kapitel 2 wurden die im Zusammenhang mit dieser Arbeit wichtigen sprachwissenschaftlichen Begriffe definiert: *Prosodie* wird als Oberbegriff für alle Themenbereiche sprachwissenschaftlicher Untersuchungen verstanden, die sich mit lautübergreifenden (suprasegmentalen) sprachlichen Eigenschaften befassen. Bei diesen prosodischen Eigenschaften handelt es sich um perzeptive Einheiten, denen akustische Korrelate gegenüberstehen. Diese im Sprachsignal meßbaren Korrelate ändern sich artikulatorisch bedingt oder aufgrund einer vom Sprecher bewußt, d.h. um dem Hörer etwas mitzuteilen, durchgeführten Variation der Sprachproduktion. Im zweiten Fall spricht man von *Intonation.* Die im Zusammenhang mit der ASE wichtige Darstellungsfunktion (intellektuelle

Bedeutung) der Intonation sowie die hierfür wichtigsten prosodischen Eigenschaften (Tonhöhe, Lautheit, zeitliche Strukturierung und Klangfarbe) wurden genauer untersucht.

Die Darstellungsfunktion zerfällt in die drei Bereiche *Markierung des Akzents, Markierung des Satzmodus* und *Gliederung der Äußerung.* Der Begriff *Akzent* wurde synonym mit *Betonung* verwendet. Jedes Wort hat - wenn es isoliert gesprochen wird - eine Wortakzentsilbe, die "auffälliger" (betont) gestaltet wird als die anderen Silben des Wortes. In kontinuierlicher Sprache werden nur einige der Wortakzentsilben betont artikuliert. Das dazugehörige Wort bzw. die Phrase wird als Träger des *Satzakzents* bezeichnet. Welche der Silben betont werden, hängt einerseits mit der rhythmischen Struktur der deutschen Sprache zusammen, andererseits (insbesondere in spontaner Sprache) mit dem Informationsgehalt des dazugehörigen Wortes bzw. der Phrase. Hierfür wurde der Begriff *Fokus* eingeführt, mit dem die wichtigste Information einer Äußerung bezeichnet, und der phonetisch durch den Satzakzent realisiert wird.

Ferner wurde im zweiten Kapitel gezeigt, mit welchen Mitteln die drei Bereiche der Prosodie realisiert werden. Dabei wurde in erster Linie auf die Mittel zur Markierung des Satzakzents eingegangen. Es wurde deutlich, daß die Akzentuierung mit der prosodischen Eigenschaft *Tonhöhe* bzw. dem Korrelat *Grundfrequenz* durch Erhöhung oder Senkung, sowohl vor der Akzentsilbe als auch danach stattfinden kann. Bei der *zeitlichen Strukturierung* bzw. dem Korrelat *Lautdauer* führt Betonung zu einer Dehnung, die durch Verkürzung der unbetonten Umgebung verstärkt werden kann. Bei der *Lautheit* wird das Korrelat *Intensität* sehr stark durch intrinsische Einflüsse verändert, die vom Menschen offensichtlich kompensiert werden. In Zusammenhang mit der ASE ist dies ein starker Nachteil, dem der Vorteil gegenübersteht, daß das Korrelat praktisch fehlerfrei berechnet werden kann. Bei der *Klangfarbe* ist die Veränderung des Korrelats *Vokalqualität* in Richtung Zentralvokal ein Indiz für Unbetontheit.

Der *Satzmodus* (z.B. Aussagesatz, Wunschsatz) kann u.a. mit intonatorischen Mitteln ausgedrückt werden. Bei der *Markierung des Satzmodus* interessierte vor allem die Unterscheidung zwischen Fragen und Nicht-Fragen, da bei informationsabfragenden Systemen Fragen und Nicht-Fragen oft allein durch die Intonation markiert werden. Die Markierung geschieht meistens durch den Tonhöhenverlauf am Ende der Äußerung

Die *Gliederung der Äußerung* geschieht mit der Tonhöhe, der zeitlichen Strukturierung und der gezielten Pausensetzung, ist insbesondere bei größeren Redebeiträgen zu beobachten und wurde im Rahmen dieser Arbeit nicht weiter untersucht.

Schließlich wurden Überlagerungen aufgezeigt, die sich daraus ergeben, daß insbesondere die prosodische Eigenschaft Tonhöhe für alle drei vorgestellten Bereiche der Intonation eine Rolle spielt: Die Art der Markierung des Satzakzents hängt stark vom Satzmodus ab.

Im dritten Kapitel wurden die verschiedenen Stichproben vorgestellt, mit denen Experimente durchgeführt wurden: Bei der *EVAR-Stichprobe* handelt es sich um Äußerungen aus dem Anwendungsbereich IC-Zugauskunft. Die *Pragmatik-Stichprobe* ist eine Teilmenge der EVAR-Stichprobe, in der die für die Pragmatik-Analyse wichtigen Wörter markiert wurden.

Die *Dialog-Stichprobe* besteht aus Aufnahmen von echten Auskunftsdialogen. Zu diesen Äußerungen wurde ein Betonungsmaß erstellt. Zu diesem Zweck mußten 15 Hörer die Silben in betont/unbetont einteilen. Jeder Silbe wurde die Zahl der "betont"-Urteile zugewiesen. Anhand dieser Betonungszahl konnte der erwartete enge Zusammenhang zwischen *betonten* und *für die Analyse wichtigen* Stellen nachgewiesen werden: Alle Wörter, die von mehr als der Hälfte der Hörer als betont eingestuft wurden (31 Prozent aller Wörter), ließen sich in ein für die Anwendung *Zugauskunft* relevantes Pragmatik-Konzept einordnen oder stellten einen metakommunikativen Dialogschritt dar.

Die Modus-Fokus-Korpora (das Material des DFG-Projekts "Modus-Fokus-Intonation") bestehen aus intonatorischen Minimalpaaren, bei denen Satzmodus und Position des Satzakzents bzw. Fokus allein durch intonatorische Merkmale indiziert werden.

Das vierte Kapitel behandelte die Extraktion von Merkmalen aus den physikalischen Korrelaten der prosodischen Eigenschaften. Hierzu wurden zunächst Silbenkerne im Sprachsignal detektiert. Im Ausgangssignal eines Bandpaßfilters, das den Frequenzbereich der ersten beiden Formanten abdeckt, wurden lokale Maxima gesucht. Genügten die Maxima gewissen Signifikanzbedingungen, so wurden sie als Silbenkerne markiert. Durch die Betrachtung von zwei weiteren Energiebändern wurden fälschlicherweise als Silbenkerne eingetragene Frikative eliminiert und silbische Nasale zur Silbenkernliste hinzugefügt. Sehr lange Silbenkerne wurden mit veränderten Signifikanzschwellen noch einmal betrachtet und u.U. aufgetrennt. Während der Analyse ist es notwendig, die Segmentierung des Sprachsignals durch das Akustik-Phonetik-Modul und das Prosodie-Modul zu synchronisieren (dies stellt sozusagen den ersten Schritt von einer datengetriebenen zu einer erwartungsgesteuerten Betonungsbeschreibung dar). Hierzu wurde ein neuartiges Segmentierungsverfahren in lautähnliche Einheiten entwickelt. Das Verfahren markiert Frames, an denen sich die Entscheidung des Lautkomponentenklassifikators ändert, als potentielle Lautanfänge und erstellt mit diesen Markierungen einen Segmentgraphen, in dem die optimale Segmentierung nach Lauten durch ein Graphsuchverfahren ermittelt wird. Durch Angleichung der Silbenkerngrenzen und der Lautgrenzen können die beiden Segmentierungen synchronisiert werden.

In Abschnitt 4.2 wurde ausführlich auf die Extraktion von Tonhöhenmerkmalen eingegangen. Nach einer Darstellung des Vorgangs der Spracherzeugung wurde anhand von zwei aus der Literatur bekannten Grundfrequenzalgorithmen, dem AMDF- und dem Seneff-Verfahren, die prinzipielle Vorgehensweise bei der Grundfrequenzbestimmung erläutert: Das Zeitsignal wird zunächst in stimmhafte und stimmlose Bereiche unterteilt. Nach verschiedenen Vorverarbeitungsschritten (z.B. Tiefpaßfilterung) wird für jeden Frame ein Grundfrequenz-Schätzwert erzeugt. Das AMDF-Verfahren arbeitet auf dem Zeitbereich und wertet lokale Minima in der Antikorrelationsfunktion aus, während das Seneff-Verfahren auf dem Frequenzbereich arbeitet und eine harmonische Analyse durchführt. Aufbauend auf diesen beiden Verfahren wurde ein neues Grundfrequenzverfahren entwickelt: An stabilen Stellen des Sprachsignals wird mit den beiden Standardverfahren ein Grundfrequenz-Schätzwert ermittelt. Der Median dieser Schätzwerte wird dazu benutzt, um aus dem

Spektrum jedes stimmhaften Frames fünf potentielle Grundfrequenz-Schätzwerte zu bestimmen. Mit der Dynamischen Programmierung wird in der Matrix der möglichen Werte die Grundfrequenzkontur bestimmt.

Aus der Kontur wurden zwei Merkmale berechnet, welche die Veränderungen der prosodischen Eigenschaft *Tonhöhe* innerhalb des Silbenkerns und im Vergleich zu den benachbarten Silbenkernen beschreiben: Die mittlere Grundfrequenz im Silbenkern im Vergleich zur mittleren Grundfrequenz der benachbarten Silbenkerne sowie die Steigung der Ausgleichsgeraden für die Grundfrequenzwerte innerhalb eines Silbenkerns. Zur Beschreibung der prosodischen Eigenschaft *zeitliche Strukturierung* wurde die Silbenkernlänge in Relation zur Länge der benachbarten Silbenkerne gesetzt. Zur Berechnung der Lautheitsmerkmale wurden das Energie-Integral und die maximale Energie des Silbenkerns mit den Werten für die benachbarten Silbenkerne in Relation gesetzt.

Mit der Theorie der vagen Mengen wurde aus den fünf Merkmalen eine Betonungsbeschreibung erstellt, welche Gegenstand des fünften Kapitels ist. Hierzu wurden zunächst die Merkmalausprägungen als betonungsrelevant eingestuft. Die einzelnen Bewertungen der Merkmale zu einer der prosodischen Eigenschaften wurden zu einer Bewertung dieser prosodischen Eigenschaft zusammengefaßt, woraus eine Gesamtbewertung erstellt wurde. Somit standen pro Silbenkern vier Bewertungen zwischen 0 (unbetont bzw. keine Betonungsmarkierung mit der prosodischen Eigenschaft X) und 1 (stark betont bzw. starke Betonungsmarkierung mit der prosodischen Eigenschaft X) zur Verfügung.

Im sechsten Kapitel wurden die Ergebnisse der eigenen experimentellen Untersuchungen vorgestellt. Das Grundfrequenzverfahren wurde dem AMDF- und dem Seneff-Verfahren gegenübergestellt. Im Vergleich zum Seneff-Verfahren konnte die Zahl der Äußerungen mit mindestens einem Grobfehler um 65 Prozent verringert werden, im Vergleich zum AMDF-Verfahren sogar um 77 Prozent.

Für die Modus-Fokus-Korpora wurden sowohl mit manuell als auch mit automatisch extrahierten Tonhöhen-Merkmalen Experimente zur Unterscheidung von Fragen und Nicht-Fragen durchgeführt. Dabei wurden Erkennungsraten bis zu 93 Prozent erzielt.

Der Algorithmus zur Silbenkerndetektion wurde für die EVAR- und die Dialog-Stichprobe ausgewertet. Während in der EVAR-Stichprobe 94 Prozent der gesprochenen Silbenkerne gefunden wurden, fiel aufgrund der veränderten Aufnahmebedingungen die Erkennungsrate für die Dialog-Stichprobe mit 87 Prozent deutlich schlechter aus.

Der Vergleich zwischen der Hörerbewertung und der automatischen Akzentzuweisung zeigte zwar einen deutlichen Zusammenhang, wies aber gleichzeitig auf die Notwendigkeit weiterer Verbesserungen hin. Hauptfehlerquelle war die zu hohe Zahl von verschmolzenen unbetonten Silbenkernen.

Der Beitrag der Betonungsbeschreibung zur Gesamtanalyse im sprachverstehenden System EVAR wurde mit zwei Klassen von Worthypothesenfiltern getestet. Das DEL-Filter verwirft Worthypothesen, bei denen die Silbenstruktur des von der Worthypothese überdeckten Zeitsignals

nicht mit der lexikalischen Silbenstruktur der Worthypothese übereinstimmt. Die PRAG-Filter schränken das Lexikon über Silbenkernbereichen ein, die automatisch als stark betont markiert wurden. Um 50 Prozent der gesprochenen Wörter zu finden, konnte die notwendige Hypothesenmenge bei der besten Kombination der beiden Worthypothesenfilter um durchschnittlich 27 Prozent verringert werden.

Die neu entwickelte Lautsegmentierung wurde zunächst ohne Verwendung der Silbenkernsegmentierung optimiert. Es gelang, die zum Auffinden des gleichen Prozentsatzes von Wörtern nötige Hypothesenmenge gegenüber dem bisher verwendeten Segmentierungsverfahren deutlich zu verringern: Um beispielsweise 80 Prozent der richtigen Wörter in der Hypothesenmenge zu erkennen sind statt 17 Hypothesen pro Segment nur noch 8 Hypothesen pro Segment notwendig. Zur Verwendung der Silbenkernsegmentierung liegen noch keine aussagekräftigen Ergebnise vor.

In Kapitel 7 wurden grundlegende Untersuchungen zu einem *Modell der intonatorischen Markierung von Modus und Fokus im Deutschen* behandelt. Dabei lag der Schwerpunkt auf der Darstellung des entwickelten Prototypen-Konzepts in *semantischen Netzen*, einem maschinellen Wissensrepräsentationsschema. Bei den Untersuchungen, die in enger Kooperation mit dem DFG-Projekt "Modus-Fokus-Intonation" durchgeführt wurden, wurde das Material mit zwei grundsätzlich verschiedenen Ansätzen untersucht: Über Perzeptionsexperimente wurden prototypische Realisierungen der im Untersuchungsmaterial möglichen Kombinationen aus Satzmodus und Position des Fokus gesucht. Über statistische Kennwerte konnten die Standardrealisierungen (Kern-Prototypen) sowie die selteneren aber akzeptablen Realisierungen (Rand-Prototypen) gefunden werden. Bei den vier untersuchten Modus/Fokus-Konstellationen (Fokus auf letzter/vorletzter Phrase und Frage/Nicht-Frage) wurden sieben Prototypen festgestellt, die als spezialisierte Konzepte realisiert werden können. Es wurde gezeigt, wie mit dem so dargestellten Modell Satzhypothesen eines sprachverstenden Systems prosodisch verifiziert werden können.

9 Literaturverzeichnis

Abkürzungen:

AIPUK	Arbeitsberichte des Instituts für Phonetik der Universität Kiel
AvglPhon	Archiv für vergleichende Phonetik
ECST	European Conference on Speech Technology
DAGA	Deutsche Arbeitsgemeinschaft für Akustik
DAGM	Deutsche Arbeitsgemeinschaft Mustererkennung
FIPKM	Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München
ICASSP	International Conference on Acoustics, Speech and Signal Processing
ICPhS	International Congress of Phonetic Sciences
IEEE	Institute of Electrical and Electronics Engineers
JASA	Journal of the Acoustical Society of Amerika
JPhon	Journal of Phonetics
Ling.Ber	Linguistische Berichte
L&S	Language and Speech
Trans. ASSP	Transactions on Acoustics, Speech, and Signal Processing
ZPhon	Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung

[ADAMS 78]
Adams, C., Munro, R.: "In Search of Acoustic Correlates of Stress: Fundamental Frequency, Amplitude and Duration in the Connected Utterance of Some Native and Non-Native Speakers of English", Phonetica 35, S.125-156, 1978.

[ADRIAENS 84]
Adriaens, L.: "A Preliminary Description of German Intonation", IPO Annual Progress Report 19, S.36-41, 1984.

[AKTAS 86]
Aktas, A., Kämmerer, B., Küpper, W., Lagger, H.: "Erkennung von phonetisch ähnlichen Wörtern in einem Spracherkennungssystem für große Wortschätze", Kleinheubacher Berichte, Band 29, S.65-69, 1986.

[ALIM 87]
Alim, O., Ahmed, R.: "A Study on the Speech of Down's Syndrome Subjects Using the Acoustic Models", Proceedings ECST, Vol.1, S.348-351, Edinburgh, 1987.

[ALTMANN 84]
Altmann, H.: "Linguistische Aspekte der Intonation am Beispiel Satzmodus", in: FIPKM 19, S.132-152, 1984.

[ALTMANN 87]
Altmann, H.: "Zur Problematik der Konstitution von Satzmodi als Formtypen", in J. Meibauer (Hg.): "Satzmodus zwischen Grammatik und Pragmatik", Niemeyer, Tübingen, S.22-56, 1987.

[ALTMANN 88]
Altmann, H. (Hg.): "Intonationsforschungen", Niemeyer, Tübingen, 1988.

[ALTMANN 89a]
Altmann, H., Batliner, A., Oppenrieder, W. (Hgg.): "Zur Intonation von Modus und Fokus im Deutschen", Niemeyer, Tübingen, 1989.

[ALTMANN 89b]
Altmann, H., Batliner, A., Oppenrieder, W.: "Das Projekt 'Modus-Fokus-Intonation'. Ausgangspunkt, Konzeption und Resultate im Überblick.", in [ALTMANN 89a], S.2-19, 1989.

[ANTONIADIS 81]
Antoniadis, Z., Strube, H.: "Untersuchungen zum <<intrinsic pitch>> deutscher Vokale", Phonetica 38, S.277-290, 1981.

[ANTONIADIS 84a]
Antoniadis, Z.: "Grundfrequenzverläufe deutscher Sätze", Dissertation, Uni Göttingen, 1984.

[ANTONIADIS 84b]
Antoniadis, Z., Strube, H.: "Untersuchungen über die Form von F_0-Verläufen", Proceedings DAGA '84, S.773-776, Darmstadt, 1984.

[ANTONIADIS 84c]
Antoniadis, Z., Strube, H.: "Dynamische Beschreibung der Satzgrundfrequenzverläufe im Deutschen", Proceedings DAGA '84, S.777-780, Darmstadt, 1984.

[ANTONIADIS 84d]
Antoniadis, Z., Strube, H.: "Untersuchungen zur spezifischen Dauer deutscher Vokale", Proceedings DAGA '84, S.793-796, Darmstadt, 1984.

[ARTEMOV 78]
Artemov, V.: "Intonation und Prosodie", Phonetica 35, 301-339, 1978.

[AULL 84]
Aull, A.M.: "Lexical Stress and its Application to Large Vocabulary Speech Recognition", Master's Thesis, MIT, 1984.

[AWAD 88]
Awad, G., Diethert, G., Faber, J., Fellbaum, K., Wolf, A.: "Einsatz der Spracherkennung für motorisch Behinderte", Proceedings ITG-Fachtagung Digitale Sprachverarbeitung, ITG-Fachbericht 105, S.241-246, Bad Nauheim, 1988.

[BAKER 87]
Baker, J.: "State-of-the-Art Speech Recognition, U.S. Research and Business Update", Proceedings ECST, Vol.1, S.440-447, Edinburgh, 1987.

[BANNERT 85]
Bannert, R.: "Fokus, Kontrast und Phrasenintonation im Deutschen", Zeitschrift für Dialektologie und Linguistik 52, S.289-305, 1985.

[BANNERT 89]
Bannert, R., Hoepelman, J., Machate, J.: "Auswertung der Fokusintonation im gesprochenen Dialog", in Burkhardt, H., Höhne, K., Neumann (Hgg.): "Mustererkennung 1989", S.536-542, Informatik-Fachberichte 219, 1989.

[BARRY 81]
Barry, W.: "Prosodic Functions Revisited Again!", Phonetica 38, S.633-644, 1981.

[BATLINER 87a]
Batliner, A., Schiefer, L.: "Stimulus Category, Reaction Time, and Order Effect - An Experiment on Pitch Discrimination", in Proceedings XIth ICPhS, Vol.5, S.46-49, Tallin, 1987.

[BATLINER 88a]
Batliner, A.: "Produktion und Prädiktion. Die Rolle intonatorischer und anderer Merkmale bei der Bestimmung des Satzmodus", in [ALTMANN 88], S.207-221, 1988.

[BATLINER 88b]
Batliner, A.: "Modus und Fokus als Dimensionen einer Nonmetrischen Multidimensionalen Skalierung", in [ALTMANN 88], S.223-241, 1988.

[BATLINER 88c]
Batliner, A.: "Der Exklamativ: Mehr als Aussage oder doch nur mehr oder weniger Aussage? Experimente zur Rolle von Höhe und Position des F_0-Gipfels.", in [ALTMANN 88], S.243-271, 1988.

[BATLINER 89a]
Batliner, A., Nöth, E., Lang, R., Stallwitz, G.: "Zur Klassifikation von Fragen und Nicht-Fragen anhand intonatorischer Merkmale", Proceedings DAGA '89, S.335-338, Duisburg, 1989.

[BATLINER 89b]
Batliner, A., Nöth, E.: "The Prediction of Focus", Proceedings ECST, Vol.1, S.210-213, Paris, 1989.

[BATLINER 89c]
Batliner, A.: "Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen", in [ALTMANN 89a], S.21-70, 1989.

[BATLINER 89d]
Batliner, A.: "Fokus, Deklination und Wendepunkt", in [ALTMANN 89a], S.71-85, 1989.

[BATLINER 89e]
Batliner, A.: "Eine Frage ist eine Frage ist keine Frage. Perzeptionsexperimente zum Fragemodus im Deutschen", in [ALTMANN 89a], S.87-109, 1989.

[BATLINER 89f]
Batliner, A.: "Wieviel Halbtöne braucht die Frage? Merkmale, Dimensionen, Kategorien", in [ALTMANN 89a], S.111-162, 1989.

[BATLINER 89g]
Batliner, A., Oppenrieder, W.: "Korpora und Auswertung", in [ALTMANN 89a], S.281-330, 1989.

[BATLINER 89h]
Batliner, A.: "Ein einfaches Modell der Frageintonation und seine Folgen", in "Akten des 24. Linguistischen Kolloquiums. Bremen 4. bis 6. September 1989", Niemeyer, Tübingen, erscheint vorraussichtlich 1990.

[BECKMANN 86]
Beckmann, M.: "Stress and Non-Stress Accent", Foris Publications, Dordrecht, 1986.

[BELASCO 53]
Belasco, S.: "The Influence of Force of Articulation of Consonants on Vowel Duration", JASA 25, S.1015-1016, 1953.

[BELLMAN 72]
Bellman, R.: "Dynamic Programming", Princeton University Press, Princeton, 1972.

[BIERWISCH 66]
Bierwisch, M.: "Regeln für die Intonation deutscher Sätze", In Studia Grammatica VII: "Untersuchungen über Akzent und Intonation im Deutschen", S.99-201, Akademie-Verlag, Berlin, 1966.

[BLEAKLEY 73]
Bleakley, D.: "The Effect of Fundamental Frequency Variation on the Perception of Stress in German", Phonetica 28, S.42-59, 1973.

[BLESSER 69]
Blesser, B.: "Perception of Spectrally Rotated Speech", PhD Thesis, MIT, 1969.

[BOLINGER 58a]
Bolinger, D.: "On Intensity as a Qualitative Improvement of Pitch Accent", Lingua 7, S.175-182, 1958.

[BOLINGER 58b]
Bolinger, D.: "A Theory of Pitch Accent in English", Word 14, S.109-149, 1958.

[BRIETZMANN 84]
Brietzmann, A.: "Semantische und Pragmatische Analyse im Erlanger Spracherkennungsprojekt", Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, 1984.

[BRIETZMANN 87]
Brietzmann, A.: "Stufenweise syntaktische Analyse mit integrierter Bewertung für die kontinuierliche Spracherkennung", Arbeitsberichte des IMMD der FAU, Band 20, No. 9, Erlangen, 1987.

[BUSSMANN 83]
Bußmann, H.: "Lexikon der Sprachwissenschaft", Alfred Kröner Verlag, Stuttgart, 1983.

[CHAFE 74]
Chafe, W.: "Language and Consciousness", Language 50, S.111-133, 1974.

[CHEN 70]
Chen, M.: "Vowel Length Variation as a Function of the Voicing of the Consonant Environment", Phonetica 22, S.129-159, 1970.

[COHEN 67]
Cohen, A., 't Hart, J.: "On the Anatomy of Intonation", Lingua 19, S.177-192, 1967.

[CORSI 82]
Corsi, R.: "Speaker Recognition: A Survey", in [HATON 82], S.277-308, 1982.

[CUTLER 83a]
Cutler, A., Ladd, D. (Hgg.): "Prosody: Models and Measurements", Springer-Verlag, Berlin, 1983.

[DELATTRE 65]
Delattre, P.: "Comparing the Phonetic Features of English, German, Spanish and French - An Interim Report", Julius-Groos-Verlag, Heidelberg, 1965.

[DE MORI 83]
De Mori, R.: "Computer Models of Speech Using Fuzzy Algorithms", Plenum Press, New York, 1983.

[DE MORI 85]
De Mori, R., Suen, C. (Hgg.): "New Systems and Architectures for Automatic Speech Recognition and Synthesis", Springer Verlag, Berlin, 1985.

[DOMMELEN 82]
Dommelen van, W.: "A Contrastive Investigation of Vowel Duration in German and Dutch", Phonetica 39, S.23-35, 1982.

[DREYFUS 87]
Dreyfus, H., Dreyfus, S.: "Künstliche Intelligenz - von den Grenzen der Denkmaschine und dem Wert der Intuition", Reinbek, Rowohlt, 1987.

[DUBNOWSKI 76]
Dubnowski, J., Schafer, R., Rabiner, L.: "Real-Time Digital Hardware Pitch Detector", IEEE Trans. ASSP-24, S.2-8, 1976.

[DUDEN 73]
Dudenredaktion, Grebe, P.: "Der Duden in 10 Bänden, Band 4: Die Grammatik", Bibliographisches Institut, Mannheim, 1973.

[DUDEN 74]
Dudenredaktion, Mangold, M.: "Der Duden in 10 Bänden, Band 6: Das Aussprachewörterbuch", Bibliographisches Institut, Mannheim, 1974.

[EHRLICH 86]
Ehrlich, U.: "Ein Lexikon für das natürlich-sprachliche System EVAR", Arbeitsberichte des IMMD der FAU, Band 19, No. 3, Erlangen, 1986.

[EHRLICH 88]
Ehrlich, U., Niemann, H.: "Using Semantic and Pragmatic Knowledge for the Interpretation of Syntactic Constituents", in [NIEMANN 88a], S.485-490, 1988.

[EHRLICH 89]
Ehrlich, U.: "Bedeutungsanalyse in einem sprachverstehenden System unter Einbeziehung kontextueller Restriktionen", Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, 1989.

[FÉRY 88]
Féry, C.: "Rhythmische und tonale Struktur der Intonationsphrase", in [ALTMANN 88], S.41-64, 1988.

[FISCHER 88]
Fischer, V.: "Variable Länge und Fortschaltzeit der Analysefenster für die automatische Lauterkennung", Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1988.

[FISCHER 89]
Fischer, V.: "Eine explizite Segmentierung von Sprachsignalen für die Lautklassifikation mit Markov-Modellen", Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1989.

[FISCHER-JØRGENSEN 64]
Fischer-Jørgensen, E.: "Sound Duration and Place of Articulation", ZPhon 17, S.175-207, 1964.

[FRY 55]
Fry. D.: "Duration and Intensity as Physical Correlates of Linguistic Stress", JASA 27, S.765-768, 1955.

[FRY 58]
Fry. D.: "Experiments in the Perception of Stress", L&S 1, S.126-152, 1958.

[GARTENBERG 87]
Gartenberg, R.: "Artikulatorische Faktoren in der Ausprägung von Intonationsmustern", Magisterarbeit, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1987.

[GLASGOW 52]
Glasgow, G.: "A Semantic Index of Vocal Pitch", Speech Monographs 19, Columbia, Miss., S.64-68, 1952.

[GLAS 75]
Glas, R.: "Das LIMAS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache", Linguistische Berichte 40, vieweg, S.63-66, 1975.

[GLINZ 65]
Glinz, H.: "Grundbegriffe und Methoden inhaltsbezogener Text- und Sprachanalyse", Düsseldorf, 1965.

[GLINZ 79]
Glinz, E., Glinz, H.: "Der Sprachunterricht im 7. und 8. Schuljahr", Zürich, 1979.

[HANSEN 89]
Hansen, J., Clements, M.: "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition", Proceedings ICASSP, S.266-269, 1989.

[HATON 82]
Haton, J. (Hg.): "Automatic Speech Analysis and Recognition", Reidel, Dordrecht, 1982.

[HECKER 71]
Hecker,M., Kreul, E.: "Description of the Speech of Patients with Cancer of the Vocal Folds. Part 1: Measures of Fundamental Frequency", JASA 49, S.1275-1282, 1971.

[HEIKE 69]
Heike, G.: "Suprasegmentale Analyse", N.G. Elwert Verlag, Marburg 1969.

[HELBIG 74]
Helbig, G., Buscha, J.: "Deutsche Grammatik", VEB Verlag Enzyklopädie, Leipzig, 1974.

[HESS 83]
Hess, W.: "Algorithms and Devices for Pitch Determination of Speech Signals", Springer Verlag, Berlin, 1983.

[HETTWER 85]
Hettwer, G.: "Zur Hörbarkeit von Grundfrequenzfehlern bei linearen Prädiktionsvokoder", Dissertation, TU Berlin, 1985.

[HEUNISCH 86]
Heunisch, S.: "Untersuchung der Brauchbarkeit verschiedener Algorithmen zur Bestimmung der Grundfrequenz für die prosodische Analyse", Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1986.

[HITZENBERGER 86]
Hitzenberger, L., Ulbrand, R., Kritzenberger, H., Wenzel, P.: "FACID Fachsprachlicher Corpus informationsabfragender Dialoge", Universität Regensburg, FG Linguistische Informationswissenschaft, 1986.

[HOUSE 53]
House, A., Fairbanks, G.: "The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels", JASA 25, S.105-113, 1953.

[HOUSE 88]
House, A.: "The Recognition of Speech by Machine - a Bibliography", Academic Press, London, 1988.

[HOUSE 87]
House, D., Bruce, G., Lacerda, F., Lindblom, B.: "Automatic Prosodic Analysis for Swedish Speech Recognition", Proc. ECST, Vol.2, S.215-218, 1987.

[HUTTENLOCHER 84]
Huttenlocher, D., Zue, V.: "A Model of Lexical Access from Partial Phonetic Information", Proceedings ICASSP, Vol. 2, S.26.4.1-26.4.4, 1984.

[INDEFREY 88]
Indefrey, H.: "Untersuchung von Algorithmen zur Grundfrequenzbestimmung von Sprachsignalen unter Verwendung des Ausgangssignals eines Laryngographen als Referenzgröße", Dissertation, Fakultät für Elektrotechnik und Informationstechnik, Technischen Universität München, 1988.

[ISAČENKO 66]
Isačenko, A., Schädlich, H.: "Untersuchungen über die deutsche Satzintonation", in Studia Grammatica VII: "Untersuchungen über Akzent und Intonation im Deutschen", S. 7-67, Akademie-Verlag, Berlin, 1966.

[JELINEK 85]
Jelinek, F.: "The Development of an Experimental Discrete Dictation Recognizer", Proceedings of the IEEE 11/73, S.1616-1624, 1985.

[JESPERSEN 32]
Jespersen, O.: "Lehrbuch der Phonetik", 5. Aufl., Teubner Verlag, Leipzig-Berlin, 1932.

[JOBST 89]
Jobst, E.: "Graphische Darstellung von Zwischenergebnissen der automatischen Sprachverarbeitung", Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1989.

[KEMPELEN 70]
von Kempelen, W.: "Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine", Faksimile-Neudruck der Ausgabe Wien 1791, Friedrich Frommann Verlag, Stuttgart, 1970.

[KIESSLING 89]
Kiessling, A.: "Ein interaktives System zur periodenweisen Bestimmung der Grundfrequenz", Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1989.

[KIM 70]
Kim, C.: "A Theory of Aspiration", Phonetica 21, S.107-116, 1970.

[KIPARSKY 66]
Kiparsky, P.: "Über den deutschen Akzent", In Studia Grammatica VII: "Untersuchungen über Akzent und Intonation im Deutschen", S.69-98, Akademie-Verlag, Berlin 1966.

[KLATT 73]
Klatt, D.: "Interaction Between Two Factors that Influence Vowel Duration", JASA 54, S.1102-1104, 1973.

[KLATT 75]
Klatt, D.: "Vowel Lengthening is Syntactically Determined in a Connected Discourse", JPhon 3, S.129-140, 1975.

[KLATT 76]
Klatt, D.: "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence" JASA 59, S.1208-1221, 1976.

[KLEIN 80]
Klein, W.: "Der Stand der Forschung zur deutschen Satzintonation", Ling.Ber. 68, S.3-33, 1980.

[KLEIN 82]
Klein, W.: "Einige Bemerkungen zur Frageintonation", Deutsche Sprache, S.289-310, 1982.

[KLINGHOLZ 88]
Klingholz, F.: "Sprachsignalstatistik zur Erfassung der Stimmgüte", Proceedings ITG-Fachtagung Digitale Sprachverarbeitung, ITG-Fachbericht 105, S.283-288, Bad Nauheim, 1988.

[KOHLER 77]
Kohler, K.: "Einführung in die Phonetik des Deutschen", Erich Schmidt Verlag, Berlin, 1977.

[KOHLER 81]
Kohler, K., Schäfer, K., Thon, W., Timmermann, G.: "Sprechgeschwindigkeit in Produktion und Perzeption", AIPUK 16, S.139-172, 1981.

[KOHLER 82]
Kohler, K., Krützmann, U., Reetz, H., Timmermann, G.: "Sprachliche Determinanten der signalphonetischen Dauer", AIPUK 17, S.3-35, 1982.

[KOHLER 83]
Kohler, K.: "Prosodic Boundary Signals in German", Phonetica 40, S.89-134, 1983.

[KOHLER 87]
Kohler, K.: "Funktionen von F0-Gipfeln im Deutschen", in [TILLMAN 87], S.133-140, 1987.

[KOMPE 89a]
Kompe, R.: "Modifikation der Worthypothesen-Bewertungen aufgrund prosodischer Information", Studienarbeit, Lehrstuhl für Informatik 5 (Musterkennung), Universität Erlangen, 1989.

[KOMPE 89b]
Kompe, R.: "Ein Mehrkanalverfahren zur Berechnung der Grundfrequenzkontur unter Einsatz der Dynamischen Programmierung", Diplomarbeit, Lehrstuhl für Informatik 5 (Musterkennung), Universität Erlangen, 1989.

[KORI 87]
Kori, S., Farnetani, E., Cosi, P.: "A Perspective on Relevance and Aplication of Prosodic Information to Automatic Speech Recognition in Italian.", Proceedings ECST, Vol.1, S.211-214, Edinburgh, 1987.

[KUHN 87]
Kuhn, T.: "Implementation von Algorithmen zum Vergleich von eindimensionalen Mustern", Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1987.

[KUHN 89]
Kuhn, T.: "Eine Suchstrategie zur kontextfreien Analyse von Worthypothesen", Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1989.

[KUNZMANN 88a]
Kunzmann, S., Kuhn, T., Niemann, H.: "An Experimental Environment for Generating Word Hypotheses in Continuous Speech", in [NIEMANN 88a], S.311-316, 1988.

[KUNZMANN 88b]
Kunzmann, S., Kuhn, T., Niemann, H.: "An Experimental Environment for the Generation and Verification of Word Hypotheses in Continuous Speech", Speech Communication 7, S.381-388, Elsevier Science Publishers B.V., North Holland, 1988.

[KUNZMANN 89a]
Kunzmann, S., Kuhn, T.: "Ordnungsoperationen zur Neubewertung von Hypothesen aus der automatischen Worterkennung", Proceedings DAGA '89, S.339-342, 1989.

[KUNZMANN 90]
Kunzmann, S.: "Die Worterkennung in einem Dialogsystem für kontinuierlich gesprochene Sprache", Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, 1990.

[LADD 83a]
Ladd, D., Silverman, K.: "Vowel Intrinsic Pitch in Paragraph Context", Proceedings Xth ICPhS, S.609, 1983.

[LADD 83b]
Ladd, D.: "Phonological Features of Intonational Peaks", Language 59, S.721-759.

[LANG 87]
Lang, R.: "Automatische Bestimmung verschiedener Satzmodi anhand von Grundfrequenzverläufen", Studienarbeit, Lehrstuhl für Informatik 5 (Musterkennung), Universität Erlangen, 1987.

[LEA 72]
Lea, W.: "Use of Syntactic Segmentation and Stressed Syllable Location in Phonemic Recognition", 84th Meeting of the Acoustical Society of America, Miami, Florida, 1972.

[LEA 73a]
Lea, W.: "Evidence that Stressed Syllables Are the Most Readily Decoded Portions of Continous Speech, JASA 55, S.436(A), 1973.

[LEA 73b]
Lea, W.: "Segmental and Suprasegmental Influences on Fundamental Frequency Contours", in Hyman, L. (Hg.): "Consonant Types and Tone", Southern California Occasional Papers in Luinguistics No. 1, 1973

[LEA 75]
Lea, W., Medress, M., Skinner, T.: "A Prosodically Guided Speech Understanding Strategy", IEEE Trans. ASSP-23, S.30-38, 1975.

[LEA 80a]
Lea, W. (Hgg.): "Trends in Speech Recognition", Prentice Hall, N. J., 1980.

[LEA 80b]
Lea, W.: "Prosodic Aids to Speech Recognition", in [LEA 80a], S.166-205, 1980.

[LECLUSE 77]
Lecluse, F.: "Elektroglottografie", Dissertation, Universität Rotterdam, 1977.

[LEHISTE 59]
Lehiste, I., Peterson, G.: "Vowel Amplitude and Phonemic Stress in American English", JASA 31, S.428-435, 1959.

[LEHISTE 70]
Lehiste, I.: "Suprasegmentals", M.I.T. Press, Cambridge, 1970.

[LEHISTE 72]
Lehiste, I.: "Timing of Utterances and Linguistic Boundaries", JASA 51, S.2018-2034, 1972.

[LIEBERMAN 60]:
Lieberman, P.: "Some Acoustic Correlates of Word Stress in American English", JASA 32, S.451-454, 1960.

[LIEBERMAN 65]
Lieberman, P.: "On the Acoustic Basis of the Perception of Intonation by Linguists", Word 21, S.40-54, 1965.

[LIEBERMAN 67]
Lieberman, P.: "Intonation, Perception, and Language", M.I.T. Press, Cambridge, 1967.

[LIEBERMAN 85a]
Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., Miller, M.: "Measures of the Sentence Intonation of Read and Spontaneous Speech in American Englisch", JASA 77 (2), S.649-657, 1985.

[LIEBERMAN 85b]
Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., Miller, M.: "Reply to Bruno H. Repp", JASA 78 (3), S.1116-1117, 1985.

[LIEBERMAN 86]
Lieberman, P.: "Alice in Declinationland - A Reply to Johan 't Hart", JASA 80 (6), S.1840-1842, 1986.

[LINDBLOM 63]
Lindblom, B.: "Spectrographic Study of Vowel Reduction", JASA 35, S.1173-1181, 1963.

[LINDNER 69]
Lindner, G.: "Einführung in die Experimentelle Phonetik", Akademieverlag, Berlin, 1969.

[LINDNER 81]
Lindner, G.: "Grundlagen und Anwendung der Phonetik", Akademieverlag, Berlin, 1981.

[LÖTSCHER 81]
Lötscher, A.: "Satzakzentuierung und Tonhöhenbewegung im Standarddeutschen", Ling.Ber. 74, S.20-34, 1981.

[LÖTSCHER 83]
Lötscher, A.: "Satzakzent und funktionale Satzperspektive im Deutschen", Niemeyer, Tübingen, 1983.

[LÖTSCHER 85]
Lötscher, A.: "Akzentuierung und Thematisierbarkeit von Angaben", Ling.Ber. 97, S.228-251, 1985.

[LÖWE 87]
Löwe, M., Schmidt, G., Wilhelm, R. (Hgg.): "Umdenken in der Informatik", Elefanten Press, Berlin, 1987.

[MAACK 37]
Maack, A.: "Phonometrische Untersuchungen über Beziehungen des Akzents zum Melodieverlauf", AvglPhon 1, S.213-221, 1937.

[MAACK 49a]
Maack, A.: "Die spezifische Lautdauer deutscher Sonanten", ZPhon 3, S.190-232, 1949.

[MAACK 49b]
Maack, A.: "Der Einfluß der Betonung auf die Lautdauer deutscher Sonanten", ZPhon 3, S.341-356, 1949.

[MAACK 53]
Maack, A.: "Die Beeinflussung der Sonantendauer durch die Nachbarkonsonanten", ZPhon 7, S.104-128, 1953.

[MAEDA 76]
Maeda, S.: "A Characterization of American English Intonation", Dissertation, M.I.T., 1976.

[MARIANI 87]
Mariani, J.: "Speech Technology in Europe", Proceedings ECST, Vol.1, S.431-439, Edinburgh, 1987.

[MARIANI 89]
Mariani, J.: "Recent Advances in Speech Processing" Proceedings ICASSP, S.429-440, 1989.

[MARKEL 72]
Markel, J.: "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. Audio Electroacoustics 20, S.367-377, 1972.

[MARTIN 79]
Martin, P.: "Automatic Location of Stressed Syllable in French", Current Issues in Linguistic Theory 9, S.1091-1094.

[MATHESIUS 29]
Mathesius, V.: "Zur Satzperspektive im modernen Englisch", Archiv für das Studium der neueren Sprachen und Literaturen 155, S.202-210, 1929.

[MCGONEGAL 77]
McGonegal, C., Rabiner, L., Rosenberg, A.: "A Subjective Evaluation of Pitch Detection Methods using LPC Synthesized Speech", IEEE Trans ASSP 25, S.221-229, 1977.

[MERCIER 88]
Mercier, G., Cozannet, A., Vaissièrre, J.: "Recognition of Speaker-dependent Continous Speech with KEAL-NEVEZH", in [NIEMANN 88a], S.459-463, 1988.

[MERMELSTEIN 75]
Mermelstein, P.: "Automatic Segmentation of Speech into Syllabic Units", JASA 58, S.880-883, 1975.

[MILLER 75]
Miller, N.: "Pitch Detection by Data Reduction", IEEE Trans. ASSP-23, S.72-79, 1975.

[MÖBIUS 87]
Möbius, B., Zimmermann, A., Hess, W.: "Untersuchungen zu mikroprosodischen Grundfrequenzvariationen im Deutschen", in [TILLMAN 87], S.102-110, 1987.

[MOHR 71]
Mohr, B.: "Intrinsic Variations in the Speech Signal", Phonetica 23, S.65-93, 1971.

[MORTON 65]
Morton, J., Jassem, W.: "Acoustic Correlates of Stress", L&S 8, S.159-181, 1965.

[MUDLER 86]
Mudler, J.: "Ein Konzept für die Nutzung prosodischer Information bei der automatischen Spracherkennung", in Hartmann, G. (Hgg.): "Mustererkennung 1986", S.134-138, Informatik-Fachberichte 125, Springer-Verlag, Berlin, 1986.

[MÜHLFELD 86]
Mühlfeld, R.: "Verifikation von Worthypothesen", Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, 1986.

[MUELLER 81]
Mueller, J.: Kurzvortrag, gehalten auf dem zweiten NATO Advanced Study Institute 'Automatic Speech Analysis and Recognition', Bonas, 1981.

[NELSON 85]
Nelson, D.: "Use of Voice Recognition to support the Collection of Product Data", Proceedings Speech Tech '85, S.268-272, New York, 1985.

[NIEMANN 74]
Niemann, H.: "Methoden der Mustererkennung", Akademische Verlagsgesellschaft, Frankfurt, 1974.

[NIEMANN 81]
Niemann, H.: "Pattern Analysis", Springer, Berlin, 1981

[NIEMANN 85]
Niemann, H., Brietzmann, A., Mühlfeld, R., Regel, P., Schukat, G.: "The Speech Understanding and Dialog System EVAR", in [DE MORI 85], S.271-302, 1985.

[NIEMANN 87a]
Niemann, H., Bunke, H.: "Künstliche Intelligenz in Bild- und Sprachanalyse", Teubner, Stuttgart, 1987.

[NIEMANN 87b]
Niemann, H.:"Entwicklung und Realisierung eines Dialogmoduls für ein System zum Verstehen kontinuierlich gesprochener deutscher Sprache", DFG-Antrag, Lehrstuhl für Informatik 5 (Mustererkennung), 1987.

[NIEMANN 88a]
Niemann, H., Lang, M., Sagerer G. (Hgg.): "Recent Advances in Speech Understanding and Dialog Systems", Springer-Verlag, Berlin, 1988.

[NIEMANN 88b]
Niemann, H., Brietzmann, A., Ehrlich, U., Posch, S., Regel, P., Sagerer, G., Schukat-Talamazzini, G.: "A Knowledge Based Speech Understanding System", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 2, No. 2, S.321-350, 1988.

[NÖTH 85]
Nöth, E.: "Einsatzmöglichkeiten der Erfahrung von Spektrogrammlesern für die automatische Lauterkennung", Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1985.

[NÖTH 87]
Nöth, E., Batliner, A., Lang, R., Oppenrieder, W.: "Automatische Grundfrequenzanalysen und Satzmodusdifferenzierung", in [TILLMAN 87], S.59-66, 1987.

[NÖTH 88a]
Nöth, E., Schmölz, S., Niemann, H.: "Prosodic Features in German Speech: Stress Assignment by Man and Machine", in [NIEMANN 88a], S.101-106, 1988.

[NÖTH 88b]
Nöth, E., Kompe, R.: "Der Einsatz prosodischer Information im Spracherkennungssystem EVAR", in Bunke, H., Kübler, O., Stucki, P. (Hgg.): "Mustererkennung 1988", S.2-9, Informatik-Fachberichte 180, Springer-Verlag, Berlin, 1988.

[NÖTH 89a]
Nöth, E., Kompe, R.: "Verbesserung der Worterkennung mit prosodischer Information", Proceedings DAGA '89, S.343-346, Duisburg, 1989.

[NOLL 67]
Noll, A.: "Cepstrum Pitch Determination", JASA 41, S.293-309, 1967.

[NORTH 82]
North, R., Lea, W.: "Application of advanced Speech Technology in Manned Penetration Bombers", Honeywell Systems & Research Center Report AFWAL, TR-82-3004.

[OPPENHEIM 75]
Oppenheim, A., Schafer, R.: "Digital Signal Processing", Prentice-Hall, Englewood Cliffs, 1975.

[OPPENRIEDER 88a]
Oppenrieder, W.: "Intonatorische Kennzeichnung von Satzmodi", in [ALTMANN 88], S.169-205, 1988.

[PETERSON 60]
Peterson, G., Lehiste, I.: "Duration of Syllable Nuclei in English", JASA 32, S.693-703, 1960.

[PIERACCINI 86]
Pieraccini, R.: "Lexical Stress and Speech Recognition", 112th Meeting of th Acoustical Society of America, Anaheim, 1986

[PIERCE 69]
Pierce, J.: "Whither Speech Recognition", JASA 4/46, S.1049-1051, 1969.

[PIERREHUMBERT 79]
Pierrehumbert, J.: "The Perception of Fundamental Frequency Declination", JASA 66, S.363-369, 1979.

[PIERREHUMBERT 80]
Pierrehumbert, J.: "The Phonology and Phonetics of English Intonation", Dissertation, M.I.T., 1980.

[PORT 81]
Port, R.: "Linguistic Timing Factors in Combination", JASA 69, S.262-274, 1981.

[RABINER 76]
Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C.: "A Comparative Study of Several Pitch Detection Algorithms", IEEE Trans. ASSP 24, S.399-413, 1976.

[RABINER 78]
Rabiner, L., Schafer, R.: "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, 1978.

[REETZ 89]
Reetz, H.: "A Fast Expert Program for Pitch Extraction", Proceedings ECST, Vol.1, S.476-479, Paris, 1989.

[REGEL 88]
Regel, P.: "Akustisch-phonetische Transkription für die automatische Spracherkennung", VDI-Verlag, Düsseldorf, 1988.

[REPP 85]
Repp, B.: "Critique of 'Measures of the Sentence Intonation of Read and Spontaneous Speech in American English' [J. Acoust. Soc. Am. 77, 649-657, (1985)]", JASA 78 (3), S.1114-1116, 1985.

[RIETVELD 75]
Rietveld, A.: "Untersuchungen zur Vokaldauer im Deutschen", Phonetica 31, S.248-258, 1975.

[ROSS 74]
Ross, M., Shaffer, H., Cohen, A., Freudberg, R., Manley, H.: "Average Magnitude Difference Function Pitch Extractor", IEEE Trans. ASSP-22, S.353-362, 1974.

[ROYÉ 83]
Royé, H.: "Segmentierung und Hervorhebung in gesprochener deutscher Standardaussprache", PHONAI Band 27, Niemeyer, Tübingen, 1983.

[RUSKE 88]
Ruske, G.: "Automatische Spracherkennung", Oldenbourg Verlag, München, 1988.

[SAGERER 88]
Sagerer, G., Kummert, F.: "Knowledge Based Systems for Speech Understanding", in [NIEMANN 88a], S.421-458, 1988.

[SAGERER 89]
Sagerer, G.: "Automatisches Verstehen von Sprache", Habilitationsschrift, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1989.

[SCHIEFER 88]
Schiefer, L., Batliner, A.: "Intonation, Ordnungseffekt und das Paradigma der Kategorialen Wahrnehmung ", in [ALTMANN 88], S.273-291, 1988.

[SCHMÖLZ 85]
Schmölz, S.: "Zur automatischen Bestimmung der Sprechgeschwindigkeit", Studienarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1985.

[SCHMÖLZ 87]
Schmölz, S.: "Zur automatischen Bestimmung der Satzbetonung", Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen, 1987.

[SENEFF 78]
Seneff, S.: "Real-time harmonic pitch detector", IEEE Trans. ASSP-26, S.358-364, 1978.

[SIEVERS 85]
Sievers, E.: "Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen", Leipzig, 1885.

[SONDHI 68]
Sondhi, M.: "New Methods of Pitch Extraction", IEEE Trans. Audio and Electroacoustics 16, S.262-266, 1968.

[SPENDER 80]
Spender, D.: "Man Made Language", Routledge & Kegan Paul, Boston, 1980.

[STALLWITZ 89]
Stallwitz, G.: "Erstellung einer symbolischen Grundfrequenzbeschreibung", Magisterarbeit, Institut für Germanistische Linguistik, Universität Erlangen, 1989.

[STETSON 51]
Stetson, R.: "Motor Phonetics", North-Holland, Amsterdam, 1951.

[STOCK 76]
Stock, E., Zacharias, C.: "Deutsche Satzintonation", Leipzig, 1976.

['T HART 86]
't Hart, J.: "Declination has not been Defeated - A Reply to Lieberman et al.", JASA 80 (6), S.1838-1840, 1986.

[THORSEN 79]
Thorson, N.: "Interpreting Raw Fundamental-Frequency Tracings of Danish", Phonetica 36, S.57-78, 1979.

[TIFFANY 59]
Tiffany, W.: "Nonrandom Sources of Variation in Vowel Quality", Journal of Speech and Hearing Research 2, S.305-317, 1959.

[TILLMANN 64]
Tillmann, H.: "Das phonetische Silbenproblem. Eine Theoretische Untersuchung", Dissertation, Universität Bonn, 1964.

[TILLMANN 80]
Tillmann, H., Mansell, P.: "Phonetik", Klett-Verlag, Stuttgart, 1980.

[TILLMAN 87]
Tillmann, H., Willée, G.(Hgg.): "Analyse und Synthese gesprochener Sprache", Georg Ohms Verlag, Hildesheim, 1987.

[TREHERN 87]
Trehern, J., Jack, M., Laver, J. Hiller, S.: "Acoustic Screening for Vocal Pathology with a Boltzmann Machine", Proceedings ECST, Vol.1, S.269-272, Edinburgh, 1987.

[TURKLE 84]
Turkle, S.: "Die Wunschmaschine - Vom Entstehen der Computerkultur", Rowohlt, Reinbek, 1984.

[UHMANN 88]
Uhmann, S.: "Akzenttöne, Grenztöne und Fokussilben. Zum Aufbau eines phonologischen Intonationssystems für das Deutsche", in [ALTMANN 88], S.65-88, 1988.

[VAISSIÈRE 82]
Vaissière, J.: "A Suprasegmental Component in a French Speech Recognition System: Reducing the Number of Lexical Hypotheses and Detecting the Main Boundary", Recherches/Acoustique, CNET, Vol. 7, S.109-125, 1982/83.

[VAISSIÈRE 83]
Vaissière, J.: "Language-Independent Prosodic Features", in [CUTLER 83a], S.53-66, 1983.

[VAISSIÈRE 88]
Vaissière, J.: "The Use of Prosodic Parameters in Automatic Speech Recognition" in [NIEMANN 88a], S.71-99, 1988.

[VON ESSEN 56]
Von Essen, O.: "Grundzüge der hochdeutschen Satzintonation", Henn, Ratingen, 1956.

[WAIBEL 86]
Waibel, A.: "Prosody and Speech Recognition", Dissertation, Carnegie-Mellon University, Pittsburgh, USA, 1986.

[WEIZENBAUM 76]
Weizenbaum, J.: "Computer Power and Human Reason", W. H. Freeman and Company, San Francisco, 1976.

[WILLIAMS 88]
Williams, B., Dalby, J.: "Feature-based Automatic Syllable and Stress Detection", Vortrag auf dem 115. Treffen der Acoustical Society of America, Seattle, eine Kurzfassung findet sich in JASA 83 (Suppl. 1), S.54, 1988.

[WINKLER 73]
Winkler, C.: "T. Die Klanggestalt des Satzes", In [Duden 73] S.637-666, 1973.

[WODARZ 71]
Wodarz, H., Wodarz-Magdics, K.: "Beiträge zu einer kontrastiven Phonetik des Deutschen und Ungarischen. I.Vokalquantität und Vokalqualität", Phonetica 24, S.116-124, 1971.

[WOHLAND 89]
Wohland, G: "Das demokratische Potential der Neuen Fabrik", FIFF KOMMUNIKATION 1/89, S.16-19, 1989.

[WUNDERLICH 88]
Wunderlich, D.: "Der Ton macht die Melodie - Zur Phonologie der Intonation des Deutschen", in [ALTMANN 88], S.1-40, 1988.

[ZWICKER 67]
Zwicker, E., Feldkeller, R.: "Das Ohr als Nachrichtenempfänger", Hirzel Verlag, Stuttgart, 1967.

Anhang

Anhang A: Liste der vereinbarten Lautklassen des Akustik-Phonetik-Moduls in EVAR

Anhang B: Liste der erkennbaren Lautkomponenten des Akustik-Phonetik-Moduls in EVAR

Anhang C: Transkriptionszeichen für die enge Transkription mit der Erlanger Kodierung

Anhang D: Liste der Äußerungen der Pragmatik-Stichprobe (siehe Kap.3.1). Die manuell bestimmten pragmatisch wichtigen Wörter sind unterstrichen.

Anhang E: Einige Anmerkungen zur Komplexität der Algorithmen.

Anhang A

Liste der vereinbarten Lautklassen des Akustik-Phonetik-Moduls in EVAR

Vokale

Offenes I	[ɪ]	I.	bist	Geschlossenes I	[i]	IH	vital
Offenes E	[ɛ]	E.	hätte	Geschlossenes E	[e]	EH	Methan
Murmellaut	[ə]	ER	halte	Helles A	[a]	A.	hat
Dunkles A	[ɑ]	AH	Wal	Abgeschwächtes A	[ɐ]	AR	Ober
Offenes O	[ɔ]	O.	Post	Geschlossenes O	[o]	OH	Moral
Offenes U	[ʊ]	U.	Pult	Geschlossenes U	[u]	UH	kulant
Ä-Laut	[æ]	AE	(engl. bat)	Offenes Ö	[ø]	Q.	Götter
Offenes Ü	[ʏ]	Y.	füllt	Geschlossenes Ü	[y]	YH	Rübe

Unsilbische Vokale

Unsilbisches I	[i̯]	IJ	Studie	Unsilbisches A	[ɐ̯]	AJ	Uhr
Unsilbisches O	[o̯]	OJ	loyal	Unsilbisches U	[u̯]	UJ	Statue
Unsilbisches Ü	[y̯]	YI	Etui				

Nasale Vokale

Nasales E	[ɛ̃]	EN	(frz. timbre)	Nasales A	[ɑ̃]	AN	(frz. penser)
Nasales O	[õ]	ON	(frz. fondue)	Nasales Ö	[œ̃]	QN	(frz. lundi)

Diphthonge

Von A nach I	[a͡ɪ]	AI	weit	Von A nach U	[a͡ʊ]	AU	Haut
Von O nach Ü	[ɔ͡ʏ]	OY	Heu				

Zusatzlaute

Hauptakzent	[ˈ]	'.	be'kommen	Langer Vokal	[ː]	:.	ka:m
Stimmritzenverschluß	[ˀ]	+.	be+achten	Pause	[ ]	-.	---

Anhang A (Fortsetzung)

Konsonanten

Stimmloses F	[f]	F.	Faß	Stimmhaftes F	[v]	V.	Was
Stimmhaftes W	[w]	W.	(engl. wind)	Stimmloses S	[s]	S.	Was
Stimmhaftes S	[z]	Z.	Hase	Stimmloses SCH	[ʃ]	SH	Schal
Stimmhaftes SCH	[ʒ]	ZH	(engl. measure)	ich-Laut	[ç]	XI	ich
ach-Laut	[x]	XA	ach	Stimmhaftes J	[j]	J.	jubeln
Stimmloses H	[h]	H.	Haare	Reibe R	[ʁ]	RA	Haare
Zäpfchen R	[ʀ]	R.		Zungenspitzen R	[r]	RR	
L-Laut	[l]	L.	Laut	Silbisches L	[l̩]	LE	Nabel

Nasale Konsonanten

M-Laut	[m]	M.	Mast	Silbisches M	[m̩]	ME	großem
N-Laut	[n]	N.	Naht	Silbisches N	[n̩]	NE	baden
NG-Laut	[ŋ]	NG	Schwung				

Verschlußlaute

Stimmloses P	[p]	P.	Pakt	Stimmhaftes B	[b]	B.	Ball
Stimmloses T	[t]	T.	Tal	Stimmhaftes D	[d]	D.	dann
Stimmloses K	[k]	K.	kalt	Stimmhaftes G	[g]	G.	Gast

Affrikate

PF-Laut	[p͡f]	PF	Pfahl	TS-Laut	[t͡s]	TS	Zahl
Stimmloses TSCH	[t͡ʃ]	C.	Cello	Stimmhaftes DSCH	[d͡ʒ]	CH	Gin

Anhang B

Liste der erkennbaren Lautkomponenten des Akustik-Phonetik-Moduls in EVAR

Lautkomponenten "stimmhaft nicht-frikativ"

[ɪ̯]	IH	vital	[ɪ]	I.	bist	[e]	EH	Methan
[ɛ]	E.	hätte	[ə]	ER	halte	[o]	OH	Moral
[ɔ]	O.	Post	[ʊ]	UH	kulant	[ʊ̯]	U.	Pult
[ʏ]	YH	Rübe	[Y]	Y.	füllt	[a]	A.	hat
[ɐ]	AR	oder	[l]	L.	Laut	[m]	M.	Mast
[n]	N.	Naht	[n̩]	NE	baden	[ŋ]	NG	Schwung
[ø]	Q.	Götter	[œ]	QH	mögen	[R]	R.	Traum

Lautkomponenten "stimmhaft-frikativ"

[z]	Z.	Hase	[v]	V.	was	[ʒ]	ZH	(engl. measure)
	CA	Echo		CI	(ich im)			

Lautkomponenten "stimmlos"

[ʁ]	RA	Haare	[ʃ]	SH	Schal	[x]	XA	ach
[ç]	XI	ich	[f]	F.	Faß	[h]	H.	Haus
[s]	S.	was						

	T.	Burst von [t]		K.	Burst von [k]		P.	Burst von [p]
	D.	Burst von [d]		G.	Burst von [g]		B.	Burst von [b]
	TH	Behauchung von [t]		KH	Behauchung von [k]		PH	Behauchung von [p]

Lautkomponenten bzgl. Pause

	-.	Pause						
	TP	Verschluß von [t]		KP	Verschluß von [k]		PP	Verschluß von [p]
	DP	Verschluß von [d]		GP	Verschluß von [g]		BP	Verschluß von [b]

Anhang C

Transkriptionszeichen für die enge Transkription mit der Erlanger Kodierung:

1) Die Transkription ist so oberflächengetreu wie möglich. Die Transkription erfolgt punktgenau, d.h. nicht mit einem festen Fortschaltraster von z.B. 10 Millisekunden, sondern im Fortschaltraster 1 Abtastpunkt. Die ersten beiden Symbole dienen der breiten Transkription, danach können beliebig viele Zeichen folgen, die zur engeren Transkription dienen.

2) Zusätzlich zu den in [REGEL 88, S.134ff] eingeführten Symbolen zur breiten Transkription von Lauten und Lautkomponenten (1. und 2. Stelle) werden folgende Kodierungen eingeführt:

   TV = Verschluß von TS

   GH = (velarer) breathy Plosiv (z.b. im Fränkischen)
   (GP Verschlußphase G Burst GH Aspiration)

   IR = hoher Schwa

   RF = Flap, geflapptes /R/

   DF = Flap, geflapptes /D/

   GL = Glide

   z.B. GLXA=Glide nach XA, kein Zungenkontakt, reduzierte Bewegung

   XX = undefinierter Laut

   SZ = Blasen

   z.B. SZB=labiales Blasen (Suppenlaut)

   SZGLXA=Blasen bei angedeuteter Mundraumkonfiguration

   FN = "Entstimmungspause" zwischen Frikativen und Nasalen oder Nasalen und Vokalen

3) Die weiteren Stellen dienen der engeren Transkription und zusätzlichen Angaben

   ..1 = linksrandige Transition

   ..2 = rechtsrandige Transition

   ..! = akzentuiert

   ..!! = stark akzentuiert

   ..S = Periodizität vorhanden
   z.B. Anfangsphase von auditiv stimmlosen Frikativen
   Verschlußphase von stimmlosen Plosiven

   ..H = aspiriert (stl) oder behaucht (sth)

   ..C = laryngal, creaky

   ..T = entstimmt (bei Sonoranten)

Anhang C (Fortsetzung)

X.Y = Sonorant X koartikuliert mit Laut Y
z.B. N.U
soll dieser Laut noch weiter charakterisiert werden, so wird die 3. Stelle wie eine 1. Stelle behandelt z.B. N.U.T

..N = nasaliert

..E = entrundet

..Z = zentralisiert

..R = retroflex

..D = dental

..F = Engebildung vorhanden

..B = labialisiert (auch gerundet)

..GL = angenäherte Zielartikulation, "Glide" nach X, reduzierte Bewegung

..P = mit Schließbewegung

..- = weiter hinten artikuliert

..+ = weiter vorne artikuliert

..* = lenis (Zwitter zwischen stimmlosem und stimmhaftem Plosiv)

..** = schwache Amplitude, keine starke Burstamplitude

4) Bemerkungen:

- Bei kurzen Vokalen wird Nasiltät nicht extra transkribiert.
- Bei zentralen nicht-tiefen Vokalen wird koartikulatorische Rundung in der Transkription nicht berücksichtigt, wenn diese nicht sehr ausgeprägt ist (sonst wird etwa Q notiert).

Anhang D

Liste der Äußerungen der Pragmatik-Stichprobe (siehe Kap.3.1). Die manuell bestimmten pragmatisch wichtigen Wörter sind unterstrichen.

Satz	Äußerung
bd2121	Wo muss ich umsteigen
bd2122	Ich will am ersten Oktober nach Bonn fahren
bd2124	Ich möchte bis vier zehn Uhr in Bonn sein
bd2127	Gibt es noch einen früheren Zug
bd2129	Ich werde den ersten nehmen
bd2130	Wann möchten Sie zurück fahren
bd2134	Was kostet die Rückfahrkarte
bd2138	Würden Sie mir bitte einen Platz reservieren
bd2140	Wo kann ich die Fahrkarte kaufen
bd2142	Wie weit ist es nach Hamburg
bd2143	Hat Fürth einen IC-Anschluß
bu2111	Von Frankfurt nach München
bu2120	Der Zug fährt nach Frankfurt
ci221f	Aber, muß ich in Münster umsteigen
ci235b	Dann will ich fünf Plätze haben
cl5550	Gibt es eine direkte Verbindung nach Stuttgart
cl5556	Was kostet das, einen Koffer aufzugeben
cl5557	Kann ich im Zug anrufen
cl5558	Und, wenn mich jemand anrufen muß
cl5560	Kann man reservieren
cl5561	Was muß ich tun,um das Sekretariat zu benutzen
cl5563	Und auf welchem Gleis kommt er an
cl5564	Weil ich nach Bonn will
cl5567	Wir wollen nach Göttingen
cl5568	Was kostet das
em5518	Er kostet zehn Mark
em5520	Wir möchten am Wochenende nach Mainz fahren
em5522	Nicht vor acht Uhr
he0263	Gibt es nach zehn noch einen Zug nach München
he229f	Aber, muß ich in Münster umsteigen
hf5535	Fahren Sie von München ab
hf5536	Sie können den IC sechs hundert zwanzig nehmen
hf5537	Sind noch Plätze frei
hf5540	Kann man Fahrräder mitnehmen
ja0250	Ich möchte am Freitag möglichst früh in Bonn sein
ja0253	Ich hätte gerne einen möglichst frühen Zug nach Bonn
ci0014_1	Gilt meine Jahreskarte am Sonntag
ci0014_2	Ich möchte nächsten Mittwoch nach Würzburg fahren
ci002f_1	Wann kommt dieser an
ci002f_2	Ist in München Aufenthalt

Anhang D (Fortsetzung)

Satz	Äußerung
ci002f_4	Ich will einen Zug mit Schlafwagen nehmen
ci0293	An welchem Werktag fährt der Intercity Nummer fünf hundert zwei und zwanzig über Nürnberg
ci236a_1	Und abends muß ich wieder nach Nürnberg kommen
ci234f_1	Welche Verbindung kann ich nehmen
ci234f_2	Hat dieser Zug auch einen Speisewagen
ci235f_1	Eine Platzbestellung notwendig
ci235f_2	Wann fährt der nächste
ci235f_3	Muß man mit einer Verspätung rechnen
ci233f	Guten Morgen, am Donnerstag Vormittag um sieben Uhr muß ich in München sein
ci237f_2	Hat dieser Zug auch einen Speisewagen und einen Telefonanschluß
hel24f_1	Führt der Zug einen Speisewagen
hel24f_2	Ja, auch an Ostern
ja245f	Guten Morgen, am Donnerstag Vormittag um sieben Uhr muß ich in München sein
ja246f_1	Welche Verbindung kann ich nehmen
ja246f_2	Hat dieser Zug auch einen Speisewagen
ja247f_1	Ist eine Platzbestellung notwendig
ja247f_2	Wann fährt der nächste
ja247f_3	Muß man mit einer Verspätung rechnen
ja247f_4	Dann will ich fünf Plätze
ja248f_1	Und abends muß ich wieder nach Nürnberg kommen
ja248f_3	Dürfen wir in München auch den nächsten Zug
ja249f	Guten Morgen, am Donnerstag Vormittag um sieben Uhr muß ich in München sein

Anhang E

Einige Anmerkungen zur Komplexität der Algorithmen:

Die für die in Kapitel 6 vorgestellten Ergebnisse verwendeten Programme wurden nicht unter dem Aspekt der Laufzeitoptimierung implementiert. Vielmehr standen hohe Flexibilität und Modularität im Vordergrund, um verschiedene Lösungsansätze schnell miteinander kombinieren zu können. Einzelne Verarbeitungsschritte wurden daher zum Teil als eigenständige Programme implementiert, die über Dateischnittstellen kommunizieren. Für eine Anwendung in einem spracherkennenden System mit nahezu Echtzeitanforderung läßt sich die Laufzeit der Algorithmen daher ohne weiteres auf die dafür notwendige Größenordnung verringern. Den größten Zeitanteil (ca. 75 Prozent) benötigt die Berechnung des Frequenzspektrums bei den Verarbeitungsschritten

- Berechnung der Energiebänder
- Tiefpaßfilterung
- Grundfrequenzbestimmung.

Gerade diese Schritte können mit Signalprozessoren deutlich schneller als Echtzeit realisiert werden, bzw. würden bei einer gleichzeitigen Digitalisierung mit höherer Abtastfrequenz (für die akustisch-phonetische Analyse) und niedriger Abtastfrequenz (für die Grundfrequenzbestimmung) ganz entfallen.

Da die zeitaufwendigen Berechnungen frameweise durchgeführt werden und pro Frame durch eine Konstante abgeschätzt werden können, verhält sich der Rechenzeitaufwand der entwickelten Algorithmen linear zur Länge des zu analysierenden Sprachfiles. Die frameweise Konstante steigt mit $n \cdot \log(n)$, wobei n die FFT-Größe angibt (typischerweise ist n in der Größenordnung zwischen 256 und 1024 Punkte).

Auf einer VAX-Station 3500 der Firma Digital Equipment (ca. 3 MIPS) unter normalem Betrieb benötigt die vorgestellte prosodische Analyse ca. 30-fache Echtzeit.

Linguistische Arbeiten

Internationalismen im modernen Deutsch, Französisch und Polnisch. Aufgezeigt in den Bereichen Sport, Musik und Mode

241 **Wilhelm Oppenrieder: Von Subjekten, Sätzen und Subjektsätzen.** Untersuchungen zur Syntax des Deutschen

242 **Joachim Liedtke: Narrationsdynamik.** Analyse und Schematisierung der dynamischen Momente im Erzählprodukt

243 **Tine Greidanus: Les constructions verbales en français parlé.** Etude quantitative et descriptive de la syntaxe des 250 verbes les plus fréquents

244 **Sebastian Löbner: Wahr neben Falsch.** Duale Operatoren als die Quantoren natürlicher Sprache

245 **Bettina Harriehausen: Hmong Njua.** Syntaktische Analyse einer gesprochenen Sprache mithilfe datenverarbeitungstechnischer Mittel und sprachvergleichende Beschreibung des südostasiatischen Sprachraumes

246 **Wolfgang Schindler: Untersuchungen zur Grammatik appositionsverdächtiger Einheiten im Deutschen.** Abgrenzung der „Apposition" unter Berücksichtigung von Attribut, Satzglied und Parenthese

247 **Bernd Pompino-Marschall: Die Silbenprosodie.** Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo

248 **Friederike Jin: Intonation in Gesprächen.** Ein Beitrag zur Methode der kontrastiven Intonationsanalyse entwickelt am Beispiel des Deutschen und Französischen

249 **Texte zu Theorie und Praxis forensischer Linguistik.** Hrsg. von Hannes Kniffka

250 **Wilfried Kuhn: Untersuchungen zum Problem der seriellen Verben.** Vorüberlegungen zu ihrer Grammatik und exemplarische Analyse des Vietnamesischen

251 **Karin Bausewein: Akkusativobjekte, Akkusativobjektsätze und Objektsprädikate im Deutschen.** Untersuchungen zu ihrer Syntax und Semantik

252 **Susanne Uhmann: Fokusphonologie.** Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie

253 **Robert B. Howell: Old English Breaking and its Germanic Analogues**

254 **Linguistische Interaktionsanalysen.** Beiträge zum 20. Romanistentag 1987. Hrsg. von Ulrich Dausendschön-Gay, Elisabeth Gülich und Ulrich Krafft

255 **Kang-Ho Lie: Verbale Aspektualität im Koreanischen und im Deutschen** mit besonderer Berücksichtigung der aspektuellen Verbalperiphrasen

256 **Michael Prinz: Klitisierung im Deutschen und Neugriechischen.** Eine lexikalisch-phonologische Studie

257 **Fragesätze und Fragen.** Referate anläßlich der 12. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Saarbrücken 1990. Hrsg. von Marga Reis und Inger Rosengren

258 **Peter Rolf Lutzeier: Major Pillars of German Syntax.** An Introduction to CRMS-Theory

259 **Elmar Nöth: Prosodische Information in der automatischen Spracherkennung.** Berechnung und Anwendung

Niemeyer