Automatische Erkennung von Satzmodus, Akzentuierung und Phrasengrenzen in einem sprachverstehendem System

Inauguraldissertation
zur
Erlangung der Doktorwürde
der
Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Volker Strom

Bonn, Januar 1998

1. Referent: Univ.–Prof. Dr. Wolfgang Hess

2. Referent: Univ.–Prof. Dr. Joachim M. Buhmann

Mündliche Prüfung am: 27. Mai 1998

Vorwort

Die vorliegende Dissertation entstand in der Zeit zwischen 1994 und 1997 am Institut für Kommunikationsforschung und Phonetik der Universität Bonn, in der ich für das vom BMBF geförderte Verbundvorhaben Verbmobil tätig war.

An dieser Stelle möchte ich allen Personen danken, die zum Gelingen dieser Arbeit beigetragen haben. An erster Stelle gilt mein Dank meinem Doktorvater Herrn Prof. Dr. Wolfgang Hess für seine wertvollen Hinweise und die gründliche Korrektur der ersten Fassung. Gleichermaßen danke ich Herrn Prof. Dr. Joachim Buhmann für die freundliche Übernahme des Koreferats.

Ein herzliches Dankeschön an meine Hilfskräfte Christina Widera und David Schlangen, an die Teilnehmer der Perzeptionsexperimente und an alle anderen Kolleginnen und Kollegen für ihre freundliche Unterstützung.

Bonn, im Januar 1998 Volker Strom

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einführung

1.1 Automatische Sprachverarbeitung

Noch vor wenigen Jahrzehnten waren Computer teure Großgeräte, zu denen nur wenige Spezialisten Zugang hatten. Der enorme Preisverfall bei gleichzeitiger Leistungssteigerung und Miniaturisierung ebnete den Weg zu immer neuen Anwendungen; mittlerweile sind Computer aus kaum einem Lebensbereich mehr wegzudenken.

Die Interaktion mit dem Computer erfolgt immer noch fast ausschließlich über Bildschirm, Tastatur und Maus. Dabei zeichneten sich in den 70er Jahren erste Erfolge bei der automatischen Spracherkennung ab, und seit etwa 10 Jahren sind sog. Diktiersysteme, also „hörende Schreibmaschinen", kommerziell erhältlich, die mit gewissen Einschränkungen eine Schreibkraft ersetzen können. Vermutlich werden Computer in einigen Jahren auch Aufgaben wie Telephonauskunft oder Flugbuchung übernehmen können; experimentelle Systeme dazu sind schon im Einsatz.

Die Vorteile sprachlicher Kommunikation liegen auf der Hand:

- Gesprochene Sprache ist die natürlichste und bequemste Form der Kommunikation zwischen Menschen; sie wird auch von jenen beherrscht, die nicht im Tippen geübt sind.

- Gesprochene Sprache ist die effektivste Art der Kommunikation: Die durchschnittliche Sprechrate beträgt etwa 150 bis 250 Wörter pro Minute, während geübte Schreibkräfte nur 100 bis 150 Wörter pro Minute erreichen [O'S87].

- Augen und Hände bleiben frei; beispielsweise trägt die sprachgesteuerte

Bedienung eines Autotelephons zur Fahrsicherheit bei. Sprachliche Kommunikation ist auch im Dunkeln möglich.

- Für viele Behinderte ist die Sprache die einzige Möglichkeit, Geräte zu steuern.

- Sprachliche Kommunikation mit Maschinen ist im Vergleich zur Kommunikation über Bildschirm und Tastatur ohne bzw. mit relativ wenig Ausrüstung und auch von jedem Telephon aus möglich.

Aus diesen Gründen begann schon in den 50er Jahren die Forschung auf dem Gebiet der **A***utomatischen* **S***prach***v***erarbeitung* (ASV). Damit ist die Verarbeitung von digitalisierten Sprachsignalen durch Computer gemeint. Nach anfänglichen Schwierigkeiten sind in den letzten Jahren erhebliche Fortschritte erzielt worden. Die Anwendungsgebiete der ASV lassen sich etwa wie folgt umreißen:

- *Spracherkennung*, wie sie in Kommandosystemen zur Gerätesteuerung oder in Diktiersystemen zur Anwendung kommt

- *Sprechererkennung*, dabei unterscheidet man zwischen Sprecherverifikation für Autorisierungszwecke und Zugangskontrollen sowie Sprecheridentifikation z.B. in der Kriminalistik;

- *Sprachkodierung* zur effizienten Übertragung und Speicherung von Sprache;

- *Sprachsynthese* für Vorleseautomaten oder zur Erzeugung der Antworten des Computers in Dialogsystemen;

- *Sprachverstehen*, das im Gegensatz zur bloßen Erkennung für Dialogsysteme (z.B. Hotelreservierung) und Sprachübersetzung nötig ist

.

Diese Liste erhebt keinen Anspruch auf Vollständigkeit, man könnte z.B. Sprachlernprogramme als separate Anwendung aufführen, und es gibt natürlich Überschneidungen: Methoden zur Sprachkodierung wurden für auch für die Erkennung eingesetzt, und Dialogsysteme enthalten Komponenten zur Erkennung und zur Synthese.

In dieser Arbeit steht die Sprachanalyse im Vordergrund, also die Spracherkennung und das Sprachverstehen. Der Unterschied zwischen einem spracherkennendem und einem sprachverstehenden System soll an der Verarbeitung des folgenden Satzes erläutert werden: *„Können Sie mir eine Zugverbindung von Bonn nach Hamburg nennen, bei der ich noch vor vier Uhr ankomme?“* Ein Diktiersystem als Beispiel eines Spracherkenners hat „nur“ die Aufgabe, diesen gesprochenen

Satz in die orthographische Darstellung zu überführen, um sie einem Textverarbeitungsprogramm zu übergeben. Ein sprachverstehendes System, in diesem Fall ein Zugauskunftssystem, muß dagegen die Intention des Fragestellers soweit zu erfassen, daß es daraus eine Datenbankanfrage erstellen kann; aus dem Resultat wird dann ein deutscher Satz generiert, der von einem Sprachsynthetisator in gesprochene Sprache umgewandelt wird. Um den semantischen Gehalt, d.h. den Bedeutungsgehalt der Anfrage erfassen zu können, muß das System über semantisches Wissen verfügen, z.B. über das Konzept „mit dem Zug fahren" (hat Abfahrts- und Ankunftsbahnhof und -zeit), und über pragmatisches Wissen, das auf den jeweiligen Diskursbereich zugeschnitten ist, z.B. daß mit vier Uhr wahrscheinlich vier Uhr nachmittags gemeint ist. Im Fall von Verständnisschwierigkeiten sollte das System auch von sich aus einen Klärungsdialog beginnen.

Auch in Spracherkennern für das Deutsche ist, jedenfalls ab einer gewissen Wortschatzgröße, rudimentäres Wissen über die deutsche Grammatik inkorporiert, nämlich in Form eines sog. *Sprachmodells* (einer stochastischen regulären Grammatik, die nur Folgen von drei oder vier Wörtern statistisch modelliert, siehe Abschnitt 7.2), um den Suchraum zu beschränken. Denn das Problem besteht in beiden Fällen in der enormen Variabilität der Sprache: Dieselbe Äußerung hat, selbst wenn sie von gleichen Sprecher zweimal unmittelbar hintereinander gesprochen wird, nie die gleiche akustische Ausprägung. Das betrifft Aussprachevarianten wie Verschleifungen, Betonung und Sprechrhythmus, aber auch Abstand zum Mikrophon und Hintergrundgeräusche. Das gilt umso mehr für sprecherunabhängige Systeme, bei denen an Problemen noch sprecherindividuelle Unterschiede hinzukommen, bedingt u.a. durch Dialekt, Alter, Geschlecht, Vokaltraktanatomie oder Gesundheit.

Ein weiteres Problem besteht in der Kontinuität der Sprache: Gesprochene Sprache wird als Folge von Wörtern oder Silben wahrgenommen, im Sprachsignal existieren jedoch keine sichtbaren Diskontinuitäten, die den Grenzen dieser Einheiten entsprechen. Beispielsweise wird der Satzteil „in München" wegen der Assimilation den [n] in „in" an das [m] von „München" als [ImYnC@n] gesprochen (zum SAMPA-Alphabet siehe Abschnitt A auf Seite A). Das heißt, die Grenze zwischen „in" und „München" geht mitten durch den Laut [m].

Die Wahrnehmung des Menschen, eines hervorragend funktionierenden Spracherkenners, beruht darauf, daß er stets sein gesamtes Wissen um die sprachlichen Phänomene zur Dekodierung des Sprachschalls einsetzt, seien sie akustischer, phonetischer, phonotaktischer, lexikalischer, prosodischer, syntaktischer, semantischer oder pragmatischer Natur. Da dies weitgehend unbewußt abläuft, wird die Komplexität des Vorgangs oft unterschätzt [ST91].

Auch sprachverstehende Systeme müssen über solches Wissen verfügen, was oberhalb der syntaktischen Ebene jedoch nur begrenzt möglich ist, da entsprechende Wissenbasen mangels automatischer Lernverfahren von Experten noch manuell erstellt werden mußsen.

Der sprachverstehende Teil, d.h. der Worterkenner, verfügt nur über Wissen bis zur lexikalischen Ebene und über rudimentäres syntaktisches Wissen. Die Folge ist, daß auch für sehr kurze Äußerungen einige hundert Worthypothesen erzeugt werden, die akustisch plausibel sind. Unter Umständen sind nicht einmal alle richtigen Wörter dabei, z.B. bei starken Verschleifungen, wenn das Sprachmodell inadäquat ist oder das Wort nicht im Aussprachelexikon enthalten ist. In traditionellen Systemen folgt die linguistische Analyse auf die akustische; die Aufgabe der höheren linguistischen Module besteht dann darin, aufgrund ihrer Wissensbasen aus den vielen Worthypothesen die richtigen auszuwählen und dabei eine angemessene semantische und pragmatische Darstellung zu erzeugen.

Eine von Menschen genutzte Informationsquelle, die oben bereits genannt wurde, ist die *Prosodie* (siehe Abschnitt 1.2. Prosodie ist etwa das, was umgangssprachlich als „Sprachmelodie" oder „Rhythmus" bezeichnet wird. Sie umfaßt u.a. die Hervorhebung von Wörtern, was man in geschriebener Sprache durch Fettdruck oder Unterstreichen ausdrücken würde, und die Gliederung der Rede, was beim Schreiben etwa der Interpunktion entspricht.

In allen bisherigen kommerziellen ASV-Systemen wird diese prosodische Information nicht genutzt, sie wird im Gegenteil „herausgerechnet", damit z.B. ein Worterkenner für Frauen- und Männerstimmen die gleichen Aussprachemodelle verwenden kann. Für den menschlichen Hörer hat die Prosodie dagegen eine großen Einfluß auf die Verständlichkeit: Äußerungen, die auf der Lautebene korrekt ausgesprochen, aber falsch oder nicht betont wurden, sind nur schwer zu erfassen; man denke z.B. an manche Englisch sprechende Franzosen oder an frühe, monoton klingende Sprachsynthetisatoren, wie sie in Spielfilmen als Klischee fortleben. Neben der Verständniserleichterung dient die Prosodie auch zur Bedeutungsunterscheidung. In gewissen Fällen kann die Intention des Sprechers nur anhand der Prosodie bestimmt werden, z.B. kann ein kurzes *„ja"*, als Reaktion auf eine Behauptung, als Infragestellung oder als Bestätigung gemeint sein (weitere Beispiele folgen im nächsten Abschnitt). Deshalb sollte diese Informationsquelle auch in der ASV genutzt werden.

Innerhalb des Verbmobil-Forschungsprojektes[1] wurden mit dem Verbmobil-Forschungsprototyp und dem experimentellen INTARC-System zum ersten Mal vollständige sprachverstehende Systeme entwickelt, die zumindest drei Aspekte der Prosodie für den Analyseprozeß nutzen, nämlich die Akzentuierung, die Phrasierung und den Satzmodus.

---

[1] Verbmobil ist ein vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie gefördertes Projekt zur Entwicklung eines mobilen Systems zur Übersetzung von spontan gesprochenen Verhandlungsdialogen.

1.2 Prosodie in der sprachlichen Kommunikation

1.2.1 Terminologie

Prosodie ist ein Phänomen *gesprochener* Sprache und bezeichnet alle Informationen, die dem Hörer über die bloße Lautfolge hinaus gegeben werden. Teilweise gibt es Entsprechungen in der Schriftsprache wie die schon erwähnte Interpunktion oder den Fettdruck; hier stehen die Aspekte der Verständniserleichterung und Bedeutungsunterscheidung im Vordergrund.

Durch die Prosodie wird aber auch Information über den Sprecher übermittelt, z.B. sein Geschlecht oder seinen emotionalen Zustand.

Zusammenfassend wird Prosodie meist als Gesamtheit der *suprasegmentalen* Eigenschaften von gesprochener Sprache definiert [Leh70, Buß90, Nöt91]. Mit Segmenten sind in diesem Zusammenhang Lautsegmente gemeint; suprasegmentale Eigenschaften sind solche, die sich auf größere Einheiten als Laute beziehen, also auf Silben, Wörter, Phrasen, Sätze, Dialogakte oder ganze Redebeiträge. In [Buß90] wird Prosodie definiert als

> **Prosodie:**
> (1) Gesamtheit sprachlicher Eigenschaften wie Akzent, Intonation, Quantität[2], Sprechpausen. Sie beziehen sich im allgemeinen auf Einheiten, die größer sind als ein einzelnes Phonem. Zur P. zählt auch die Untersuchung von Sprechrhythmus und Sprechpausen.
> (2) [Auch: Prosodik]. Untersuchung der P.(1).

Der Begriff *Intonation* wird oft synonym zu Prosodie gebraucht, soll in der vorstehenden Definition und in dieser Arbeit jedoch im engeren Sinn verstanden werden als zeitlicher Tonhöhenverlauf gesprochener Sprache. Die Prosodie hat in der sprachlichen Kommunikation verschiedene *Funktionen*, die man in folgende Gruppen einteilen kann:

- Akzentuierung,
- Phrasierung,
- Markierung des Satzmodus,
- paralinguistische Funktion (z.B. Ausdruck von Emotionen),

[2] Unter Quantität versteht man in der Sprachwissenschaft i.a. die Dauer sprachlicher Einheiten.

- indexikalische Funktion (z.B. Dialekt, siehe Abschnitt 1.2.5)

Zur Verständniserleichterung und Auflösung von Mehrdeutigkeiten am wichtigsten sind dabei die ersten drei Gruppen: Durch Phrasierung wird die Rede gegliedert, durch Akzentuierung werden wichtige Wörter hervorgehoben, und durch Intonation kann der Satzmodus markiert werden, wenn dies nicht schon durch syntaktische Mittel geschieht.

Die *Funktionen* der Prosodie werden durch unterschiedliche *formale* prosodische Mittel ausgedrückt:

- *Intonation* (Tonhöhenverlauf),
- *Dehnung* von Lauten bzw. Silben,
- Änderung der *Lautheit*,
- Änderung der *Sprechgeschwindigkeit*,
- Setzen von *Sprechpausen*.

Der wahrgenommenen Tonhöhe entspricht auf der akustischen Ebene die Grundfrequenz des Sprachsignals (siehe Abschnitt 5.1), der Lautheit entspricht ihre Energie. Die Mittel Dehnung, Sprechgeschwindigkeit und Pausensetzung lassen sich unter dem Begriff *zeitliche Strukturierung* zusammenfassen.

In [Kie97] werden noch weitere prosodische Mittel angegeben:

- *Stimmqualität*; meist unterscheidet man dabei nur zwischen Stimmhaftigkeit, Stimmlosigkeit und Stille (Sprechpausen), aber auch spezielle Stimmbildungsphänomene wie Laryngalisierungen (siehe Abschnitt 5.2.1) gehören dazu, individuelle Eigenheiten wie eine heisere oder rauhe Stimme, und auch Emotionen färben auf die Stimmqualität ab.
- *Klangfarbe*; sie hängt ab von der Stimmqualität und der Vokaltraktanatomie des Sprechers.
- *Stimmlage*; man kann sie zur Intonation rechnen.
- *Tempo* ist gleichbedeutend mit Sprechgeschwindigkeit.
- *Rhythmus* meint die zeitliche Struktur von Akzenten.

Die prosodischen Funktionen werden teilweise durch die gleichen prosodischen Mittel ausgedrückt, d.h. es gibt keine eindeutige Beziehung zwischen Funktion und Form. Beispielsweise kann die Dehnung einer Silbe ihre Akzentuierung bedeuten oder eine Phrasengrenze ankündigen. Es können aber auch beide Funktionen gleichzeitig beabsichtigt sein, z.B. kann durch Dehnung der phrasenfinalen Silbe ihre Akzentuierung und gleichzeitig die folgende Phrasengrenze angezeigt werden. Auf diese Schwierigkeiten wird in Abschnitt 2.1 näher eingegangen. Im folgenden werden die prosodischen Funktionen näher erläutert.

1.2.2 Akzentuierung

Unter Akzentuierung versteht man die Hervorhebung von Silben oder Wörtern durch Steigerung der *Lautheit* (Verstärkung des Atemdrucks), Änderung der *Tonhöhe* und/oder durch *Dehnung* [Buß90]. Akzentuierte Silben sind meist deutlicher artikuliert, in unbetonten Silben ist der Vokal dagegen meist abgeschwächt zum neutralen Vokal [@] (siehe Anhang A) oder liegt zumindest näher an ihm, er kann auch ganz reduziert sein: Das Wort „reden", mit dem lexikalischen Akzent auf der ersten Silbe, kann artikuliert werden als [re:d@n] oder [re:d=n].

Man kann Akzente nach den prosodischen Mitteln unterscheiden: Steht die Lautheit im Vordergrund, spricht man von *dynamischen Akzent* (engl.: stress), ist es die Tonhöhenänderung, spricht man vom *musikalischen Akzent* (engl.: pitch accent)[3].

Man unterscheidet auch Wortakzent, Phrasenakzent und Satzakzent, je nach der vom Akzent betroffenen sprachlichen Einheit. Der *Wortakzent*, auch *lexikalischer Akzent*, gibt an, welche Silbe eines Wortes am stärksten hervorgehoben ist, wenn es isoliert geäußert wird. Die Begriffe *Hauptakzent* (auch: Primärakzent) und *Nebenakzent* differenzieren dabei unterschiedliche Grade der Hervorhebung.

Es gibt im Deutschen nur wenige Beispiele dafür, daß der Akzent schon auf der Wortebene bedeutungsunterscheidend ist, wie in:

umfahren vs. um**fah**ren (1.1)

übersetzen vs. über**set**zen (1.2)

unterstellen vs. unter**stel**len (1.3)

Tenor vs. Te**nor** (1.4)

[3] Für Akzente, die hauptsächlich durch Dehnung markiert sind, gibt es keine entsprechende Bezeichnung.

Der lexikalische Akzent kann aber dazu dienen, ein Kompositum zu unterscheiden von seinen als Einzelwörtern gesprochenen Bestandteilen wie in

fünf und **zwan**zig vs. **fünf**undzwanzig (1.5)

da bei Komposita fast immer das erste Teilwort den lexikalischen Akzent trägt.

In fließender Rede sind jedoch nicht alle lexikalischen Akzente markiert. *Funktionswörter* sind in der Regel nicht akzentuiert (Ausnahme ist z.B. der *Kontrastakzent*, siehe Ende des Abschnitts). Unter Funktionswörtern versteht man Wörter, die primär grammatikalische Bedeutung tragen anstatt lexikalischer und vor allem syntaktisch-strukturelle Funktionen erfüllen. Hierzu gehören u.a. Artikel, Präpositionen und Konjunktionen [Buß90]. Akzentuiert sind in der Regel nur *Inhaltswörter* wie Nomina und Verben, insbesondere die, die im *Fokus* stehen, d.h. die „wichtigste Information" tragen[4].

Der *Phrasenakzent* und der *Satzakzent* bezeichnen das am stärksten hervorgehobene Wort innerhalb der Phrase bzw. des Satzes; in der Regel entspricht es dem semantischen Fokus, es sei denn, dieser ist bereits durch die Wortstellung, also durch grammatikalische Mittel ausgedrückt. Wenn der Fokus vom Sprecher durch prosodische Mittel gekennzeichnet wird, ist der Phrasen-, Satz- oder *Fokusakzent* auf Satzebene bedeutungsunterscheidend und damit für ein sprachübersetzendes System wie Verbmobil relevant.

Betrachte dazu folgende Äußerung aus Verbmobil-Korpus (das Beispiel ist [BBK95] entnommen) und ihre alternativen Übersetzungen:

Dann sollten wir noch ein Treffen im März ausmachen. (1.6)

Then we need another meeting date in March. (1.7)

Then we still need a meeting date in March. (1.8)

Wenn in 1.6 das Wort „noch" akzentuiert ist, wäre 1.8 eine angemessene Übersetzung. Liegt der Akzent dagegen auf dem Wort „März", der Default-Position (nach dem *rightmost principle* ist im Deutschen normalerweise das am weitesten rechts liegende Inhaltswort einer Phrase akzentuiert), wäre 1.7 die richtige Übersetzung.

Innerhalb des akzentuierten Wortes liegt der Akzent auf der Silbe, die den lexikalischen Akzent trägt, außer es handelt sich um einen *Kontrastakzent* wie in:

„Er ist Dir nicht **unter**legen, sondern **über**legen!" (1.9)

[4]Die Unterteilung Funktionswort/Inhaltswort ist jedoch im strikten Sinne nicht haltbar[Buß90, Seite 118]. Im Verbmobil-Korpus sind oft Präpositionen wie „vor" und „nach" in Zeitangaben akzentuiert, weil sie inhaltlich von hoher Bedeutung sind, während die Verben „vereinbaren" und „ausmachen" im Zusammenhang mit Terminvereinbarungen meist unbetont sind.

1.2.3 Phrasierung

Die Phrasierung dient der Gliederung der Rede; Phrasengrenzen sind Einschnitte im Redefluß, die in etwa der Interpunktion in der Schriftsprache entsprechen. So wie durch Interpunktion (und auch durch Absätze, Überschriften, etc.) Schriftsprache in übersichtliche Einheiten gegliedert wird, wodurch in vielen Fällen erst der Sinn eindeutig wird, dient in gesprochener Sprache die Prosodie der Verständniserleichterung und zur Auflösung von Mehrdeutigkeiten. Ein vielzitiertes Beispiel zur Illustration ist:

„Strauß sagte: Kohl wird nie Kanzler.“ vs. (1.10)
„Strauß — sagte Kohl — wird nie Kanzler.“

Zu welcher Verwirrung falsch gesetzte Phrasengrenzen führen, verdeutlicht folgendes Scherzrätsel:

Ich habe zwanzig Finger an jeder Hand, fünfundzwanzig an Händen und Füßen[5]. (1.11)

In folgendem Beispiel aus dem Verbmobil-Korpus stellen die vertikalen Linien mögliche Satz- bzw. Nebensatzgrenzen dar:

ja | zur Not | geht's | auch | am Samstag | (1.12)

In der Schriftsprache können diese Linien z.T. durch Punkt, Komma oder Fragezeichen ersetzt werden, wodurch sich mindestens 36 unterschiedliche, zumindest syntaktisch korrekte Alternativen ergeben [Kom95]. Zwei Varianten mit den entsprechenden englischen Übersetzungen sind:

1.	Ja?	Zur Not geht's?	Auch am Samstag?
	Really?	It's possible if necessary?	Even on Saturday?
2.	Ja.	Zur Not.	Geht's auch am Samstag?
	Yes.	If necessary.	Would Saturday be possible as well?

Die beiden Varianten sind nicht als Reaktion auf die gleiche Äußerung denkbar, daher kann die semantische Ambiguität in diesem Beispiel auch aus dem bisherigen Dialogverlauf aufgelöst werden (zu Beispielen, wo das nicht möglich ist, siehe den nächsten Abschnitt). Allerdings gliedern die Phrasengrenzen in den beiden Varianten der Äußerung 1.12 die Wortfolge in Sätze, die für sich übersetzt werden können.

[5] Ich habe zwanzig Finger: an jeder Hand fünf, und zwanzig an Händen und Füßen.

Die verständniserleichternde Funktion prosodischer Phrasengrenzen wird auch am sogenannten *garden path problem* deutlich: Man betrachte die beiden Äußerungen:

Ich schlage vor, wir treffen uns am Dienstag vormittags. (1.13)

Ich schlage vor, wir treffen uns am Dienstag. Vormittags würde es mir gut passen. (1.14)

Diese Äußerungen sind zwar eindeutig verschieden, aber bis zum Wort „vormittags" von der Lautfolge her identisch. Wenn die Phrasengrenze (bzw. Satzgrenze) nach „Dienstag" nicht deutlich markiert ist, ist die Interpretation nach der Äußerung von „vormittags" aber mehrdeutig, weil der Hörer zu diesem Zeitpunkt noch nichts von der Fortsetzung weiß. Die Mehrdeutigkeit wird zwar durch den weiteren Verlauf der Äußerung aufgelöst, die prosodisch markierte Grenze disambiguiert allerdings früher. Dies spielt deshalb eine Rolle, weil beim menschlichen Hörer der Verstehensprozeß (wie beim INTARC-System der Analyseprozeß, siehe Abschnitt 7.1) bereits während der Produktion der Äußerung beginnt.

1.2.4 Satzmodus

Syntaktisch korrekte Fragen sind im Deutschen und im Englischen entweder durch ein Fragepronomen oder ein Verb zu Beginn des Satzes gekennzeichnet. Erstere werden als Ergänzungs- oder W-Fragen bezeichnet; ein Beispiel aus dem Verbmobil-Korpus ist:

Wann wär's Ihnen denn recht? (1.15)

Letztere werden als Entscheidungsfragen bezeichnet, weil man darauf mit ja oder nein antworten kann wie in

Treffen wir uns dann in meinem Büro? (1.16)

außer es handelt sich um eine Alternativfrage wie

Treffen wir uns dann in meinem oder in Ihrem Büro? (1.17)

Von der Prosodie her interessant sind zum einen die in Spontansprache häufigen Ellipsen, zum anderen die assertiven Fragen, d.h. Fragen, die syntaktisch Aussagen sind, aber intonatorisch als Fragen markiert sind.

Ein Beispiel einer assertiven Frage aus dem Verbmobil-Korpus ist

Dann kommen Sie zu mir ins Büro? (1.18)

mit steigender Intonation am Ende. Syntaktisch als Frage markiert wäre

Kommen Sie dann zu mir ins Büro? (1.19)

aber in der Umgangssprache haben Entscheidungsfragen oft Verb-Zweit-Stellung aufweisen und sind dafür intonatorisch als Fragen markiert.

Ellipsen sind unvollständige Sätze wie „ja“ oder „auch am Samstag“ im Beispiel 1.12. Deshalb gibt die Syntax keinen Aufschluß darüber, ob z.B. „ja“ als Frage oder Aussage (Bestätigung) gemeint ist, dies kann nur die Intonation.

Entscheidungsfragen sind meist auch intonatorisch als Frage markiert (Ausnahmen sind vor allem Alternativfragen), während W-Fragen mit steigender oder fallender Intonation realisiert sein können; da die Frage bereits syntaktisch als solche markiert ist, bleibt dies dem Sprecher überlassen (und ist auch stark sprecherabhängig).

Das Altmannsche Satzmodussystem [Alt93] für das Deutsche orientiert sich an formalen Kriterien: Als Fragen gelten diejenigen Äußerungen, die formal, d.h. syntaktischen oder intonatorisch als Frage gekennzeichnet sind. Möglich wäre auch, sich an der „semantischen Fragehaltigkeit“ zu orientieren wie in [KLP+94]. Dann würden auch indirekte Fragen wie *„Ich würde gerne wissen, ob ...“* nicht mehr zu den Aussagen zählen, obwohl sie formal Aussagen sind. Das Problem dabei ist jedoch, daß man fast jeder Äußerung in einem Dialog eine gewisse Fragehaltigkeit zusprechen kann, da sie eine Aufforderung an den Dialogpartner impliziert, seine Meinung dazu zu äußern [Bat94].

Im Altmannschen Satzmodussystem wird neben steigender und fallender Intonation auch die progrediente (weiterführende) Intonation unterschieden, die durch gleichbleibende oder nur leicht steigende Intonation am Satzende gekennzeichnet ist. Sie entspricht in der Schriftsprache meist dem Komma; Weiterführung kann aber auch am Satzende auftreten, etwa bei Rückmeldungen, wenn also z.B. eine Zeit- oder Ortsangabe wiederholt wird, um zu signalisieren: „Ja, ich habe verstanden“.

1.2.5 Paralinguistische und indexikalische Funktionen

Während die prosodische Markierung von Akzenten, Phrasengrenzen und des Satzmodus zur Bedeutungsunterscheidung und Verständniserleichterung dient und mittlerweile auch in der ASV augenutzt wird, kann Prosodie emotionale Grundzustände wie Wut, Trauer, Angst und Freude ausdrücken, oder Haltungen dem Hörer gegenüber wie Zustimmung, Ablehnung, Bewunderung und Verachtung. Ein besonderer Fall ist die Ironie, mit der die wörtliche Bedeutung einer Äußerung nicht nur modifiziert, sondern in ihr Gegenteil verwandelt werden kann.

Diese *paralinguistischen Funktionen* der Prosodie drücken meist vorübergehende Zustände oder Haltungen des Sprechers aus, während länger andauernde Charakteristika des Sprechers unter dem Begriff *indexikalische Funktionen* zusammengefaßt werden können. Dazu zählen in erster Linie Alter und Geschlecht, aber auch Dialekt und soziale Herkunft.

Diese Aspekte spielen noch eine untergeordnete Rolle in der ASV, aber auch hier sind Anwendungen denkbar: Die indexikalischen Aspekte könnten zur Sprecherverifikation oder -adaption eingesetzt werden; die Analyse von Emotionen in einem Auskunftssystem könnte dazu dienen, bei aufkeimender Verärgerung des Fragestellers das Gespräch an einen menschlichen Bearbeiter weiterzuleiten.

1.3 Das Verbmobil-Projekt

Das langfristige Ziel des Verbundprojektes VERBMOBIL [Wah93] ist die Entwicklung eines mobilen Systems zur Übersetzung von Verhandlungsdialogen in sog. face-to-face Situationen. Beteiligt sind daran 22 Universitäten bzw. Forschungszentren und 7 Unternehmen; gefördert wird das Projekt vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Das Szenario sieht vor, daß sich ein Deutscher und z.B. ein Japaner gegenübersitzen (Gestik und Mimik sollen aber nicht berücksichtigt werden) und Termine vereinbaren oder Geschäftsreisen planen. Beide sollen Englisch zumindest passiv beherrschen. Der Dialog wird in Englisch geführt und vom Verbmobil-System grob verfolgt; bei Bedarf soll das Verbmobil-System einen Dialogbeitrag von der jeweiligen Landessprache ins Englische übersetzen. Langfristig soll auf die Zwischensprache Englisch verzichtet werden, es sollen weitere Sprachenpaare hinzukommen, ein Ergebnisprotokoll soll erstellt werden können, und anderes.

Der Verbmobil-Forschungsprototyp als Ergebnis der ersten Projektphase (1993 - 1996) beschränkt sich auf die Domäne Terminverhandlungen und die Übersetzung vom Deutschen ins Englische.

Ein Ausschnitt aus einem einfachen Beispieldialog ist:

A: *I guess we should meet in September. How about Friday the first of September.*

B: *(Mausklick) Montag wäre mir lieber (Mausklick).*

Vm: *I would prefer Monday.*

A: *OK, so Monday the third. That's fine with me. What about 11 o'clock?*

B: *(Mausklick) Gut, wir treffen uns dann in meinem Büro. (Mausklick)*

Vm: OK, so then we meet in my office.

Das System setzt zwar kooperative Sprechweise voraus, z.B. soll Dialekt vermieden werden, allerdings soll es auch frei formulierte Alltagssprache, sog. *Spontansprache* verarbeiten können. Es müssen daher Phänomene verarbeitet werden können wie elliptische (also im Sinne der Duden-Grammatik unvollständige) Sätze, umgangssprachliche Wortstellung, Satzabbrüche und Selbstkorrekturen, gefüllte Pausen *(„ähm")*, Schmatzen, Husten, Lachen, Störgeräusche wie Papierrascheln und Telephonklingeln.

Im Gegensatz zum Mensch-Maschine-Dialog, der meist kurze, einfach gebaute Sätze verlangt, können beim Mensch-(Maschine)-Mensch-Dialog auch relativ lange Redebeiträge mit komplizierterer Struktur auftreten wie:

> *<ähm> Dienstag würde mir gut passen , <ähm> das heißt , Moment , allerdings erst<Z> nachmittags . das wird dann wahrscheinlich 'n bißchen schwierig . Dienstag , mittwochs<Z> <äh> +/is=/+ sieht das bei mir +/sch=/+ schwierig aus . da hab' ich tagsüber Termine . <A> <ähm> wie sieht das bei Ihnen am Donnerstag aus ?*

oder

> *Ja, dann <A> können wir ja wieder ein/- <a> +/ vielleicht nehmen wir der *Einfachhalt /+ <Lachen> vielleicht nehmen wir der Einfachheit halber den<Z> dritten ?*

(<Z> steht für Zögern, <A> für Atmen, /- kennzeichnet einen abgebrochenen Satz, = ein abgebrochenes Wort. Satzteile, die später selbst korrigiert werden, sind in +/.../+ geklammert).

Abbildung 1.1 zeigt die Komponenten den Verbmobil-Prototyps: Wird vom Deutschen ins Englische übersetzt, folgt auf die Wort- bzw. Spracherkennung die Prosodieerkennung, das Ergebnis wird an die syntaktische und semantische Analyse weitergegeben. Die eigentliche Übersetzung erfolgt im Transfer-Modul, aus dessen Ausgabe wird eine englische Äußerung generiert und vom Synthesemodul gesprochen. Vom Englischen ins Deutsche wird nur grob übersetzt: Vom *Keywordspotter* werden nur wichtige Wörter, sog. Schlüsselwörter erkannt; die Übersetzung erfolgt aufgrund des bisherigen Dialogverlaufs und der erkannten Schlüsselwörter.

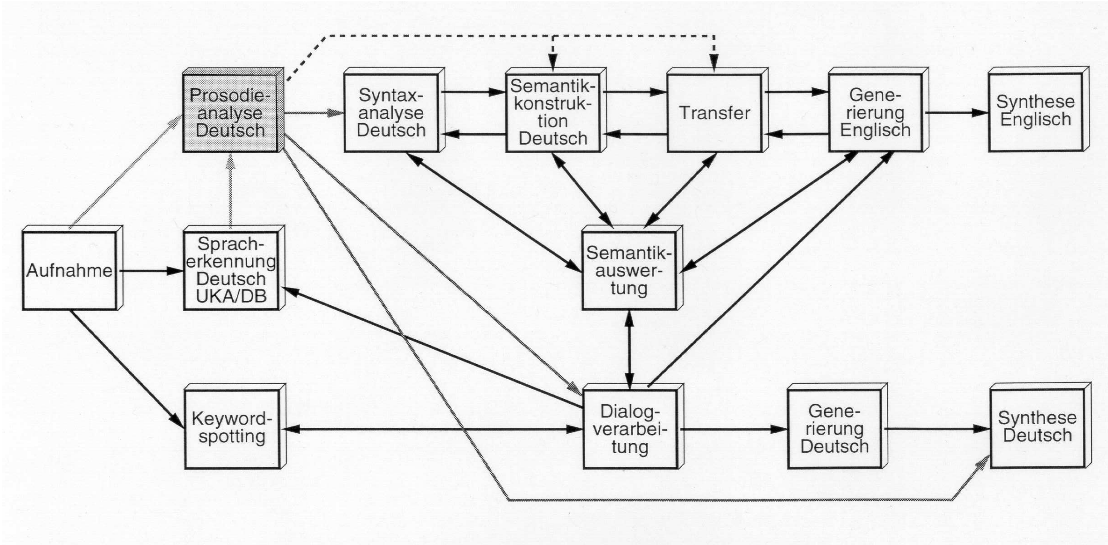

Abbildung 1.1: *Übersicht des Verbmobil-Forschungsprototyps. Durchgezogene Pfeile stellen die Schnittstellen dar, gestrichelte Pfeile zusätzlichen Datenfluß vom Prosodiemodul aus.*

1.4 Das Architektur-Teilprojekt

Innerhalb des Verbmobil-Projektes wurde ein experimentelles System entwickelt, INTARC (integrated architecture, siehe Kapitel 7), das ebenfalls ein Übersetzer spontan gesprochener deutscher Sprache ins Englische ist. Der Schwerpunkt lag auf der Entwicklung einer neuartigen Architektur, die durch inkrementelle, interaktive und parallele Verarbeitung gekennzeichnet ist.

Inkrementelle Verarbeitung bedeutet schritthaltende Verarbeitung: Während der Verbmobil-Prototyp noch am seriellen Modell ausgerichtet ist (erst wenn die Äußerung beendet ist, beginnt die Worterkennung; wenn diese abgeschlossen ist, beginnt die Syntaxanalyse, usw.), sollen im INTARC-System alle Komponenten *parallel* arbeiten, noch während die Äußerung andauert. Sie beginnen natürlich nicht genau gleichzeitig, die Inkrementgröße ist jedoch relativ klein im Vergleich zur Dauer eines Dialogbeitrags: Zwischen Worterkennung und Syntaxanalyse beträgt sie eine Worthypothese, zwischen Prosodieerkennung und Syntaxanalyse nur eine Silbe, und zwischen Syntax- und Semantikanalyse eine syntaktische Phrase. Die Bedeutung liegt in der Verarbeitung sehr langer Redebeiträge im Hinblick auf echtzeitfähige Systeme.

Für die inkrementelle Analyse mußten zum einen völlig neue Algorithmen entwickelt werden. Zum anderen besteht ihr Preis darin, daß der fehlende rechte Kontext zu einer erheblichen Vergrößerung des Suchraumes z.B. bei der Syntaxanalyse führt. Die Parallelverarbeitung kann dies nur zu einem kleinen Teil ausgleichen. Dem Problem wurde vor allem mit hoher *Interaktivität* begegnet, d.h. mit massivem (asynchronen) Nachrichtenaustausch zwischen den Modulen, um z.B. semantische Restriktionen möglichst frühzeitig in die Syntaxanalyse einfließen zu lassen. Eine weitere Möglichkeit, den Suchraum bei der Syntax- und Semantikanalyse zu begrenzen, bietet die Prosodieerkennung. Beispielsweise unterstützen prosodisch detektierte Phrasengrenzen die Syntaxanalyse bei der Gliederung der Rede in syntaktische Einheiten.

Diese Arbeit entstand bei der Entwicklung des Akzent- Phrasengrenzen- und Satzmoduserkenners für das INTARC-System (in der letzten INTARC-Version wurde der Akzenterkenner nicht mehr verwendet, dafür enthält sie einen Fokuserkenner als eigenständige Entwicklung [Pet95, EK96], die nicht Teil dieser Arbeit ist).

Das INTARC-System wurde im Vergleich zum Verbmobil-System mit wesentlich geringerem Personalaufwand entwickelt. Dies zeigt sich in der reduzierten Funktionalität der einzelnen Komponenten, z.B. hat die Grammatik eine geringere Abdeckung, die englische Generierung ist tabellengesteuert, die Sprachsynthese ein kommerzielles Produkt. Es wurde jedoch ein vollständiges Übersetzungssystem mit Spracheingabe und Sprachausgabe erstellt, die Kommunikationssoftware wurde bereits in den Verbmobil-Forschungsprototyp übernommen, und andere Methoden, z.B. die Phrasengrenzenerkennung ohne Wortinformation, werden in zukünftige Verbmobil-Systeme einfließen.

1.5 Beitrag dieser Arbeit zum wissenschaftlichen Fortschritt

Mit der Verbmobil-Korpus liegen erstmals größere Mengen prosodisch etikettierter deutscher Spontansprache vor. Der Verbmobil-Forschungsprototyp und das INTARC-System sind die weltweit ersten sprachverstehenden Systeme, die automatisch detektierte prosodische Ereignisse wie Akzente und Phrasengrenzen zur linguistischen Analyse einsetzen. Während das an der Universität Erlangen etwa zeitgleich entwickelte Prosodiemodul [Kie97, Kom97] für den Forschungsprototyp dem Worterkenner *nach*geschaltet ist, also neben dem Sprachsignal auch Worthypothesen als Eingabe erhält, arbeitet das INTARC-Prosodiemodul unabhängig von der Worterkennung und zudem inkrementell. Dies geschah nicht nur im Hinblick auf den alternativen linguistischen Worterkenner in früheren INTARC-Versionen, der dem Prosodiemodul *nach*geschaltet war, sondern auch, um zu untersuchen, wie gut Prosodieerkennung ohne Wortinformation möglich ist. Um dafür eine Obergrenze abschätzen zu können, wurde ein Teil des Verbmobil-

Korpus so verfremdet, daß die prosodischen Charakteristika erhalten bleiben, die Verständlichkeit aber verloren ging; in dieser verfremdeten Sprache versuchten 11 Testpersonen, Akzente und Phrasengrenzen herauszuhören.

Mit dem INTARC-Prosodiemodul wurden Detektoren für Akzente und Phrasengrenzen entwickelt, wobei an starken Phrasengrenzen auch der Satzmodus klassifiziert wurde. Die inkrementelle Verarbeitung erforderte die Entwicklung neuer Verfahren zur Merkmalberechnung, z.B. zur Grundfrequenzinterpolation und zur robusten Dekomposition der Grundfrequenz (die Zerlegung der Tonhöhenbewegung in lokale und globale Schwankungen).

Da ohne Worthypothesen die zu klassifizierenden Einheiten, nämlich Wörter und Wortgrenzen, nicht festliegen, wurde ein Verfahren zur Silbenkerndetektion reimplementiert, verbessert und an die Aufgabenstellung angepaßt.

Die den Detektoren zugrundeliegenden Klassifikatoren wurden mit den prosodisch etikettierten deutschen Sprachdaten trainiert und getestet. Es wurde auch untersucht, inwieweit sich die prosodisch detektierten Phrasengrenzen zur Dialogaktsegmentierung einsetzen lassen.

Das Prosodiemodul wurde in das INTARC-System integriert, wobei ein neues Verfahren zur Kopplung mit dem Syntaxmodul erarbeitet wurde. In weiteren Versuchsreihen wurde die Auswirkung auf das Syntax- und das Semantikmodul untersucht.

1.6 Überblick

Die Arbeit ist folgendermaßen gegliedert: **Kapitel 2** befaßt sich mit der Problematik der Prosodieerkennung in der ASV und stellt bisherige Arbeiten zur Akzent- Phrasengrenzen- und Satzmoduserkennung vor.

Das grundlegende Vorgehen bei der Prosodieerkennung mit Methoden der Mustererkennung wird in **Kapitel 3** dargestellt.

Kapitel 4 befaßt sich mit den manuell erstellten Etiketten zu den verwendeten Sprachdaten, die zum Training des Prosodieerkenners benutzt wurden.

Kapitel 5 geht nach einer kurzen Einführung in den Mechanismus der menschlichen Sprachproduktion detailliert auf die entwickelten Verfahren zur Extraktion prosodischer Merkmale ein.

In **Kapitel 6** werden die Erkenner für Satzmodus, Akzente und Phrasengrenzen beschrieben und Erkennungsraten angegeben.

In **Kapitel 7** wird das INTARC-System vorgestellt und die Integration des Phrasengrenzen- und Satzmoduserkenners in das System sowie ihre Auswirkung auf die Performanz behandelt.

Kapitel 8 werden Perzeptionsexperimente an verfremdeter Sprache beschrieben, in denen menschliche Hörer die Aufgabe hatten, ohne Wortinformation Akzente und Phrasengrenzen zu erkennen.

Einen Ausblick auf die Fortführung dieser Arbeit gibt **Kapitel 9**, ihre wesentlichen Ergebnisse werden in **Kapitel 10** noch einmal zusammengefaßt.

Kapitel 2

Prosodie in der ASV

Da die Prosodie in der zwischenmenschlichen Kommunikation auf nahezu allen sprachlichen Ebenen eine unterstützende Rolle spielt, liegt es nahe, prosodische Information auch in der ASV zu verwenden. Obwohl diese Idee nicht neu ist, siehe z.B. [Lea73, LMS75], gibt es bislang nur sehr wenige sprachverstehende Systeme, die eine Prosodieerkennung in den Analyseprozeß integrieren. In Abschnitt 2.1 werden Probleme bei der Prosodieerkennung dargestellt, Abschnitt 2.2 gibt einen Überblick zur Entwicklung der Prosodieerkennung bis zum heutigen Stand der Technik.

2.1 Probleme

Verglichen mit der Worterkennung steht die Prosodieerkennung noch am Anfang der Entwicklung. Ein wesentlicher Grund dafür ist, daß es kein festes Inventar von Klassen gibt, das den Wörtern in einem Lexikon entspricht. Beispielsweise besitzt die Prosodie gliedernde Funktion nicht nur auf der Ebene von Sätzen und Teilsätzen, sondern auch auf der Ebene von Dialogakten, Redebeiträgen oder (bei Lesesprache) ganzen Abschnitten [Swe97]. Auch ist nicht klar, ob es verschiedene Klassen von prosodischen Grenzen gibt [SBP+92] oder ob es sich um ein Kontinuum von „schwacher Grenze“ zu „starker Grenze“ handelt [dPS94]. Daher werden Anzahl und Art der Klassen einer prosodischen *Funktionsdomäne* von der geplanten Weiterverarbeitung in anderen Modulen bestimmt, also von der Anwendung. Unter einer prosodischen *Funktionsdomäne* wird nach [Kie97] eine Menge von prosodischen Klassen verstanden, die zusammen einen bestimmten prosodischen Problemkreis bilden. Tabelle 2.1 zeigt die drei derzeit in der ASV wichtigsten Funktionsdomänen.

Diese drei Funktionsdomänen können jedoch auch durch andere Klassen

Funktionsdomäne	Beispiele möglicher Klassen
Akzentuierung	akzentuierte Silbe, nicht akzentuierte Silbe
Phrasierung	starke Grenze, schwache Grenze, keine Grenze
Satzmodus	Frage, Aussage, Weiterführung

Tabelle 2.1: *Beispiele prosodischer Funktionsdomänen und der darin enthaltenen Klassen.*

gebildet werden; die prosodischen Etiketten zum Verbmobil-Korpus unterscheiden z.B. vier Akzentstufen und vier Phrasengrenzentypen.

In der ASV noch kaum untersucht wurde bisher z.B. der Bereich der Emotionen [DPW96] (siehe Abschnitt 1.2.5), da zur Zeit noch nicht klar ist, welche Emotionen anhand prosodischer Ausdrucksmittel unterschieden werden können und wie sich mögliche Klassen, z.B. Angst oder Ärger, in einem ASV-System sinnvoll weiterverarbeiten lassen.

Wenn die Funktionsdomänen und ihre Klassen durch die beabsichtigte Anwendung festgelegt sind, besteht das nächste Problem darin, daß es zwischen den *funktionalen* Klassen und den *formalen* Klassen keine eindeutige Abbildung gibt [CL83]: Die formalen Klassen werden durch die prosodischen Ausdrucksmittel bestimmt, die sich in akustischen Parametern wie Grundfrequenz- oder Energieverlauf widerspiegeln. Beispielsweise können Sätze mit steigendem Tonverlauf am Satzende eine formale Klasse bilden, die funktional der Klasse der Fragesätze zuzuordnen ist, aber umgekehrt hat nicht jeder Fragesatz am Ende einen steigenden Tonverlauf. Die Gründe für das Fehlen einer eindeutigen Beziehung zwischen Form und Funktion sind vielfältiger Natur (vergleiche auch [Kie97]):

- *Fakultativität der prosodischen Mittel*: Eine bestimmte linguistische Funktion *kann* durch prosodische Mittel markiert sein, *muß* es aber nicht, wenn diese Funktion bereits durch andere, z.B. grammatische Mittel ausgedrückt wird. Beispielsweise können Fragen durch Verb-Erst-Stellung als solche gekennzeichnet werden und deshalb steigenden oder fallenden Tonverlauf am Satzende haben. In diesem Zusammenhang spricht man auch von *freien Varianten* [Bat89a, S. 155], wenn prosodische Mittel (wie die Intonation) unterschiedliche Ausprägungen einnehmen können, weil sie den Sinngehalt einer Äußerung ohnehin nicht nicht beeinflussen.

- *Gegenseitige Beeinflussung der prosodischen Mittel*: Eine bestimmte Funktion kann durch verschiedene prosodische Mittel ausgedrückt werden. Beispielsweise kann Akzentuierung durch Dehnung und/oder durch Änderung der Tonhöhe markiert sein. Oft stehen die prosodischen Mittel in einer sog. *trading relation* zueinander [Bat89a, S. 159], d.h. die schwächere Ausprägung des einen Mittels kann durch stärkere Ausprägung des anderen ausgeglichen werden [Fry58]. Gegenseitige Beeinflussung liegt auch vor, wenn zeitlich aufeinanderfolgende prosodische Markierungen fließend ineinander übergehen, z.B. die Tonhöhenbewegung aufgrund von Akzentuierung und Phrasierung.

- *Mehrdeutigkeit der prosodischen Mittel*: Verschiedene Funktionen können durch die gleichen prosodischen Mittel gekennzeichnet werden. Beispielsweise kann ein ansteigender Tonverlauf am Satzende eine Frage indizieren, wie in *„Fünfzehn Uhr vier?“*, aber auch einen Akzent[1] auf der letzten Silbe, wie in *„Fünfzehn Uhr* **vier***“*.

- *Sprecherabhängigkeit*: Die Wahl und die Ausprägung der prosodischen Mittel ist individuell sehr verschieden. Dies geht z.T. auf die Überlagerung durch paralinguistische und indexikalische Funktionen zurück (wie Emotion und Dialekt, siehe Abschnitt 1.2.5), jedoch ist deren Markierung selbst ebenfalls sprecherabhängig.

Desweiteren bestimmt auch die Signalqualität — wie in der ASV allgemein — den Schwierigkeitsgrad der Prosodieerkennung. Dabei spielen Störgeräusche wie Telephonklingeln oder Hintergrundstimmen ebenso eine Rolle wie Lautstärkeschwankungen durch wechselnden Abstand zum Mikrophon. Bei Telephonsprache bewirkt die Bandbegrenzung auf 300–3400 Hz, daß die Grundfrequenz im Sprachsignal oft nicht mehr vorhanden ist; sie ist zwar anhand der Harmonischen[2] rekonstruierbar, die automatische Grundfrequenz-Analyse wird dann aber schwieriger. Selbst bei hoher Signalqualität ist die Grundfrequenz-Analyse keineswegs trivial, siehe Abschnitt 5.2.1, sie stellt aber eine wesentliche Voraussetzung für die Prosodieerkennung dar. Auch die Messung der Silben- oder Silbenkerndauer ist fehlerbehaftet.

In Abschnitt 1.2 wurde ausgeführt, daß die Prosodie in der sprachlichen Kommunikation eine wichtige Rolle spielt. Deshalb sollte sie trotz der hier genannten Schwierigkeiten auch in der ASV als zusätzliche Informationsquelle genutzt werden.

---

[1] Die Mehrdeutigkeit besteht zumindest dann, wenn das Wort *„vier“* mit Talakzent (statt dem Standard-Gipfelakzent) gesprochen wird statt einem (Standard-) Gipfelakzent, so daß der Ton in zur Mitte des Worts hin abfällt und zum Ende hin wieder ansteigt.

[2] Die Harmonischen sind die ganzzahligen Vielfachen der Grundfrequenz.

2.2 Von den Anfängen bis zum Stand der Technik

Die Untersuchung prosodischer Phänomene in der Linguistik und der Phonetik begann mit der Phänomenologie: An wenigen prototypische Beispielen, die sorgfältig ausgewählt oder extra dafür konstruiert worden sind (wie in [Fry58] oder [NS63]), wurde untersucht, welche prosodischen Phänomene überhaupt auftreten und mit welchen Mitteln die prosodischen Funktionen ausgedrückt werden.

Die Verwendung der Prosodie in der ASV ist dagegen ein relativ junges Forschungsgebiet. Gegenüber der klassischen linguistischen und phonetischen Forschungsausrichtung liegt hier der Schwerpunkt auf der statistischen Modellierung von Merkmalvektoren anhand größerer Stichproben aus realen (im Gegensatz zu künstlich erzeugten) Sprachdaten. Dies wurde möglich durch Methoden zur automatischen Segmentierung von Sprache (in Wörter, Silben oder Phoneme) und die automatische Extraktion prosodischer Merkmale. Die manuelle prosodische Etikettierung stellt aber immer noch ein Problem dar, zum einen wegen der in Abschnitt 2.1 genannten Schwierigkeiten, zum anderen wegen des hohen Zeitaufwands, so daß die bisher verwendeten Stichproben immer noch relativ klein sind [Rey98].

Grundlegende Untersuchungen zur automatischen Erkennung von Akzenten, Phrasengrenzen und des Satzmodus werden in den Abschnitten 2.2.1 bis 2.2.3 aufgeführt, einen gründlicheren Literaturüberblick geben [Kie97, Kom97, Wig92].

Mit Ausnahme des EVAR-Systems und der Verbmobil-Systeme Forschungsprototyp und INTARC gibt es bisher noch keine sprachverstehenden Systeme, die Prosodie zur linguistischen Analyse einsetzen. Dies mag daran liegen, daß sich der Einsatz der Prosodie erst ab einer gewissen Komplexität des Systems lohnt: Die gliedernde Funktion prosodischer Phrasengrenzen beispielsweise kann erst dann wirkungsvoll ausgenutzt werden, wenn ein sprachverstehendes System längere Äußerungen (d.h. länger als ein Satz) überhaupt verarbeiten kann; die disambiguierende Funktion von Akzenten wirkt sich in sprachverstehenden Systemen weit stärker aus als in spracherkennenden Systemen; die Satzmodusbestimmung anhand der Prosodie ist vor allem wichtig bei elliptischen Äußerungen, wie sie in der schwieriger zu verarbeitenden Spontansprache häufig auftreten. Die Nutzung prosodischer Information zahlt sich also erst in komplexeren Systemen voll aus, wie sie seit einigen Jahren existieren. Abschnitt 2.2.4 zeigt, wie eine Satzmodusklassifikation in das EVAR-System integriert wurde, und stellt kurz die Neuerungen dar, die im Verbmobil-Projekt hinsichtlich der Prosodieintegration geleistet wurden.

2.2.1 Akzenterkennung

Akzentuierung wirkt sich u.a. auf die Vokalqualität [Lin63] und die Silbendauer [Kla76] aus. Implizite Modellierung von Akzentuierung in einem HMM-Worterkenner durch separate Modelle für akzentuierte und nicht-akzentuierte Vokale [ADD92], evtl. mit zusätzlicher Dauermodellierung für beide Fälle [Bis92], brachte leichte Verbesserungen der Wortfehlerrate.

In [Wai88] und in [Aul84] wurden Akzentmuster explizit benutzt, um die Lexikongröße bei der Einzelworterkennung zu reduzieren; der Ansatz wurde jedoch nicht in den Worterkenner integriert.

In [NK89] wurden Akzentmerkmale verwendet, um mit ihnen inkompatible, von einem HMM-Erkenner erzeugte Worthypothesen zu verwerfen. Dadurch konnte bei gelesener Sprache die Worterkennungsrate um 5-10 Prozentpunkte (abhängig von der Worthypothesendichte) gesteigert werden.

In [Cam92, Cam95] wurden Energie- und Dauermerkmale zur Akzentdetektion eingesetzt, wobei die Phonemsegmentierung manuell erfolgte. Für eine Sprecherin konnten 81% der akzentuierten Silben detektiert werden bei 5% Einfügungen.

In [WO92] werden in einer von vier Radiosprechern gelesenen Stichprobe akzentuierte Silben detektiert, wobei neben akustischen auch linguistische Merkmale verwendet werden, z.B. eine Markierung für den lexikalischen Akzent. Es wird eine Erkennungsrate von 86% angegeben.

Mit einer Neuimplementierung des Intonationsmodells von Fujisaki (siehe Abschnitt 5.2.2) gelang es in [Geo93], 91% der Akzente in 51 gelesenen japanischen Sätzen allein anhand des Grundfrequenzverlaufs zu erkennen. Allerdings wurde nicht gesagt, wie genau die sog. Akzentkommandos mit den akzentuierten Silben zeitlich korrelieren.

Die genannten Arbeiten unterscheiden sich in mindestens einem der folgenden Punkte vom Akzenterkenner, der in dieser Arbeit vorgestellt wird und

- Spontansprache verarbeitet,
- Akzenthypothesen inkrementell ausgibt: die Verzögerung beträgt weniger als eine Silbe, d.h. es werden keine globalen Merkmale verwendet,
- keine Wortinformation benötigt: die zu klassifizierenden Einheiten (Silben, Wörter) müssen nicht vorgegeben sein, es werden keine Informationen zur Dauer der Phonemsegmente verwendet und keine linguistischen Merkmale.

2.2.2 Phrasengrenzenerkennung

Auch die Untersuchungen zur Phrasengrenzendetektion fanden bisher überwiegend anhand gelesener Sprache statt. Klassifikationsbäume dienen in [Wig92] dazu, um mit Lautdauer-, Grundfrequenz- und Energiemerkmalen, aber auch mit linguistischen Merkmalen Phrasengrenzen zu klassifizieren. In [WO92] wird eine Erkennungsrate von 77% für das Zweiklassenproblem angegeben, in [WO94] 91.5%.

Ein alternatives Vorgehen ist in [SK92, SN94] beschrieben: Mithilfe prototypischer Grundfrequenzkonturen für einzelne Phrasen wird die Grundfrequenzkontur einer Äußerung in Phrasen segmentiert. Durch DP-Suche (**D**ynamische **P**rogrammierung) können 88% der Phrasengrenzen detektiert werden.

In [WH92] wurde ein erstes Ergebnis für Spontansprache anhand der ATIS-Stichprobe [PSZ91] erzielt: Klassifikationsbäume wurden eingesetzt, um mit linguistischen und prosodischen Merkmalen (Akzenten, Intonationsgrenzen, die manuell bestimmt wurden) Phrasengrenzen zu klassifizieren. Als Erkennungsrate wird 90% angegeben.

Die Gruppe um Mari Ostendorf untersuchte die mögliche Anwendung prosodisch detektierter Phrasengrenzen in der Syntaxanalyse. Dabei werden sieben verschiedene Phrasengrenzentypen unterschieden. Die Grammatik wurde so erweitert, daß an jeder Wortgrenze nur bestimmte Phrasengrenzentypen auftreten dürfen. Es wurden gelesene, mehrdeutige Sätze untersucht. Wenn in der gesprochenen Wortkette die Wortgrenzen automatisch klassifiziert werden, kann damit die Anzahl der syntaktischen Analysen um 25% reduziert werden [BP90, PWOB90, OPBW90].

Später verfolgte die gleiche Gruppe einen flexibleren Ansatz, der ohne eine speziell erweiterte Grammatik auskommt und darin dem (unabhängig davon entwickelten) Ansatz in Abschnitt 7.3 ähnelt: Syntaktische Ableitungsbäume werden nachträglich neu bewertet, indem die akustische Bewertungen des Worterkenners und die Wort-Bigrammbewertungen einerseits kombiniert werden mit den prosodischen Bewertungen für die Phrasengrenzen- und Akzentklassen und den Wahrscheinlichkeiten eines akustisch-prosodischen Modells andererseits. Für mehrdeutige, gelesene Sätze wurden je zwei syntaktische Ableitungen manuell ausgewählt; das Modell entschied sich dann in 73% der Fälle für die richtige Ableitung. Abgesehen von der Tatsache, daß diese Untersuchungen an gelesener Sprache durchgeführt wurden[3], ist diese Methode so nicht in einen Syntaxparser integrierbar, weil die nachträgliche Neubewertung bereits eine vollständige

[3] Für das spontansprachliche ATIS-Korpus wurde bereits über Versuche publiziert, bei denen die n bestbewerteten Wortketten mit denselben Methoden nachträglich neu bewertet wurden [POSHF91].

Ableitung voraussetzt.

2.2.3 Satzmodusklassifikation

In [Bat89b] werden Untersuchungen an elizierten[4] Frage/Nichtfrage-Minimalpaaren beschrieben. Durch Diskriminanzanalyse von Grundfrequenzmerkmalen ließ sich mit sprecherweisem *leave one out* bei 355 Äußerungen von 6 Sprechern eine Erkennungsrate von 91% erzielen.

In [BOPSH90] wurde der Satzmodus von Äußerungen klassifiziert, die aus einem ein- oder zweisilbigen, vollständig stimmhaften Wort bestehen und von 5 Sprechern gelesen wurden. Dabei wurden neben Fragen, Aussagen und Weiterführungen noch Befehle und Ausrufe unterschieden. Die verwendeten Grundfrequenz- und Energiemerkmale wurden vektorquantisiert und mit HMM klassifiziert. Dabei wurden Erkennungsraten von bis zu 89% erreicht.

In [DZ90] wurde ein Klassifikationsbaum trainiert, um zwischen hohen und tiefen Grenztönen anhand von Grundfrequenzmerkmalen zu unterscheiden. Mit dem spontansprachlichen VOYAGER-Korpus [SZ90] wurde für 431 Äußerungen eine Erkennungsrate von 90% erzielt.

Keiner dieser Ansätze wurde bisher in ein sprachverstehendes System integriert.

2.2.4 Integration von Prosodie in ASV-Systeme

Das EVAR-System

Das erste sprachverstehende System, in das eine Komponente zur Prosodieerkennung integriert wurde, ist das experimentelle Zugauskunftssystem EVAR (**E**rkennen, **V**erstehen, **A**ntworten, **R**ückfragen) [EFK+92, MKE+94]. Seine Aufgabe ist es, einen natürlichen Dialog in gesprochener Sprache zu führen, in dem sich ein Benutzer über Intercity-Verbindungen informieren kann.

In realen Dialogen treten häufig elliptische Äußerungen auf, bei denen der Benutzer nur eine Uhrzeit wiederholt. Hier wird die Satzmodusklassifikation zur Dialogsteuerung eingesetzt [Ott93, KKK+93]: Unterschieden werden steigende, fallende und progrediente Intonation entsprechend den Satzmodi Frage, Aussage, Weiterführung (bzw. Rückmeldung).

- Wiederholt der Benutzer eine Uhrzeit mit interrogativer Intonation, wird sie vom System wiederholt.

[4]Die Testsätze waren in modussteuernde Kontexte eingebettet, d.h. die Versuchspersonen hatten die Aufgabe, sich aufgrund eines Kontextsatzes oder einer Situationsbeschreibung in eine Situation hineinzudenken und die Testsätze entsprechend zu intonieren.

- Terminale Intonation wird vom System als Bestätigung interpretiert, deshalb erfolgt keine Reaktion.

- Progrediente Intonation wird vom System interpretiert als „Ja, ich habe verstanden“, die Reaktion hängt dann davon ab, ob die Uhrzeit vollständig wiederholt wurde, ob nur die Stunden oder nur die Minuten wiederholt wurden: Werden nur die Stunden wiederholt, wird dies interpretiert als „Ich habe die Stunden verstanden, aber nicht die Minuten

30 Uhrzeiten wurden jeweils mit den drei verschiedenen Intonationen gelesen, so daß sich 360 Äußerungen ergaben. Davon wurden 15 wegen groben Fehlern bei der Grundfrequenzanalyse aussortiert, ebenso 23 Äußerungen, bei denen Perzeptionsexperimente auf falsche Intonation schließen ließen. Ein Normalverteilungsklassifikator wurde mit verschiedenen Grundfrequenzmerkmalen trainiert und getestet; bei sprecherweisem *leave one out* ergab sich eine Erkennungsrate von 88%. Wurde der Klassifikator mit allen Sprechern trainiert und mit 200 anderen, von „naiven“ Sprechern geäußerten Uhrzeiten getestet, ergab sich eine Erkennungsrate von 71%; die Fehlklassifikationen gehen vor allem auf Äußerungen mit progredienter Intonation zurück. Auf einer kleinen Teststichprobe, die realen Dialogen entnommen wurde (5 Äußerungen waren interrogativ, 7 terminal und 17 progredient), wurden mit dem gleichen Klassifikator 7 der progredienten und alle terminalen und interrogativen Äußerungen richtig klassifiziert [KKK+93].

Verbmobil

Der innerhalb des Verbmobil-Projekts entwickelte Forschungsprototyp und das zeitgleich entstandene INTARC-System sind die ersten Systeme, die prosodisch detektierte Phrasengrenzen gewinnbringend in der linguistischen Analyse einsetzen [NNK+97, SEG+97].

Im Forschungsprototyp bewirken prosodisch detektierte Phrasengrenzen aufgrund einer speziell dafür erweiterten Grammatik eine Verkürzung der Analysezeit um 98% und eine Reduktion der Zahl der syntaktischen Ableitungen um 96% [NNK+97].

In Gegensatz dazu ist die Grammatik im INTARC-System auf der syntaktischen Ebene noch nicht um prosodischen Phrasengrenzen erweitert worden. Die Prosodie greift während der Syntaxanalyse vielmehr indirekt durch eine zusätzliche prosodische Bewertung von Wortgrenzen ein, siehe Abschnitt 7.3.1. Dabei wird durch Favorisieren der richtigen Worthypothesen der Suchraum eingeschränkt bzw. die Worterkennungsrate (hier gemessen an den syntaktisch analysierbaren Anfangsstücken von Wortketten) erhöht. Erst auf der semantischen Ebene wurde die Grammatik um Restriktionen bezüglich der prosodischen Phrasengrenzen und des Satzmodus erweitert, um die Anzahl der Lesarten zu reduzieren, siehe Abschnitt 7.3.2.

Während im Forschungsprototyp Akzente derzeit noch nicht sinnvoll verarbeitet werden können, dienen im INTARC-System Fokusakzente zur flachen semantischen Analyse: Falls die tiefe Analyse fehlschlägt, erfolgt anhand der besten Wortkette und des Fokusakzents eine grobe dialogaktbasierte Übersetzung [EK96, SEG+97], siehe auch Abschnitt 7.1.

Kapitel 3

Ausgewählte Methoden der Mustererkennung

Das Problem der Spracherkennung und der hier behandelten Prosodieerkennung ist der Spezialfall eines Mustererkennungsproblems. Dieses Kapitel, das auf [Nie83, Nie90] basiert, gibt einen kurzen Überblick über die in dieser Arbeit gewählten Methoden und insbesondere den Normalverteilungsklassifikator.

Unter einem *Muster* versteht man eine meßbare Größe, die im allgemeinen als Funktion $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^m$ dargestellt wird. Für den Problemkreis Sprachverarbeitung sind diese Muster die Sprachsignale $f(t) \in \mathbb{R}, t \in \mathbb{R}$, die den Schalldruck als Funktion der Zeit angeben.

Nach [Nie83] besteht ein System zur Mustererkennung aus den Schritten *Aufnahme* des Musters, *Vorverarbeitung*, *Merkmalgewinnung* und *Klassifikation*.

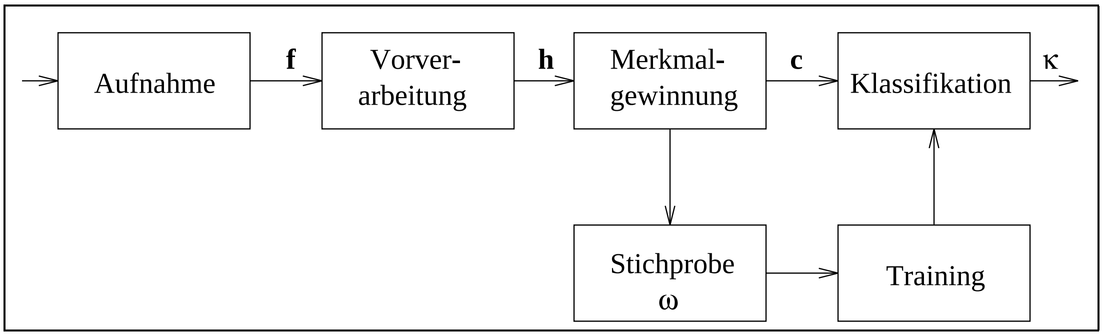

Abbildung 3.1: *Allgemeine Struktur eines Mustererkennungssystems, nach [Nie90].*

Abbildung 3.1 zeigt das allgemeine Schema: Bei der Aufnahme erfolgt die Digitalisierung des Musters, damit es von Rechner bearbeitet werden kann. Danach folgt die Vorverarbeitung des digitalisierten Musters $\mathbf{f}(\mathbf{x})$ zur Verbesserung der

Signalqualität hinsichtlich der interessierenden Eigenschaften; dies erfolgt mit relativ einfachen Mitteln, z.B. einem Filter zur Rauschunterdrückung, und kann auch zum nächsten Schritt gezählt werden.

Aus dem resultierenden Muster $\mathbf{h}(\mathbf{x})$ werden die hinsichtlich der Klassifikation relevanten Eigenschaften extrahiert, die sogenannten *Merkmale*. Die Merkmalgewinnung kann als Transformation des Musters $\mathbf{h}(\mathbf{x})$ in den Merkmalvektor $\mathbf{c} \in \mathbb{R}^n$ aufgefaßt werden. Der wesentliche Aspekt dabei ist die Datenreduktion, wobei gleichzeitig die relevante Information möglichst erhalten bleiben soll. In der Sprachverarbeitung liegt n typischerweise zwischen 10 und 100. Die Bestimmung der Merkmale erfolgt in der Regel nach heuristischen Methoden.

Die Klassifikation schließlich ist die Abbildung des Merkmalvektors auf einen Klassennamen oder Klassenindex

$$\mathbf{c} \rightarrow \kappa \in \{1, ..., k\} \tag{3.1}$$

wobei k Klassen $\Omega_1, ..., \Omega_k$ unterschieden werden.

Die Klassifikation kann auf Regeln basieren, die ein menschlicher Experte aufgestellt hat, oder auf Entscheidungsregeln, die aufgrund einer *Teststichprobe* ω während des *Trainings* automatisch gelernt wurden.

Falls keine Klassen vorgegeben sind und es zunächst darauf ankommt, möglichst homogene Klassen zu bilden (z.B. Tarifklassen bei einer Haftpflichtversicherung), können die Klassen durch Analyse von Häufungsgebieten ermittelt werden; dies wird als *unüberwachtes Lernen* bezeichnet. Im Fall der Prosodieerkennung sind die Klassen vorgegeben in Form einer *handklassifizierten* bzw. *handetikettierten* Teststichprobe (die Handklassifikation ist in Bild 3.1 nicht dargestellt). Die Ableitung der Entscheidungsregeln daraus wird als *überwachtes Lernen* bezeichnet.

3.1 Aufnahme und Vorverarbeitung

Zu Beginn dieses Kapitels wurde ein Sprachsignal durch den Schalldruck als Funktion der Zeit definiert. Damit diese kontinuierliche Funktion $s(t)$ im Rechner bearbeitet werden kann, muß sie in eine Folge von diskreten Abtastwerten $s_n = s(nT_a), n \in \mathbb{N}$ umgewandelt werden mit dem zeitlichen *Abstastintervall* T_a. Dazu wird das Schalldrucksignal zunächst mit Mikrophon und Verstärker in ein elektrisches Signal umgewandelt und dann im Analog-Digital-Wandler digitalisiert.

Voraussetzung zur Abtastung ist die spektrale Begrenzung des Eingangssignals auf die Hälfte der *Abtastfrequenz* bzw. *Abtastrate* $f_a = 1/T_a$. Dann kann nach dem Abtasttheorem [Sch88, Seite 50] das kontinuierliche Signal aus dem abgetasteten exakt rekonstruiert werden. Die Bandbegrenzung auf $f_a/2$ wird durch vorherige Tiefpaßfilterung in für praktische Zwecke ausreichendem Maß sichergestellt.

Die Quantisierung der Amplitudenwerte erfordert, daß diese Werte einen gewissen Bereich nicht verlassen. Dazu wird der Verstärker so eingestellt, daß das Signal nicht *übersteuert* ist. Der Aussteuerungsbereich wird in gleichgroße Intervalle Q unterteilt. Der dadurch entstehende Quantisierungsfehler hängt im wesentlichen von der Größe der Quantisierungsstufe Q ab, in geringerem Maß von der Verteilungsdichtefunktion des Sprachsignals, und kann mithilfe des Quantisierungstheorems [Sch88, Seite 104] abgeschätzt werden.

Das Verbmobil-Korpus ist mit 16 kHz abgetastet und mit 16 bit quantisiert. Im folgenden wird unter einem Muster bzw. Sprachsignal die diskrete Folge der Abtastwerte s_n verstanden. Die Folge ist stets endlich; sie kann einem ganzen Redebeitrag entsprechen oder auch nur einem Ausschnitt von z.B. 10 ms Dauer.

Die Vorverarbeitung zielt nach [Nie83] darauf ab, mit relativ geringem Aufwand die Qualität von Mustern zu verbessern, so daß der Aufwand der nachfolgenden Verarbeitungsschritte verringert oder die Klassifikationsleistung erhöht wird. Bei der Grundfrequenzanalyse beispielsweise liegt die interessierende Information im unteren Frequenzbereich. Durch nochmalige Tiefpaßfilterung und Abtastung, das sog. *Downsampling*, wird der Aufwand der Analyse verringert. Man könnte auch die Grundfrequenzanalyse selbst zur Vorverarbeitung rechnen, weil in dieser Arbeit dafür ein fertiges Verfahren benutzt wurde, siehe Abschnitt 5.2.1.

Einige Autoren zählen auch die Merkmalgewinnung zur Vorverarbeitung, hier soll es umgekehrt sein: Alle Verarbeitungsschritte nach der Aufnahme zählen zur Merkmalgewinnung, die in Kapitel 5 ausführlich behandelt wird. Es gibt jedoch eine Ausnahme: Bei einigen Karlsruher Dialogen wich der Mittelwert des Sprachsignals erheblich von der Nullinie ab, offenbar wegen einer Fehlbedienung oder Dejustierung der Aufnahmegeräte. Daher wurden alle Turns vorab mittelwertbereinigt.

3.2 Merkmalgewinnung

Die Merkmalgewinnung besteht in der Transformation eines Musters, hier des abgetasteten Sprachsignals s_n, in den Merkmalvektor $\mathbf{c}$. Durch diese Transformation soll zum einen die Datenmenge reduziert werden, zum anderen soll die hinsichtlich der Klassifikation relevante Information möglichst erhalten bleiben. Die Merkmale sind daher so zu wählen, daß sie die Klassen möglichst gut trennen: Muster aus gleichen Klassen sollen ähnliche Merkmale haben, Häufungsgebiete im Merkmalsraum bilden, während Muster aus verschiedenen Klassen möglichst entfernte Gebiete im Merkmalsraum einnehmen sollten.

Es gibt numerische Gütemaße für diese Kriterien, in der Regel ist es aber nicht aufwendiger, die Güte direkt durch ein Klassifikationsexperiment zu bestimmen, wodurch man auch gleich die unmittelbar anschaulichen Erkennungsraten erhält, die in Abschnitt 3.3.1 definiert werden.

Es gibt allerdings keine Möglichkeit, systematisch gute Merkmale zu finden, es sei denn, man beschränkt sich auf eine lineare Transformation. Im allgemeinen findet man mit heuristischen Methoden und Expertenwissen über den Problemkreis bessere Merkmale. Im Fall der Prosodieerkennung beziehen sich diese Merkmale auf die Dauer (von Silben, Pausen, etc., allgemein: auf die zeitliche Struktur), die Intonation (den Grundfrequenzverlauf), und die Lautheit (den Energieverlauf). Die in der Literatur vorgeschlagenen Merkmale unterscheiden sich nur durch die konkrete Parametrisierung dieser drei Größen.

Bei dem hier beschriebenen Prosodieerkenner beschreiben die Merkmale zunächst den Grundfrequenz- und den Energieverlauf. Durch die Berechnung der mittleren Grundfrequenz und Energie für jeden *Frame* (Sprachsignalabschnitt fester Länge, hier 160 Abtastwerte entsprechend 10 ms, siehe dazu auch Abschnitt 3.4) wird die Datenmenge reduziert. Die durch Interpolation der Grundfrequenz, Dekomposition und zeitliche Ableitung gebildeten *Basismerkmale* (siehe Abschnitt 5.5) sind bereits die Grundlage der Akzentklassifikation. Das Dauerkriterium kommt erst durch die Nachbearbeitung ins Spiel.

Zur Phrasengrenzen- und Satzmodusklassifikation werden mithilfe der Energie Silbenkerne detektiert. Die Klassifikation erfolgt dann für jedes Fenster aus vier Silben. Dabei werden die Basismerkmale nur an bestimmten Punkten des Fensters verwendet, wodurch sich eine weitere Datenreduktion ergibt. Die Dauer der Silbenkerne und ihre Abstände ergeben zeitliche Merkmale.

3.3 Klassifikation

Die Klassifikation ist die Abbildung des Merkmalvektors auf einen Klassennamen oder Klassenindex:

$$\mathbf{c} \rightarrow \kappa \in \{1, ..., k\}$$

wobei k Klassen $\Omega_1, ..., \Omega_k$ unterschieden werden. Die optimale Entscheidungsregel, die die Fehlerwahrscheinlichkeit minimiert, wählt die Klasse Ω_λ mit der maximalen *a posteriori Wahrscheinlichkeit* $p(\Omega_\lambda|\mathbf{c})$ aus. Sie gibt die Wahrscheinlichkeit der Klasse Ω_λ bei gegebenem Merkmalvektor $\mathbf{c}$ an.

Der *statistische Klassifikator* schätzt diese a posteriori Wahrscheinlichkeiten direkt aus der handklassifizierten Trainingsstichprobe. Mit der Bayes-Regel ergibt sich

$$p(\Omega_\lambda|\mathbf{c}) = \frac{p_\lambda p(\mathbf{c}|\Omega_\lambda)}{\sum_{\kappa=1}^{k} p_\kappa p(\mathbf{c}|\Omega_\kappa)} \quad . \tag{3.2}$$

Da der Nenner bei gegebenem $\mathbf{c}$ konstant ist, müssen für jede Klasse λ nur noch die a priori Wahrscheinlichkeit p_λ und die Verteilungsdichte $p(\mathbf{c}|\Omega_\lambda)$ bestimmt werden.

Der wichtigste Vertreter des statistischen Klassifikators ist der Normalverteilungsklassifikator; dabei wird angenommen, daß die Merkmale klassenweise normalverteilt sind. Dann ist

$$p(\mathbf{c}|\Omega_\lambda) = p(\mathbf{c}|\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda) \tag{3.3}$$

$$= \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_\lambda|}} \, \mathrm{e}^{-\frac{1}{2}(\mathbf{c} - \boldsymbol{\mu}_\lambda)_t \boldsymbol{\Sigma}_\lambda^{-1} (\mathbf{c} - \boldsymbol{\mu}_\lambda)} \quad . \tag{3.4}$$

Die Klassifikation erfolgt dann mit der Entscheidungsregel:

$$\kappa = \underset{1 \le \kappa \le k}{\operatorname{argmax}} \{p(\mathbf{c}|\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)\} \tag{3.5}$$

Einen allgemeineren Ansatz verfolgen die *verteilungsfreien Klassifikatoren* mit der Entscheidungsregel

$$\kappa = \underset{1 \le \kappa \le k}{\operatorname{argmax}} \{d_\lambda(\mathbf{c}, \mathbf{a})\} \tag{3.6}$$

mit den Diskriminanzfunktionen $d_\lambda(\mathbf{c}, \mathbf{a})$. Die Aufgabe besteht dann darin, eine Familie von parametrischen Funktionen d_λ zu bestimmen und die tatsächlichen Parameter $\mathbf{a}_1, ..., \mathbf{a}_k$ zu optimieren. Sei

$$\boldsymbol{\delta} = \Big(\delta_1(\mathbf{c}), ..., \delta_k(\mathbf{c})\Big) \text{ mit } \delta_\kappa(\mathbf{c}) = \begin{cases} 1 \text{ falls } \mathbf{c} \in \Omega_\kappa \\ 0 \text{ sonst} \end{cases} \tag{3.7}$$

die ideale Diskriminanzfunktion, dann kann der Fehler

$$\varepsilon = E\{(\boldsymbol{\delta} - \boldsymbol{d})^2\} \tag{3.8}$$

als Optimierungskriterium dienen. Wenn $\mathbf{d}$ hinreichend komplex ist, dann ist die Funktion $\mathbf{d}^*$, die den Fehler in 3.8 minimiert, identisch mit dem Vektor der a posteriori Wahrscheinlichkeiten [Nie90]:

$$\mathbf{d}^* = \Big(p(\Omega_1|\mathbf{c}), ..., p(\Omega_k|\mathbf{c})\Big) \quad . \tag{3.9}$$

Meist wird ein polynomialer Klassifikator verwendet (jede Funktion, die stetig ist und n-te Ableitungen hat, kann durch eine Taylorreihe angenähert werden), wobei man sich oft auf den quadratischen Klassifikator beschränkt. Dieser ist in seiner Struktur identisch mit dem Normalverteilungsklassifikator; allerdings werden seine Parameter anders berechnet, so daß es sich um einen anderen Klassifikatortyp handelt. Sein Vorteil ist, daß keine Annahmen über die klassenbedingten Verteilungen $p(\Omega_\kappa)$ (z.B. Normalverteilung) gemacht werden müssen. Insbesondere führt die Hinzunahme ungeeigneter Merkmale zu keiner Verschlechterung des Klassifikationsergebnisses wie beim Normalverteilungsklassifikator. Dafür berechnet er nur im Idealfall a posteriori Wahrscheinlichkeiten, in der Praxis gilt oft nicht einmal $0 \leq d_\lambda(\mathbf{c}, \mathbf{a}) \leq 1$. Da im INTARC-System (siehe Kapitel 7) keine harten Klassifikatorentscheidungen an andere Module weitergegeben werden sollten, sondern Hypothesen mit Konfidenzmaßen, stellt sich das Problem der Metrik, wenn diese Konfidenzen z.B. im Syntaxmodul mit Wahrscheinlichkeitsmaßen verrechnet werden sollen.

Der statistische und der verteilungsfreie Klassifikator gehören beide zur Gruppe der *parametrischen Klassifikatoren.* Ihnen ist gemeinsam, daß die statistischen Eigenschaften der klassenbedingten Verteilungen $p(\Omega_\kappa)$ durch eine Familie parametrischer Funktionen explizit oder implizit modelliert werden. Die Parameter werden aus der handklassifizierten Lernstichprobe geschätzt, ihre Anzahl ist aber unabhängig von der Stichprobengröße. Dies erlaubt eine Klassifikation mit geringem Aufwand an Speicherplatz und Rechenleistung.

Bei den nichtparametrischen Klassifikatoren werden keine Kenntnisse über die statistischen Eigenschaften der Merkmale vorausgesetzt. Der Preis dafür ist allerdings, daß ein großer Teil der Lernstichprobe gespeichert werden muß (es gibt Verfahren zur Verdichtung der Stichprobe). Ein wichtiger Vertreter ist der Nächster-Nachbar-Klassifikator: Er ordnet einen Merkmalvektor die Klasse des nächstgelegenen Merkmalvektors aus der Lernstichprobe zu. Diese Regel läßt sich verallgemeinern, indem man unter den m nächsten Nachbarn die häufigste Klasse κ bestimmt; dies kann man als Schätzung der bedingten Dichte $p(\mathbf{c}|\Omega_\kappa)$ auffassen. Wegen der hohen Anforderung an Speicherplatz und Rechenleistung kommt dieser Klassifikatortyp nur für Voruntersuchungen an kleinen Stichproben in Frage.

Zur Klassifikation wird oft sehr erfolgreich ein *Mehrschichtperzeptron* (*multilayer perceptron* MLP) eingesetzt, ein künstliches Neuronales Netz (KNN) mit spezieller Topologie [Rip96]. MLPs können hier nur sehr knapp beschrieben werden, es sollen aber einige Eigenschaften genannt werden, um anschließend motivieren zu können, welcher Klassifikatortyp in der vorliegenden Arbeit verwendet wurde.

Ein KNN besteht besteht aus einer großen Zahl von einfachen Verarbeitungseinheiten (Neuronen), die hochgradig miteinander vernetzt sind. Ein MLP aus

n Schichten hat eine Eingabeschicht, $n-1$ verborgene Schichten und eine Ausgabeschicht. Die Eingabeschicht besteht aus einem Neuron für jedes Merkmal. Jedes Neuron einer Schicht ist mit allen Neuronen der nächsten Schicht verbunden (und nur mit diesen). Die *Aktivierung* eines Eingabeneurons ist identisch mit dem Zahlenwert des entsprechenden Merkmals. Außerdem gibt es ein spezielles Eingabeneuron mit der konstanten Aktivierung 1, das ausnahmsweise mit den Neuronen aller folgenden Schichten verbunden ist. Die Aktivierung aller übrigen Neuronen wird durch einen Wert zwischen 0 und 1 modelliert; er ergibt sich aus der *gewichteten Summe* der Eingabeaktivierungen gefolgt von einer *Schwellwertoperation*. Da der Trainingsalgorithmus nur differenzierbare Funktionen zuläßt, wird die Schwellwertfunktion durch eine Sigmoidfunktion angenähert. Die Klassifikation erfolgt durch Bestimmung des Ausgabeneurons mit maximaler Aktivierung. Das Lernen des MLPs besteht in der Adaption der Gewichtungsfaktoren an die gewünschte Ausgabe; sie werden z.B. durch den *Back-Propagation-Algorithmus* iterativ optimiert, von dem man zeigen kann, daß er die Fehlerfunktion zu einem lokalen Minimum führt.

MLPs können als verteilungsfreie Klassifikatoren betrachtet werden. Einige Autoren glauben, daß sie Schätzwerte der a posteriori Wahrscheinlichkeiten berechnen, falls sie genügend Gewichte enthalten und das Training zu einem globalen Minimum der Fehlerfunktion führt; andere meinen, daß auch unter diesen Voraussetzungen die Aktivierungen an den Ausgabeneuronen sich nicht wie Wahrscheinlichkeiten verhalten. Hinsichtlich der Verknüpfbarkeit mit Wahrscheinlichkeiten, die in anderen Modulen berechnet werden, stellt sich daher das gleiche Problem wie bei den verteilungsfreien Klassifikatoren.

In [Kom97] wurden Experimente zur Phrasengrenzenerkennung mit verschiedenen Klassifikatoren angestellt: verschiedenen MLPs, polynomialen Klassifikatoren (quadratisch und kubisch) und einem Normalverteilungsklassifikator mit Mischverteilungen. Zunächst erwiesen sich die MLPs und der quadratische Klassifikator untereinander als etwa gleichwertig und besser als der Normalverteilungsklassifikator; der kubische Klassifikator war wegen numerischer Schwierigkeiten nicht mehr trainierbar. Nach Hinzunahme weiterer Merkmale traten schon beim quadratischen Klassifikator numerische Probleme auf. Der Normalverteilungsklassifikator verbesserte sich nicht, während sich die Erkennungsrate der MLPs weiter erhöhte.

Das größte Problem ist die aufwendige Trainingsphase. Der Back-Propagation-Algorithmus konvergiert zwar gegen ein lokales Minimum, aber er konvergiert langsam, da die partiellen Ableitungen der Fehlerfunktion nur die Richtung angeben, in der sich die Gewichte ändern müssen, nicht aber die optimale Schrittweite. Diese kann heuristisch beeinflußt werden durch die *Lernrate* und einen Trägheitsfaktor, der das Oszillieren der Gewichte verhindert. Alternativ kann der Quick-Propagation-Algorithmus benutzt werden, der eine grobe Schätzung der optimalen Schrittweite benutzt. Er konvergiert oft schneller, aber er konvergiert nicht notwendigerweise.

Neben der Lernrate und dem Trägheitsfaktor müssen die Anzahl der verborgenen Schichten und ihre Dimensionierung experimentell ermittelt werden. Für eine gegebene Merkmalsextraktion und Handklassifikation sind daher mehrere Klassifikationsexperimente mit Training und Test durchzuführen, deren Dauer die eines vergleichbaren Experiments mit einem Normalverteilungsklassifikator um Größenordnungen übersteigt. Forschungsschwerpunkte dieser Arbeit waren jedoch u.a. die Entwicklung geeigneter Merkmale und die Suche nach verbesserten Referenzetiketten für diese Merkmalvektoren (die Referenzetiketten wurden in verschiedener Weise aus den wortbezogenen Handetiketten abgeleitet). Deshalb wurden vor allem aus Zeitgründen keine Untersuchungen mit MLPs angestellt. Hinzu kommt das Problem der Interpretierbarkeit der Ausgabeaktivierungen (siehe oben) und die — im Vergleich zum Normalverteilungsklassifikator — mangelnde Robustheit selbst gegenüber kleinen Änderungen der statistischen Eigenschaften der Merkmale [Kie93], wie sie z.B. durch veränderte Aufnahmebedingungen hervorgerufen werden (das INTARC-System wurde an verschiedenen Standorten vorgeführt).

Die numerischen Probleme der Erlanger Gruppe mit dem Polynomklassifikator ließen es nicht lohnenswert erscheinen, mit diesem Klassifikator zu experimentieren, zumal auch er keine a posteriori Wahrscheinlichkeiten berechnet.

Daher wurden alle Klassifikationsexperimente in dieser Arbeit mit dem Normalverteilungsklassifikator durchgeführt, auch wenn Performanzverluste dadurch hingenommen werden mußten, daß einige Merkmale nur in sehr grober Näherung normalverteilt sind. Dem Problem, daß sich bei Hinzunahme schlechter Merkmale auch die Erkennungsrate verschlechtert, wurde durch ein Merkmalsauswahlverfahren begegnet.

3.3.1 Qualitätsmaße

Wenn ein Klassifikator trainiert ist, möchte man seine Qualität messen. Dies geschieht durch Klassifikation einer handklassifizierten Teststichprobe und Bestimmung der *Erkennungsrate* RR. Sie ist definiert als

$$RR = \frac{\text{Anzahl korrekt klassifizierte Testmuster}}{\text{Anzahl Testmuster}} \cdot 100\% \qquad (3.10)$$

und gibt den Prozentsatz korrekt klassifizierter Testmuster an. Mehr Aufschluß gibt die Verwechslungs- bzw. Konfusionsmatrix, die alle klassenabhängigen

Erkennungsraten

$$RR(\Omega_k) = \frac{\text{Anz. korrekt klassifizierte Testmuster der Klasse } \Omega_k}{\text{Anzahl Testmuster der Klasse } \Omega_k} \cdot 100\% \quad (3.11)$$

enthält. Oft wird noch die *mittlere Erkennungsrate* $avRR$

$$avRR = \frac{1}{k} \sum_{\kappa=1}^{k} RR(\Omega_k) \quad (3.12)$$

angegeben; wenn alle Klassen gleich häufig auftreten, ist sie identisch mit der Erkennungsrate, denn die Erkennungsrate kann auch ausgedrückt werden als gewichtete mittlere Erkennungsrate

$$RR = \frac{1}{k} \sum_{\kappa=1}^{k} p_\kappa RR(\Omega_k) \quad (3.13)$$

mit den a priori Wahrscheinlichkeiten der Klassen p_κ als Gewichten.

Der Normalverteilungsklassifikator entscheidet sich nach Gleichung 3.5 für die Klasse mit der maximalen a posteriori Wahrscheinlichkeit; dies wird im folgenden als Klassifikation mit der Bayes-Regel bezeichnet. Der hier verwendete Klassifikator kann jedoch auch so betrieben werden, daß er sich für die Klasse mit dem maximalen Wert der Likelihood-Funktion $p(\mathbf{c}|\Omega_\kappa)$ entscheidet, also ohne Berücksichtigung der a priori Wahrscheinlichkeiten p_κ (vergleiche mit Gleichung 3.2). Dies wird als Klassifikation mit der *Maximum-Likelihood-Regel* (ML-Regel) bezeichnet.

Beispielsweise sind die Phrasengrenzen **B3** (siehe Abschnitt 4.1.2) viel seltener als normale Wortgrenzen **B0**. Die Bayes-Regel optimiert die RR und entscheidet sich daher im Zweifelsfall eher für die häufigere Klasse **B0**. Für bestimmte Anwendungen kann es aber günstiger sein, sich eher für **B3** zu entscheiden, z.B. wenn eingefügte **B3** weniger stören als ausgelassene. Dann ist es besser, die ML-Regel anzuwenden, womit die mittlere Erkennungsrate optimiert wird. Die schlechtest mögliche Erkennungsrate ist die maximale p_κ, die schlechtest mögliche mittlere Erkennungsrate bei k Klassen ist dagegen $1/k$.

Üblicherweise wird die handklassifizierte Stichprobe in eine Lern- und eine Teststichprobe aufgeteilt. Wenn die Stichprobe ausreichend groß ist, hat die gemessene Erkennungsrate für Test- ungleich Lernstichprobe etwa den gleichen Wert wie für Test- gleich Lernstichprobe. Für die erforderliche Stichprobengröße gibt es nur grobe Richtwerte; in [Nie83] werden 1000 bis 10000 Muster pro Klasse angegeben, eine andere Faustregel besagt, daß pro zu trainierendem Parameter mindestens 10 Muster nötig sind [Nie89]. Im Fall des Normalverteilungsklassifikators müssen bei n Merkmalen für jede Klasse n a priori Wahrscheinlichkeiten, n Mittelwerte und $n^2/2$ Kovarianzen trainiert werden.

Man kann den Normalverteilungsklassifikator derart vereinfachen, daß von der Kovarianzmatrix nur die Diagonale verwendet wird; das läuft auf die Annahme hinaus, daß die Merkmale statistisch voneinander unabhängig sind. Damit würde man die Anzahl der zu trainierenden Parameter beträchtlich reduzieren, allerdings sind die in dieser Arbeit gewählten Merkmale teilweise bewußt so gewählt, daß Abhängigkeiten bestehen.

Wenn die Stichprobe nicht groß genug ist, um die Dichten der $p(\Omega_\kappa|\mathbf{c})$ schätzen zu können, wird die Erkennungsrate bei Test- gleich Lernstichprobe höher sein als die Erkennungsrate bei Test- ungleich Lernstichprobe. Die Differenz zwischen beiden Werten ist daher ein Maß für die Angemessenheit der Stichprobengröße.

Ein dritter Experimentiermodus ist der des *leave one out*. Dabei wird der Klassifikator mit allen N Mustern bis auf eines trainiert und mit dem ausgelassenen getestet. Dies wird N-mal wiederholt, wobei man jedesmal ein anderes Muster ausläßt. Im Endeffekt wird die Stichprobe dadurch auf $N-1$ Trainings- und N davon verschiedene Testmuster vergrößert. Beim sprecherweisen *leave one out* werden in jedem Durchgang alle Muster eines Sprechers beim Training ausgelassen und zum Testen verwendet.

3.4 Musteranalyse

Bisher wurde davon ausgegangen, daß ein Muster $\mathbf{f}$ bzw. ein Merkmalvektor $\mathbf{c}$ auf genau eine Klasse κ abgebildet wird. In [Nie83] wird dies als Klassifikation von einfachen Mustern bezeichnet, weil das Muster als Ganzes klassifiziert wird. Dagegen geht man bei der *Musteranalyse* davon aus, daß sich ein komplexes Muster aus mehreren einfachen Mustern zusammensetzt. Die Analyse besteht dann in einer Abbildung einer Folge von T Merkmalvektoren in eine Folge von m Klassennamen.

Bei der Worterkennung z.B. sind die m Klassennamen die Wörter der Äußerung[1]. In fließender Rede sind die Wortgrenzen im Sprachsignal nicht vorhanden, deshalb wird das Sprachsignal in Abschnitte fester Länge, sog. *Frames* oder (Sprachsignal-) Analysefenster unterteilt, die so kurz sind, daß Variationen im Zeitbereich von Phonemen noch erfaßt werden. Die Framedauer ist aber ein Kompromiß zwischen Zeit- und Frequenzauflösung: Plosive können kürzer dauern als 5 ms, dagegen ist bei einer Grundfrequenz von 80 Hz (typische Untergrenze bei Männerstimmen) nur in Frames von 12.5 ms Dauer eine komplette Grundschwingung enthalten.

In dieser Arbeit wurden überlappende Frames der Breite 16 ms verwendet, die

[1] Tatsächlich ist die Ausgabe eines Worterkenners ein Worthypothesengraph, der aber als kompakte Darstellung alternativer Wortfolgen aufgefaßt werden kann.

um 10 ms gegeneinander verschoben sind. Dies sind auch in der Spracherkennung übliche Werte.

In jedem Fall ist bei der Musteranalyse die Anzahl der Merkmalvektoren T erheblich größer als die Anzahl m der Klassennamen, da die Grenzen zwischen den einfachen Mustern erst während der Analyse bestimmt wird.

In dieser Arbeit werden Akzente und Phrasengrenzen detektiert, wobei zu bestimmten Phrasengrenzen noch der Satzmodus klassifiziert wird. Akzente und Phrasengrenzen beziehen sich auf Wörter bzw. Wortgrenzen oder auf Silben bzw. Silbengrenzen. Diese Einheiten bzw. Einheitengrenzen sind im Signal nicht vorhanden. Ihre Bestimmung während der Analyse erfolgt mit zwei unterschiedlichen Ansätzen:

Bei der Akzenterkennung (siehe Abschnitt 6.2) wird statt nach akzentuierten Silben nach den Vokalen in akzentuierten Silben gesucht. Aufgrund von Merkmalen, die gut zwischen Vokalen und Nichtvokalen trennen, ist dies auf Frame-Ebene möglich. Durch eine Nachbearbeitung werden die so klassifizierten Frames zu größeren Bereichen zusammengefaßt.

Bei der Phrasengrenzenerkennung (siehe Abschnitt 6.3) werden zunächst die Silbenkerne, die in etwa den Vokalen in der Silbenmitte entsprechen, explizit bestimmt. Jeder Bereich zwischen zwei Silbenkernen enthält eine Silbengrenze, die eine Phrasengrenze sein kann; die zu klassifizierenden Einheiten werden also vorab durch die Silbenkerndetektion bestimmt.

Das Prosodiemodul des Verbmobil-Forschungsprotoyps [Kie97, Kom97] erkennt im selben Sprachmaterial Akzente und Phrasengrenzen. Eingabe dieses Moduls ist neben Grundfrequenz- und Energieverlauf der Worthypothesengraph; jedes Wort (bzw. jede Wortgrenze) darin wird prosodisch klassifiziert. Ein Problem dabei ist, daß zur Klassifikation ein gewisser Kontext benötigt wird, in einem Graph aber für jede Kante verschiedene Kontexte in Frage kommen. Das Problem wurde gelöst, indem für jedes Wort nur der akustisch am besten bewertete Kontext berücksichtigt wurde. Der Vorteil bei diesem Ansatz ist, daß mit den Worthypothesen nicht nur zusätzliche Information vorliegt, sondern auch die zu klassifizierenden Einheiten von vornherein feststehen.

Kapitel 4

Etikettierung

Die in dieser Arbeit entwickelten Prosodiedetektoren basieren auf statistischen Klassifikatoren. Wie in Kapitel 3 ausgeführt, sind zum Trainieren der Klassifikatoren etikettierte Sprachdaten in ausreichender Menge nötig.

Der erste Satzmodusklassifikator im INTARC 1.2-System wurde mit einem Teil des Phondat-II-Zugauskunftskorpus trainiert, der sowohl in Bonn als auch später an der TU Braunschweig [BR94] prosodisch etikettiert wurde. Auf diese Stichprobe und ihre Etiketten wird in Abschnitt 4.1 nur kurz eingegangen; sie diente für Voruntersuchungen zur Akzentdetektion.

Später wurde an der TU Braunschweig begonnen, Teile des Verbmobil-Korpus zu etikettieren. Die Prosodiemodule in den späteren INTARC-Versionen wurden mit den jeweils verfügbaren Daten trainiert und getestet; das waren zuletzt 27 Dialoge bzw. 716 Turns, die zusammen 82.4 Minuten Sprache ergeben (später wurden 6 weitere Dialoge etikettiert, die aber nicht mehr verwendet wurden). Diese Etiketten werden in Abschnitt 4.1.2 beschrieben. Dort werden auch einige Angaben zur Konsistenz dieser Etiketten zitiert.

Die prosodischen Etiketten für das Verbmobil-Korpus basieren auf einer Wortsegmentierung, die automatisch erstellt und manuell korrigiert wurde. Die Zeitpunkte der Wortgrenzen waren nicht sehr zuverlässig, darüber hinaus wurden für Training und Test des Akzentdetektors und später auch des Phrasengrenzendetektors die Zeitpunkte der Silbengrenzen sowie die genauen Vokalpositionen benötigt. Dafür wurden mehrere, schrittweise verbesserte automatische Phonemsegmentierungen erstellt, die in Abschnitt 4.2 beschrieben sind.

4.1 Manuelle prosodische Etikettierung

4.1.1 Etiketten zum Phondat-Korpus

Für Voruntersuchungen wurde ein kleiner Teil des Phondat-II-Zugauskunftskorpus prosodisch etikettiert, und zwar die Turns des Sprechers SAT. Es handelt sich (im Gegensatz zu den spontanen Äußerungen des Verbmobilkorpus) um 200 vollständige, syntaktisch korrekte Sätze, die von jedem Sprecher abgelesen wurden.

Für 60 der 200 Sätze war eine manuelle Phonemsegmentierung vorhanden, für 134 Sätze eine manuelle Phonem- oder Wortsegmentierung; die Voruntersuchungen zur Satzmodusklassifikation wurden an diesen 134 Turns durchgeführt.

Am IKP wurden von einer Phonetikerin in den 60 phonemsegmentierten Sätzen die Akzente etikettiert; unterschieden wurde dabei sowohl zwischen Akzenten mit steigendem und fallendem Intonationsverlauf als auch zwischen starken und schwachen Akzenten. Zusammen mit der Default-Klasse für nichtakzentuierte Silben ergaben sich daher fünf Klassen.

Intonationsverlauf und Stärke wurden dabei in erster Linie perzeptiv, also nach dem Höreindruck etikettiert, auch wenn die Etikettiererin den Grundfrequenzverlauf vor Augen hatte. Das „steigend“ oder „fallend“ bezog sich also nicht unbedingt auf den Grundfrequenzverlauf im Silbenkern, wie sich später zeigte (siehe Abschnitt 6.2.1).

Etwa zeitgleich wurden an der TU Braunschweig ebenfalls Akzentetiketten erstellt, wobei Primär- und Nebenakzente unterschieden wurden, so daß sich hier insgesamt drei Akzentstufen ergeben. Diese Etiketten wurden rein perzeptiv vergeben.

An der TU Braunschweig wurde untersucht, wie gut sich diese drei Klassen perzeptiv unterscheiden lassen, indem 5 Versuchspersonen 480 Turns etikettierten und anschließend die Übereinstimmung gemessen wurde [Rey94]: zunächst wurde die Übereinstimmung für ein bestimmtes Etikett *label* zwischen zwei Versuchspersonen bestimmt:

$$corr_{1,2,label} = \frac{n_{corr(1,2),label}}{(n_{1,label} + n_{2,label})/2} \tag{4.1}$$

Dabei ist $n_{corr(1,2),label}$ die Anzahl der übereinstimmenden Etiketten und $n_{1,label}$ bzw. $n_{2,label}$ die Gesamtanzahl der von Person 1 bzw. 2 vergebenen Etiketten.

Für den Primärakzent ergab sich zwischen den 5 Versuchspersonen eine mittlere Übereinstimmung von 72%, für den Nebenakzent nur 40%. Die nach dem ToBI-Standard [SBP+92] gemessene Übereinstimmung für die Akzentetiketten betrug 80%.

Die Bonner Akzentetiketten wurden von nur einer Person erstellt, daher gibt es dazu keine derartigen Untersuchungen. In Tabelle 6.5 auf Seite 84 werden die Bonner mit den Braunschweiger Etiketten verglichen, wobei die 5 bzw. 3 Klassen auf die beiden Klassen „akzentuiert“ (**A**) und „nicht-akzentuiert“ (**NA**) vergröbert sind. Es zeigt sich, daß die Bonner Etikettiererin sich eher für **A** entschied, aber nur 1.7% der Braunschweiger **A** in Bonn als **NA** eingestuft wurden und nur 0.7% der Bonner **NA** in Braunschweig als **A**. Die dazu komplementären Zahlen (17.9% und 40.3%) zeigen jedoch wie die Braunschweiger Untersuchungen zur Konsistenz, daß die Grenze zwischen **A** und **NA** aufgrund der Perzeption nicht eindeutig zu ziehen ist.

4.1.2 Etiketten zum Verbmobil-Korpus

Die prosodische Etikettierung des Verbmobil-Korpus erfolgte auf drei Ebenen: auf der *funktionalen Ebene*, der *Ton-Ebene* und der *Phrasierungs-Ebene*.

Auf der funktionalen Ebene wurden Satzmodus und Akzente etikettiert; beim Satzmodus wurden FRAGE und NICHT-FRAGE unterschieden, bei den Akzenten NEBENAKZENT, PHRASENAKZENT und EMPHASE/KONTRAST (mit steigender Prominenz).

Auf der Phrasierungs-Ebene wurde die prosodische Strukturierung etikettiert: Es wurde unterschieden zwischen Grenzen zwischen Intonationsphrasen **B3**[1] und intermediären Grenzen **B2** innerhalb von Intonationsphrasen; außergrammatische Grenzen erhalten das spezielle **B9**-Etikett, und alle anderen Wortgrenzen gelten als **B0**-Grenze. Diese Etiketten sind, abgesehen von der Numerierung, vergleichbar denen im ToBI-System (**To**ne and **B**reak **I**ndices) [SBP+92].

Auf der Ton-Ebene wird die Intonation etikettiert, und zwar an allen Akzenten und Phrasengrenzen. Diese Etiketten orientieren sich ebenfalls stark am ToBI-System.

Funktionale Ebene

Auf der funktionalen Ebene werden die Akzentuierung und der Satzmodus etikettiert. Es werden (ohne die Default-Stufe) drei Akzentstufen unterschieden:

- Der *Primärakzent* (**PA**) liegt auf dem am stärksten hervorgehobenen Wort innerhalb einer Intonationsphrase (die durch **B3** begrenzt ist, siehe unten). Innerhalb des Wortes liegt das Akzent auf der betonten Silbe.
- Mit dem *Nebenakzent* (**NA**) werden alle weiteren hervorgehobenen Wörter markiert (wieder an der lexikalisch betonten Silbe).

[1] „B“ steht für „break“.

- Bei besonders starker Hervorhebung kann statt des Primärakzents *Emphase/Kontrast* (**EK**) vergeben werden.

Diese Etiketten lassen Rückschlüsse auf die Fokusstruktur zu, wenn auch der Fokus (das Wort, das für den Hörer die wichtigste Information trägt) durch die Wortstellung statt durch prosodische Mittel ausgedrückt werden kann.

Akzente werden zwar auch auf der Ton-Ebene etikettiert, allerdings werden dort keine Prominenzstufen unterschieden. Hohe Prominenz kann auch durch starke Dehnung ausgedrückt werden, die auf der Tonebene nicht beschrieben werden kann.

An allen **B3**-Grenzen wird der Satzmodus markiert, wobei *Fragen* (**?**) von *Nicht-Fragen* (kein Eintrag) unterschieden werden. Fragen, die mit einem Verb oder einem Fragepronomen beginnen, sind in der Basistransliteration schon als solche zu erkennen. Anders verhält es sich bei assertiven Fragesätzen wie „dreißigster geht auch nicht, dreißigster Mai?", da das Verb an zweiter Stelle steht und sie als Frage nur am hohen Grenzton erkennbar sind. Gleiches gilt für verblose Ellipsen bzw. freie Phrasen, die in spontaner Sprache oft vorkommen: Die Interpunktion der Basistransliteration, die nach Duden gesetzt werden sollte, ist in diesem Punkt nicht zuverlässig genug, so daß die Fragen noch einmal gesondert etikettiert werden mußten.

Phrasierungs-Ebene

Es werden vier Arten von Wortgrenzen unterschieden:

- Die *volle Intonationsphrasengrenze* **B3** markiert einen Einschnitt im Redefluß. Der Einschnitt kann in einer (Atem-) Pause bestehen oder auch nur in einer charakteristischen Intonationsbewegung. Meistens ist die letzte oder sind die beiden letzten Silben der Intonationsphrase gedehnt.

- Als *intermediäre Phrasengrenzen* **B2** werden Intonationsbewegungen innerhalb einer Intonationsphrase gekennzeichnet, z.B. (vergleiche mit Abbildung 4.1) „Schön hervorragend, (**B3**) dann lassen Sie uns (**B2**) doch noch ein' Termin ausmachen. <Pause> (**B3**) Wann wär's Ihnen denn recht? (**B3**)"

- Als *irreguläre Phrasengrenze* **B9** wird eine Wortgrenze etikettiert, wenn der normale Redefluß gestört ist, z.B. wenn in „Also ich dachte noch, (**B3**) in (**B9**) der nächsten Woche, (**B3**) ..." das „in" verzögert ist, oder wenn wie in „ich würde (**B9**) <eräh> (**B9**) können Sie mir einen anderen vorschlagen? (**B3**)" ein Satz abgebrochen und neu begonnen wird.

Ton-Ebene

Die Ton-Etiketten stellen eine symbolische Beschreibung des Intonationsverlaufs dar. Die Etiketten in Verbmobil basieren auf dem Tonsequenz-Ansatz [Pie80] und wurden für das Englische in [SBP+92] angewendet. Die erste Anpassung an das Deutsche erfolgte an der TU Braunschweig [BR94], später einigten sich verschiedene Gruppen auf ein Deutsches ToBI [GRB+97].

Die ToBI-Etiketten beschreiben den Intonationsverlauf durch eine Folge von hohen Tönen (**H**) und tiefen Tönen (**L** für „low"). Die Intonation wird an jedem Akzent (markiert durch '∗') durch einen Ton oder eine Kombination aus zwei Tönen angegeben, ebenso an jeder Phrasengrenze: an den **B3**-Grenzen durch eine Kombination aus Phrasenton (auch „Minuston", markiert durch '**-**') und Grenzton (markiert durch '%'), an den **B2**- und **B9**-Grenzen nur durch den Phrasenton. Welche Kombinationen in Verbmobil verwendet wurden, zeigt Tabelle 4.1.

Töne an akzentuierten Silben	
H∗	Standard-Gipfelakzent; bei ihm wird der Ton im Silbenkern hoch wahrgenommen bzw. fallend, wenn eine tiefe Phrasengrenze folgt
!H∗	Gipfelakzent mit Downstepping (tiefer als vorhergehender Gipfel)
L∗	normaler Talakzent; das Tal liegt innerhalb der akzentuierten Silbe
L+H∗	starker Anstieg innerhalb der akzentuierten Silbe, Gipfel eher am Ende der akzentuierten Silbe
L+!H∗	wie **L+H∗** mit Downstepping
H+!H∗	früher Gipfel, Fall bereits vor der akzentuierten Silbe auf mittleres oder tiefes Niveau
L∗+H	später Gipfel, Gipfel auf der folgenden Silbe
Töne an **B3**-Grenzen	
L-L%	tief auslaufende Phrasengrenze (terminal)
H-H%	hoch auslaufende Phrasengrenze (progredient oder terminal)
L-H%	von unten stark steigende Phrasengrenze (meist interrogativ)
H-L%	steigende und dann leicht fallende Phrasengrenze (meist progredient)
Töne an **B2**- und **B9**-Grenzen	
H-	hoher Phrasenton
L-	tiefer Phrasenton

Tabelle 4.1: *Die an ToBI angelehnten Ton-Etiketten in Verbmobil.*

Da die Töne nur an Akzenten und Phrasengrenzen etikettiert werden, liegt mit den Ton-Etiketten auch eine grobe funktionale Etikettierung vor. In [Kom97] wird argumentiert, daß für die linguistischen Module in Verbmobil *nur* die funktionale Klasse relevant ist, während formale Ausprägungen keine Rolle spielen, sofern der Klassifikator diese lernen kann. Der einzige Unterschied bestehe beim Satzmodus in der Unterscheidung zwischen interrogativen, terminalen und progredienten Phrasengrenzen.

Da die Verbmobil-Etiketten funktional nur zwischen Frage und Nicht-Frage unterscheiden, wurden in dieser Arbeit die terminalen von den progredienten Phrasengrenzen anhand der Ton-Etiketten zusätzlich unterschieden, um den Satzmodusklassifikator zu trainieren, siehe Abschnitt 6.3.

Die zweite Anwendung der Ton-Etiketten lag bei der Phrasengrenzenklassifikation: Da ein Normalverteilungsklassifikator und nicht wie in [Kom97] ein MLP verwendet wird, ist es besser, die formalen Ausprägungen durch entsprechende Unterklassen zu modellieren; die Ton-Etiketten dienten hierzu als explizite Clusterung.

Abbildung 4.1 auf der nächsten Seite zeigt als Beispiel eine Äußerung aus dem Verbmobil-Korpus mit Sprachsignal, Grundfrequenz, Silbengrenzen, Phonemsegmentierung, und allen prosodischen Etiketten.

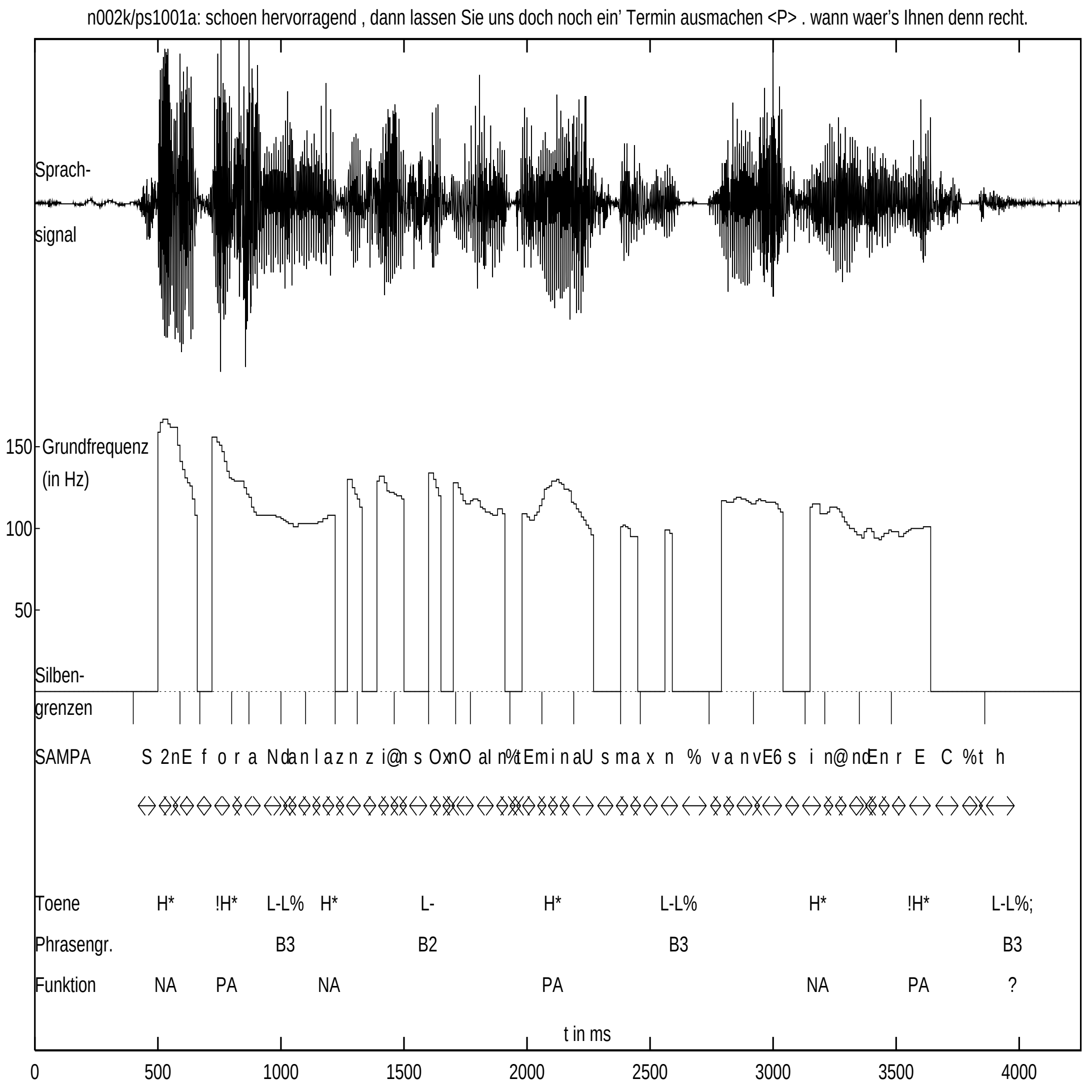

Abbildung 4.1: *Ein Beispiel zu den prosodischen Etiketten: Sprachsignal und Grundfrequenz, Silbengrenzen und Phonemsegmentierung in* Sampa *(mit % als Symbol für Pausen), darunter die drei Ebenen der prosodischen Etikettierung: die an ToBI angelehnten Ton-Etiketten, die Phrasengrenzen (ohne die Default-B0-Grenzen) und die Funktionalen Etiketten zur Akzentuierung und zum Satzmodus.*

Untersuchungen zur Konsistenz

Die Konsistenz der prosodischen Etiketten zur Verbmobil-Stichprobe wurde — wie bei den Phondat-Etiketten — ebenfalls gemessen, an 233 Turns, aber diesmal nur zwischen zwei Etikettierern.

Die nach dem ToBI-Standard gemessene Übereinstimmung beträgt für die Funktionalen Etiketten 91%, für die Phrasengrenzen-Etiketten 94% und für die Ton-Etiketten an Phrasengrenzen 85%.

Die mit Gleichung 4.1 ermittelte Übereinstimmung bei **PA** beträgt 86%, bei **NA** 32%, bei **B3** 90% und bei **B2** 44%. Die Übereinstimmung der Ton-Etiketten schwankt zwischen 35% für **L-H%** und 75% für **L-L%**.

An dieser Stelle sei darauf hingewiesen, daß diese Zahlen nur die Konsistenz zwischen zwei Etikettieren angeben, nicht aber die „Konsistenz“ mit den Sprachdaten: So sind z.B. die Phrasengrenzen-Etiketten stark von syntaktischen Vorerwartungen beeinflußt; einige **B3**-Grenzen sind intonatorisch so schwach markiert, daß sie nur „gehört“ werden, weil dort eine starke syntaktische Grenze vorliegt.

Der Erlanger Phrasengrenzen-Erkenner kann solche Vorerwartungen mit einem statistischen Sprachmodell nachbilden. Der hier beschriebene Phrasengrenzen-Erkenner, der über keine Wortinformation verfügt und nur anhand akustisch-prosodischer Merkmale entscheidet, ist eher vergleichbar mit jemandem, der eine ihm unbekannte Sprache prosodisch etikettieren soll. Auf diese Problematik wird in Kapitel 8 weiter eingegangen.

4.2 Automatische Phonemsegmentierung

Die prosodische Etikettierung baut auf einer automatischen Wortsegmentierung auf, die die Zeitzuordnung der Wortgrenzen liefert. Dazu wurde ein HMM-Worterkenner gezwungen, die gesprochene Wortkette zu erkennen [Leh94]. Grundlage dafür war die Basistransliteration: neben der orthographischen Verschriftung, die teilweise auch Aussprachevarianten berücksichtigt, sind in ihr auch Pausen, Atmen, Schmatzen, Störgeräusche u.a. gekennzeichnet. Die Basistransliteration wurde an der TU Braunschweig teilweise geändert; hauptsächlich wurden Pausen ergänzt, da der Braunschweiger Worterkenner diese offensichtlich nicht selbst einfügen konnte, und Fehler in der Basistransliteration korrigiert, z.B. unvollständige oder fehlende Wörter als solche gekennzeichnet.

Zum Training und Testen der hier beschriebenen Prosodiedetektoren war jedoch eine feinere Segmentierung notwendig; neben den Wortgrenzen wurden auch die Zeitpunkte der Silbengrenzen und Silbenkerngrenzen gebraucht, so daß eine möglichst genaue Phonemsegmentierung angestrebt wurde.

Eine manuelle Phonemsegmentierung zu erstellen, wäre viel zu aufwendig gewesen, deshalb wurde zunächst vom Hamburger Projektpartner[2] eine automatische Phonemsegmentierung erzeugt [Hü94], ebenfalls mit Hilfe eines HMM-Worterkenners, der auf dem HTK[3]-Toolkit basiert. Grundlage war allerdings die kanonische Aussprache, weil die mit Worterkennung befaßten Gruppen alle ein kanonisches Ausspracheworterbuch benutzten. Der Erkenner mußte also z.B. *„Wann wäre es Ihnen denn recht“* statt *„Wann wär's Ihnen denn recht“* erkennen, also zwei Silben mehr, als tatsächlich gesprochen wurden.

Zwar sollten Aussprachevarianten wie *„wär's“* in der Basistransliteration enthalten sein, und ihre Aufnahme in das Aussprachelexikon ist mit relativ geringem Aufwand möglich. Das Problem war eher, daß die Basistransliteration, zumindest für die Karlsruher Dialoge, nicht sehr zuverlässig ist, was die Aussprachevarianten betrifft. Daher wurde in Bonn eine neue Transkription erstellt, die zumindest alle Elisionen berücksichtigt. Das Verbmobil-Aussprachelexikon [GE95] aus Bielefeld, das auch Silbengrenzen enthält, wurde entsprechend erweitert. Die automatische Phonemsegmentierung erfolgte dann in Bonn mit einem vom Münchner Projektpartner[4] zur Verfügung gestellten Worterkenner [WS94], der ebenfalls auf HMMs basiert. Ein weiteres Problem war nun die Abbildung der prosodischen Etiketten auf die Silbengrenzen bzw. Silbenkerne, da die Zeitpunkte nicht mehr übereinstimmten. Im Prinzip erfolgte die Zuordnung zur nächstgelegenen Grenze mit Hilfe einiger weiterer Heuristiken. Grobe Abweichungen wurden automatisch detektiert und manuell behandelt. Diese recht aufwendige Prozedur mußte auch für die unten beschriebenen späteren Phonemsegmentierungen durchgeführt werden.

Der Münchner Phonemsegmentierer hatte allerdings Schwierigkeiten bei der Pausendetektion, auch paßten die Plosiv-Modelle nicht für alle Sprecher. Einige Fehler konnten mit Hilfe der kanonischen Segmentierung automatisch korrigiert werden, aufgrund anderer (automatisch entdeckter) grober Fehler mußten einige Turns aussortiert werden.

Zu einem späteren Zeitpunkt wurde von einer Projektparterin[5] an der Uni Bielefeld ein Aussprachelexikon entwickelt, das zu jedem Wort automatisch erzeugte Aussprachevarianten enthält, nicht nur was Elisionen betrifft, es enthält z.B. auch *„Tach“* statt *„Tag“* [Kir95]. Der Hamburger Partner erweiterte seinen HMM-basierten Segmentierer dergestalt, daß die am besten passende Aussprachevariante gewählt wurde, und erzeugte so eine Phonemsegmentierung, die näher an der akustischen Realität lag [Kir95]. Allerdings gingen bei der Erzeugung der Aus-

[2] Kai Hübener
[3] Hidden Markov Model Toolkit von Entropic
[4] Florian Schiel
[5] Katrin Kirchhoff

sprachevarianten die Markierungen für die Silbengrenzen verloren. Nach Aussage der Bielefelder Partnerin lassen sich bei gewissen Varianten keine definitiven Silbengrenzen mehr angeben. Daher wurde versucht, die Silbengrenzen nach heuristischen Regeln zwischen zwei Silbenkerne zu setzen, leider oft mit unerwünschtem Resultat.

Danach wurde deshalb vom Hamburger Partner mit einem verbesserten HMM-Erkenner und mithilfe des Bonner „Elisionenlexikons" und der zugehörigen Transliteration eine weitere Phonemsegmentierung erzeugt [Jos96], die wesentlich akkurater ist als die vorher in Bonn mit dem Münchner Erkenner erzeugte. Mit ihrer Hilfe wurden schließlich die besten Erkennungsraten erzielt.

Kapitel 5

Prosodische Merkmale

In Kapitel 3 wurden die Grundbestandteile eines Systems zur Musterklassifikation genannt: Aufnahme, Vorverarbeitung, Merkmalgewinnung und Klassifikation. Ziel der Merkmalgewinnung ist es, eine möglichst kompakte Beschreibung des zu klassifizierenden Musters zu erlangen. Muster aus der gleichen Klasse sollen zu ähnlichen Werten der Merkmale führen, zu Häufungen im Merkmalsraum, während Muster aus verschiedenen Klassen möglichst entfernte Gebiete im Merkmalsraum einnehmen sollen.

Gesucht ist daher eine Transformation des Musters $\mathbf{f}$ in einen Merkmalvektor $\mathbf{c}$, die dieser Anforderung möglichst gut entspricht. Es gibt allerdings keine systematische Methode, die optimale Transformation zu finden. Daher müssen Merkmale mit heuristischen Methoden und Expertenwissen über den Problemkreis, hier die Prosodie, bestimmt werden.

Akzentuierung und Phrasierung spiegeln sich in den akustischen Kategorien Intonation, Lautheit und zeitliche Strukturierung wider. Daher sollten Merkmale zur Trennung prosodischer Klassen eine geeignete Beschreibung des Grundfrequenzverlaufs, des Energieverlaufs und der Dauerverhältnisse leisten.

Abschnitt 5.1 gibt zunächst einen kurzen Überblick über die physiologischen Mechanismen der Sprachproduktion, um die Verbindung zu den meßbaren Größen *Grundfrequenz* und *Energie* herzustellen.

Abschnitt 5.2 behandelt die Grundfrequenz-Analyse und die daraus abgeleiteten Merkmale: Zunächst wurde die Grundfrequenz mit dem Fujisaki-Modell in zwei Komponenten zerlegt. Durch die Dekomposition wird vor allem das Herausarbeiten der lokalen Grundfrequenzbewegung, relativ zur globalen Bewegung möglich (Abschnitt 5.2.2). Wegen offensichtlicher Mängel des Modells wurde die Dekomposition dann mit einer Filterbank bewerkstelligt (Abschnitt 5.2.3) deren Voraussetzung die vorherige Interpolation der Grundfrequenz ist (Abschnitt 5.2.3).

In Abschnitt 5.3 werden Energiemerkmale beschrieben, die zusammen mit den Grundfrequenzmerkmalen die frameweise berechneten *Basismerkmale* bilden und zur Akzentdetektion herangezogen wurden.

Die Energiemerkmale werden auch zur Silbenkerndetektion verwendet (Abschnitt 5.3.1). Sie erlauben die Berechnung von Dauermerkmalen (Abschnitt 5.4) und die Bildung der komplexen Merkmale (Abschnitt 5.5), die zur Phrasengrenzendetektion eingesetzt wurden.

5.1 Sprachproduktion

Dieser Abschnitt, der sich an [Koh77] orientiert, gibt einen sehr kurzen Überblick über die physiologischen Mechanismen der menschlichen Sprachproduktion, um die Verbindung zu den meßbaren Größen *Grundfrequenz* und *Energie* herzustellen.

Die Erzeugung menschlicher Sprachlaute setzt sich aus drei Prozessen zusammen: der Luftstromerzeugung (Respiration), der Stimmbildung (Phonation) und der Lautbildung (Artikulation).

Im Deutschen werden Sprachlaute beim Ausatmen gebildet. Dabei durchströmt die aus der Lunge verdrängte Luft zunächst im Kehlkopf die Stimmritze (Glottis), die durch zwei Muskeln, die Stimmlippen gebildet wird.

Sind die Stimmlippen geöffnet, durchströmt die Luft sie turbulent, was z.B. beim Flüstern als Rauschen zu hören ist. Dies wird als stimmlose Anregung bezeichnet.

Werden die Stimmlippen jedoch angespannt, beginnen sie zu schwingen, d.h. sich periodisch zu öffnen und zu schließen. Hier spricht man von stimmhafter Anregung. Die Periode des Schwingungzyklus definiert die Grundperiode in einem stimmhaften Sprachsignalabschnitt, ihr Kehrwert die *Sprachgrundfrequenz*. Sie wird als Tonhöhe wahrgenommen und hängt im wesentlichen von der Spannung der Stimmlippen, dem Luftdruck in der Lunge und anatomischen Unterschieden ab, die vor allem beim Vergleich von Frauen-, Männer- und Kinderstimmen deutlich werden. Die Sprachgrundfrequenz bewegt sich ungefähr zwischen 50 und 800 Hz [Hes83].

Die Lautbildung erfolgt auf dem weiteren Weg des Luftstroms von der Stimmritze bis zur Mundöffnung und/oder den Nasenlöchern. Der Vokaltrakt, siehe Abbildung 5.1, kann durch die Stellung der Artikulatoren, namentlich der Zunge, des weichen Gaumens, des Unterkiefers und der Lippen in seiner Form verändert werden. Der Vokaltrakt kann als akustische Röhre betrachtet werden, die durch die an der Stimmritze entstehende Luftschwingung, die glottale Welle, angeregt wird. Dabei werden die Obertöne der glottalen Welle an den Resonanzfrequenzen hervorgehoben, indem die übrigen Frequenzen gedämpft werden. Die Reso-

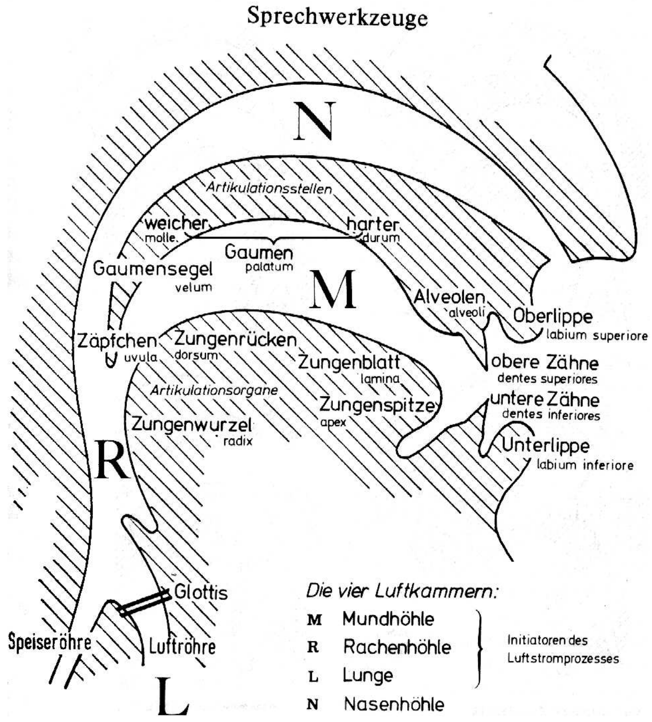

Abbildung 5.1: *Die Sprechwerkzeuge, aus [Buß90, Seite 101]. Rachen- und Mundhöhle zusammen werden auch als Vokaltrakt bezeichnet.*

nanzfrequenzen, die in diesem Zusammenhang als *Formanten* bezeichnet werden, liegen im Mittel 1 kHz auseinander, beginnend bei etwa 500 Hz. Ihre tatsächlichen Werte werden durch die Stellung der Artikulatoren bestimmt. Unterschiedliche Formantfrequenzen werden als unterschiedliche Klangfärbung wahrgenommen, z.B. zwischen den Vokalen /a/, /e/, /o/ und /u/.

Bezieht man den Nasenraum, der bei der Artikulation von /n/, /m/ und /N/ eine Rolle spielt, in die Betrachtung ein, kommt man von der akustischen Röhre auf ein T-Rohr mit dem weichen Gaumen als Verbindungsstelle; die Wirkung einer der Teilröhren kann durch Antiresonanzen modelliert werden. Antiresonanzen entstehen auch bei der Artikulation von Frikativen: Beispielsweise wird das /x/ in „a**ch**“ durch eine Engstelle zwischen Zunge und Gaumen gebildet; die turbulente Luftströmung an der Engstelle stellt nun die Anregung dar, die auf dem weiteren Weg durch Resonanzen ausgeformt wird, während der Teil des Vokaltrakts zwischen Stimmritze und der Engstelle Antiresonanzen beisteuert [MG76].

Die Details sollen hier nicht weiter betrachtet werden. Im Hinblick auf die Silbenkern- und Akzentdetektion ist hier lediglich von Bedeutung, daß sich die für verschiedene Laute unterschiedlichen Resonanzen und Antiresonanzen auf die spektrale Energieverteilung auswirken und daher Merkmale zur Unterscheidung von Vokalen und Nichtvokalen auf diese Verteilung abzielen.

5.2 Grundfrequenzmerkmale

Abschnitt 5.2.1 behandelt kurz die Analyse der Sprachgrundfrequenz. Unter der Vielzahl möglicher Algorithmen wurde in dieser Arbeit der beste verfügbare verwendet, der zum kommerziell erhältlichen Programmpaket ESPS (**E**ntropic **S**ignal **P**rocessing **S**ystem der Entropic Research Laboratory Inc.) gehört. Er wird neben zwei anderen Verfahren kurz vorgestellt und mit ihnen hinsichtlich der Fehlerrate verglichen.

Um aus der ermittelten Grundfrequenzkontur Merkmale zu erhalten, müssen die wesentlichen Eigenschaften der Kontur durch wenige Parameter beschrieben werden. Häufig werden dazu Regressionsgeraden verwendet [LKJ$^+$85] [Ott93][Hub88][Kie97], als Merkmale können dann Steigung, Höhe (Schnittpunkt mit dem linken Rand des Analysefensters) dienen. Um eine globale Beschreibung zu erhalten, wird die Regressionsgerade durch die gesamte Kontur gelegt, für eine lokale Beschreibung, z.B. des Grundfrequenzverlaufs am Satzende, durch entsprechend kleinere Abschnitte, z.B. durch den letzten stimmhaften Bereich wie in [Kie97].

Aufgrund der am IKP geleisteten Vorarbeiten bot sich zur Parametrisierung jedoch zunächst das Intonationsmodell von Fujisaki [FHO79] an, das ursprünglich für die japanische Sprachsynthese entwickelt wurde und eine Grundfrequenzkontur als Überlagerung eines Basiswertes, einer Phrasenkomponente und einer Akzentkomponente beschreibt. Die Phrasenkomponente modelliert den globalen Grundfrequenzverlauf, während die Akzentkomponente die Tonbewegung im Bereich von Akzentgruppen[1] beschreibt. Das Modell erhebt den Anspruch, bei der Analyse einer gegebenen Grundfrequenzkontur durch Fehlerminimierung die resultierenden Modellparameter, die sog. *Akzent-* und *Phrasenkommandos* direkt den linguistischen Konzepten *Akzent* und *Phrase* zuordnen zu können.

Abschnitt 5.2.2 befaßt sich mit den Grundzügen des Modells, der früheren Bonner Implementierung von Pätzold [Pät91], auf der die Arbeit von Möbius beruht [Möb93] und mit der selbst entwickelten, inkrementell und vollautomatisch arbeitenden Implementierung.

Die Implementierung des Fujisaki-Modells zeigte jedoch, daß dieses Modell zur Analyse der deutschen Prosodie nicht geeignet ist, selbst wenn seine Parameter nur als Merkmale verwendet werden, siehe auch Abschnitt 6.1. Deshalb

[1] Eine Akzentgruppe besteht aus einer betonten Silbe, gefolgt von 0 bis n unbetonten Silben.

wurde ein alternativer Ansatz zur Parametrisierung der Grundfrequenzkontur verfolgt, der ebenfalls die Idee der Dekomposition beinhaltet. Die Dekomposition erfolgt hier jedoch mit einer weit weniger störanfälligen Digitalfilterbank. Voraussetzung dafür ist eine Interpolation der Grundfrequenzkontur in stimmlosen Bereichen, die speziell für die Erfordernisse dieser Digitalfilterbank entwickelt wurde. Diese alternative Dekomposition der Grundfrequenzkontur wird in Abschnitt 5.2.3 beschrieben.

Die Ausgabe der Filterbank besteht aus den Komponenten der Grundfrequenzkontur, die sich nur langsam ändern. Die Reduktion auf wenige Parameter, d.h. die eigentliche Merkmalgewinnung, erfolgt durch Abgreifen dieser Komponenten an Stellen, die durch die Silbenkerndetektion bestimmt werden, oder durch Mittelung in bestimmten stimmhaften Bereichen, wie die Abschnitte 5.5 und 6.1 zeigen werden.

In Abschnitt 6.1 werden aus den zwei Parametrisierungen zwei Merkmalgruppen extrahiert und hinsichtlich der Satzmodusklassifikation verglichen. In allen übrigen Klassifikationsexperimenten werden dann keine Parameter des Fujisaki-Modells mehr verwendet.

5.2.1 Grundfrequenz-Analyse

Die Grundperiode T_0 wurde in Abschnitt 5.1 als Dauer des glottalen Zyklus definiert, die Grundfrequenz als ihr Kehrwert $F_0 = 1/T_0$. Die Grundperioden sind auch nach der Ausformung des Anregungssignals im Vokaltrakt, der Aufprägung der Formanten, im resultierenden Sprachsignal meist noch gut sichtbar, wenn auch nicht mehr so deutlich; die erste Formantfrequenz F_1 interagiert am stärksten mit der Grundfrequenz F_0 und kann sogar niedriger als diese sein.

Alle Verfahren zur Grundfrequenz-Analyse basieren auf der Annahme, daß in stimmhaften Bereichen die Grundperioden einigermaßen regelmäßig sind und sich nur allmählich ändern. Es gibt Frequenzbereichsverfahren, die die mittlere Grundfrequenz innerhalb eines Frames schätzen, und Zeitbereichsverfahren, die die genaue Lage der Perioden bestimmen (sog. Pitchmarker). Wenn die Annahmen zutreffen, was für etwa 90% aller Frames der Fall ist, stellt die Bestimmung der Periodendauer trotz dem Einfluß der Formanten kein großes Problem dar.

Schwierigkeiten bereiten die 10%, in denen die Stimmlippen zwar schwingen, aber unregelmäßig. Dies kann auf ein Phänomen des Stimmeinsatzes zurückzuführen sein, auf einen verschliffenen Glottalverschluß. Oft kann an solchen Stellen eine Verdopplung der Periodendauer oder sogar Vervierfachung beobachtet werden, die Stimmlippen können aber auch völlig unregelmäßig schwingen. Nach [Leh70] werden solche Phänomene *Laryngalisierungen* genannt.

Darüber hinaus führen auch die Verschlußpausen bei den stimmhaften Plosiven /b/, /d/ und /g/ zu einer Irregularität im Sprachsignal, die sich ähnlich problematisch auswirkt wie eine Laryngalisierung [Str93a].

Nach [Hes83] ist die Grundfrequenzbestimmung in irregulären Signalabschnitten oft auch manuell nicht möglich. Deshalb überrascht es nicht, daß diese Stellen auch automatischen Verfahren Probleme bereiten. Bei Verfahren, die nicht ausschließlich lokal (auf einem Frame) arbeiten, können sich Fehler auch in reguläre Signalbereiche fortpflanzen.

Ein Verfahren nach Hess [Hes83] in verschiedenen Konfigurationen wurde verglichen mit dem kommerziellen, zum Programmpaket ESPS[2] gehörenden Verfahren [SD83] und einem in Erlangen entwickelten Verfahren [KKN+92]. Das Erlanger Verfahren berücksichtigt in einer Konfiguration sogar irreguläre Signalbereiche, wobei gegenüber anderen Konfigurationen eine Senkung der F_0-Fehlerrate erzielt werden konnte [NDK+94]. Die niedrigste Fehlerrate hatte jedoch das ESPS-Verfahren [Str93b].

Für das Gesamt-Verbmobil-System galt die Vorgabe, keine kommerziellen Produkte zu inkorporieren. Die experimentellen INTARC-Systeme waren jedoch frei von dieser Einschränkung, so daß hier auf das beste verfügbare Verfahren zurückgegriffen werden konnte.

Dieses Verfahren soll hier kurz skizziert werden: Der erste Schritt besteht in der *inversen Filterung* des Sprachsignals. Dabei werden die Resonanzen des Vokaltrakts mit Hilfe der *Linearen Prädiktion* aus dem Sprachsignal geschätzt [MG76]. Das dazu inverse Filter kompensiert den Einfluß des Vokaltrakts, so daß das invers gefilterte Sprachsignal eine Approximation des Anregungssignals[3] darstellt.

Nach einer weiteren Bearbeitung mit einem zeitvariablen Filter, das den „Kontrast“ zwischen stimmlosen und stimmhaften Bereichen verstärken soll (nach eigenen Untersuchungen aber nicht viel nützt), werden in der Autokorrelierten die Maxima als F_0-Kandidaten bestimmt, wobei ihre Höhe als Gütemaß dient. Die Auswahl der jeweils besten Kandidaten erfolgt schließlich durch dynamische Programmierung. Die Kostenfunktion beinhaltet das o.g. Gütemaß, die F_0-Änderung (wobei auch Oktavsprünge erlaubt sind) und die Änderung der Stimmhaftigkeit; sie kann bei Programmstart durch verschiedene Parameter beeinflußt werden.

Das ESPS-Verfahren wurde mit den Standard-Einstellungen betrieben. Die Schwelle für die nachträgliche Stimmhaft-Stimmlos-Entscheidung wurde auf 0.95 festgelegt (wobei stimmlose Frames dann mit dem Grundfrequenzwert 0 codiert wurden). Zuletzt wurde dieses Signal mit einem Medianfilter der Breite 3 geglättet, hauptsächlich, um an den Rändern stimmhafter Bereiche ein Hin- und

[2] Entropic Signal Processing System, Entropic Inc.

[3] In [MG76] wird die glottale Welle als tiefpaßgefiltertes Anregungssignal dargestellt.

Herschalten der Stimmhaft-Stimmlos-Entscheidung zu vermeiden.

5.2.2 Dekomposition der Grundfrequenzkontur mit dem Fujisaki-Modell

Das von Fujisaki und seinen Mitarbeitern entwickelte Modell basiert auf grundlegenden Arbeiten von Öhman und Lindqvist [ÖL66] [Öhm67]. Fujisaki und Nagashima konnten zeigen, daß sich Intonationskonturen natürlichsprachlicher japanischer Äußerungen mit einer Vorversion des Modells recht gut nachbilden lassen [FN69]. Die erste vollständig entwickelte Version ist in [FHO79] beschrieben, sie wurde später hinsichtlich der Steuerparameter modifiziert [Fuj88].

Nach jener Version wird eine F_0-Kontur durch Überlagerung dreier Komponenten modelliert: Einem *Basiswert* F_{min}, der als sprecher– oder äußerungsabhängig betrachtet wird, einer *Akzentkomponente*, um die lokalen F_0–Bewegungen im Zeitbereich von Silben nachzubilden und einer *Phrasenkomponente*, um die globalen Bewegungen im Zeitbereich größerer syntaktischer Einheiten nachzubilden.

Die Komponenten sind Impulsantworten rekursiver Filter zweiten Grades; die Erzeugung eines Ausschlags in der Akzentkomponente durch einen positiven Anschalt– und negativen Ausschaltimpuls wird üblicherweise durch einen entsprechenden Rechteckimpuls dargestellt. Bild 5.2 zeigt oben ein ein Phrasen– und zwei Akzentkommandos, in der Mitte die resultierenden Phrasen– und Akzentkomponenten. Die Summe aus den beiden Komponenten und der F_{min} ergibt den Modellverlauf, im Bild unten zu sehen.

Die Modell–F_0 als Überlagerung von Basiswert, Phrasen– und Akzentkomponente wird formuliert als:

$$
\begin{aligned}
F_0(t) &= F_{min} + \sum_{j=1}^{J} A_{pj}\, G_{pj}(t - T_{0j}) + \sum_{k=1}^{K} A_{ak} \Big[G_{ak}(t - T_{1k}) - G_{ak}(t - T_{2k}) \Big] \\
&\text{mit} && (5.1) \\
G_{pj}(t) &= \alpha_j t \cdot \mathrm{e}^{-\alpha_j t}\;\; \delta_{-1}(t) \qquad \text{Form der Phrasenkomponenten} && (5.2) \\
G_{ak}(t) &= 1 - (1 - \beta_k t) \cdot \mathrm{e}^{-\beta_k t}\;\; \delta_{-1}(t) \qquad \text{Form der Akzentkomponenten} && (5.3)
\end{aligned}
$$

mit dem Dirac-Sprung

$$\delta_{-1}(t) = \begin{cases} 0 & : \quad t < 0 \\ 1 & : \quad t \geq 0 \end{cases}$$

Parameter in dem Modell sind die Dämpfungsfaktoren α_j und β_k, die F_{min}, die J Amplituden und Zeitpunkte der Phrasenkommandos $\langle A_{pj}, T_{0j} \rangle$, und die K Amplituden und An– und Ausschaltzeitpunkte der Akzentkommandos $\langle A_{ak}, T_{1k}, T_{2k} \rangle$.

Die Gleichungen weichen in zwei Punkten von der Darstellung bei Fujisaki ab: Zum einen wird auf die logarithmische Skalierung der Grundfrequenz verzichtet, weil das Modell nur an einem Sprecher untersucht wurde. Zum anderen ist $G_{pj}(t)$ geändert: Im Original ist $G_{pj}(t) = \alpha_j^2 t \cdot \mathrm{e}^{-\alpha_j t}\ \delta_{-1}(t)$, vermutlich weil dann beim empfohlenen Wert $\alpha \approx 3$ gilt: $\max G_{pj}(t) \approx 1$. Dadurch werden zwar die Phrasen–Amplituden A_{pj} anschaulicher, dafür bekommt $G_{pj}(t)$ dann die Dimension $\frac{1}{t}$, während $G_{ak}(t)$ dimensionslos bleibt. Außerdem führt der Term α^2 zu unschönen Gleichungen bei der — hier nicht gezeigten — Berechnung des zum Phrasensteuerungsmechanismus inversen Filters.

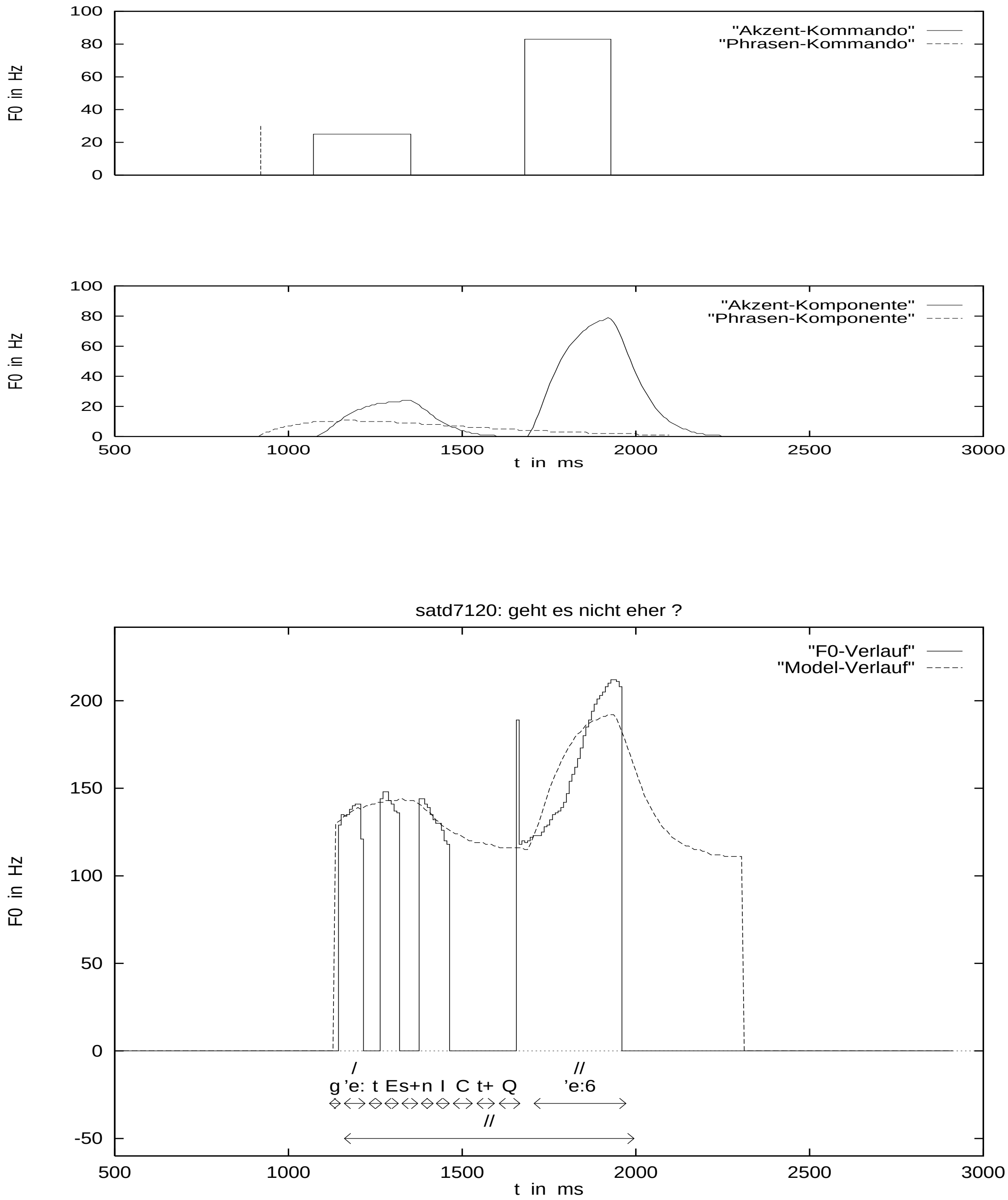

Abbildung 5.2: *Beispiel zum Fujisaki–Modell: Oben ein Phrasen– und zwei Akzentkommandos, in der Mitte die resultierenden Phrasen– und Akzentkomponenten, unten die zusammengesetzte Modell–F0 und die ESPS–F0 (die Modell–F0 ist vor dem ersten stimmhaften Bereich und 600 ms nach dem letzten auf 0 gesetzt).*

Aufgabe ist nun, die Parameter so zu wählen, daß ein gegebener F_0–Verlauf durch den Modellverlauf möglichst gut angenähert wird. Das Gütekriterium in den Arbeiten von Fujisaki ist zunächst der mittlere quadratische Abstand (RMS).

In [Geo93] wird Gleichung 5.1 nach den Parametern abgeleitet, um durch Nullsetzen die optimale Lösung zu finden. Da die $3J + 4K + 1$ Gleichungen analytisch nicht auszuwerten sind, werden sie durch ihre Taylorentwicklungen erster Ordnung ersetzt. Auch jene Implementierung arbeitet inkrementell, d.h. die F_0–Werte werden nach und nach eingelesen, wobei die Parameter reoptimiert werden. J und K sind zunächst 1 und werden bei Überschreiten einer Fehlerschwelle erhöht.

Allerdings werden fast alle Parameterkonfigurationen weiterverfolgt, die zu lokalen Minima gehören (Suchstrahlbreite 10), und die letztliche Auswahl erfolgt durch ein Backtracking, das dann nicht mehr inkrementell ist. Das stückweise Abarbeiten der F_0 hat mehr den Zweck, zu bestimmen, wann neue Kommandos nötig werden, d.h. wann J und K erhöht werden müssen [Geo93, sect. 3.2].

Eine andere Möglichkeit ist der Einsatz eines Suchverfahrens. Hier besteht die Schwierigkeit in der Größe des Suchraums mit $3J + 4K + 1$ Dimensionen. Selbst wenn man die Dämpfungsfaktoren α und β konstant läßt, wie von Fujisaki und Möbius vorgeschlagen, bleiben noch $2J+3K+1$ Dimensionen. Zudem ist J und K nicht gegeben: Die gleiche F_0–Kontur kann mit unterschiedlichen Anzahlen von Phrasen– und Akzentkommandos nachgebildet werden, wobei die Alternativen nahezu gleich gut im Sinne des RMS sind.

Um die Optimierung auf die Spitze zu treiben, könnte man eine F_0–Kontur durch viele impulsförmige Akzentkommandos nachbilden. Eine linguistische Interpretation dieser Kommandos ist dann aber nicht mehr möglich: Falls sich der Wortakzent und die damit verbundene intonatorische Einheit Akzentgruppe in den Akzentkommandos widerspiegeln sollen, darf für jede Akzentgruppe nur ein Kommando vergeben werden [Möb93, Seite 80].

Um die Modellparameter in linguistische Kategorien abbilden zu können, müssen Nebenbedingungen eingeführt werden. In der Arbeit von Möbius und Pätzold waren die Nebenbedingungen im Datenmaterial enthalten: Phrasengrenzen und Akzentgruppengrenzen wurden von Hand markiert, so daß nicht nur die Anzahl der Phrasen– und Akzentkommandos gegeben war, sondern auch deren ungefähre Lage. Mit Hilfe dieser Grenzen werden in Pätzolds Implementierung die übrigen Modellparameter durch Intervallschachtelung gefunden.

Inkrementell arbeitende Implementierung

Die neue Implementierung wurde aus zwei Gründen nötig: Zum einen arbeitet Pätzolds Implementierung nicht vollautomatisch (Phrasengrenzen und Akzentgruppengrenzen mußten vorher von Hand markiert werden, dies kann auch nicht durch einen einfachen Präprozessor ersetzt werden). Zum anderen verbietet die Forderung nach inkrementeller Arbeitsweise, daß die F_0–Kontur als Ganzes betrachtet wird. Vielmehr muß die Analyse lokal arbeiten, d.h. sie darf in der F_0–Kontur nur wenig voraussehen (um den Signal–Nachlauf der Prosodiekomponente gering zu halten), und Phrasen– und Akzentkommandos zu weiter zurückliegenden Signalabschnitten sollten nicht mehr geändert werden.

Die global optimalen Modellparameter im Sinne des RMS zu finden, ist dann nicht mehr möglich. Entscheidend war aber vielmehr, ob die Modellierung akzeptabel ist, sowohl im Sinne des Fehlermaßes als auch nach linguistischen Kriterien.

Es soll also der Fehler minimiert werden, und gleichzeitig müssen Randbedingungen eingehalten werden, um den linguistischen Kriterien gerecht zu werden. Beispielsweise wird in der neuen Implementierung angestrebt, nur ein Akzentkommando pro Akzentgruppe zu vergeben, auch wenn mit zwei Kommandos eine bessere Approximation möglich wäre. Dies wird verhindert, indem Grenzwerte für Dauer und Amplitude von Akzentkommandos sowie ein Mindestabstand zwischen Kommandos eingehalten werden müssen.

Es folgt nun eine stichpunktartige Beschreibung der eigenen Implementierung. Sie wurde für die erste Implementierung des Satzmodus-Klassifikators im INTARC-1.2-System entworfen, siehe Abschnitt 6.1. Das INTARC-1.2-System war für einen Teil des Phondat-II-Zugauskunftskorpus ausgelegt; das bedeutet hier, daß jeder Turn in der Regel aus nur einem Satz bestand, so daß das Satzende (die Grenze der zu klassifizierenden Einheit) in fast allen Fällen durch die Pausenschwelle von 600 ms (siehe unten) gefunden werden konnte.

1. Die Dämpfungsfaktoren (siehe Gleichung 5.1 auf Seite 54) α und β werden konstant gewählt. Zu optimieren bleiben F_{min}, die Phrasenkommandos $\langle A_{pj}, T_{0j} \rangle$ und die Akzentkommandos $\langle A_{ak}, T_{1k}, T_{2k} \rangle$.

2. Für jeden Satz wird nur ein Phrasenkommando vergeben (als Satzende gilt eine Pause von mehr als 600 ms Pause); T_0 wird so gewählt, daß das Maximum der Phrasenkomponente zu Beginn des ersten stimmhaften Bereichs in Satz erreicht wird. Als Baisiwert F_{min} wird der bisher kleinste Grundfrequenzwert benutzt; falls er kleiner wird, werden die Amplituden der bisherigen Akzentkommandos und des Phrasenkommandos neu berechnet.

3. Die F_0–Kontur wird stückweise eingelesen, immer bis zum nächsten ausgeprägten lokalen Minimum/Maximum oder zum Ende eines stimmhaften Bereichs. Dann erfolgt die Neuberechnung der Parameter, und das nächste Konturstück wird eingelesen, u.s.w. bis Satzende. Bei Erreichen des nächsten Satzanfangs wird der Algorithmus reinitialisiert.

4. Die nachgeladenen F_0–Werte werden durch Modifizieren des letzten Akzentkommandos modelliert, bis der Fehler zu groß wird oder Randbedingungen verletzt werden (s.u.). Dann wird die letzte Modifikation zurückgenommen, und ein neues Akzentkommando wird vergeben.

5. Die Zeitpunkte T_1 und T_2 und die Amplitude des letzten Akzentkommandos werden durch ein Suchverfahren (Koordinatenabstieg) optimiert, die Startwerte für T_1 und T_2 sind heuristisch vorgegeben.

6. Fehlermaß ist der mittlere quadratische Abstand zwischen Eingabe– und Modell–F_0; ein neues Akzentkommando wird vergeben, wenn der Fehler eine dynamisch angepaßte relative oder absolute Schwelle überschreitet, oder wenn Randbedingungen verletzt werden, die sich auf die Parameter-Wertebereiche beziehen (insbesondere gibt es keine negativen Amplituden).

7. Nach der Vergabe des dritten Akzentkommandos werden die ersten beiden Akzentkommandos und das Phrasenkommando ausgegeben, alle folgenden Akzentkommandos werden ausgegeben, sobald sie festliegen, d.h. immer wenn ein neues Akzentkommando vergeben wird, wird das bis dahin letzte ausgegeben. Falls frühere Kommandos aufgrund einer gefallenen Modell–F_{min} geändert werden, hat das auf die Ausgabe keinen Einfluß mehr.

8. Der Signalnachlauf in ms ist demnach so lang wie die aktuelle Akzentgruppe. Dieser Wert ist für spätere Versionen relevant, in der auch Akzentereignisse an andere Komponenten gemeldet werden. Da der Satzmodus erst am Satzende bestimmt wird, ist hier der Signalnachlauf durch die Pausenschwelle von 600 ms bestimmt.

Das Programm kann durch zahlreiche Optionen beim Start beeinflußt werden. Er ist soweit wie möglich modularisiert, aber das Steuermodul, in dem alle Rand– und Ausnahmebedingungen als Regeln codiert sind, ist ziemlich groß. Um das Zusammenwirken der Regeln nachvollziehen zu können, ist zum Ausgleich ein interaktiver Betrieb möglich, d.h. die Analyse kann an definierten Stellen angehalten werden, um Parameter zu ändern, und um die Regelentscheidungen in Textform und alle Zwischenergebnisse in graphischer Form via `gnuplot`[4] zu betrachten (die Bilder 5.2 und 5.3 sind vom Programm erzeugt).

Abbildung 5.3 zeigt die Analyse eines längeren Turns mit dem Fujisaki-Modell. Obwohl es unter den „guten“ Beispielen ausgewählt wurde, zeigt es eine prinzipielle Schwäche des Modells: Die im Japanischen nicht vorkommenden

[4] `gnuplot` ist ein PD–Graphikprogramm.

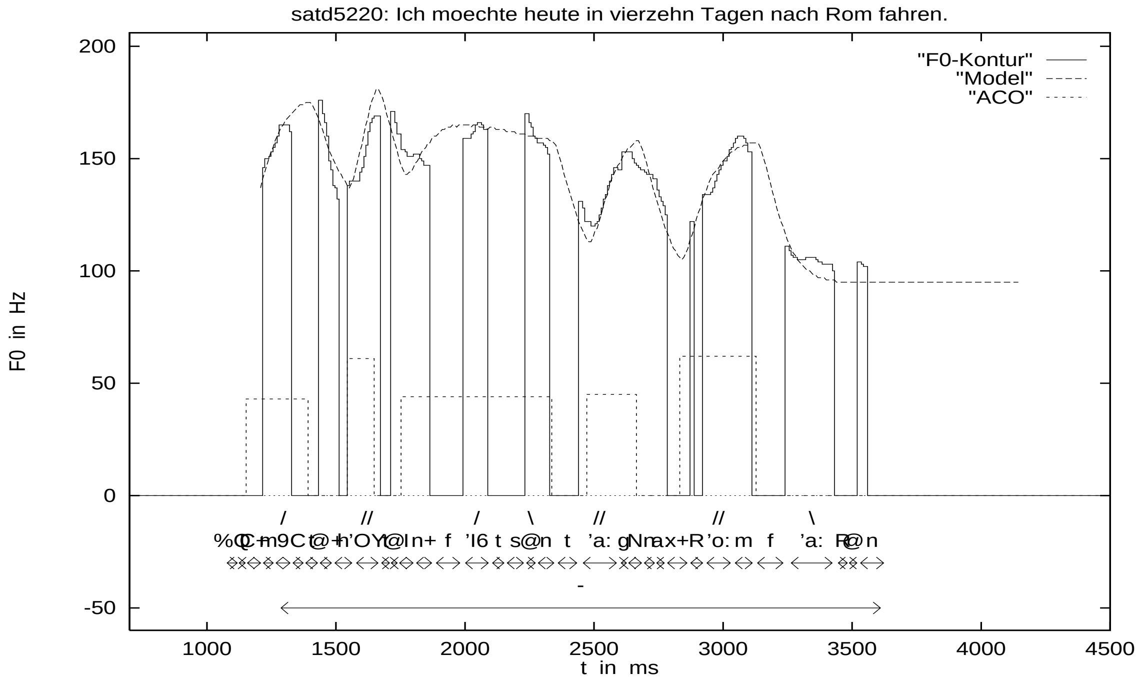

Abbildung 5.3: *Der Satz* `satd5220`, *analysiert nach dem Modell von Fujisaki.* `Model` *bezeichnet die Modell-Grundfrequenz, die vor dem ersten und 600 ms nach dem letzten stimmhaften Bereich auf 0 gesetzt ist.* `ACO` *bezeichnet die Akzentkommandos. Das Phrasenkommando ist hier nicht dargestellt. Oberhalb der* SAMPA*-Transkription sind die Handetiketten für Akzente zu sehen als einfache oder doppelte Schrägstriche für schwach/stark steigend/fallend.*

Akzente mit fallender Grundfrequenz können nicht explizit durch Akzentkommandos dargestellt werden. Vielmehr muß das vorhergehende Akzentkommando modifiziert werden, wie hier beim Brückenakzent über „Rom fahren". Dies ist mit ein Grund für die selbst in diesem Beispiel relativ schlechte Zeitauflösung der Akzentkommandos.

Für die meisten Sätze des Phondat-II-Zugauskunftskorpus sieht die Parametrisierung noch unbefriedigender aus. Zwar lassen sich einzelne ungünstige Modellierungen durch leichtes Ändern der Schwellwerte verhindern, das Problem liegt jedoch darin, daß sich eine gegebene Grundfrequenzkontur mit verschiedenen Modellparametern etwa gleich gut annähern läßt im Sinne des Fehlermaßes; für welche sich der Suchalgorithmus entscheidet, hängt empfindlich von diesen Schwellwerten und den konkreten Grundfrequenzwerten ab.

5.2.3 Dekomposition der Grundfrequenzkontur mit einer Filterbank

Wegen der praktischen Schwierigkeiten bei der Implementierung des Intonationsmodells von Fujisaki, die bereits auf Schwächen des Modells hinweisen, und der damit erzielten bescheidenen Satzmodus-Erkennungsraten (vergleiche Abschnitt 6.1) wurde eine alternative Merkmalsextraktion entwickelt, die sich an der bewährten Parametrisierung durch Regressionsgeraden orientiert (vergleiche die einleitenden Bemerkungen zu Abschnitt 5.2).

Die Forderung nach inkrementeller Verarbeitung schließt globale Merkmale aus, die sich z.B. aus einer Regressionsgeraden durch die gesamte Äußerung ergeben. Möglich gewesen wären Regressionsgeraden durch Ausschnitte der Grundfrequenzkontur mit fester Länge, wobei im Sinne der inkrementellen Verarbeitung das Analysefenster allmählich über die Grundfrequenzkontur geschoben wird, breite Fenster für das globale Verhalten, schmale für das lokale.

Einen ganz ähnlicher Effekt läßt sich durch Tiefpaßglättung erreichen, wenn die Grundfrequenzkontur vorher in stimmlosen Bereichen interpoliert wird, so daß sich ein glatter Verlauf ergibt: Das Tiefpaßfilter eliminiert die lokalen Schwankungen der Grundfrequenz, so daß die Amplitude des gefilterten Signals und die Amplitude ihrer zeitlichen Ableitung in etwa der Höhe und Steigung einer Regressionsgeraden für das globale Verhalten entsprechen, deren Parameter sich mit Fortschreiten des Analysefensters ebenfalls nur allmählich ändern. Die Grenzfrequenz entspricht dabei in etwa der Fensterbreite.

Bei Tiefpaßfilterung mit Butterworthfiltern [Sch88] ist das „Fenster“ genau genommen nach links offen, aber je höher die Grenzfrequenz liegt, desto geringer ist der Einfluß weiter zurückliegender Signalwerte, siehe die Ausführungen zur Gruppenlaufzeit weiter unten. Im Gegensatz dazu kommt bei der Berechnung der Regressionsgeraden eine Mittelung über ein Fenster fester Breite zum Tragen, die ebenfalls Tiefpaßcharakteristik hat.

Um eine Beschreibung des lokalen Verhaltens der Grundfrequenz zu erhalten, wird die interpolierte Grundfrequenzkontur bandpaßgefiltert, wobei die Grenzfrequenzen den Grad der Lokalität bestimmen. Die bandpaßgefilterte Grundfrequenz gibt dann die Tonhöhe relativ zur globalen Tonbewegung an. Da sie später ebenfalls nur an einigen Stellen abgetastet wird, wird auch von ihr die zeitliche Ableitung berechnet, um an diesen Stellen Information über die Tendenz der relativen Tonbewegung zu erhalten.

Zunächst wurde die Grundfrequenzkontur in zwei Komponenten zerlegt: das Band von 0 bis 1.1 Hz[5], welches die globale Tonbewegung wiedergibt, und das Band von 1.4 bis 3.2 Hz, welches das Steigen und Fallen der Tonhöhe im Zeit-

[5]Hier handelt es sich um 3 dB Grenzfrequenzen.

bereich von Silben widerspiegelt. Der Einfluß höherfrequenter Anteile, die von mikroprosodischen Effekten oder Fehlern in der F_0–Berechnung herrühren, ist dadurch weitgehend eliminiert.

Die Digitalfilter wurden als Potenz- oder Butterworth-Filter implementiert [Sch88]. Tatsächlich wurden statt eines Tief- und Bandpasses zwei Tiefpässe verwendet, wobei der Bandpaß durch Subtraktion des Zeitsignals zum Band 0 – 1.1 Hz vom Zeitsignal zum Band 0 – 3.2 Hz emuliert wurde. Das ist ohne weiteres möglich, wenn die Zeitsignale synchronisiert sind.

Bei jedem kausalen Digitalfilter ist das Ausgabesignal gegenüber dem Eingabesignal phasen- bzw. zeitverschoben. Im allgemeinen hängt die Phase von der Frequenz ab, nur bei linearphasigen Filtern ist sie konstant. Ein übliches Maß dafür, wie schnell die einzelnen Frequenzanteile das Filter passieren, ist die *Gruppenlaufzeit*, die als Ableitung der Phase nach der Frequenz definiert ist. Um die *Verzögerung in Abtastwerten* zu erhalten, wird die stetige (d.h. sprungbereinigte) Phase durch die normierte Frequenz dividiert.

Die vom Autor implementierten Filter kompensieren diese Verzögerung durch Abschneiden von entsprechend vielen Signalwerten zu Beginn und Ergänzen von Signalnullen am Ende. Die Verzögerung in Abtastwerten wird während des Filterentwurfs berechnet und, da sie ebenfalls eine Funktion der Frequenz ist, über alle Frequenzen des Durchlaßbereichs gemittelt.

Abbildung 5.4 zeigt das Dämpfungsverhalten für die beiden Filter mit den o.a. Grenzfrequenzen.

Die Grenzfrequenzen wurden zunächst auf visueller Basis eingestellt. Nach der Entwicklung des ersten Satzmodusklassifikators wurden drei Bänder verwendet, deren Grenzfrequenzen mit einem Suchverfahren (Koordinatenabstieg mit dynamisch angepaßten Schrittweiten) hinsichtlich der Akzenterkennungsraten optimiert wurden. Abbildung 6.1 auf Seite 88 zeigt die drei Bänder für einen Beispielsatz.

Die Werte der Grundfrequenzkontur und ihrer ersten Komponente sind nur in sehr grober Näherung normalverteilt, was für die Klassifikation mit einem Normalverteilungsklassifikator von Nachteil ist. Andere Autoren stellen die Grundfrequenz oft auf einer Halbton- oder logarithmischen Skala dar. Das Experiment zur Satzmodusklassifikation in Abschnitt 6.1.2 wurde mit logarithmierten Grundfrequenzwerten wiederholt, wobei sich die Erkennungsrate verschlechterte. Da logarithmierte Werte offensichtlich noch weiter von der Normalverteilung abweichen, wurden im folgenden die Grundfrequenzwerte in Hz als Merkmale verwendet.

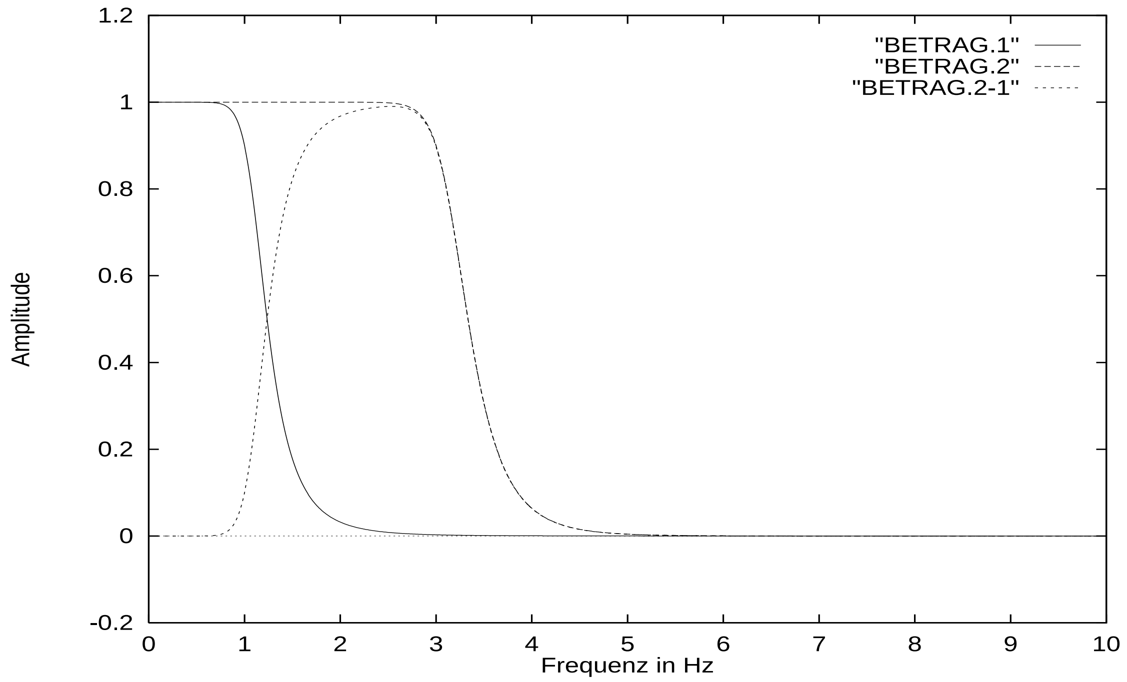

Abbildung 5.4: *Betrag der Übertragungsfunktionen der zwei Tiefpaßfilter in der Filterbank (Butterworth-Filter 6. und 12. Grades) und der Differenzübertragungsfunktion.*

Grundfrequenz-Interpolierer

Die Grundfrequenz ist nur in stimmhaften Bereichen definiert, in stimmlosen Bereichen wird sie durch einen sonst nicht vorkommenden Wert codiert, meist mit Null. Das resultierende Grundfrequenz-„Signal“ hat daher an jedem Stimmhaft-Stimmlos-Übergang einen Sprung. Würde man ein solches Signal mit einem Digitalfilter bearbeiten, hätte jeder Sprung einen stark ausgeprägten Einschwingvorgang zur Folge, der die gewünschte Filterausgabe für die „stetigen“ Abschnitte der Grundfrequenz zu sehr überlagern würde. Deshalb muß die Grundfrequenz in stimmlosen Bereichen interpoliert werden, um ein möglichst glattes Grundfrequenzsignal zu erhalten.

Die einfachste Möglichkeit ist die lineare Interpolation, gefolgt von einer Tiefpaßfilterung, um Knickstellen zu glätten. Der Tiefpaß sollte nicht zu rigide sein, um die berechneten Grundfrequenzwerte an den Rändern stimmhafter Bereiche möglichst wenig zu verändern.

Bei Voruntersuchungen zur Akzentdetektion wurden auch zwei andere Interpolationsverfahren eingesetzt und hinsichtlich der Akzenterkennungsrate verglichen. Kubische Splines führten zu schlechteren Ergebnissen, weil der Splineverlauf in stimmlosen Bereichen empfindlich von den konkreten Werten an den Rändern der umgebenden stimmhaften Bereiche[6] abhängt. In [HE94] ist ein für F_0–Konturen modifiziertes Spline–Verfahren beschrieben, das nach mehreren Iterationsschritten zu einer stabilen Interpolation gelangt. Zu einem späteren Zeitpunkt fand mit dem Autor ein Softwaretausch statt, und auch dieses Verfahren wurde in den Vergleich einbezogen. Das weiter unten beschriebene, selbst entwickelte iterative Verfahren erwies sich jedoch für die hier gewünschte Anwendung, die nachfolgende Bandpaßfilterung, als günstiger.

Auch das Intonationsmodell von Fujisaki eignet sich zur Interpolation, wenn man die linguistisch motivierten Randbedingungen wegläßt, so daß eine Grundfrequenzkontur auch durch eine dichte Folge impulsförmiger Akzentkommandos modelliert werden kann. Mit Merkmalen aus der so interpolierten Grundfrequenzkontur konnten Akzente besser erkannt werden. Das lag zum Teil an der relativen Unempfindlichkeit gegenüber Grundfrequenzfehlern wegen Phänomenen Stimmeinsatzes. Doch auch nach Aussortieren aller Turns mit entsprechenden Fehlern (für die Voruntersuchungen wurde nur eine sehr kleine Stichprobe verwendet) schnitt die Fujisaki-Interpolation besser ab. Ein wesentlicher Unterschied gegenüber der linearen Interpolation besteht darin, daß bei Fujisaki die Modell-Grundfrequenz in stimmlosen Bereichen ein Tal bildet, so daß die Grundfrequenzgipfel an akzentuierten Silben stärker hervorgehoben werden. Da auch im Deutschen die meisten Akzente hoch sind, ist dies möglicherweise ein Grund für das bessere Abschneiden der Interpolation mit dem Fujisaki-Modell.

Diese Voruntersuchungen führten dazu, die lineare Interpolation so zu modifizieren, daß sie wie die Fujisaki-Interpolation umso stärker zur Talbildung in einem stimmlosen Bereich neigt, je größer dieser Bereich ist. Dies wurde mit einem iterativen Verfahren erreicht, das im folgenden beschrieben wird.

Die Grundidee war, zwei Interpolationsverfahren schrittweise aneinander anzunähern. Bei dem einen handelt es sich um lineare Interpolation in stimmlosen Bereichen und der Bildung einer Rampe vor dem ersten und nach dem letzten Bereich, bei dem anderen um Interpolation durch Tiefpaßglättung. Bild 5.5 zeigt die Wirkung beider Verfahren.

Die Talbildung in längeren stimmlosen Bereichen kann iterativ erreicht werden. Den ersten Schritt zeigt Bild 5.6: Die „angepaßte“ Kurve `A1` ist in stimmhaften Bereichen identisch mit der originalen Grundfrequenzkontur. In stimmlosen Bereichen wird die tiefpaßgefilterte Grundfrequenzkontur des letzten Iterationsschrittes (hier die Kurve `I0` in Bild 5.5) an den Rändern des stimmlosen

[6] Kritisch sind Phänomene des Stimmeinsatzes, besonders bei periodengenauer F_0–Berechnung. Nicht alle diese Fehler können durch Median-Glättung verhindert werden. Siehe als Beispiel Abbildung 5.2 auf Seite 56.

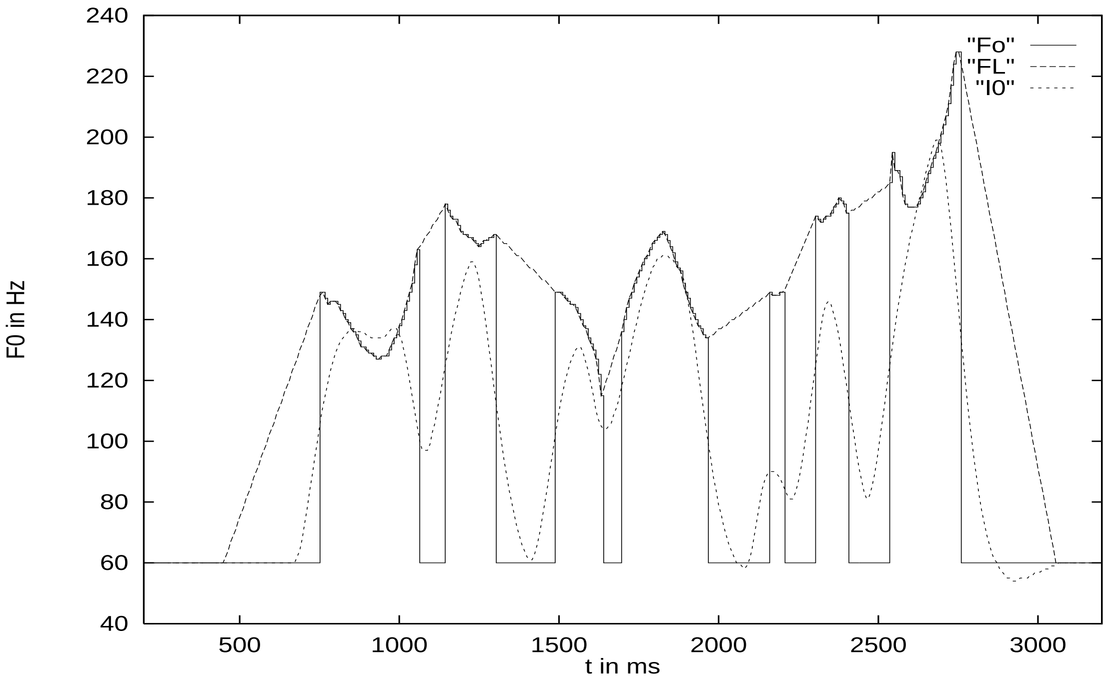

Abbildung 5.5: *Die initialen Signale des F0-Interpolierers:* Fo *ist die originale Grundfrequenzkontur, wobei die Stimmlos-Linie auf 60 Hz angehoben ist,* FL *ist die linear interpolierte Grundfrequenzkontur, während* I0 *durch Tiefpaßfilterung der Grundfrequenzkontur entsteht. Im Laufe der Iteration soll in stimmlosen Bereichen ein Kompromiß zwischen* FL *und* I0 *gebildet werden; gleichzeitig soll die interpolierte Grundfrequenzkontur in stimmhaften Bereichen der originalen Grundfrequenzkontur möglichst nahe kommen.*

Bereiches „hochgezogen", um Stetigkeit zu erreichen. Dieses Hochziehen erfolgt durch Differenzbildung aus tiefpaßgefiltertem und linear interpoliertem Signal, Gewichtung der Differenz mit einem halben Cosinus (so daß die Ränder mit 1, die Mitte mit 0 gewichtet werden), und Addition dieser gewichteten Differenz auf das tiefpaßgefilterte Signal.

Dieses angepaßte Signal, `A1` in Bild 5.6, ist an den Stimmhaft/Stimmlos–Übergängen stetig, aber nicht glatt. Daher wird es tiefpaßgeglättet, und in Bild 5.6 ergibt sich die Kurve `I1`. Dies ist die Ausgabe des ersten und die Eingabe des zweiten Iterationsschrittes.

Der Effekt von fünf Iterationsschritten ist in den Bildern 5.7 und 5.8 zu sehen: In kurzen stimmlosen Bereichen nähert sich die interpolierte Kurve stark an die linear interpolierte an, in längeren stimmlosen Bereichen fällt die interpolierte

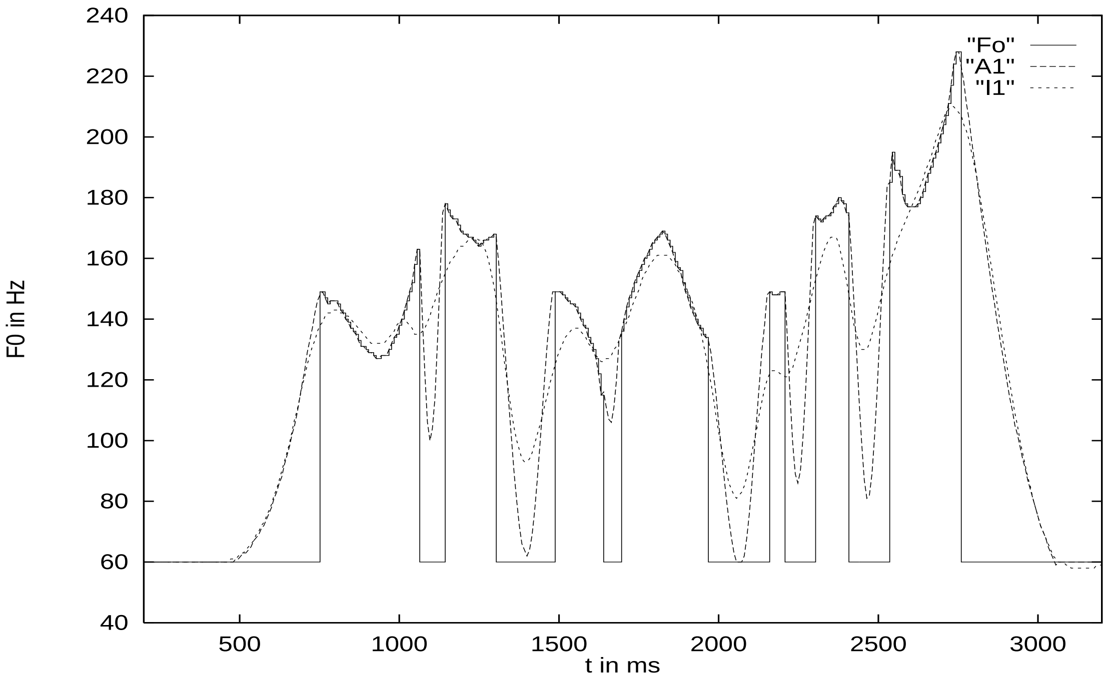

Abbildung 5.6: A1 *entsteht durch Anpassen von* I0 *an* FL, *d.h. in stimmhaften Bereichen durch Gleichsetzen mit* Fo *(siehe Bild 5.5), und in stimmlosen Bereichen durch Subtraktion der hanning-gewichteten Differenz aus* FL *und* I0. *Tiefpaßfilterung des so entstandenen* A1 *ergibt* I1.

Kurve ab, weil der Effekt der ersten Tiefpaßfilterung länger erhalten bleibt. Quantitativ wird dieses Verhalten durch die Anzahl der Iterationsschritte und durch die Filtergrenzfrequenzen gesteuert.

Die Ein– und Ausgabe des Interpolierers zeigt Bild 5.9. Die Sprünge 300 ms vor dem ersten und nach dem letzten stimmlosen Bereich, das ist gerade die Breite der Rampenfunktion, stören hier bei der weiteren Verarbeitung nicht, weil die Sprungantwort der Filterbank dort, wo dann die Merkmale berechnet werden, wieder abgeklungen ist. Für andere Anwendungen könnte man auf das Anheben und Absenken der Nullinie verzichten und stattdessen die Rampen breiter wählen. Das würde aber die Zeitverzögerung des Interpolierers erhöhen, die es für die Prosodiekomponente ebenfalls zu minimieren galt.

Abschließend noch einige Worte zu den Tiefpaßfiltern und zum Zeitverhalten des Interpolierers: Es handelt sich um Butterworth–Filter zweiten Grades, deren Dämpfungsverhalten aus Bild 5.10 ersichtlich ist. Die Duchlaßbereiche nehmen im Lauf der Iteration linear zu, damit die interpolierten Kurven I0

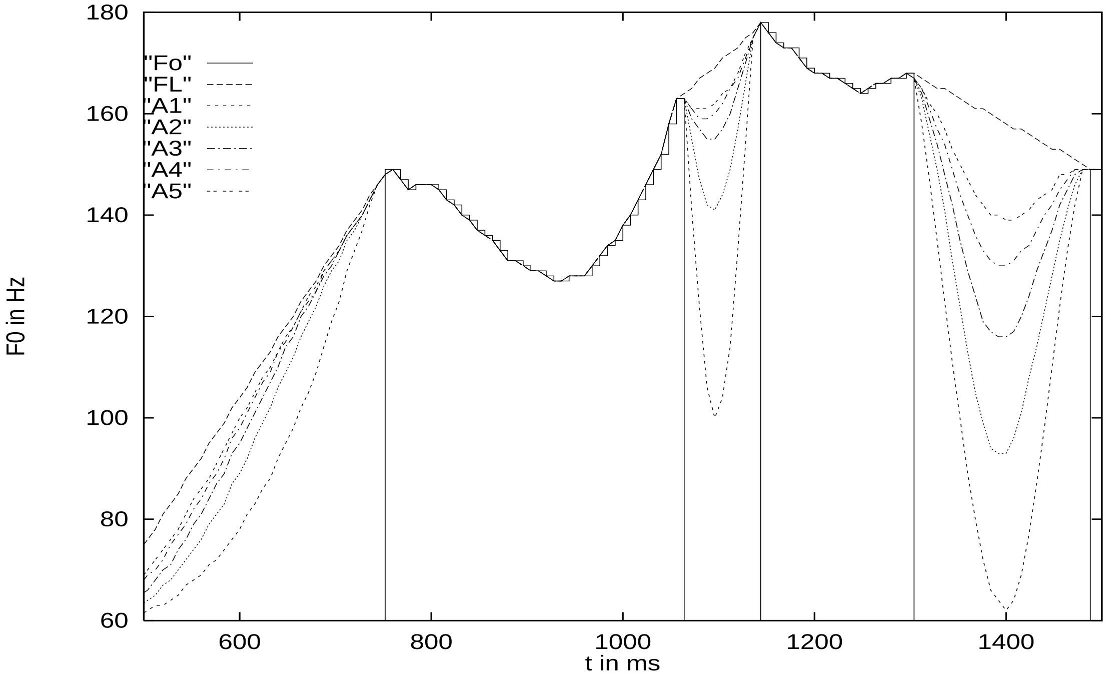

Abbildung 5.7: *Änderung der „angepaßten Kurven" im Laufe der 5 Iterationsschritte, zu sehen an einem Ausschnitt aus Bild 5.6.*

bis I5 an dem Rändern stimmhafter Bereiche immer weniger durch die Tiefpaßglättung verfälscht werden (gegenüber der originalen F_0); das wird möglich, weil die angepaßten Kurven A1 bis A5 im Lauf der Iteration glatter werden[7].

Die Zeitverzögerung der Filter, die sich aus der mittleren Phasenverschiebung in den Durchlaßbereichen ergibt, addiert sich für alle Filter auf 392 ms. Die Zeitverzögerung des linearen Interpolierers ist 0 in stimmhaften Bereichen, und in stimmlosen Bereichen gerade so lange wie der stimmlose Bereich, jedoch maximal 300 ms, das ist die Breite der Rampen.

[7]Würde man gleich in der ersten Stufe mit einem breiten Durchlaßbereich arbeiten, würde der Einschwinganteil, der von den noch sehr scharfen Knicken in der Kurve A1 herrührt, nicht genügend gedämpft und sich durch alle Iterationen fortpflanzen.

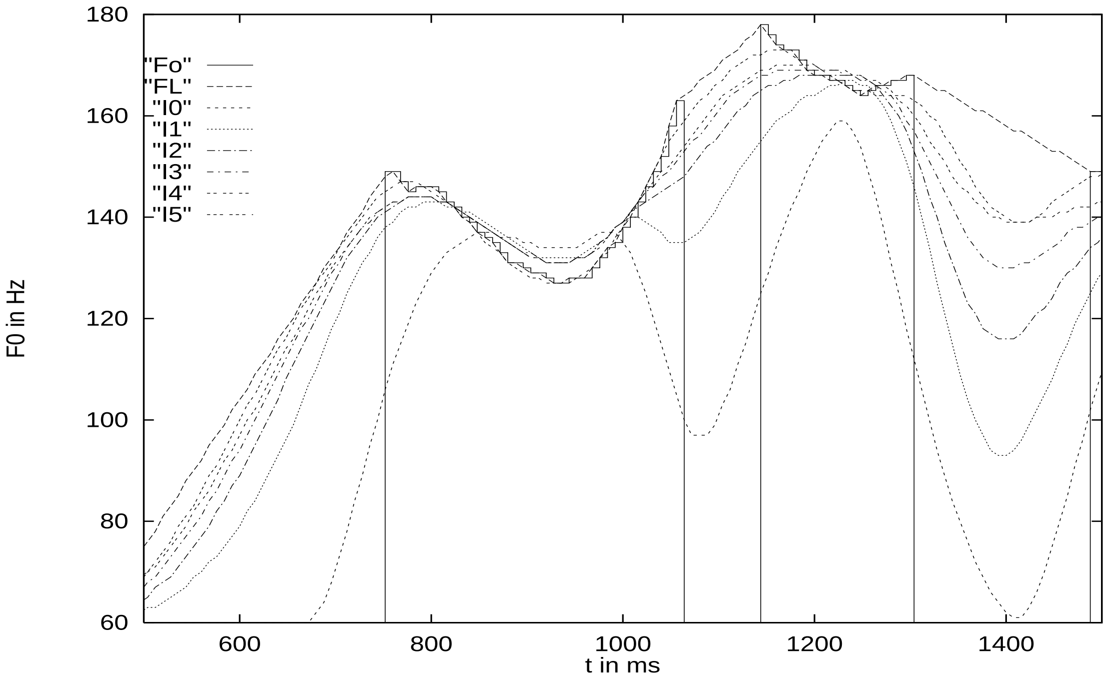

Abbildung 5.8: *Interpolierter Grundfrequenzverlauf nach 1 bis 5 Iterationsschritten, zu sehen an einem Ausschnitt aus Bild 5.6.*

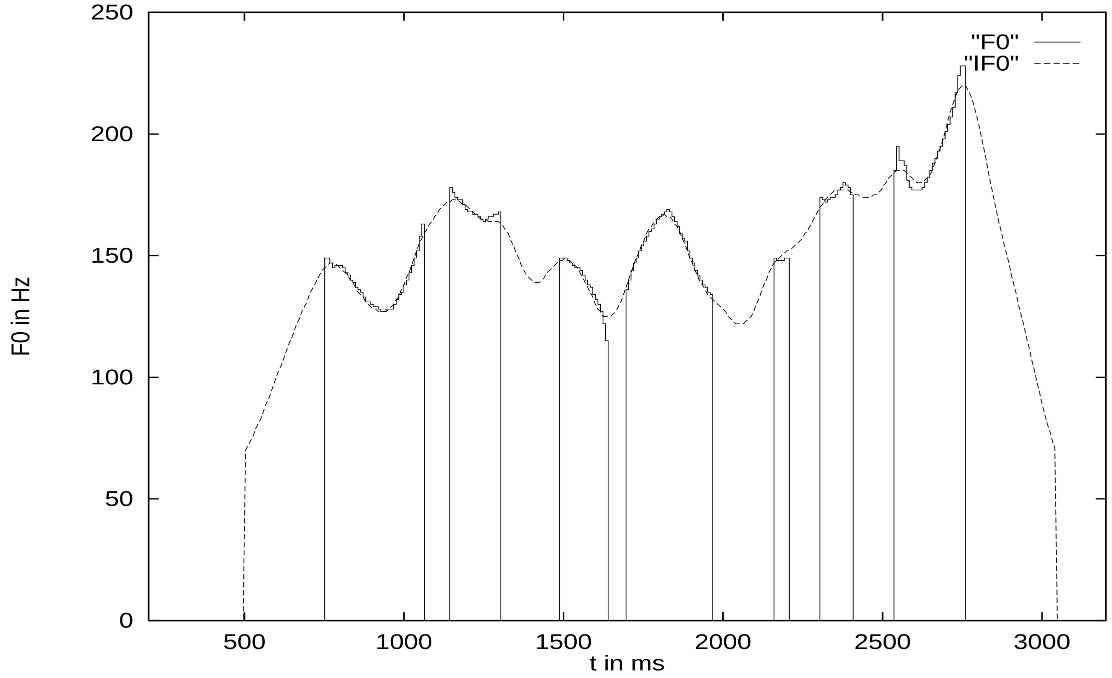

Abbildung 5.9: *Der originale Grundfrequenzverlauf* F0 *und der interpolierte Grundfrequenzverlauf* IF0 *(identisch mit* I5 *aus Bild 5.8, nur die Stimmlos–Linie ist wieder auf 0 Hz abgesenkt, vgl. Bild 5.5).*

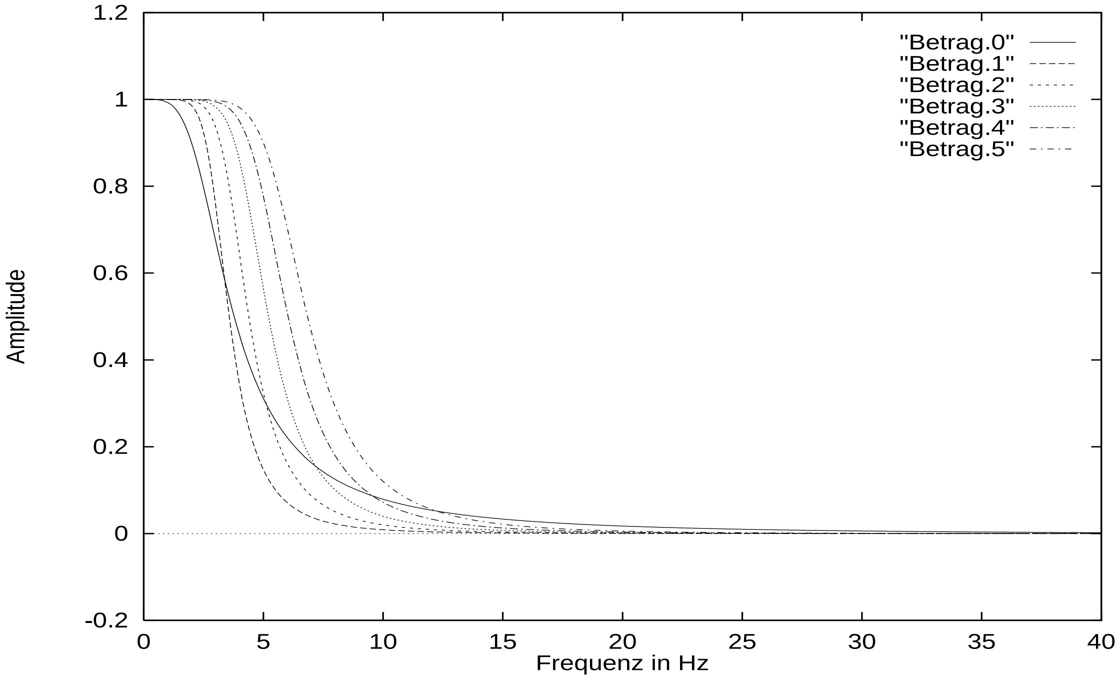

Abbildung 5.10: *Betrag des Frequenzgangs der fünf Tiefpaßfilter im F0-Interpolierer (Butterworth-Filter zweiten Grades).*

5.3 Energiemerkmale

Zur Beschreibung des Energieverlaufs werden aus dem Sprachsignal für jedes Analysefenster drei Energiemerkmale berechnet, die in [Nöt91] für die Silbenkerndetektion verwendet werden. Dazu wird die Kurzzeit-DFT[8] über 256 Abtastwerte bzw. 16 ms zu drei Frequenzbändern zuammengefaßt:

- 100 - 300 Hz (nasales Band)
- 300 - 2300 Hz (sonorantes Band)
- 2300 - 5000 Hz (frikatives Band)

Anschließend der zeitliche Verlauf dieser Merkmale mit einem Medianfilter der Breite 5 geglättet.

Diese Energiemerkmale bilden zusammen mit den Grundfrequenzmerkmalen die *Basismerkmalvektoren* zur Akzent- und Phrasengrenzendetektion, außerdem werden mit ihnen Silbenkerne detektiert als Voraussetzung zur Phrasengrenzendetektion, siehe Abschnitt 5.5.

5.3.1 Silbenkerndetektion

Der Begriff *Silbenkern* stammt aus der *Schallfülletheorie* nach [Sie85], die z.B. in [Koh77] beschrieben ist. In der Regel entspricht der Silbenkern dem Vokal in der Silbenmitte, Grenzfälle sind die silbischen Konsonanten: In Silben, in denen der Reduktionsvokal `/@/` Silbenträger ist, kann bei Elision des `/@/` der folgende Konsonant zum Silbenträger werden (z.B. reden → red**n**).

Der Silbenkerndetektor ist eine Reimplementierung des in [Nöt91] beschriebenen Verfahrens. Es arbeitet mit den oben genannten drei Energiemerkmalen. Im wesentlichen werden „ausgeprägte" lokale Maxima im sonoranten Band gesucht, d.h. ihr Energiewert muß über einer globalen Schwelle liegen. Als zum Silbenkern gehörig zählen alle Frames in der Umgebung des Maximums[9], deren Energie im sonoranten Band sowohl über einer Schwelle relativ zu diesem Maximum als auch über der Energie im frikativen Band liegt.

Die Grenzfrequenzen des sonoranten Bandes geben den Wertebereich der ersten beiden Formantfrequenzen von Vokalen wieder. Kritisch ist die untere

[8] Diskrete Fouriertransformation

[9] Die Maxima in einer Folge aus Energiewerten $\langle e_i \rangle$ werden bestimmt mit der Bedingung $e_i < e_{i+1} > e_{i+2}$. Wegen der Medianglättung können aber drei oder mehr aufeinanderfolgende Werte exakt gleich sein. Um dieses Problem (nahezu sicher) zu umgehen, werden die Energiewerte nach der Median-Glättung noch einer linearen Tiefpaßglättung unterzogen, deren Effekt gering ist, aber zur eindeutigen Bestimmung der Signalmaxima ausreicht.

Grenze, da die erste Formantfrequenz von Nasalen und Liquiden nur knapp darunter liegt.

Auslassungsfehler bei silbischen Konsonanten (z.B. red**n**) sollen vermieden werden, indem im nasalen Band nach ausgeprägten Maxima gesucht wird, die zu keinem Maximum im sonoranten Band gehören. Um Einfügungsfehler bei Frikativen zu vermeiden (der niedrigste Formant des `/S/` liegt bei etwa 2000 Hz) wird die Energie im sonoranten Band mit der im frikativen Band verglichen.

Die Aufspaltung „überlanger Silbenkerne“ [Nöt91] durch Absenken der Energieschwelle relativ zum lokalen Maximum innerhalb dieses Silbenkerns (mit der relativen Energieschwelle wird die Dauer der Silbenkerne beeinflußt) wurde ebenfalls implementiert. Nach der Optimierung der Schwellwerte trat dieser Fall jedoch nicht mehr auf.

Die Schwellwerte und die Grenzfrequenzen der Energiebänder wurden mit einem Optimierungsverfahren[10] an die Karlsruher Dialoge angepaßt; die oben genannten Grenzfrequenzen änderten sich dabei nur geringfügig.

Als Erweiterung des in [Nöt91] beschriebenen Verfahrens wurden die konstanten Schwellwerte durch dynamisch angepaßte Schwellwerte ersetzt. Die damit erzielte Verbesserung war zwar gering, aber bei ungleichmäßig ausgesteuerten Sprachsignalen würde sich diese Erweiterung auszahlen.

5.4 Dauermerkmale

Akzentuierung und Phrasierung spiegeln sich nicht nur im Verlauf der Grundfrequenz und der Energie wieder, sondern auch in der Silbendauer: Sowohl betonte Silben als auch phrasenfinale Silben werden gedehnt [BE88, Cam94], wobei sich beide Effekte überlagern. Verschiedene Untersuchungen darüber, welche der drei akustischen Aspekte, Dauer, Grundfrequenz- oder Energieverlauf, am stärksten mit der Akzentuierung korrelieren, ergeben kein klares Bild [Bec86, Seite 173]. Für die Erkennung der Phrasengrenzen in der Verbmobil-Stichprobe kommt auch [Kie97] zu dem Schluß, daß zwischen den entsprechenden Merkmalgruppen keine eindeutige Hierarchie besteht, sondern daß sie sich vielmehr ergänzen.

Wenn die Wortkette bzw. ein Worthypothesengraph vorliegt, lassen sich aus der Zeitzuordnung der Laute problemlos Silben- und Silbenkerndauern bestimmen. Die absolute Dauer eines Silbenkerns hängt jedoch nicht nur davon ab, ob die Silbe akzentuiert ist oder nicht, sondern auch von der Sprechgeschwindigkeit und davon, ob es sich um einen kurzen oder langen Vokal oder einen Diphthong handelt. Bei Kenntnis der gesprochenen Wörter und der intrinsischen Lautdauern (Mittelwerte und Standardabweichungen werden anhand einer Stichprobe vorab

[10] Koordinatenabstieg mit dynamisch angepaßten Schrittweiten

bestimmt) kann zunächst die Sprechgeschwindigkeit geschätzt werden (was allerdings inkrementell nur bedingt möglich ist, weil dafür ein größerer Zeitraum benötigt wird, z.B. eine Phrase oder der ganze Turn). Damit läßt sich dann die normierte Silben- oder Silbenkerndauer berechnen, bei der die akzentbedingte Dehnung deutlicher hervortritt.

Wird jedoch auf Wortinformation verzichtet, kann auch keine Normierung auf lautintrinsische Dauern vorgenommen werden. Selbst die (nicht-inkrementelle) Bestimmung der Sprechgeschwindigkeit allein aus den detektierten Silbenkernen ist nicht möglich, wie Voruntersuchungen gezeigt haben. Deshalb werden zur Phrasengrenzendetektion die absoluten Silbenkerndauern in ms als Merkmale herangezogen (siehe nächster Abschnitt).

Bei der Akzentdetektion, die sich nicht auf die Silbenkerndetektion stützt, kommt das Dauerkriterium nur indirekt, bei der Nachbearbeitung der klassifizierten Frames, in Form der „Mindestlänge akzentuierter Bereiche“ zum Tragen (siehe Abschnitt 6.2).

Trotzdem tragen auch nicht-normierte Dauermerkmale zur Verbesserung der Erkennung bei, wie das Resultat des Merkmalauswahlsverfahren in Abschnitt 6.3 zeigen wird.

5.5 Komplexe Merkmale

Die Akzentklassifikation ist bereits für jeden Frame möglich; dies erfolgt mit den oben genannten Grundfrequenz- und Energiemerkmalen, die für jeden Frame berechnet und im folgenden *Basismerkmale* genannt werden. Die Basismerkmale benachbarter Frames sind sich relativ ähnlich wegen der Glättungsschritte: Medianfilterung bei den Energiemerkmalen, Tiefpaßglättung bei den Grundfrequenzkomponenten, Bildung der Ableitungen durch Regressionsgeraden.

Da Grenzen im Gegensatz zu den Kernen akzentuierter Silben keine Ausdehnung besitzen, müssen vor der Klassifikation von Phrasengrenzen erst die in Frage kommenden Stellen ermittelt werden, damit dort als Merkmalvektor eine Beschreibung der umgebenden Grundfrequenz- und Energiekontur erstellt werden kann.

Im Falle der ersten Implementierung des Satzmodusdetektors war dies einfach: Die Turns des Phondat-Korpus bestanden in der Regel aus einem Satz, für die Ausnahmen wurde nach längeren Sprechpausen gesucht. Zur Bildung des zusammengesetzten Merkmalvektors wurden die Basismerkmale im ersten, vorletzten und letzten stimmhaften Bereich komponentenweise gemittelt, siehe Abschnitt 6.1.2.

Im Verbmobilkorpus, dessen Turns meist aus mehreren satzwertigen Phrasen[11] bestehen, können anhand von Sprechpausen (und dem END-OF-FILE) nicht mehr alle Satzgrenzen gefunden werden. Sollen auch Phrasengrenzen gefunden werden, die keine Satzgrenzen sind, müssen alle Wortgrenzen klassifiziert werden.

Liegt die Wortkette mit zeitlicher Lautzuordnung vor, ist es einfach, daraus die Zeitpunkte der Wortgrenzen zu bestimmen. Neben dem Grundfrequenz- und Energieverlauf an diesen Stellen ist es dann auch einfach, die (normierten) Dauern der umgebenden Silben zu bestimmen.

Da hier keine Wortinformation benutzt wird, müssen diese Stellen aus der Silbenkerndetektion abgeleitet werden: Zwischen zwei Silbengrenzen (und nach dem letzten Silbenkern im Turn) muß eine Silbengrenze kommen, die potentiell eine Wortgrenze und damit eine Phrasengrenze ist.

Klassifiziert wird also jede Silbengrenze, genauer jeder Bereich zwischen zwei Silbenkernen. Um auszudrücken, daß es sich um einen Bereich handelt, und daß die Silbenkerndetektion fehlerhaft sein kann, werden im weiteren die detektierten Silbenkerne als SILBENKERNE geschrieben und die Bereiche zwischen ihnen als SILBENGRENZEN, im Gegensatz zu den tatsächlichen Silbengrenzen. Bei korrekter Silbenkerndetektion überdeckt eine SILBENGRENZE genau eine Silbengrenze.

Zur Grenzdetektion wird ein Fenster aus — wenn möglich — vier SILBEN betrachtet. Um die Grundfrequenz- und Energiekontur innerhalb des Fensters zu beschreiben, genügt es, die sich nur allmählich ändernden Basismerkmale an geeigneten Stellen auszuwählen.

Beim Fenstertyp **4** (siehe Abbildung 5.12) werden die Basismerkmale in der Mitte der vier detektierten SILBENKERNE ausgewählt und zu einem großen Merkmalvektor zusammengestellt. Die Dauern der SILBENKERNE und ihre Abstände (in ms) ergeben 7 weitere Merkmale. Abbildung 5.11 veranschaulicht das Vorgehen.

Für Fenster aus weniger als 4 SILBEN enthält der Merkmalvektor entsprechend weniger Komponenten. Bei n Basismerkmalen und einem k-SILBEN-Fenster ist die Anzahl der Komponenten

$$m = kn + k + (k-1) \quad . \tag{5.4}$$

Abbildung 5.12 zeigt die übrigen Fenstertypen sowie deren Abfolge: Für die erste SILBENGRENZE werden Merkmale innerhalb der ersten 3 SILBEN bestimmt, für die letzte SILBENGRENZE innerhalb der letzten drei SILBEN der Äußerung (die durch die Äußerungsende-Schwelle von 500 ms definiert ist). Zur Klassifikation

[11] Unter satzwertigen Phrasen werden hier Konstruktionen ohne finites Verb verstanden, die außerhalb von Sätzen stehen können, z.B. Satzwörter wie „ja“ in „Ja? Zur Not geht's?“ oder Ellipsen wie „Auch am Dienstag?“; vergleiche [Buß90].

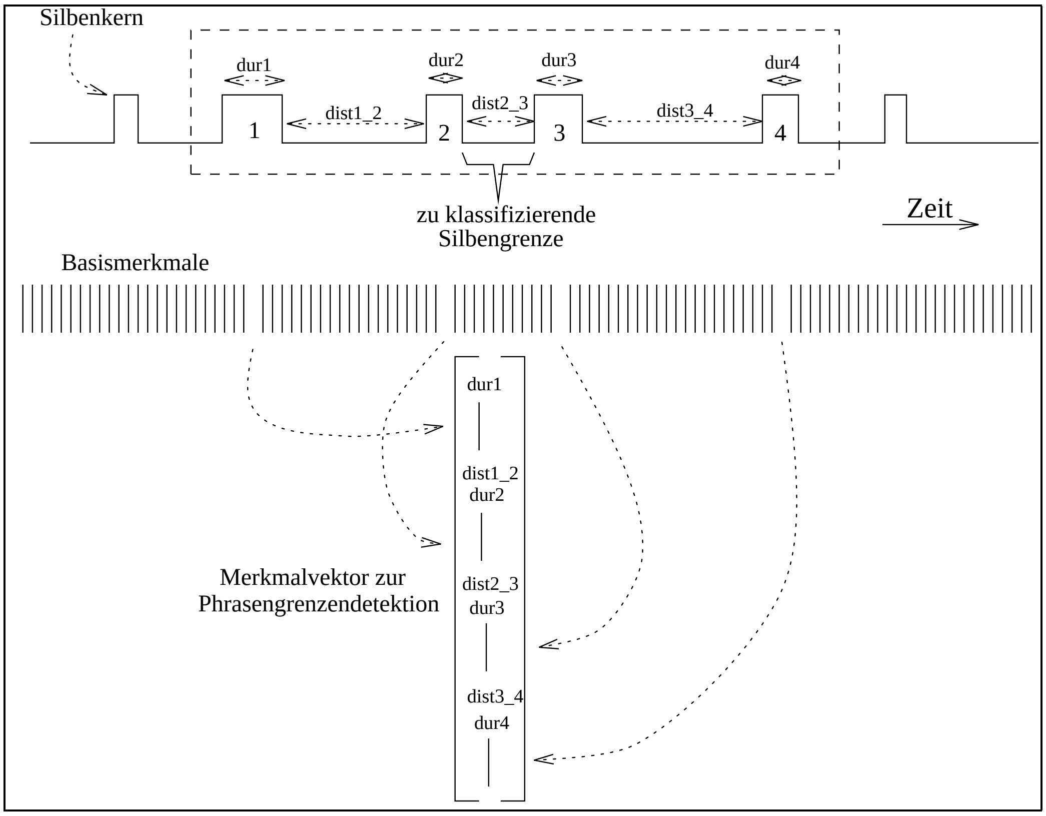

Abbildung 5.11: *Der Phrasengrenzen- und Satzmodusdetektor betrachtet (wenn möglich) immer Fenster aus 4 Silben. Die Basismerkmalvektoren mitten in den 4 detektierten Silbenkernen sowie 7 Zeitmerkmale bilden den Merkmalvektor, mit dem die Grenze zwischen dem 2. und 3. Silbenkern klassifiziert wird.*

des Äußerungsendes selbst werden Merkmale innerhalb der letzten beiden SILBEN bestimmt (Fenstertyp **2b**). Für die Sonderfälle ein- und zweisilbiger Äußerungen sind zusätzlich die Fenstertypen **1** und **2** nötig.

Die normale Fensterfolge ist **3a, 4, 4, ... 4, 3b, 2b**[12]. Sobald das Ende des 3. SILBENKERNS erreicht ist, wird die erste Grenzhypothese ausgegeben (Fenster **3a**) und nach jedem Ende eines weiteren SILBENKERNS eine weitere Grenzhypothese. Kommt 500 ms nach einem SILBENKERN kein weiterer, werden die Fenster **3b** und **2b** unmittelbar hintereinander angewendet.

Zu einem späteren Zeitpunkt wurde ein weiterer Fenstertyp **3c** eingeführt, der in mehr als zweisilbigen Äußerungen das Fenster **2b** ersetzt. Damit konnte der

[12] Falls ein Fenstertypname den Buchstaben *a* oder *b* enthält, bedeutet das, daß die SILBENGRENZE, die mit diesem Fenster klassifiziert wird, links bzw. rechts von der Fenstermitte liegt.

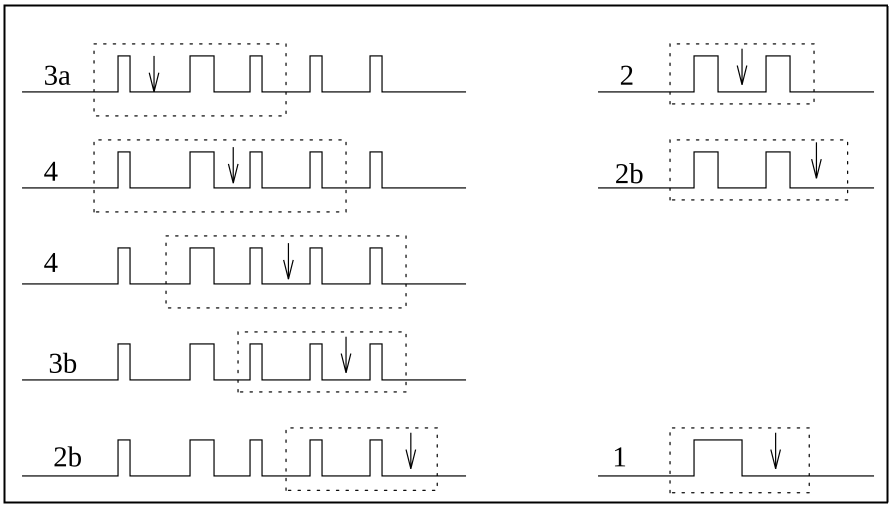

Abbildung 5.12: *Anfang und Ende einer Äußerung werden besonders behandelt, daher sind 6 Fenstertypen notwendig. In der linken Bildhälfte sieht man die normale Fensterfolge:* **3a, 4, 4, ... 4, 3b, 2b**. *Für den Fall ein- oder zweisilbiger Äußerungen (rechte Bildhälfte) sind die Fenstertypen* **1** *und* **2** *nötig. Die nach unten gerichteten Pfeile markieren die Silbengrenzen, auf die sich die Hypothesen beziehen.*

linke Kontext an diesen für die Phrasengrenzendetektion wichtigen Stellen von zwei auf drei SILBEN erhöht werden. Da aber für jeden Fenstertyp ein eigener Klassifikator trainiert werden muß und die eher zu kleine Stichprobe sich noch auf die Fenstertypen aufteilt (für die selten auftretenden Fenstertypen **1**, **2** und **2b** blieben nur entsprechend wenige Stichprobenelemente), wurde die Anzahl der Fenstertypen möglichst klein gehalten. Aus diesem Grund wurden auch keine Experimente mit größeren Kontexten, z.B. mit 6 oder 8 SILBEN, angestellt.

Kapitel 6

Klassifikation

In diesem Kapitel werden die Detektoren für Satzmodus, Akzente und Phrasengrenzen beschrieben. Sie basieren auf den in Kapitel 5 eingeführten Merkmalen, den Etiketten aus Kapitel 4 und verwenden durchweg Normalverteilungsklassifikatoren. Für jeden Detektor und jede Stichprobe (Phondat und Verbmobil) werden die Ergebnisse einiger Klassifikationsexperimente präsentiert.

Abschnitt 6.1 beschreibt den ersten Satzmodusklassifikator, wie er in das INTARC-1.3-System integriert wurde. Er war für die Phondat-Stichprobe ausgelegt, in der fast alle Turns aus genau einem Satz bestehen, somit konnten die Satzgrenzen, auf die sich die Modushypothesen beziehen, relativ leicht bestimmt werden. Die Referenz-Klassen wurden aus der Orthographie bestimmt. Zwei unterschiedliche Merkmalsätze wurden verwendet, die „Fujisaki-Merkmale“ und die — wie der Vergleich zeigt — besseren „Filterbankmerkmale“.

Die Filterbankmerkmale zusammen mit den Energiemerkmalen ergeben die Basismerkmale, die zur Akzentklassifikation dienen. Dem in Abschnitt 6.2 vorgestellten Akzentdetektor liegt ein Normalverteilungsklassifikator zugrunde, der für jeden Frame zwischen Silbenkernen und Nicht-Silbenkernen unterscheidet und gleichzeitig den Silbenkernen eine Akzentstufe zuordnet. Diese Klassifikator-Entscheidung wird auf die Oberklassen „akzentuiert/nicht-akzentuiert“ abgebildet und anschließend nachbearbeitet, wobei die Silbenkerndauer eine Rolle spielt. Der Akzentdetektor wurde zunächst für die Phondat-Stichprobe entwickelt, für die Etiketten aus Bonn und aus Braunschweig vorhanden waren, und beim Umstellen auf die Verbmobil-Stichprobe in seiner Struktur beibehalten.

Abschnitt 6.3 behandelt den Phrasengrenzen- und Satzmodusdetektor, der für die Verbmobil-Stichprobe entwickelt wurde, deren Turns meist aus mehreren Phrasen oder Sätzen bestehen. Beiden Detektoren liegt derselbe Normalverteilungsklassifikator zugrunde. An den **B2**-, **B3**- und **B9**-Grenzen wird der handetikettierte Verlauf der Intonation unterschieden, an den **B3**-Grenzen zusätzlich drei verschiedene Satzmodi, so daß sich 13 Klassen ergeben. Phrasengrenzentyp und Satzmodus werden durch Abbildung in die entsprechenden Oberklassen ermittelt. Im Gegensatz zur Akzentdetektion geht der Grenzdetektion die Silbenkerndetektion voraus: Die Silbenkerne definieren das Analysefenster zur Bildung der komplexen Merkmalvektoren aus Abschnitt 5.5 und sie legen mit den SILBENGRENZEN (vergleiche Abschnitt 5.5) die zu klassifizierenden Einheiten fest.

In Abschnitt 6.4 wird untersucht, inwieweit sich die detektierten **B3**-Grenzen zur Dialogaktsegmentierung eignen. Diese Anwendung war im Verbmobil-Prototyp für Situationen geplant, in denen keine Worthypothesen vorliegen und in denen deshalb der Erlanger Phrasengrenzenerkenner nicht eingesetzt werden kann.

6.1 Satzmodus

Für das System INTARC 1.2 wurde ein erster Satzmodus-Klassifikator implementiert, der zum einen auf dem Intonationsmodell von Fujisaki (siehe Abschnitt 5.2.2) und zum anderen auf der in Abschnitt 5.2.3 beschriebenen Dekomposition der Grundfrequenz mit einer Digitalfilterbank beruht.

Für die Verbmobil-Stichprobe wurde ein völlig neues Verfahren implementiert, das die Satzmodus- und die Phrasengrenzenerkennung in einen Klassifikator integriert, siehe Abschnitt 6.3.

Das INTARC-1.2-System war für 134 Turns des Phondat-II-Zugauskunftskorpus ausgelegt. 130 Turns bestanden aus nur einem Satz, 4 weitere aus 2 Sätzen. Das Satzende und damit die zu klassifizierende Einheit wurde durch eine sehr einfache Methode, nämlich durch eine Pausenschwelle von 600 ms bestimmt. Die 4 zweisätzigen Turns wurden korrekt zerlegt, 4 weitere wurden fälschlicherweise in zwei „Sätze“ zerlegt[1]; in diesen Fällen wurde dem ersten Satzteil der Modus des gesamten Satzes zugewiesen. Damit ergeben sich 142 Sätze.

Unterschieden wurden die Satzmodi AUSSAGE, W-FRAGE und E-FRAGE. W-FRAGEN, auch Ergänzungsfragen genannt, beginnen mit einem Fragepronomen, z.B. mit *„Wann“*; als Antwort wird in diesem Fall eine Zeitangabe erwartet. Entscheidungsfragen oder E-FRAGEN können dagegen mit *„ja“* oder *„nein“* beantwortet werden (vgl. Abschnitt 1.2.4). Der Satzmodus wurde aus der Orthographie bestimmt, was problemlos möglich war, da es sich im Gegensatz zu den spontanen Äußerungen des Verbmobilkorpus durchweg um vollständige, syntaktisch korrekte Sätze handelte.

[1]Nach Aussage der Projektpartner waren Einfügungsfehler weniger kritisch als Auslassungsfehler.

E-Fragen haben in der Regel am Ende einen steigenden Grundfrequenzverlauf, Aussagen einen fallenden. Bei W-Fragen kann der Ton am Ende ansteigen, wenn auch nicht so stark wie bei E-Fragen, oft fällt er aber genauso stark ab wie bei Aussagen. Deshalb sind W-Fragen am schwierigsten zu erkennen. Manchmal ist ein starker Anstieg auf dem Fragepronomen zu beobachten, deshalb ist der Grundfrequenzverlauf zu Beginn des Satzes ein guter Indikator.

6.1.1 Fujisaki-Merkmale

Bei Erreichen des Satzendes werden einige Parameter des Fujisaki-Modells (siehe Abschnitt 5.2.2) als Merkmale herangezogen: Untersucht wurden zunächst Amplitude und Dauer des ersten, vorletzten und letzten Akzentkommandos (in allen Sätzen ergaben sich mindestens zwei Akzentkommandos), Amplitude des Phrasenkommandos sowie die zeitlichen Abstände der Kommandos untereinander und zum Satzende. Mit Hilfe der linearen Diskriminanzanalyse wurden dann die sechs besten Merkmale ausgewählt:

- Dauer des ersten und letzten Akzentkommandos
- Zeit vom Ende des letzten Akzentkommandos zum Äußerungsende
- Amplitude des Phrasenkommandos und des letzten Akzentkommandos
- Dauer der Äußerung

Mit Lern- gleich Teststichprobe, vollbesetzter Kovarianzmatrix und Klassifikation mit der Maximum–Likelihood–Regel ergibt sich eine mittlere Erkennungsrate von 60.6 %, siehe Tabelle 6.1.

		klassifiziert als		
	Anzahl	Aussage	W-Frage	E-Frage
AUSSAGE	88	55.7 %	31.8 %	12.5 %
W-FRAGE	24	20.0 %	57.7 %	23.3 %
E-FRAGE	30	8.3 %	8.3 %	83.3 %

Tabelle 6.1: *Verwechslungsmatrix des Klassifikators zur Satzmodusbestimmung nur mit Fujisaki-Merkmalen bei Klassifikation mit* **Maximum–Likelihood–Regel**. *Erkennungsrate ist 60.1 %, mittlere Erkennungsrate ist 60.6 %.*

6.1.2 Filterbankmerkmale

Zur Gewinnung der Filterbankmerkmale wird die interpolierte Grundfrequenzkontur durch die in Abschnitt 5.2.3 beschriebene Filterbank in zwei Komponenten zerlegt. Die erste Komponente gibt die globale Tonbewegung wieder, die zweite das Steigen und Fallen der Tonhöhe im Zeitbereich von Silben. Zusätzlich wird noch die Ableitung der zweiten Komponente berechnet; damit erhält man neben der Amplitude - der relativen Tonhöhe zu einem Zeitpunkt - auch Information über die Tendenz zu diesem Zeitpunkt.

Die Parametrisierung der F_0–Kontur erfolgt durch Mittelung dieser drei Signale in stimmhaften Bereichen[2], d.h. für jeden stimmhaften Bereich ergeben sich drei Mittelwerte. Am Satzende werden aus den Mittelwerten des ersten, vorletzten und letzten stimmhaften Bereichs sechs der neun Werte als Merkmale ausgewählt.

Mit Lern- gleich Teststichprobe, vollbesetzter Kovarianzmatrix und Klassifikation mit der Maximum–Likelihood–Regel ergibt sich eine mittlere Erkennungsrate von 81.1 %, siehe Tabelle 6.2.

		klassifiziert als		
	Anzahl	Aussage	W-Frage	E-Frage
AUSSAGE	88	78.4 %	18.2 %	3.4 %
W-FRAGE	24	23.3 %	73.3 %	3.3 %
E-FRAGE	30	4.2 %	4.2 %	91.7 %

Tabelle 6.2: *Verwechslungsmatrix des Klassifikators zur Satzmodusbestimmung nur mit Filterbankmerkmalen bei Klassifikation mit* **Maximum–Likelihood–Regel**. *Erkennungsrate ist 79.6 %, mittlere Erkennungsrate ist 81.1 %.*

[2] Bei der in [Ott93] beschrieben Satzmodusklassifikation wird zur Merkmalsextraktion eine Regressionsgerade durch die gesamte Äußerung und durch den letzten stimmhaften Bereich gelegt. Problematisch an diesem Ansatz ist die Empfindlichkeit gegenüber Stimmhaft–Stimmlos–Fehlern: Wenn z.B. der letzte stimmhafte Bereich durch einen Fehler in zwei Bereiche aufgespalten wird, kann das auf die Lage der Regressionsgeraden durch den letzten stimmhaften Bereich erheblichen Einfluß haben. Die hier verwendeten Amplituden der Bandpaßausgaben und deren Ableitungen sind unabhängig von der Stimmhaft–Stimmlos–Entscheidung, sie kommt erst bei der anschließenden Mittelung zum Tragen, wobei sich Fehler offensichtlich weniger stark auswirken.

6.1.3 Gesamtergebnis

In einem weiteren Klassifikationsexperiment zum Satzmodus wurden die 6 Fujisaki-Merkmale und die 6 Filterbankmerkmale zu einem Merkmalvektor zusammengefaßt. Zunächst wurde als Teststichprobe wieder die Lernstichprobe genommen und die Kovarianzmatrix voll besetzt.

Bei der Klassifikation mit der Bayes-Regel ergibt sich eine Erkennungsrate von 88.0 % (siehe Tabelle 6.4), mit der Maximum–Likelihood–Regel eine mittlere Erkennungsrate von 87.1 % (siehe Tabelle 6.3).

Vergleicht man diesen Wert mit der mittleren Erkennungsrate aufgrund der Fujisaki-Merkmale allein (61 %) und der Filterbank-Merkmale allein (81 %), kann man schließen, daß die sich die Parameter des Fujisaki-Modells auch indirekt, als Merkmale verwendet, nicht gut zur Beschreibung der deutschen Intonation eignen.

		klassifiziert als		
	Anzahl	Aussage	W-Frage	E-Frage
AUSSAGE	88	86.4 %	10.2 %	3.4 %
W-FRAGE	24	16.7 %	83.3 %	0.0 %
E-FRAGE	30	4.2 %	4.2 %	91.7 %

Tabelle 6.3: *Verwechslungsmatrix des Klassifikators zur Satzmodusbestimmung bei Klassifikation mit* **Maximum–Likelihood–Regel**. *Erkennungsrate ist 86.6 %, mittlere Erkennungsrate ist 87.1 %.*

		klassifiziert als		
	Anzahl	Aussage	W-Frage	E-Frage
AUSSAGE	88	90.9	5.7	3.4
W-FRAGE	24	23.3	76.7	0.0
E-FRAGE	30	8.3	0.0	91.7

Tabelle 6.4: *Verwechslungsmatrix des Klassifikators zur Satzmodusbestimmung bei Klassifikation mit* **Bayes–Regel**. *Erkennungsrate ist 88.0 %, mittlere Erkennungsrate ist 86.4 %.*

Wegen des Experimentiermodus Lern- gleich Teststichprobe sind diese Erkennungsraten sicherlich zu optimistisch. Wegen der geringen Stichprobengröße stellt sich der Klassifikator zu stark auf die Stichprobe ein. Die unabhängig von der Stichprobe existierenden Verteilungsdichten der Merkmale können nicht zuverlässig geschätzt werden: Als Faustformel gilt, daß die Stichprobe mindestens zehn mal größer als die Anzahl der zu trainierenden Parameter sein sollte. Das sind bei 12 Merkmalen und voller Kovarianzmatrix 12^2 Parameter plus 12 Mittelwerte, mal 3 Klassen, insgesamt 468 Parameter. Damit ist die Stichprobe um den Faktor 33 zu klein.

Nimmt man die statistische Unabhängigkeit der Merkmale voneinander an, wird die Kovarianzmatrix zu einer Diagonalmatrix. Dann sind nur noch 720 Parameter zu trainieren, und die Stichprobe ist nur noch um den Faktor 5 zu klein. Die resultierende Erkennungsrate ist dann 75.4 % (wieder mit ML–Regel). Allerdings sind die Merkmale teilweise bewußt so gewählt, daß Abhängigkeiten bestehen.

Eine andere Möglichkeit, den Einfluß der Stichprobengröße zu untersuchen, ist die Methode des „leave one out“: Dabei wird beim Trainieren ein Satz ausgelassen, und mit diesem dann getestet. Das wird für alle Sätze wiederholt. Dabei ergab sich hier eine Erkennungsrate von 68.3 %. Dieser Wert ist allerdings zu pessimistisch, da sich bei diesem Verfahren die zu geringe Stichprobengröße nachteilig auswirkt.

6.2 Akzentuierung

Das INTARC-1.3-System enthielt zwei alternative Worterkenner: einen HMM-Worterkenner und den linguistischen Worterkenner BELLE [ACBD+95]. BELLE erstellt aus akustisch-phonetischen Ereignissen eine Silben-Lattice[3], wobei phonotaktische Regeln zur Anwendung kommen. Anschließend wird die Silben-Lattice in eine Morphem-Lattice umgewandelt, aus der schließlich die Worthypothesen gebildet wurden. Bei der Umwandlung der Silben- in die Morphem-Lattice durch den *morphologisch-prosodischen Parser* wurden die prosodisch detektierten Akzente eingesetzt. Allerdings sind die Arbeiten an BELLE vorzeitig eingestellt worden, so daß es keine Ergebnisse darüber vorliegen, inwieweit die Akzente die Worterkennung verbessert haben.

Die in den Abschnitten 5.2.3 und 5.3 beschriebenen Basismerkmale, die für jeden Frame berechnet werden, bildeten die Grundlage der Akzentklassifikation. Die in Abschnitt 5.5 eingeführten komplexen Merkmale, die aufgrund der Silbenkerndetektion aus den Basismerkmalen zusammengesetzt werden, kommen erst bei der Phrasengrenzen- und Satzmodusdetektion zum Einsatz. Die Akzentdetektion arbeitet also unabhängig von der Silbenkerndetektion, da zunächst jeder Frame klassifiziert wird. Zusammenhängende Frames, die zur gleichen

[3]Die Silben-Lattice ist einem Worthypothesengraph vergleichbar, nur sind hier Silben die Einheiten.

Klasse gehören, werden anschließend zu einer Akzenthypothese mit Anfangszeit, Endzeit und einem Konfidenzmaß zusammengefaßt.

Die Akzentetiketten beziehen sich auf Silben. Da sich die Ausprägung der Energiemerkmale innerhalb von Silben stark ändert, vor allem an den Vokalgrenzen, werden alle Akzentetiketten dem Vokal der Silbe zugeordnet, und eine zusätzliche Klasse NICHT-VOKALE wird eingeführt. Die so entstehenden Klassen werden dann auf die beiden Oberklassen „Vokal in einer akzentuierten Silbe ja/nein" abgebildet, kurz: „akzentuierter Vokal ja/nein". Abschnitt 6.2.2 wird noch einmal auf diese Abbildung zurückkommen.

Die Ergebnisse der frameweisen Klassifikation sind hier nicht von Belang, da es nicht darauf ankam, gleichzeitig die genauen Vokalgrenzen zu detektieren. Die Auswertung erfolgte vielmehr silbenweise: Wenn in einer nicht-akzentuierten Silbe *kein* Frame als akzentuierter Vokal klassifiziert wurde, ist diese Silbe als richtig klassifiziert gewertet worden. Wenn in einer akzentuierten Silbe *mindestens ein* Frame als akzentuierter Vokal klassifiziert wurde, ebenfalls.

Zusammenhängende Frames der Oberklasse „akzentuierter Vokal" wurden zu Akzenthypothesen zusammengefaßt. Die so entstehenden „akzentuierten Bereiche" wurden hinsichtlich ihrer Dauer und ihrer Lage zu den Silbengrenzen untersucht. Dabei stellte sich heraus, daß die silbenweise ermittelte Erkennungsrate noch verbessert werden kann, wenn akzentuierte Bereiche kürzer als k Frames wieder auf nicht-akzentuiert gesetzt werden. Für den Sprecher SAT der Phondat-Stichprobe liegt der optimale Wert für k bei 11, für die Verbmobil-Stichprobe bei 9 (was bei 8 ms bzw. 10 ms Fortschaltintervall 88 ms bzw. 90 ms entspricht).

Da die Vokalgrenzen oft mit den Silbengrenzen zusammenfallen, hätte eine Verschmälerung der akzentuierten Bereiche mit dem oben genannten Auswertungsverfahren zu einer weiteren Verbesserung geführt. Darauf wurde jedoch verzichtet, um dem Empfänger der Akzenthypothesen (dem linguistischen Worterkenner BELLE) die Dauerinformation nicht vorzuenthalten.

6.2.1 Akzenterkennung in der Phondat-Stichprobe

Die erste Auswertung des oben beschriebenen Akzentdetektors erfolgte mit den 60 Sätzen des Sprechers SAT aus dem Phondat-II-Zugauskunftskorpus, für die sowohl prosodische Etiketten als auch eine manuelle Phonemsegmentierung vorhanden war, siehe Abschnitt 4.1.1. Die 60 Sätze bestanden aus 879 Silben bzw. 29288 Frames (8 ms).

Die Bonner prosodischen Etiketten unterscheiden nicht-akzentuierte Silben von akzentuierten mit stark/schwach steigender/fallender Intonation, also 5 Klassen. Die Braunschweiger prosodischen Etiketten unterscheiden 3 Akzentstufen (EMPHASE/KONTRAST wurde in der Phondat-Stichprobe noch nicht vergeben), so daß sich 3 Klassen ergeben.

			automatisch klassifiziert als		Klasse gem. TUBS Labels		Klasse gem. IKP Labels	
		Anz.	akz.	n.-akz.	akz.	n.-akz.	akz.	n.-akz.
TUBS Labels	akzentuiert	242	64.0	36.0			98.3	1.7
	nicht-akz.	637	10.3	89.7			17.9	82.1
	RR/avRR		83.0 / 76.9				86.6 / 90.2	
IKP Labels	akzentuiert	340	51.5	48.5	59.7	40.3		
	nicht-akz.	539	10.9	89.1	0.7	99.3		
	RR/avRR		74.5 / 70.3		84.0 / 79.5			

Tabelle 6.5: *Vier Verwechslungsmatrizen zur Akzentklassifikation für den Sprecher SAT des Phondat-Korpus (Lern- gleich Teststichprobe): Die zwei linken Verwechslungsmatrizen beziehen sich auf den Akzenterkenner, wobei zum Training und Testen einmal die Braunschweiger* (TUBS) *und zum zweiten die Bonner* (IKP) *prosodischen Etiketten verwendet wurden. Die anderen beiden Verwechslungsmatrizen zeigen den Grad der Übereinstimmung der prosodischen Etiketten untereinander.* RR *bezeichnet die Erkennungsrate,* avRR *die mittlere Erkennungsrate. Alle Angaben (außer in Spalte* Anz.*) sind in Prozent.*

Es wurden zu jedem Frame 3 Energie- und 6 Grundfrequenzmerkmale berechnet. Die 3 bzw. 5 Akzentetiketten wurden auf die Vokale abgebildet, eine weitere Klasse für NICHT-VOKAL eingeführt. Damit wurde jeweils ein Normalverteilungsklassifikator dimensioniert. Da es sich um einen Vorversuch handelte, wurden zum Testen die gleichen Daten verwendet. Nach der Klassifikation wurden die 4 bzw. 6 Klassen auf die beiden Oberklassen „Vokal in akzentuierter Silbe ja/nein" abgebildet, und akzentuierte Bereiche kürzer 11 Frames auf nicht-akzentuiert zurückgesetzt. Tabelle 6.5 zeigt das Ergebnis bei silbenweiser Auswertung. Darin sind auch die Bonner Etiketten, bei denen 39 % der Silben als akzentuiert eingestuft wurden, den Braunschweiger Etiketten gegenübergestellt, bei denen nur 28 % der Silben als akzentuiert gelten.

Werden für Training und Test die Bonner Etiketten zugrundegelegt, ist die Erkennungsrate niedriger als bei Zugrundelegung der Braunschweiger Etiketten. Dies mag zunächst verwundern, weil die feinere Etikettierung, die neben der Akzentstufe auch den Verlauf der Grundfrequenz berücksichtigt, auf eine explizite Clusterung hinausläuft. Nach der Verwechslungsmatrix für das 5-Klassen-Problem (hier nicht gezeigt) werden jedoch oft steigende mit fallenden Akzenten verwechselt; nach Auskunft der Etikettiererin wurde der Intonationsverlauf perzeptiv etikettiert, und eine visuelle Inspektion ergab schon an wenigen Beispielen, daß der Intonationsverlauf beschränkt auf den Silbenkern oft ein anderes Bild ergibt.

In einem weiteren Experiment wurden die Bonner Handetiketten dem tatsächlichen Grundfrequenzverlauf angepaßt: „steigend“ wurde in „fallend“ geändert, wenn die Grundfrequenz im Silbenkern fallend war, und umgekehrt; das Urteil „schwach“ oder „stark“ (fallend oder steigend) wurde aber beibehalten, weil es die Stärke des wahrgenommenen Akzents ausdrückt und nicht nur von der Steilheit des Grundfrequenzverlaufs abhängt, sondern auch von der Lautheit und der Dehnung der Silbe. Die frameweise Auswertung zeigte, daß zwar weniger Verwechslungen innerhalb der Klasse „akzentuiert“ auftraten, dafür mehr „nicht-akzentuierte Frames“ als „akzentuiert“ klassifiziert wurden, so daß sich die Akzentuiert/Nicht-Akzentuiert-Klassifikation insgesamt von 82.7 % auf 68.6 % verschlechterte.

Die Clusterung der Oberklassen „akzentuiert“ und „nicht-akzentuiert“ mit dem K-Means-Algorithmus in 4 bzw. 2 Unterklassen brachte ebenfalls eine Verschlechterung der Akzentuiert/Nicht-Akzentuiert-Klassifikation auf 65.6 % (wieder für Frames). Dieser überraschende Umstand konnte aus Zeitgründen nicht weiter untersucht werden.

6.2.2 Akzenterkennung in der Verbmobil-Stichprobe

Der Akzenterkenner für die Verbmobil-Stichprobe ist, von den Trainingsdaten abgesehen, identisch mit dem für die Phondat-Stichprobe. Ein Normalverteilungsklassifikator bildet für jeden Frame den Merkmalvektor in eine von fünf Unterklassen ab (Vokal mit einer der vier Akzentstufen sowie die Klassen NICHT-VOKAL), die anschließend auf die zwei Oberklassen „akzentuierter Vokal ja/nein“ abgebildet werden. Schließlich werden benachbarte Frames der Klasse „akzentuierter Vokal“ zu Akzenthypothesen zusammengefaßt, wenn sie die Mindestlänge akzentuierter Bereiche nicht unterschreiten.

Die Abbildung der Unter- in die Oberklassen soll jetzt näher betrachtet werden. In Abschnitt 3.3.1 wurde gesagt, daß die Klassifikation mit der Bayes- oder der ML-Regel erfolgen kann, je nachdem, ob man die Erkennungsrate oder die mittlere Erkennungsrate maximieren möchte. Bei der ML-Regel spielen die a priori Wahrscheinlichkeiten der Klassen keine Rolle, während die Bayes-Regel sich eher für die häufigere Klasse entscheidet. Im Falle des Zwei-Klassen-Problems ist das die Klasse „kein akzentuierter Vokal“.

Die Braunschweiger Etiketten der Verbmobil-Stichprobe stufen nur etwa jede vierte Silbe als akzentuiert ein. Diese Etiketten sollen die *Fokus*struktur beschreiben, die auf semantischer oder Dialogebene relevant ist. Für den linguistischen Worterkenner BELLE ist jedoch eher der prosodische *Wort*akzent, den weit mehr Silben tragen, relevant. Insofern sind die Akzentetiketten für Verbmobil bzw. der damit trainierte Akzenterkenner gar nicht geeignet zur Unterstützung eines Worterkenners.

Geht man jedoch davon aus, daß sich Wortakzent und Fokusakzent nur in ihrer *Prominenz* unterscheiden, kann man einen Akzenterkenner auch mit den selteneren Fokusakzenten trainieren und den Klassifikator anschließend durch einen Gewichtungsfaktor dazu bringen, sich eher für die Klasse „akzentuiert" zu entscheiden.

Strittig ist nach wie vor die Frage, ob Akzentuierung kategorial oder graduell ist. Die Prosodiesteuerung des Bonner Sprachsynthesesystems [PH97] modelliert Akzentuierung als Silbenprominenz. Auch der Vergleich der Bonner Akzentetiketten, die rein perzeptiv vergeben wurden, mit den Braunschweiger Etiketten, die sich neben der Perzeption auch auf die linguistische Funktion (den Fokus) beziehen, läßt dies vermuten: Tabelle 6.5 zeigt, daß nur 1.7 % der TUBS-Akzente keine IKP-Akzente sind, und nur 0.7 % der IKP-Nicht-Akzente sind TUBS-Akzente.

Da für die Verbmobil-Stichprobe nur die Braunschweiger Etiketten vorhanden waren, blieb auch gar nichts anderes übrig, als mit der Prominenzhypothese zu arbeiten. Mit den Entwicklern von BELLE wurde vereinbart, den Akzenterkenner auf die mittlere Erkennungsrate hin zu optimieren[4].

Bei der frameweisen Klassifikation mit der ML-Regel wird zwar die mittlere Erkennungsrate für das 5-Klassen-Problem maximiert, nicht jedoch die mittlere Erkennungsrate hinsichtlich der zwei Oberklassen. Deshalb erfolgte die Abbildung auf die zwei Oberklassen durch Summierung der a posteriori Wahrscheinlichkeiten über jede Oberklasse, Gewichtung einer der beiden Summen und Entscheidung für die Klasse mit dem größeren Wert.

Der Gewichtungsfaktor wurde zusammen mit der Schwelle für die Mindestlänge eines akzentuierten Bereichs durch ein Suchverfahren (Koordinatenabstieg mit dynamisch angepaßten Schrittweiten) so eingestellt, daß bei silbenweiser Auswertung der Akzentklassifikation die mittlere Erkennungsrate maximal wurde.

Tabelle 6.6 zeigt das Ergebnis für die Teststichprobe bei silbenweiser Auswer-

[4] Wäre zu diesem Zeitpunkt der Silbenkerndetektor fertiggestellt gewesen, hätte man – analog zur Phrasengrenzendetektion – für jede Silbe als Ganzes eine a posteriori Wahrscheinlichkeit für die Klasse „akzentuiert" berechnen und den Entwicklern von BELLE die Optimierung der Entscheidungsschwelle überlassen können. Bei dem hier beschriebenen Verfahren fallen dagegen zwei harte Entscheidungen: bei der Abbildung auf der Oberklasse „akzentuierter Vokal" und bei der Nachbearbeitung durch die Schwelle für die Mindestlänge eines akzentuierten Bereichs. Die Arbeiten an BELLE wurden jedoch vorzeitig eingestellt.

	klassifiziert als		
	akzentuiert	nicht-akz.	rel. Häuf.
akzentuiert	66.5	33.5	25.4
nicht-akz.	23.5	76.5	74.6

Tabelle 6.6: *Verwechslungsmatrix des Klassifikators zur Akzentdetektion der Verbmobil-Stichprobe. Die Erkennungsrate beträgt 74.0 %, die mittlere Erkennungsrate 71.5 %.*

tung (die frameweise Erkennungsrate bei 5 Klassen hätte mit der Bayes-Regel 72.2 % betragen, mit der ML-Regel aber nur 48.9 %). Die Erkennungsrate ist zwar gerade so groß wie die relative Häufigkeit der nicht-akzentuierten Silben, was sich auch mit einem trivialen „Klassifikator" hätte erreichen lassen. Es galt jedoch, die mittlere Erkennungsrate zu maximieren, und diese liegt deutlich über dem Zufallsniveau.

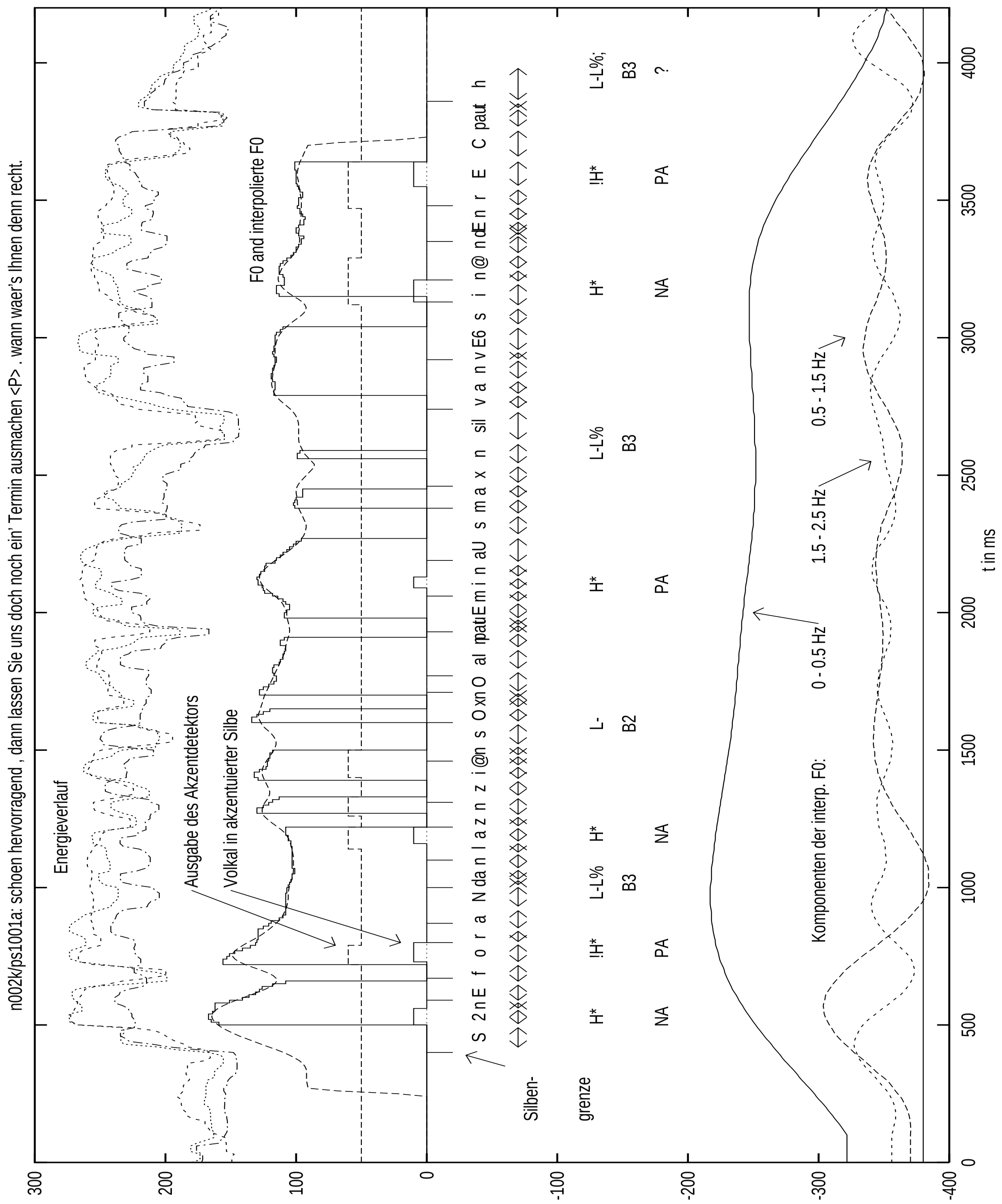

Abbildung 6.1: *Übersicht zur Akzentdetektion: Ein Beispielsatz mit allen Etiketten, den daraus abgeleiteten Referenz-Akzent-Etiketten, einigen Merkmalen (3 Energiemerkmale, interpolierte Grundfrequenz und ihre Komponenten) und der Detektor-Ausgabe.*

6.3 Phrasierung und Satzmodus

Bei den Phrasengrenzen werden die Typen **B0, B2, B3** und **B9** unterschieden, siehe Abschnitt 4.1.2. Die B3-Grenzen können optional ein funktionales Etikett für „Frage" tragen (`FUN: ?`). Ursprünglich war auch ein funktionales Etikett für „Aussage" vorgesehen (`FUN: .`), dies wurde jedoch nicht verwirklicht. Deshalb wird, einem Vorschlag von Anton Batliner folgend, der Satzmodus an den B3-Grenzen, die kein funktionales Etikett tragen, aus dem Grenzton abgeleitet:

- Trägt eine B3-Grenze das funktionale Etikett `FUN: ?`, dann handelt es sich um eine Frage.
- Trägt eine B3-Grenze kein funktionales Etikett und ist der Grenzton tief (`L%`), dann handelt es sich um eine eine abgeschlossene Nicht-Frage, eine Aussage.
- Trägt eine B3-Grenze kein funktionales Etikett und ist der Grenzton hoch (`H%`), dann ist die Grenze weiterführend.

In dieser Definition wird die Unterscheidung zwischen „weiterführend" und „abgeschlossen" also intonatorisch getroffen und nicht nach syntaktischen (oder semantischen) Kriterien. Allerdings sind die hier definierten „Aussagen" meist auch Aussagen im syntaktischen Sinn; was circa einmal pro Dialog vorkommt, ist z.B. eine „Weiterführung" am Turnende:

```
n004k/nak3k005a:ja ich w"urd' dann sagen, da"s wir uns mehr in
der Mitte der Woche treffen,
```

wobei „Weiterführung" als ',' (Komma) dargestellt ist.

In die Merkmale zur Grenzdetektion geht mehr Kontext aus dem Grundfrequenzverlauf ein als bei der Akzentdetektion: In den meisten Fällen enthält das Analysefenster vier Silben, siehe Abschnitt 5.5 auf Seite 73. Daher bietet es sich an, das Ton-Etikett, das ebenfalls den Tonverlauf beschreibt, in die Klassendefinitionen aufzunehmen. Zusammen mit dem optionalen Etikett für „Frage" an dem B3-Grenzen gibt es 13 zulässige Kombinationen von Phrasengrenzen- und Ton-Etiketten[5], die in Tabelle 6.7 zu sehen sind. Jede dieser 13 Klassen kann eindeutig einem Phrasengrenzentyp (Spalte **PG**) und zugleich einer Satzmodusklasse (Spalte **Modus**) zugeordnet werden; somit basieren die Detektoren für Phrasengrenzen und Satzmodus auf demselben Klassifikator.

[5] Im Dialog `g071a` tritt viermal der Grenzton `!H-H%` auf; dies war in den Dialogen `n001k` bis `n019k` nirgends der Fall, deshalb ist dieser Ton hier nicht berücksichtigt.

Nr.	PG	Ton	Fun.	Modus
0	B0			Nicht-B3
1	B2	H%		Nicht-B3
2	B2	L%		Nicht-B3
3	B3	H-H%		weiterführend
4	B3	H-H%	?	Frage
5	B3	H-L%		Aussage
6	B3	H-L%	?	Frage
7	B3	L-L%		Aussage
8	B3	L-L%	?	Frage
9	B3	L-H%		weiterführend
10	B3	L-H%	?	Frage
11	B9	H%		Nicht-B3
12	B9	L%		Nicht-B3

Tabelle 6.7: *Die 13 Grenzklassen, bestehend aus den Kombinationen der prosodischen Etiketten für Phrasengrenzen (***PG***), für die Intonation gemäß ToBI (***Ton***) und das an den B3-Grenzen optional auftretende funktionale Etikett (***Fun.***) für* Frage. *Die Spalte* **Modus** *zeigt den aus* **Ton** *und* **Fun.** *abgeleiteten Satzmodus.*

Die Merkmalvektoren zum Training des Grenzdetektors beziehen sich auf SILBENGRENZEN statt auf Silbengrenzen, siehe Abschnitt 5.5. Bei korrekten Entscheidungen des Silbenkerndetektors überspannt eine SILBENGRENZE genau eine Silbengrenze, und der Merkmalvektor bekommt dann das Etikett dieser Silbengrenze: B2, B3, B9 oder, falls keines davon vorhanden ist, B0.

Im Fall eines Einfügungsfehlers entsteht eine SILBENGRENZE ohne eine entsprechende Silbengrenze. Der zugehörige Merkmalvektor wird dann ein B0-Etikett erhalten. Dies hat auf die Erkennungsrate des Grenzdetektors insofern einen Einfluß, als die Häufigkeit der Klasse B0 künstlich erhöht wird.

Im Fall eines Auslassungsfehlers überspannt eine SILBENGRENZE zwei Silbengrenzen. Der zugehörige Merkmalvektor erhält in diesem Fall das Etikett der ersten, d.h. der am weitesten links liegenden Silbengrenze. Falls der Klassifikator für diesen Merkmalvektor das gleiche Etikett liefert, zählt diese SILBENGRENZE als richtig klassifiziert, aber daß die zweite Silbengrenze nicht klassifiziert wird, zählt bei SILBEN-weiser Auswertung nicht als Fehler.

Kurz gesagt, bei SILBEN-weiser Auswertung zählen die Fehler der Silbenkerndetektion nicht mit. Tabelle 6.8 zeigt das dadurch entstehende Ergebnis.

	klassifiziert als				r.H.		klassifiziert als				r.H.
	B0	B2	B3	B9			0	W	A	F	
B0	87.5	2.2	8.4	1.9	76.8	0	91.4	1.9	6.1	0.6	84.1
B2	56.8	32.9	7.3	3.0	4.2	W	53.0	30.9	7.7	8.5	4.6
B3	26.0	1.9	67.3	4.7	15.9	A	30.2	1.3	65.4	3.1	8.9
B9	34.8	1.2	16.4	47.6	3.1	F	2.1	5.3	33.0	59.6	2.4

Tabelle 6.8: *Verwechslungsmatrizen zur Phrasengrenzen- und Satzmodusklassifikation im* INTARC-*1.3-System bei* SILBEN-*weiser Auswertung (alle Angaben in Prozent). Die Erkennungsrate für Phrasengrenzen und Satzmodus beträgt 80.76 bzw. 85.50 %, die mittlere Erkennungsrate 58.84 bzw. 61.90 %.* **0** *bezeichnet Nicht-***B3**, **W** *Weiterführung,* **A** *Aussage und* **F** *Frage. In den Spalten r.H. sind die relativen Häufigkeiten der Klassen angegeben.*

	klassifiziert als				rel. Häufigk.
	Nicht-B3	weiterf.	Aussage	Frage	
Nicht-B3	94.6	1.7	3.3	0.4	78.2
weiterf.	64.2	26.1	5.5	4.2	6.3
Aussage	56.7	2.9	33.9	6.5	12.4
Frage	23.0	6.6	7.9	62.5	3.1

Tabelle 6.9: *Verwechslungsmatrix des Klassifikators zur Satzmodusdetektion im* INTARC-*1.3-System bei* wortweiser *Auswertung (alle Angaben in Prozent). Erkennungsrate ist 81.78 %, mittlere Erkennungsrate ist 54.28 %.*

Bei wortweiser Auswertung werden die klassifizierten SILBENgrenzen wieder den Wortgrenzen zugeordnet, hier zählen die Fehler der Silbenkerndetektion also mit. Tabelle 6.9 gibt das entsprechende Ergebnis für den Satzmodus wieder. Der Vergleich mit der Verwechslungsmatrix für den Satzmodus in Tabelle 6.8 zeigt, daß die Klasse Nicht-**B3** seltener ist, weil es weniger Wortgrenzen als Silbengrenzen gibt. Umgerechnet auf das Zweiklassenproblem **B3**/Nicht-**B3** ergibt sich eine Erkennungsrate von 84.0 % und eine mittlere Erkennungsrate von 70.2 %.

Die in [Kie97, Seite 191] genannt beste Erkennungsrate für das Zweiklassenproblem von 83.3 % bei einer mittleren Erkennungsrate von 86.6 % ist mit diesem Ergebnis nur bedingt vergleichbar. Zum einen wird in den dort verwendeten Merkmalen Wortinformation erfaßt in Form von normierten Dauermerkmalen und lexikalischen Merkmalen (z.B. ob eine Silbe betonbar ist). Zum anderen ist dort die Klasse **B3** mit einer a priori Wahrscheinlichkeit von 11.4 % nur etwa halb so häufig wie hier (21.8 %). Das liegt daran, daß dort die turnfinale **B3**-Grenze nicht in die Auswertung einbezogen wurde, weil sie bei nicht-inkrementeller Analyse schon anhand des END-OF-FILE „detektiert" werden kann (bis auf die sehr seltenen Fälle, in denen ein Turn mit einem Satzabbruch endet, also einer **B9**-Grenze). Bei inkrementeller Analyse ist ihre Detektion jedoch nicht trivial: Häufig übersieht der Silbenkerndetektor den turnfinalen Silbenkern; in diesem Fall wird auch die turnfinale Grenze nicht klassifiziert.

In Abschnitt 5.5 wurden verschiedene Fenstertypen zur Merkmalberechnung eingeführt: In den meisten Fällen betrachtet der Phrasengrenzendetektor ein Fenster aus vier Silben. An den Äußerungsgrenzen, die hier durch einen Abstand von mehr als 500 ms zwischen zwei Silbenkernen definiert sind, werden die kleineren Fenster **3a**, **3b** und **2b** verwendet; für die Ausnahmefälle ein- und zweisilbiger Äußerungen gibt es die Fenster **1** und **2**.

Da die Merkmale für jeden Fenstertyp unterschiedlich aufgebaut sind, mußte auch für jeden Fenstertyp ein eigener Normalverteilungsklassifikator trainiert werden. Die in den Tabellen 6.9 und 6.8 angegebenen Erkennungsraten sind daher ein zusammenfassendes Ergebnis.

Um die Anzahl der zu trainierenden Parameter an die Stichprobengröße anzupassen, wurden nicht alle der sich nach Gleichung 5.4 (Seite 74) ergebenden Merkmale verwendet. Vielmehr wurde für jeden Fenstertyp durch ein Merkmalauswahlsverfahren nach [Nie83] eine angemessene Teilmenge bestimmt.

Im Falle des 4er-Fensters ergeben sich bei 11 Basismerkmalen 51 Merkmale. Zunächst wurde dasjenige Merkmal bestimmt, mit dem alleine die höchste Erkennungsrate erreicht werden kann. Dieses wurde festgehalten, und unter den 50 verbleibenden Merkmalen dasjenige bestimmt, das in Kombination mit dem festgehaltenen die höchste Erkennungsrate bewirkt, usw.. Alternativ hätte man auch mit allen Merkmalen beginnen und eines nach dem anderen herausstreichen können. Es zeigte sich allerdings, daß ab einer gewissen Anzahl von Merkmalen die Erkennungsrate wieder sinkt, so daß beim Normalverteilungsklassifikator die Verwendung aller Merkmale nicht optimal ist.

Auf diese Weise ergab sich eine Rangliste der Relevanz der 51 Merkmale. Das Verfahren ist natürlich suboptimal, da die Merkmale nicht unabhängig voneinander sind. Es wäre jedoch zu aufwendig gewesen, für alle möglichen Teilmengen ein Klassifikationsexperiment durchzuführen. Stattdessen wurden zwei Durchläufe mit dem oben beschriebenen Verfahren begangen: Erst wurde mit der Bayes-Regel klassifiziert, wodurch die Erkennungsrate optimiert wurde, dann wurde mit der ML-Regel klassifiziert, womit die mittlere Erkennungsrate optimiert wurde. Im zweiten Durchgang wurden solche Merkmale früher ausgewählt, die zur Trennung der selteneren Klassen beitragen. Mithilfe beider Ranglisten wurden die 30 besten Merkmale ausgewählt.

Diese Prozedur wurde ebenso für die anderen Fenstertypen durchgeführt. Da sich die Stichprobe auf die Fenstertypen entsprechend der Häufigkeit ihrer Anwendung verteilt, ist die Stichprobengröße bei den selteneren Fenstertypen entsprechend kleiner, so daß sich schon bei weniger als 30 Merkmalen ein Sättigungseffekt einstellt oder die Erkennungsrate wieder verschlechtert. So wurden z.B. für die Fenstertypen **3a** und **3b** nur 24 bzw. 13 von 38 Merkmalen verwendet.

Gruppiert man die Merkmale nach Grundfrequenz-, Energie- und Dauermerkmalen, ergibt sich hinsichtlich der Relevanz kein eindeutiges Bild aus den Ranglisten; ein Ergebnis, das auch in [Kie97] gefunden wurde. Bemerkenswert ist aber die Tatsache, daß auch Dauermerkmale, die nicht auf Sprechgeschwindigkeit oder intrinsische Lautdauern normiert sind, zur Trennung der Klassen beitragen. So sind z.B. die Dauer des phrasenfinalen Silbenkerns und sein Abstand zum nächsten Silbenkern beim Fenster **2b** bereits unter den 8 wichtigsten von 25 Merkmalen.

Für das INTARC-2-System wurde der Phrasengrenzendetektor in zwei Punkten verbessert: Zum einen wurde er um einen siebten Fenstertyp **3c** erweitert, der in mehr als zweisilbigen Äußerungen (also der Mehrzahl der Äußerungen) den Fenstertyp **2b** ersetzt, siehe Abschnitt 5.5). Zur Klassifikation der Äußerungsenden kann daher ein Kontext von drei statt zwei Silben berücksichtigt werden, wodurch sich vor allem die Erkennungsrate für die **B3**-Grenzen erhöhte. Umgerechnet auf das Zweiklassenproblem **B3**/Nicht-**B3** ergibt sich damit eine Erkennungsrate von 91.6 % und eine mittlere Erkennungsrate von 80.4 %.

Den größeren Beitrag zur Steigerung der Erkennungsrate leistete jedoch eine akkuratere automatische Phonemsegmentierung, siehe Abschnitt 4.2. Die die Abbildung der prosodischen Handklassifikation von Wörtern auf Silben und anschließend auf SILBEN basiert auf der automatischen Phonemsegmentierung. Liegt die Phonemsegmentierung näher an der akustischen Realität, trifft dies auch auf die abgeleiteten Handetiketten zu. Sie sind damit konsistenter, bzw. weniger fehlerbehaftet, was schließlich durch die höhere Erkennungsrate in Tabelle 6.10 bestätigt wurde.

6.4 Dialogaktgrenzen

Dieser Abschnitt hat die Dialogaktsegmentierung mit den prosodisch detektierten **B3**-Grenzen zum Thema. Zunächst wird auf der Zweck der Dialogaktsegmentierung im Verbmobil-Prototyp eingegangen. Dann wird der Phrasengrenzenerkenner aus Abschnitt 6.3 anhand der zwei Klassen „Dialogaktgrenze ja/nein“ ausgewertet. Der Erlanger Phrasengrenzenerkenner (siehe Abschnitt 1.5) wurde mit denselben Daten ausgewertet, eine verbesserte Version später noch an anderen

	klassifiziert als				r.H.		klassifiziert als				r.H.
	B0	B2	B3	B9			0	W	A	F	
B0	90.8	1.9	6.1	1.3	84.3	0	94.0	2.4	3.6	0.0	91.3
B2	40.0	55.0	4.2	0.8	4.1	W	47.9	47.0	5.1	0.0	3.7
B3	35.1	0.6	63.7	0.6	8.7	A	27.1	1.0	71.8	0.0	4.7
B9	24.4	0.0	4.1	71.5	2.9	F	41.2	0.0	5.9	52.9	0.3

Tabelle 6.10: *Die Erkennungsrate für Phrasengrenzen bzw. Satzmodus im* Intarc-2-*System beträgt 86.38 bzw. 91.16 %, die mittlere Erkennungsrate beträgt 70.23 bzw. 66.45 %.*

Daten. Was die Auswertung der Erlanger **B3**-Grenzen betrifft, handelt es sich um Voruntersuchungen, da sie mithilfe eine Worthypothesengraphs erzeugt wurden, der in der hier beabsichtigten Anwendung nicht zur Verfügung steht. Trotzdem werden die Erkennungsraten des Bonner und des Erlanger Phrasengrenzenerkenners vergleichend gegenübergestellt.

Das Verbmobil-System wurde in der ersten Projektphase so konzipiert, daß es Terminverhandlungsdialoge zwischen einem japanischen und einem deutschen Dialogpartner unterstützt. Für beide Partner wurde eine zumindest passive Kenntnis des Englischen vorausgesetzt, so daß die Terminverhandlung größtenteils auf Englisch als gemeinsamer Sprache stattfinden kann. Nur wenn sich ein Partner nicht sicher ist, wird auf Anfrage eine Übersetzung erstellt. Demnach hat das System zwei Betriebsarten:

- die *tiefe Verarbeitung* muttersprachlicher (in der ersten Phase: deutscher) Äußerungen, die Worterkennung, syntaktische und semantische Analyse, Dialogverarbeitung, Transfer, Sprachgenerierung und Sprachsynthese beinhaltet, und

- die *flache Verarbeitung* englischer Beiträge, bei der die Dialogschritte aufgrund erkannter Schlüsselwörter *(keywords)* grob verfolgt werden.

Bei der flachen Verarbeitung ist statt eines Verbundworterkenners nur ein Schlüsselworterkenner aktiv; aufgrund seiner Ausgaben wird eine Dialogaktklassifikation vorgenommen und in das Dialoggedächtnis eingetragen, um Rückbezüge oder Mehrdeutigkeiten in späteren Dialogschritten auflösen zu können.

Es wurden zunächst 18 Dialogakte unterschieden, die auf das Szenario „Terminabsprache“ zugeschnitten sind [JKM+95]; sie umfassen u.a. Request (Aufforderung), Suggest (Vorschlag), Accept (Zustimmung) und Reject

(Ablehnung). Sie wurden später zu 42 Unterkategorien verfeinert, z.B. durch DATE (Termin), DURATION (Dauer) und LOCATION (Ort). Daneben gibt es Dialogakte für Begrüßung, Vorstellung, Dank, Abschied, usw..

Für jede Dialogaktklasse wurden ca. ein bis zwei Dutzend charakteristische Schlüsselwörter bestimmt [Mas95]; für REJECT sind das u.a. *nein*, *schlecht*, *leider*, *ausgebucht*, *Arzt*, *Urlaub*, *keine_Zeit* und *ganz_schlecht* (lange Wörter oder Wortfolgen werden leichter erkannt als kurze).

Schlüsselwörter für SUGGEST sind u.a. *vorschlagen*, *sagen_wir*, *wie_wär's*, *zur_Not*, sowie Wochentage und Zeitangaben. Für das Englische wurden ähnliche Wortlisten erstellt.

Das Problem liegt nun darin, daß ein Turn aus mehreren Dialogakten bestehen kann. Ohne Kenntnis der Dialogaktgrenzen, mit den Keywords allein, kann es zu Mehrdeutigkeiten kommen, wie bei

1. Nein, am Dienstag habe ich schon einen Termin

2. Nein. Aber wie wär's mit einem Dienstags-Termin?

wobei die Schlüsselwörter hier „nein“ und „Dienstag“ sind: (1) wäre ein REJECTDATE, (2) REJECTDATE gefolgt von SUGGESTDATE. Es wurde daher schon Anfang 1995 überlegt, die Grenze in (1) mithilfe der Prosodie zu detektieren. Der Erlanger Prosodieerkenner kam dafür nicht in Frage, da er lückenlose Worthypothesengraphen benötigt. Der in dieser Arbeit beschriebene Bonner Prosodieerkenner arbeitet dagegen unabhängig von der Worterkennung, was sich in dieser Situation als Stärke erweist.

Zunächst wurde in Zusammenarbeit mit der Erlanger Gruppe untersucht, wie gut die starken prosodischen Phrasengrenzen, die **B3**-Grenzen, zur Dialogaktsegmentierung geeignet sind. Da noch nicht genügend englische Daten zur Verfügung standen, wurden diese Voruntersuchungen mit deutschen Daten durchgeführt, mit 26 prosodisch etikettierten Dialogen, in denen die Dialogaktgrenzen manuell annotiert wurden. Da dies fast das gesamte prosodisch etikettierte Material darstellt, sind in diesen Daten auch die Trainingsdaten sowohl der Erlanger als auch des Bonner Prosodieerkenners enthalten.

Diese 26 Dialoge bestehen aus 698 Turns, sie enthalten 1446 Dialogakte (2.07 pro Turn) und 1979 handetikettierte **B3**-Grenzen. 90.5 % der Dialogaktgrenzen sind **B3**-Grenzen, und umgekehrt sind 66.1 % der **B3**-Grenzen Dialogaktgrenzen ($1308/(1308 + 671)$, siehe Tabelle 6.11).

Der Erlanger Prosodieerkenner wurde in verschiedenen Konfigurationen (auf sie soll nicht weiter eingegangen werden) mit dem Bonner Erkenner verglichen, wobei die Zeile „BN“ dem Phrasengrenzendetektor aus dem INTARC-1.3-System

	DAGs erkannt		eingefügt		Akkuratheit
	abs	in %	abs	in %	in %
Handlab.	1308	90.5	671	46.4	44.05
ER1	1231	85.1	2680	185.3	-100.21
ER2	1267	87.6	1674	115.8	-28.15
ER2(0.55)	1249	86.3	1692	117.0	-30.64
ER2(0.6)	1237	85.5	1356	93.8	-8.23
ER2(0.7)	1204	83.2	1063	73.5	9.75
BN	992	68.6	855	59.1	9.47
BN neu	980	67.7	519	35.9	31.88

Tabelle 6.11: *Erkennung der 1446 Dialogaktgrenzen anhand der prosodischen* **B3**-*Grenzen. Die erste Zeile gibt das Ergebnis für die prosodischen Handetiketten an, die weiteren für die automatischen Etiketten aus Erlangen (ER) und Bonn (BN). Die Akkuratheit berechnet sich als (erkannt-eingefügt)/1446; auch die übrigen Prozentzahlen beziehen sich auf die 1446 Dialogaktgrenzen.*

entspricht und die „BN neu“ dem verbesserten Phrasengrenzendetektor aus dem INTARC-2-System (vergleiche die Tabellen 6.8 und 6.10).

Das Ergebnis des Vergleichs zeigt Tabelle 6.11: Das Erlanger Modul erkennt zwar mehr Dialogaktgrenzen, fügt aber vergleichsweise mehr Grenzen ein, wie die Spalte „Akkuratheit“ zeigt.

		mit Gewicht 1		mit Gewicht 0.1	
		klassifiziert als			
	Anz.	DAG	¬DAG	DAG	¬DAG
DAG	1446	43.3	56.7	64.8	35.2
¬DAG	9302	6.4	93.6	17.8	82.2
RR / avRR		86.8 / 68.4		79.9 / 73.5	

Tabelle 6.12: *Erkennung der 1446 Dialogaktgrenzen (DAGs) anhand der mit dem Bonner Erkenner detektierten prosodischen* **B3**-*Grenzen, ausgewertet an 10741 Wortgrenzen, wobei die Klassifikation aufgrund der gewichteten a posteriori Wahrscheinlichkeit der Klasse* **B3** *erfolgte.*

Das Optimum hinsichtlich der Dialogakt*klassifikation* liegt vermutlich weder bei einer möglichst hohen Erkennungsrate noch bei einer möglichst hohen Akkuratheit. Nach Aussagen der Erlanger Kollegen stören Einfügungsfehler nicht so sehr wie Auslassungsfehler. Allerdings kann für das Bonner Modul die Entscheidungsschwelle beliebig verschoben werden. Tabelle 6.12 zeigt zwei Ergebnisse mit unterschiedlichen Gewichtungsfaktoren für die a posteriori Wahrscheinlichkeit der Klasse **B3**; die Ergebnisse sind diesmal auf Wortgrenzen bezogen.

Zu einem späteren Zeitpunkt führte die Erlanger Gruppe mit ihrem Erkenner eine weitere Untersuchung zur Dialogaktsegmentierung durch [MKH+96], die sich von der hier beschriebenen in folgenden Punkten unterscheidet:

- Die Auswertung erfolgte auf einer kleineren Teststichprobe, die die Trainingsstichprobe nicht mehr enthielt.

- Der Grenzenerkenner (ein Multi-Layer-Perzeptron, kurz MLP, siehe Abschnitt 3.3) wurde mit einem statistischen Sprachmodell (Language Model, kurz LM, siehe Abschnitt 7.2) kombiniert, das Wortsequenzen $w_{i-2}w_{i-1}w_iv_iw_{i+1}w_{i+2}$ betrachtet, wobei w_i das aktuelle Wort und v_i entweder DAG oder ¬DAG ist. Zum Trainieren des LMs werden als v_i die von Hand annotierten Dialogaktgrenzen verwendet, während des Tests können DAGs aufgrund eines Kontextes von ±2 Wörtern vorhergesagt werden. Die „Erkennungsrate“ des LMs allein ist mit 90.7 % schon höher als die des MLPs mit 84.4 % (diese Erkennungsrate basiert auf akustischen Merkmalen, inklusive Wortinformation).

- Das MLP wurde nicht mit **B3**-Grenzen, sondern mit DAGs trainiert. Die Erkennungsrate des MLPs allein verschlechterte sich zwar dadurch auf 83.6 %, die Erkennungsrate des MLPs in Kombination mit dem LM erhöhte sich aber von 91.3 % auf 92.2 %.

Bei den in [MKH+96] angegebenen Erkennungsraten sind die 453 turnfinalen Grenzen nicht berücksichtigt. Zum besseren Vergleich mit Tabelle 6.11 sind diese Zahlen hier so umgerechnet, daß die turnfinalen Grenzen mit eingehen.

Von 1115 DAGs werden 1016 (91.0 %) erkannt, von den übrigen 7317 Wortgrenzen 6819 (93.2 %) richtig klassifiziert. Die entspricht einer Erkennungsrate von 92.9 % und einer mittleren Erkennungsrate von 92.2 %.

Andererseits stehen den 1016 erkannten DAGs immer noch 498 eingefügte gegenüber, dies entspricht einer Akkuratheit von 46.5 %. Diese Verhältnis ist zwar besser als beim Bonner Modul, aber diese Verbesserung beruht hauptsächlich auf dem Einsatz des LMs.

Darüber hinaus darf nicht vergessen werden, das alle angegebenen Resultate des Erlanger Moduls darauf beruhen, daß zusätzliche Wortinformation vorliegt, die sowohl zur Verbesserung der akustischen Merkmale beiträgt als auch die Anwendung eines LMs erst möglich macht. Die Ergebnisse beruhen allerdings auf der *gesprochenen* Wortkette. Mit anderen Worten, sie gehen von der Annahme einer zu *100 % korrekten Worterkennung* aus und sind daher mit dem auf realen Daten gemessenen Bonner Ergebnis nicht vergleichbar.

Die Dialogaktsegmentierung anhand prosodischer Grenzen war in Verbmobil ursprünglich für englische Redebeiträge gedacht. Ein erster Test auf amerikanischen Daten (mit dem auf Karlsruher Dialogen trainierten Phrasengrenzendetektor zeigte, daß trotz des Übergangs auf eine andere Sprache immerhin noch knapp 60 % der DAGs erkannt werden, wobei sich die Anzahl der Einfügungen allerdings mehr als verdoppelt hat (die untersuchten amerikanischen Dialoge zeichneten sich durch abgehackte Sprechweise mit langen Pausen aus).

Dies läßt vermuten, daß die auf rein akustischer Basis mit NVK detektierten Phrasengrenzen in gewisser Weise „robuster" sind, als die, die mit MLP und Merkmalen detektiert werden, die auch linguistische Information enthalten: Die Erlanger Grenzdetektoren erzielen höhere Erkennungsraten durch stärkere Anpassung an die Problemstellung, dafür lassen sich die vom Bonner Erkenner gefundenen Grenzen ohne größere Einbußen eher auf verschiedene Arten interpretieren, z.B. als „prosodische" **B**-Grenzen, in denen auch syntaktische Vorerwartungen stecken, oder als Dialogaktgrenzen, die auf der auf der Ebene der Diskurssemantik liegen.

Kapitel 7

Einsatz der Phrasengrenzen im INTARC-System

Ziel der Entwicklung von Prosodiedetektoren innerhalb des Architektur-Teilprojektes in Verbmobil war es nicht in erster Linie, Erkennungsraten gemessen an Handetiketten zu maximieren, sondern die Performanz des INTARC-Systems zu verbessern. So hätten z.B. die Entwickler des Syntax-Parsers vom Prosodiemodul am liebsten syntaktische Phrasengrenzen empfangen, während der Phrasengrenzendetektor aber nur die im Sprachsignal prosodisch realisierten Phrasengrenzen finden kann (siehe dazu Kapitel 8); die durch die prosodischen Etiketten markierten Phrasengrenzen liegen wegen der Vorerwartungen der Etikettierer vermutlich irgendwo dazwischen.

In diesem Kapitel wird die Verwendung der akustisch detektierten prosodischen Phrasengrenzen im INTARC-System beschrieben. Abschnitt 7.1 gibt zunächst eine Übersicht über das System. Nach einem Abschnitt über statistische Sprachmodelle folgt in Abschnitt 7.3 eine Beschreibung des Syntax- und des Semantikparsers, die mit der gleichen Grammatik arbeiten. Grammatik und Parser können nur grob skizziert werden, da ihre detailierte Beschreibung den Rahmen dieser Arbeit sprengen würde; der Schwerpunkt liegt vielmehr auf der Kopplung mit dem Phrasengrenzendetektor. In einer Reihe von Experimenten wurde untersucht, inwieweit die detektierten prosodischen Phrasengrenzen das Syntaxparsing und die Semantikkonstruktion verbessern können. Die Ergebnisse sind, daß sich mit Hilfe der Prosodie im Syntaxparser die Worterkennungsrate erhöht und/oder die Anzahl der analysierten Kantenpaare (als Maß für die Laufzeit) abnimmt (Abschnitt 7.3.1), und daß im Semantikparser die Anzahl der Lesarten reduziert wird.

7.1 Systemübersicht

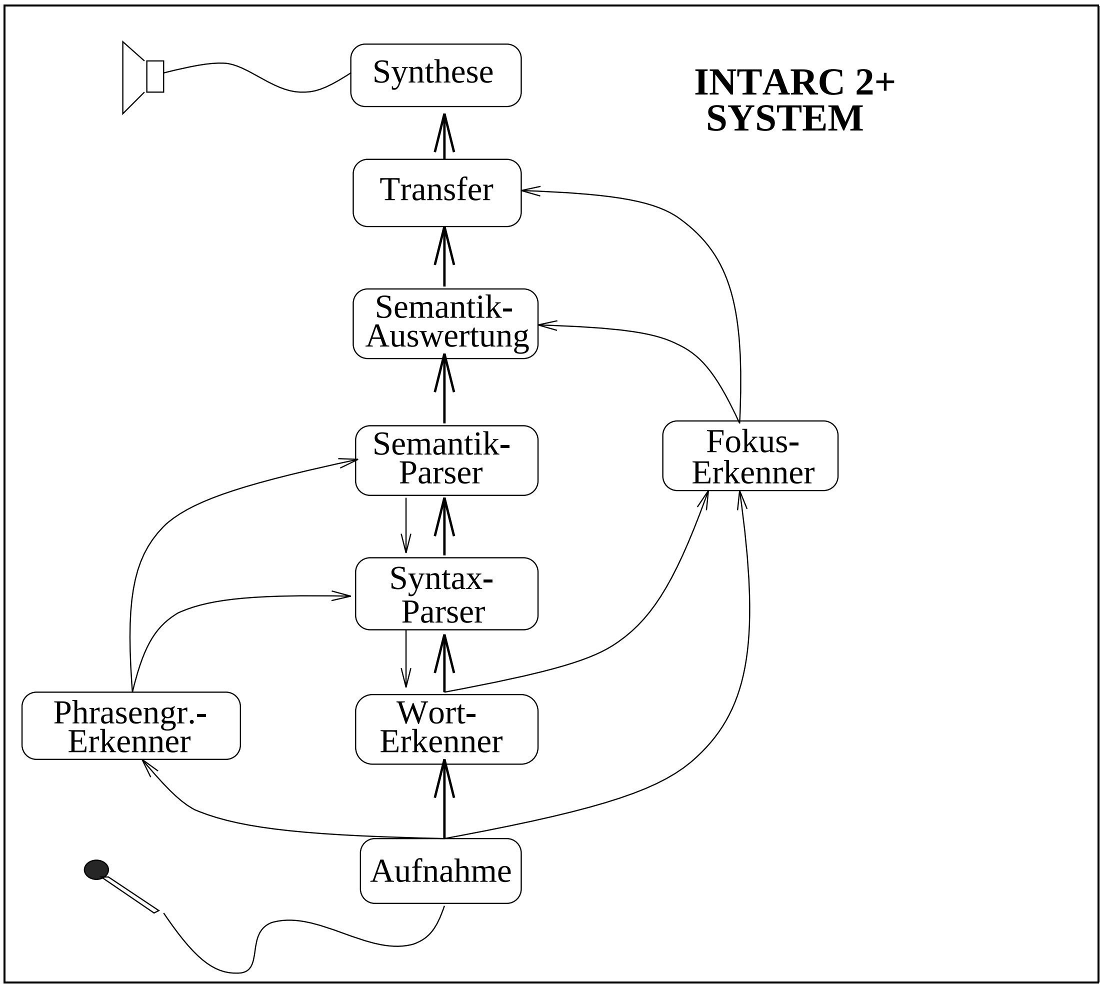

Abbildung 7.1: *Übersicht des* INTARC-*Systems (Version 2+)*

Wie der Verbmobil-Prototyp ist auch das experimentelle INTARC-System ein Übersetzer von gesprochenem Deutsch ins Englische [SEG+97].

Nach der Sprachaufnahme durch eine Gradient-Box beginnt parallel die Wort-, Phrasengrenzen- und Fokuserkennung, siehe Abbildung 7.1. Die Erkennung des Fokusakzents erfolgt mit einem regelbasierten Ansatz [Pet95], der alternativ zu dem in Kapitel 6.2 beschriebenen Akzenterkenner entwickelt wurde; auf ihn wird hier nicht weiter eingegangen. Eingabe des Phrasengrenzen- und Fokusdetektors ist neben dem Sprachsignal die Sprachgrundfrequenz, die hier nicht eingezeichnet ist.

Der Worterkenner [HJH96] ist ein modifizierter Viterbi-Decoder, der auf dem HTK[1]-Toolkit basiert. Die Modifikation besteht darin, daß nur die Vorwärtssuche ausgeführt wird, so daß die Rückwärtssuche Teil des Parsing-Algorithmus wird. Bei jedem Erreichen eines HMM-Endzustandes wird eine Worthypothese an den Syntaxparser verschickt. Der so inkrementell übertragene Worthypothesengraph kann selbst bei kurzen Äußerungen aus mehreren hundert Worthypothesen bestehen.

Aufgabe des Syntaxparsers ist es, im Worthypothesengraph jene Pfade zu finden, die eine grammatikalisch akzeptable Äußerung bilden, und diese als n beste Ableitungsbäume an die Semantikmodule weiterzuleiten. Dies ist als Strahlsuche implementiert, wobei in die Bewertungsfunktion fünf probabilistische Faktoren eingehen: die akustische Bewertung des Worterkenners, die Wort-Bigrammbewertung, vom Phrasengrenzenerkenner die a posteriori Wahrscheinlichkeiten für jede Grenzklasse, die Bewertung des prosodischen Bigramms und ein Grammatikfaktor. Diese Faktoren werden weiter unten erläutert.

Die Semantikmodule erstellen zu jedem Ableitungsbaum die passende semantische Darstellung, wobei syntaktisch korrekte, aber semantisch nicht passende Analysen verworfen werden. Syntax- und Semantikparser basieren auf der gleichen Grammatik im HPSG-Formalismus, die sprechaktbezogene Information über den Satzmodus enthält. Der prosodisch bestimmte Satzmodus wird in der Semantikauswertung bei der Auflösung von Kontextbezügen und zur Dialogaktklassifikation verwendet.

Aufgrund des Dialogaktes und der zugehörigen Inhaltswörter, z.B. TerminVorschlagOrt und Ortsname, wird eine schablonenhafte Übersetzung angefertigt, die anschließend mit einer kommerziellen Sprachsynthese gesprochen wird[2].

Die Fokusdetektion dient zur Steigerung der Robustheit: Die Fokuspositionen werden der bestbewerteten Wortkette zugeordnet. Falls ein Inhaltswort im (prosodisch detektierten) Fokus steht, wird es in einen Dialogakt klassifiziert, wobei ein endlicher Automat mit probabilistischen Präferenzen zur Anwendung kommt. Aus dem so bestimmten Dialogakt (z.B. TerminVorschlagOrt wird eine schablonenhafte Übersetzung erzeugt, wobei fehlende Information (z.B. der Ortsname) aus der besten Wortkette extrahiert wird. Falls die tiefe linguistische Analyse fehlschlägt, kann so auf eine „flache" Übersetzung zurückgegriffen werden, die in 30% der Fälle noch akzeptabel ist (in fast allen anderen Fällen reicht die Information für eine Übersetzung nicht aus).

[1]**H**idden Markov Model **T**ool**k**it von Entropic

[2]Im Verbmobil-Prototyp erfolgt eine detailgetreuere, aber wesentlich aufwendigere Übersetzung, an der neben dem eigentlichen Transfermodul ein eigenes Dialogmodul, und zur Erzeugung der englischen Satzes ein Generierungsmodul beteiligt sind. Siehe auch Abbildung 1.1 auf Seite 14.

In einem herkömmlichen System werden die Arbeitsschritte sequentiell ausgeführt, d.h. erst wird für den kompletten Turn ein Worthypothesengraph erzeugt, dann beginnt die Analyse des Syntaxparsers, das fertige Resultat wird an das Semantikmodul weitergereicht, usw. Wesentliches Kennzeichen des INTARC-Systems ist die inkrementelle Verarbeitung: die Analyse soll beginnen, noch bevor der Turn zu Ende gesprochen wurde, d.h., alle Module arbeiten parallel und geben ihre Ergebnisse möglichst früh an andere Module weiter; darüber hinaus verarbeiten einige Module auch Hypothesen von höheren Modulen. Die prosodischen Phrasengrenzen tragen auch wesentlich dazu bei, den Suchraum beim Parsen zu begrenzen, um das Fehlen des rechten Kontextes teilweise zu kompensieren. Zur Illustration sollen die beiden Äußerungen

- *Ok, treffen wir uns am Dienstag nachmittags.*
- *Ok, treffen wir uns am Dienstag. Nachmittags habe ich aber keine Zeit*

dienen, die bis *nachmittags* identisch sind; wenn nur die Wortkette vorliegt, ist die Interpretation an diesem Punkt mehrdeutig. Die Ambiguität besteht in der vollständigen Äußerung nicht mehr, aber durch prosodisch detektierte Phrasengrenzen kann die Ambiguität schon vorher aufgelöst werden.

7.2 Statistische Sprachmodelle

In den klassischen HMM-basierten Worterkennern wird zur Einschränkung des Suchraums üblicherweise ein n-Gramm als statistisches Sprachmodell (engl. *language model*, kurz *LM*) eingesetzt [Jel89]. Eine n-Gramm-Grammatik, die als stochastische reguläre Grammatik interpretiert werden kann, gibt die Wahrscheinlichkeit eines Satzes S, bestehend aus den N Wörtern $w_1...w_n$ (und den Symbolen s_A und $w_{N+1} = s_E$ für Satzanfang und -ende) an als

$$P(S) = \prod_{i=1}^{N+1} P(w_i|s_A, w_1, w_2...w_{i-1}) \tag{7.1}$$

wobei die bedingten Wahrscheinlichkeit $P(w_k|w_1...w_{k-1})$ nur aus den n vorhergehenden Wörtern geschätzt wird:

$$P(w_k|w_1...w_{k-1}) \approx P(w_k|w_{k-n+1}...w_{k-1}) \quad . \tag{7.2}$$

Die bedingten Wahrscheinlichkeiten werden aus einer Stichprobe geschätzt. In der Praxis beschränkt man sich auf Bi- oder Trigramme, da es schon bei einem Wortschatz von nur 1000 Wörtern eine Milliarde mögliche Sequenzen von drei Wörtern gibt. Da dennoch ein großer Teil der Sequenzen in der Stichprobe nicht auftritt und somit eine Wahrscheinlichkeit von 0 erhalten würde, behilft man sich durch Glättung: Trigramm-Wahrscheinlichkeiten werden aus Bigramm-, Unigramm- und notfalls Zerogramm-Wahrscheinlichkeiten interpoliert, wobei von den Wahrscheinlichkeitswerten der gesehenen Sequenzen ein Teil abgezogen wird, so daß sich die Wahrscheinlichkeiten wieder auf 1 addieren [Kat87].

Ein LM eignet sich auch zur Vorhersage von Akzenten und Phrasengrenzen. Ein solches Modell wurde zunächst an der TU Braunschweig entwickelt, um das prosodische Etikettieren zu erleichtern, wenn schon ein Teil etikettiert ist [Leh94, Leh96]. Da der Umfang des etikettierten Materials relativ gering ist (für das letzte LM mit prosodischen Etiketten wurden ca. 600 Turns verwendet bei einer Wortschatzgröße von etwa 2000), wurden nicht die Wörter selbst, sondern 55 grammatikalische Wortkategorien[3] als Einheiten verwendet. Dies ist möglich, weil viele prosodische Phrasengrenzen mit syntaktischen Konstituentengrenzen einhergehen [BKK+93]; ebenso kann man bei Akzentuierungsmustern von Wörtern auf Wortarten abstrahieren. Zu jeder Wortart gibt es 16 Varianten entsprechend den 4 Akzentstufen (bei mehrsilbigen Wörtern bezieht sich die Akzentstufe auf die Silbe mit dem lexikalischen Akzent) und der 4 Phrasengrenzenklassen (die sich auf die Grenze *nach* dem Wort beziehen). Insgesamt ergeben sich damit 880 Symbole, d.h. Wortkategorien *mit* prosodischen Etiketten.

Nachdem in einem Satz aus N Wörtern die Wörter in Wortkategorien umgewandelt wurden, gibt es für die Zuordnung von Akzenten und Phrasengrenzen 16^N Möglichkeiten, von denen mit Hilfe des LMs die mit der maximalen Wahrscheinlichkeit bestimmt werden kann. Bei längeren Sätzen ist das praktisch nicht mehr durchführbar; man beschränkt sich dann auf eine suboptimale Zuordnung, die durch Strahlsuche gefunden wird: Beginnend beim ersten Wort werden nur die Zweige weiterverfolgt, die relativ zur besten Bewertung eine gewisse Schwelle nicht unterschreiten. Bei Erreichen des letzten Wortes wird die beste Symbolfolge durch Rückverfolgung des Pfades ermittelt.

Das Erlanger Prosodiemodul enthält neben dem akustischen Erkenner ein prosodisches LM (mit 150 statt 55 Wortkategorien), wobei die „Erkennungsrate" des LMs allein schon höher ist als die des akustischen Erkenners [Kom95] (vergleiche auch mit Abschnitt 6.4); in den meisten Veröffentlichungen darüber wird die kombinierte Erkennungsrate angegeben.

Das INTARC-System enthält ebenfalls ein prosodisches LM, das allerdings im Syntaxparser integriert ist und zur Kopplung zwischen Phrasengrenzenerkennung und Syntaxanalyse dient (siehe den nächsten Abschnitt). Zuerst wurde hier das LM der TU Braunschweig verwendet, später wurde am IKP ein eigenes entwickelt.

[3] Darin sind auch spontansprachliche Phänomene wie nichtartikulatorische Geräusche enthalten.

7.3 Syntax- und Semantik-Parser

Syntax- und Semantikparser basieren auf der gleichen Grammatik im HPSG-Formalismus, wobei der kontextfreie Anteil für den Syntax-Parser in eine kontextfreie stochatische Grammatik kompiliert wurde: Die Grammatik-Analyse ist also strenggenommen nicht in einen syntaktischen und einen semantischen Anteil aufgespalten, vielmehr dient der Syntaxparser mit seiner kubischen Komplexität als Filter für den Semantikparser; so daß die volle Unifikation der dorthin durchgereichten Analysebäume auch bei exponentieller Komplexität in tragbarer Rechenzeit möglich wird [DKK95, KK96b]. Um die beiden Parser während der parallelen Verarbeitung zu synchronisieren, muß, da die gleiche Grammatik verwendet wird, nur kommuniziert werden, welche Regeln (bzw. Regelindices) erfolgreich und welche nicht erfolgreich angewendet werden konnten. Jeder Parser kann damit den Zustand des jeweils anderen rekonstruieren.

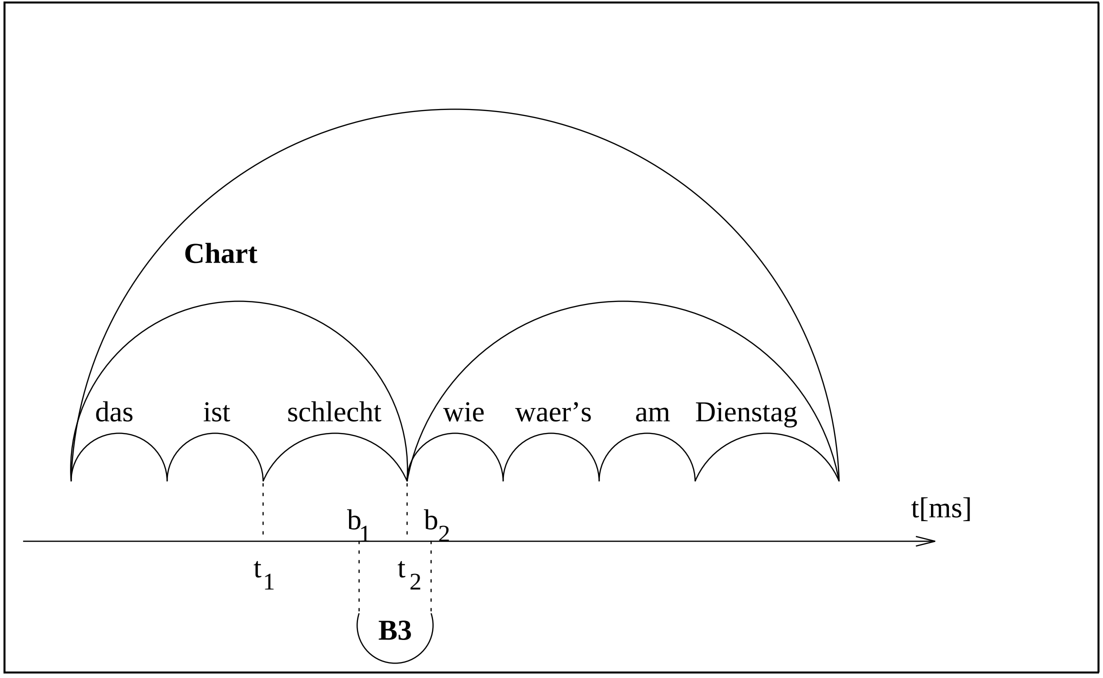

Abbildung 7.2: *Zuordnung der prosodisch detektierten Phrasengrenzen auf die wortüberspannenden Kanten im Syntax- und Semantikparser, nach [KK96b]: Die* **B3***-Grenzhypothese, die sich auf das Zeitintervall* $[b1, b2]$ *bezieht, wird der Wortgrenze nach „schlecht" zugeordnet nach der Bedingung* $t_1 \leq b_1 \leq t_2$.

7.3.1 Syntaxparser

Der Syntaxparser [Web95] erhält als Eingabe Worthypothesen (zur Kopplung Erkenner/Parser [HW94a, HW94b]) und prosodische Phrasengrenzenhypothesen. Die Zuordnung der Grenzhypothesen, die sich auf das Zeitintervall zwischen zwei Silbenkernen beziehen, auf die wortüberspannenden Kanten erfolgt aufgrund der Anfangs- und Endzeiten der Worthypothesen, siehe Abbildung 7.2.

Der Syntaxparser versucht, von links nach rechts Ableitungsbäume gemäß der stochastischen kontextfreien Grammatik zu erzeugen; partielle Ableitungen werden ebenfalls als Kanten dargestellt. Die Suche nach den n besten Ableitungsbäumen besteht im sukzessiven Kombinieren von Kantenpaaren zu neuen Kanten, wobei die Anwendung möglicher Grammatikregeln durch Strahlsuche begrenzt wird. Als Bewertungskriterium für die jeweils hypothetisierte Wortsequenz dient dabei eine Linearkombination aus drei (bzw. fünf) Faktoren:

1. Decoder-Faktor bestehend aus
   (a) akustischer Bewertung des Worterkenners
   (b) Bigramm-Bewertung,
2. Prosodie-Faktor bestehend aus
   (a) akustischer Bewertung des Phrasengrenzenerkenners (die a posteriori Wahrscheinlichkeiten für die vier Grenzklassen)
   (b) Bewertung des prosodischen Trigramms,
3. Grammatik-Faktor.

Der Decoder-Faktor ist das wohlbekannte Produkt aus der akustischen Bewertung (d.h. der HMM-Wahrscheinlichkeit der akustischen Merkmalvektoren, falls die hypothetisierte Wortsequenz gesprochen wurde) und der Bigramm-Bewertung für diese Wortfolge (vergleiche Abschnitt 7.2).

Der Prosodiefaktor für ein Kantenpaar ist das maximale Produkt aus der Wahrscheinlichkeit des prosodischen Bigramms und den a posteriori Wahrscheinlichkeiten der Phrasengrenzenklassen. Maximiert wird über die 16 möglichen Kombinationen von Phrasengrenzenklassen zwischen den beiden Kanten und dem (ersten) Wort, das durch die zweite Kanten überspannt wird. Betrachte als Beispiel die Grenze zwischen *ist* und *schlecht* in Abbildung 7.2: Nach der Abbildung der Wörter auf die Wortkategorien[4], hier *finites Hilfsverb* und *prädikatives*

[4] Bei mehrdeutigen Wörtern wie *überlegen*, das gem. den Kategorien in [Leh96] ein prädikatives Adjektiv, ein finites Verb oder ein Infinitiv sein kann, erzeugt der Syntaxparser entsprechend drei verschiedene Kanten

Adjektiv, wird für alle Kombinationen, z.B. **B0** nach *ist* und **B3** nach *schlecht*, die Wahrscheinlichkeit des prosodischen Bigramms[5] für diese Folge aus Wortkategorien und Phrasengrenzen mit den a posteriori Wahrscheinlichkeiten für die beiden Phrasengrenzenhypothesen multipliziert. Das Maximum wird durch lokale Viterbi-Suche bestimmt und ergibt den Prosodiefaktor. Falls für eine Wortgrenze keine Phrasengrenzenhypothese vorliegt, wird **B0** mit Wahrscheinlichkeit 1 angenommen. In seltenen Fällen überspannt eine Phrasengrenzenhypothese mehr als eine Wortgrenze; dann ergibt sich die Zuordnung zu einer Wortgrenze ebenfalls aus der lokalen Viterbi-Maximierung. Lokal bedeutet hier, daß im Sinne der links-rechts-inkrementellen Verarbeitung die Grenze nach dem zweiten Wort, im Beispiel die nach *schlecht*, bei der späteren Suche nicht mehr verwendet wird, wenn also ein Kantenpaar mit der Grenze zwischen *schlecht* und *wie* zusammenfällt.

Der Grammatik-Faktor ist die normierte Wahrscheinlichkeit des Grammatik-Modells dafür, daß eine bestimmte neue Analysekante für das Eingabe-Kantenpaar erzeugt wird.

Eine günstige Gewichtung der Faktoren wurde experimentell ermittelt. Als Gütemaße dienten die Worterkennungsrate und die Zahl der Kantenpaare, die während der Analyse untersucht wurden und damit ein Maß für die Laufzeit darstellen; diese Anzahl kann durch die Strahlweite beeinflußt werden.

Die Worterkennungsrate wurde bezogen auf die 10 besten Analysebäume. Diese Worterkennungsrate des Syntaxparsers ist höchstens so hoch wie die des Worterkenners, kann aber wesentlich niedriger sein, da nur die Wörter als erkannt gewertet wurden, die Teil einer gültigen Ableitung von Beginn des Turns an sind; Wörter rechts davon, die nicht in die Ableitung integriert werden können, zählen alle als Auslassungsfehler.

Durch diese Art der Koppelng kann die Prosodieerkennung zwar nicht zwischen verschiedenen syntaktischen Ableitungen für dieselbe Worthypothesenfolge disambiguieren, weil sich dann auch die Prosodiefaktoren nicht unterscheiden. Für diese Art von Disambiguierung müßten die prosodischen Phrasengrenzen in der Grammatik selbst modelliert werden. Der Prosodiefaktor wirkt aber unterstützend bei der Auswahl zwischen verschiedenen Worthypothesen.

Zum Testen wurden zwei verschiedene Stichproben verwendet: einmal vier Verbmobil-Dialoge (Testset 2) und einmal Dialoge mit einfacherer semantischer Struktur, die für das INTARC-System aufgenommen wurden (Testset 1).

Bei ersten Experimenten auf dem Testset 1 konnte bei gleicher Worterkennungsrate eine Reduktion der Kantenpaare um 20% erreicht werden, wenn die

[5] Die prosodisch detektierten Akzente werden hier nicht verwendet, deshalb werden im prosodischen Bigramm die Wahrscheinlichkeiten für die vier Akzentklassen aufsummiert.

Gewichtung für den Prosodiefaktor von 0 auf 4.5 angehoben wurde (viele der Kanten sind konstant, nämlich jene, die die Eingabe-Worthypothesen repräsentieren, und die gerade aktiven, leeren Kanten. Zählt man nur die durch die Suche dynamisch erzeugten Kanten, beträgt die Reduktion 65%).

Es zeigte sich in folgenden Experimenten, daß die detektierten Phrasengrenzen von den Erwartungen des prosodischen Bigramms zu stark abwichen, welches mit den handetikettierten Phrasengrenzen trainiert wurde: Für eine gegebene Folge von Worthypothesen wird der Prosodiefaktor dann besonders hoch, wenn die entsprechende Folge von Wortkategorien vom prosodischen Bigramm hoch bewertet wird und gleichzeitig die dabei vorhergesagten Phrasengrenzen mit den akustisch detektierten Phrasengrenzen gut übereinstimmen. Damit der Prosodiefaktor für die richtige Wortkette besonders hoch wird, muß das prosodische Bigramm die tatsächlich realisierten Phrasengrenzen vorhersagen können. Die handetikettierten Phrasengrenzen hingegen sind stark durch linguistische Vorerwartungen geprägt, wie sich auch in Kapitel 8 zeigen wird.

Deshalb wurde in Bonn ein prosodisches Bigramm trainiert, das stärker an die akustischen Gegebenheiten angepaßt ist. Dazu wurden die gleichen Wortklassen wie für das Braunschweiger Bigramm verwendet, statt der handetikettierten Phrasengrenzen aber die automatisch detektierten Phrasengrenzen.

Erst mit diesem prosodischen Bigramm konnten deutlich bessere Ergebnisse erzielt werden: Für den Testset 1 reduzierte die Prosodie die Anzahl der Kantenpaare um 40% bei gleichzeitiger Erhöhung der Worterkennungsrate von 84% auf 86%. Für den schwierigeren Testset 2, die Verbmobil-Dialoge, erhöhte die Prosodie bei gleicher Anzahl von Kantenpaaren die Worterkennungsrate von 48.2% auf 53.2%.

7.3.2 Semantikparser

Der Semantikparser bzw. die Semantik-Konstruktion ist ein Bottom-Up-Chart-Parser [Gö88, Win83], der zu den Ableitungsbäumen vom Syntaxparser die passende semantische Darstellung erzeugt, wobei er die semantische Information aus der HPSG-Grammatik bezieht (das Modul Semantik-Auswertung dient zur Auflösung von Referenzen, Erkennung von Dialogakten und Verwaltung des Dialoggedächtnis). Dies geschieht durch Rekonstruktion der Chart [KKSW96]: Sobald der Syntaxparser für einen satzwertigen linken Teilpfad im Worthypothesengraph, wie *„das ist schlecht“*, eine überspannende Chart erzeugen kann, wird diese an den Semantikparser gesendet. Falls der Semantikparser diese Chart aufgrund semantischer Beschränkungen nicht rekonstruieren kann, fordert er vom Syntaxparser die am nächstbesten bewertete Chart für diese Wortfolge an.

Da die Grammatik nicht Sätze, sondern Turns, also ganze Dialogbeiträge beschreibt (ein Redebeitrag besteht meist aus mehreren Sätzen oder satzwertigen Äußerungen), besteht ein Hauptproblem in der korrekten Segmentierung in Sätze bzw. Äußerungssegmente. Dazu benutzt der Semantikparser die prosodisch detektierten Phrasengrenzen. Die Grammatik wurde so erweitert, daß segmentverbindende Regeln eine Phrasengrenze erfordern, während segmentinterne Regeln Phrasengrenzen ausschließen. Dies wird für Heuristiken zur verzögerten Auswertung von segmentübergreifenden Regeln ausgenutzt. Des weiteren benutzt der Semantikparser den prosodischen Satzmodus, um den syntaktischen Satzmodus zu disambiguieren [KK96a]; dies ist vor allem für die Dialogaktklassifikation von Bedeutung. Ein Beispiel aus dem Verbmobil-Testdialogen ist: *„sagen wir gleich vierzehn Uhr?“*. Dies könnte auch als Aussage gesprochen sein; es würde sich zwar in beiden Fällen um einen Terminvorschlag handeln, der Unterschied ist jedoch bei der Sprachsynthese der englischen Übersetzung relevant.

Die Anzahl der Parsing-Hypothesen reduziert sich um 65.4%, wenn „ideale”Phrasengrenzenhypothesen, d.h. die handetikettierten Phrasengrenzen, als Eingabe verwendet werden. Dadurch reduziert sich die Anzahl der Lesarten um 41.9%.

Da die prosodisch detektierten Phrasengrenzen nicht immer mit grammatischen Phrasengrenzen einhergehen, wurde der Semantikparser um einen *Recovery-Mechanismus* erweitert, um Hypothesen, die durch die Prosodie ausgeschlossen wurden, reaktivieren zu können [KK96a]. Mit diesem Recovery-Mechanismus reduzieren die prosodisch detektierten Phrasengrenzen die durchschnittliche Anzahl der Lesarten eines Turns um 24.7%.

Kapitel 8

Vergleich der Erkennungsleistung mit dem menschlichen Hörer

Ziel der Arbeit war es, die prosodische Ausprägung der linguistischen Konzepte Akzent, Phrasengrenze und Satzmodus so zu modellieren, daß eine automatische Klassifikation möglich wird. Das Problem besteht darin, daß die linguistische Funktion z.B. der Phrasierung als Gliederung in syntaktische Einheiten, mit der prosodischen Realisierung, z.B. der Form des Grundfrequenzverlaufs, in keiner eindeutigen Beziehung steht.

Die automatische Klassifikation erfordert zunächst eine Beschreibung der prosodischen Realisierung in Form von Merkmalen. In Kapitel 6 wurde gesagt, daß die Merkmale so zu wählen sind, daß sie die Klassen möglichst gut trennen. Schlechte Klassifikationsergebnisse können also auf inadäquate Merkmale zurückzuführen sein. Wirklich schwierig wird es jedoch, wenn die Klassen selbst „inkonsistent" sind, wenn sich z.B. akzentuierte Silben von nicht-akzentuierten Silben anhand ihrer prosodischen Realisierung *allein* nicht trennen lassen, weil dann auch mit den optimalen Merkmalen keine guten Klassifikationsergebnisse zu erreichen sind.

Es stellt sich die Frage, wie gut sich die gegebenen prosodischen Klassen überhaupt trennen lassen, wenn nur akustisch-prosodische Information zur Verfügung steht, aber keine Wortinformation und andere linguistische Information. Dies zu beantworten ist mit statistischen Methoden allein nicht möglich, etwa mit der Analyse von Häufungsgebieten, da sie schon eine bestimmte Merkmalsextraktion voraussetzt.

In einer abschließenden Untersuchungsreihe wurde deshalb getestet, wie gut der Mensch mit seinen perzeptiven Fähigkeiten als universeller, leistungsstarker Mustererkenner diese Aufgabe bewältigt. Dazu wurde ein Teil der Verbmobil-Stichprobe *delexikalisiert*, d.h. so verändert, daß die Verständlichkeit verloren geht, die prosodischen Charakteristika aber erhalten bleiben, und anschließend Mitarbeitern des IKP vorgespielt mit der Aufgabe, Phrasengrenzen und Akzente zu markieren. Wenn nun die menschlichen Hörer Phrasengrenzen und Akzente besser erkennen als der automatische Erkenner, muß sich der Erkenner noch verbessern lassen. Wenn die Hörer nicht besser abschneiden, legt dies die Vermutung nahe, daß anhand der „reinen Prosodie" keine wesentlich bessere Trennung der Klassen erreichbar ist.

In den folgenden Abschnitten werden zwei Methoden zur Delexikalisierung und die zugehörigen Perzeptionsexperimente beschrieben. An beiden Experimenten nahmen jeweils 11 Hörer teil, die in 20 Turns die Phrasengrenzen und in 22 Phrasen die Akzente markierten. Um den Aufwand erträglich zu halten, wurden nur **B3** von Nicht-**B3**-Grenzen und akzentuierte von nicht-akzentuierten Silben unterschieden. Die Erkennungsleistung der Hörer wird mit der des automatischen Erkenners verglichen.

8.1 Delexikalisierung

Zweck der Delexikalisierung von gesprochener Sprache ist die Zerstörung der Wortinformation unter Beibehaltung der prosodischen Information. Die Sprache soll also unverständlich werden, die prosodischen Eigenschaften wie Intonationsverlauf, Sprechrhythmus aber hörbar bleiben. Um die Hörer nicht zu überfordern soll das Ergebnis möglichst angenehm und „sprachähnlich" klingen.

In der Literatur finden sich verschiedene Methoden zur Delexikalisierung: Bei der spektralen Inversion [LW76] wird das Vorzeichen jedes zweiten Abtastwertes invertiert, was einer Spiegelung des Spektrums entspricht, so daß die hohen Frequenzen auf tiefe abgebildet werden und umgekehrt. Vor der Inversion erfolgt eine Hochpaß-Filterung, danach eine Tiefpaß-Filterung; dieses Signal wird mit dem tiefpaß-gefilterten originalen Signal überlagert.

In [Leh79], [Kre82] und [Sch84] wurden zur Delexikalisierung rigide Tiefpaß- bzw. Bandpaßfilter eingesetzt. Alle diese Verfahren wurden implementiert und akustisch beurteilt, aber verworfen, weil sie die segmentale Information nicht vollständig auslöschen. Da die Teilnehmer der Perzeptionsexperimente mit der Verbmobil-Stichprobe teilweise vertraut waren, war die Gefahr der Wiedererkennung von einzelnen Turns zu groß.

In [dPS94] wird zum gleichen Zweck LPC[1]-Resynthese vorgeschlagen, wobei die Formanten vorher auf neutrale Werte gesetzt werden, die dem schwachen Vokal (siehe Anhang A) entsprechen. Der heikle Punkt daran ist das Formant-

[1]Linear Predictive Coding, neben dem Psola-Verfahren die gebräuchlichste Methode zur Sprachsynthese.

Tracking; der zur Verfügung stehende Formant-Tracker brachte keine akzeptablen Ergebnisse und manuelle Korrektur wäre zu aufwendig gewesen. Aus diesem Grund wurde ein alternatives Verfahren entwickelt.

8.2 Sägezahn-Signale

Bei diesem Verfahren zur Delexikalisierung wird das Sprachsignal in stimmhaften Bereichen durch ein Sägezahnsignal ersetzt, in stimmlosen durch Stille. Dazu wird jede Grundperiode durch einen Sägezahn gleicher Länge ersetzt und die Amplitude so eingestellt, daß die Energie gleich ist. Dafür sind die Markierungen der Grundperioden nötig. Dazu wurde der Pitchmarker des Bochumer Projektpartners [Rin93] verwendet, der für diesen Zweck ausreichte. Die Sägezahnsignale klingen relativ sprachähnlich und hören sich an wie Brummen; gelegentliche Fehler des Pitchmarkers stören nicht besonders.

In einer Voruntersuchung wurden stimmlose Bereiche wahlweise auch durch weißes Rauschen gleicher Energie ersetzt, in der Hoffnung, noch sprachähnlichere Signale zu erhalten. Die Perzeption von Akzenten und Phrasengrenzen wurde dadurch aber nur geringfügig beeinflußt, und die meisten Hörer empfanden das Rauschen als eher störend.

8.2.1 Versuchsablauf

Die Sägezahnsignale wurden den Versuchsteilnehmern rein auditiv präsentiert. Zusätzliche visuelle Information wie der Grundfrequenzverlauf hätte zwar das Setzen der Akzent- und Phrasengrenzenmarkierungen erleichtert, das Hörerurteil aber möglicherweise beeinflußt. Daher wurden die Markierungen durch Tastendruck gesetzt mit der Möglichkeit zur akustischen Kontrolle: Sobald der Versuchsteilnehmer die erste Phrasengrenze wahrnimmt, schneidet er das Signal per Tastendruck an dieser Stelle. Die Signalteile vor und nach dem Schnitt können beliebig oft angehört werden, der Schnitt kann in kleinen Schritten verschoben werden, oder der ganze Vorgang kann so lange wiederholt werden, bis die Phrasengrenze bestätigt wird. Die erste Phrase wird dann endgültig abgetrennt, und man fährt fort mit dem Rest des Signals. Die Reaktionszeit spielt keine Rolle, außer für die Dauer der Sitzung.

Wenn eine **B3**-Grenze innerhalb eines stimmlosen Bereichs lag, wurde ein Schnitt innerhalb dieses Bereichs als korrekt gewertet, ansonsten wurde eine Entfernung von bis zu 50 ms zugelassen.

Akzente wurden ebenfalls per Tastendruck markiert. Zur auditiven Kontrolle wurde das Signal an der entsprechenden Stelle mit einem kurzen Piepton überlagert. Auch hier bestand die Möglichkeit, vor der Bestätigung die Markierung beliebig oft zu verschieben bzw. den Vorgang zu wiederholen.

Akzentmarken wurden als korrekt gewertet, wenn sie innerhalb einer akzentuierten Silbe lagen.

Die 11 Versuchsteilnehmer waren Mitarbeiter des IKP und hatten Erfahrung im prosodischen Etikettieren. In Anlehnung an die Prozedur beim prosodischen Etikettieren des Verbmobil-Materials in Braunschweig (siehe Abschnitt 4.1) wurden Phrasengrenzen und Akzenten in getrennten Sitzungen bearbeitet. Sowohl das Etikettieren von Phrasengrenzen als auch von Akzenten wurde in je zwei Sitzungen unterteilt, denen eine Einweisung und eine Trainingsphase voranging. Jede Sitzung dauerte ungefähr eine dreiviertel Stunde.

8.2.2 Phrasengrenzen

Für das Etikettieren der Phrasengrenzen wurden 20 Turns (von 3 Frauen und 12 Männern) ausgewählt, wobei die Ausgewogenheit der Grenz-Etiketten **B2**, **B3** und **B9** und der damit verbundenen Ton-Etiketten ausschlaggebend war.

Diese Turns enthalten 58 **B3**-Grenzen, wobei die 20 äußerungsfinalen **B3**-Grenzen nicht berücksichtigt wurden, weil deren Erkennung sowohl für den Hörer als auch für den Phrasengrenzendetektor trivial ist, wenn man die Forderung nach Inkrementalität fallen läßt. 16 der verbleibenden 38 **B3**-Grenzen tragen das Ton-Etikett L-L%, 11 ein H-H%, 7 ein H-L% und 4 ein L-H%. Die 20 Turns enthalten darüber hinaus 12 **B2**- und 8 **B9**-Grenzen.

Der Phrasengrenzendetektor erkannte in den Originalversionen dieser Turns 21 der 38 **B3**-Grenzen und fügte 11 Grenzen ein. Dem Detektor wurde die Originalversionen statt der delexikalisierten präsentiert, um eine aufwendige Anpassung der Vorverarbeitung zu vermeiden; dies betrifft weniger die Grundfrequenzbestimmung als die Silbenkerndetektion. Was das betrifft, gelingt dem Hörer die Umstellung von Sprachsignalen auf sprachähnliche Sägezahnsignale fast mühelos. In der CD-ROM-Version zu [SW96] sind illustrative Hörbeispiele enthalten.

Die Hörer erkannten in den delexikalisierten Versionen der Turns im Durchschnitt 23.0 **B3**-Grenzen und fügten 10.3 Grenzen an Stellen ein, an denen weder eine **B2**-, eine **B3**- oder **B9**-Grenze vorhanden war. Die Einfügungsfehler der Hörer wurden weniger streng gewertet, weil in der Einweisung der Begriff „Phrasengrenze" lediglich als „Einschnitte im Redefluß" definiert wurde (bei der Auswertung der Akzentmarken im nächsten Abschnitt hätte man den Hörern gegenüber auf ähnliche Art großzügig sein können, indem z.B. Auslassungen von Nebenakzenten nicht als Fehler gewertet werden. Für das wesentliche Ergebnis dieser Versuchsreihe ist dies jedoch nicht von Belang.).

Bei einer genaueren Analyse zeigte sich, daß alle **B3**-Grenzen, die mit einer Pause einhergehen, von den Hörern erkannt wurden, und fast alle vom Detektor. Die etikettierten Phrasen- und Grenztöne stehen jedoch in keinem systematischen Zusammenhang zu den Erkennungsfehlern.

Da nun variable Einheiten verwendet werden, nämlich stimmlose Bereiche oder Umgebungen von ± 50 ms, und diese Einheitengrenzen nur an Phrasengrenzen klar definiert sind, läßt sich keine Erkennungsrate mehr angeben. Eine Alternative wäre gewesen, als Einheit den Bereich von einer Silbenkernmitte zur nächsten Silbenkernmitte zu nehmen, mit je einer offenen Einheit am Turnanfang und -ende. Dies hätte aber eine manuelle Korrektur der automatischen Phonemsegmentierung erfordert. Um dennoch eine Erkennungsleistung in eine Zahl kondensieren zu können, wird hier das in der Worterkennung gebräuchliche Maß der *Akkuratheit* verwendet:

$$\text{Akkuratheit} = \frac{\text{Anz. erkannte B3} - \text{Anz. eingefügte B3}}{\text{Anzahl B3}} \tag{8.1}$$

Die Akkuratheit ist dann, wenn keine **B3** eingefügt werden, gleich der Erkennungsrate, sonst niedriger und kann leicht negativ werden. Ein „Detektor“, der nichts ausgibt, hat eine Akkuratheit von 0, für eine bestimmte Anwendung kann aber ein Detektor mit negativer Akkuratheit trotzdem besser sein.

Für den Phrasengrenzendetektor ergibt sich eine Akkuratheit von 26%, die Hörer erreichen eine Akkuratheit von durchschnittlich 33%, siehe auch Abbildung 8.2 im nächsten Abschnitt.

Da die prosodische Realisierung einer **B3** oft der einer **B9** ähnlich ist, wurden in einer weiteren Auswertung beide Grenztypen zusammengefaßt. Von diesen 46 **B[39]** erkannten die Hörer im Durchschnitt 26 Grenzen, womit die Akkuratheit nur leicht auf 36% ansteigt. Der Detektor fand 2 der 8 zusätzlichen Grenzen, fügte aber weitere ein, so daß seine Akkuratheit auf 22% absinkt.

8.2.3 Akzente

Für das Etikettieren der Akzente wurden 22 Phrasen (von 3 Sprecherinnen und 8 Sprechern) ausgewählt, die 51 akzentuierte Silben enthalten, 24 Hauptakzente und 27 Nebenakzente. Der Akzentdetektor erkannte davon 33 bei 18 Einfügungen, was einer Akkuratheit von 29% entspricht. Die Hörer erkannten im Durchschnitt 27.6 Akzente und fügten 16.6 ein; dies ergibt eine Akkuratheit von 21%.

Die Hauptakzente werden sowohl vom Detektor als auch von den Hörern zuverlässiger erkannt. Aber es scheint wieder kein Zusammenhang zu den Ton-Etiketten (hoher oder tiefer Akzent, Position innerhalb der Silbe) zu bestehen.

8.3 Sinnleere Sprachsignale

Das im vorangehenden Abschnitt dargestellte Verfahren zur Delexikalisierung hat mit den in der Literatur beschriebenen gemeinsam, daß die (Phonem-)Segmentgrenzen zerstört werden und damit auch die Grenzen der Einheiten verloren gehen, an denen man Phrasengrenzen- und Akzent-Etiketten festmachen kann (bei den Sägezahn-Signalen bleiben lediglich die Grenzen zwischen stimmhaften und stimmlosen Bereichen erhalten). Deshalb mußte das für Phonetiker ungewohnte Verfahren der Etikettierung per rechtzeitigem Tastendruck entwickelt werden. Ohne ausgeprägtes Rhythmusgefühl und gute Reaktion lassen sich die Etiketten zwar auch korrekt setzen, aber nur mit mehr Zeit und Mühe.

Um herauszufinden, ob dies auf das Testergebnis einen Einfluß hatte, wurde eine Methode zur Delexikalisierung entwickelt, bei der die Segmentgrenzen erhalten bleiben. Diese ermöglichte die *Verschriftung* der delexikalisierten Äußerungen und damit das Markieren der Akzente und Phrasengrenzen an der Verschriftung. Hierzu wurde der am IKP vorhandene PSOLA-Synthesizer [PHH97] verwendet. Eingabe des Synthesizers sind die Phonemkette (in SAMPA) sowie Dauer-, Energie- und Grundfrequenzwerte. Die Phonemkette und die Phonemdauern wurden aus der automatischen Phonemsegmentierung gewonnen (siehe Abschnitt 4.2), die Grundfrequenz mithilfe des ESPS[2]-Paketes. Vor der Resynthese wurde jedes Phonem gegen eines aus der gleichen Klasse zufällig ausgetauscht, also z.B. Vokale gegen Vokale und stimmhafte Plosive gegen stimmhafte Plosive. Die Klassen wurden so gewählt, daß die deutsche Phonotaktik erhalten bleibt.

Die Energie hätte sich am Synthesizer auch manipulieren lassen. Darauf mußte jedoch verzichtet werden, weil die Zeitauflösung der Phonemsegmentierung nicht ausreichte, um brauchbare Energiewerte zu erhalten.

Beispielsweise wurde die Phrase (genauer: die in Klammern angegebene SAMPA-Verschriftung):

auf jeden Fall noch im April `[aUfjed@nfalnOxImapRIl]`

permutiert zu

eus röh bem se me fon ni klur `[OYsR2:b@msERmEfOnIklUR]`

wobei die Intonation und die Dauerverhältnisse bei der Resynthese erhalten blieben. Die permutierte Phonemkette wurde wieder in deutsche Orthographie übertragen, wobei die Silben zur besseren Lesbarkeit durch Leerzeichen getrennt wurden, siehe das Beispiel oben.

Die so verschrifteten Turns bzw. Phrasen wurden den Hörern ausgedruckt

[2]Entropic's Signal Processing System, Entropic Research Laboratory, Inc.

vorgelegt, die zugehörigen Sprachdateien konnten beliebig oft abgehört werden. Das Markieren von Phrasengrenzen und Akzenten erfolgte wieder in getrennten Sitzungen.

8.3.1 Phrasengrenzen

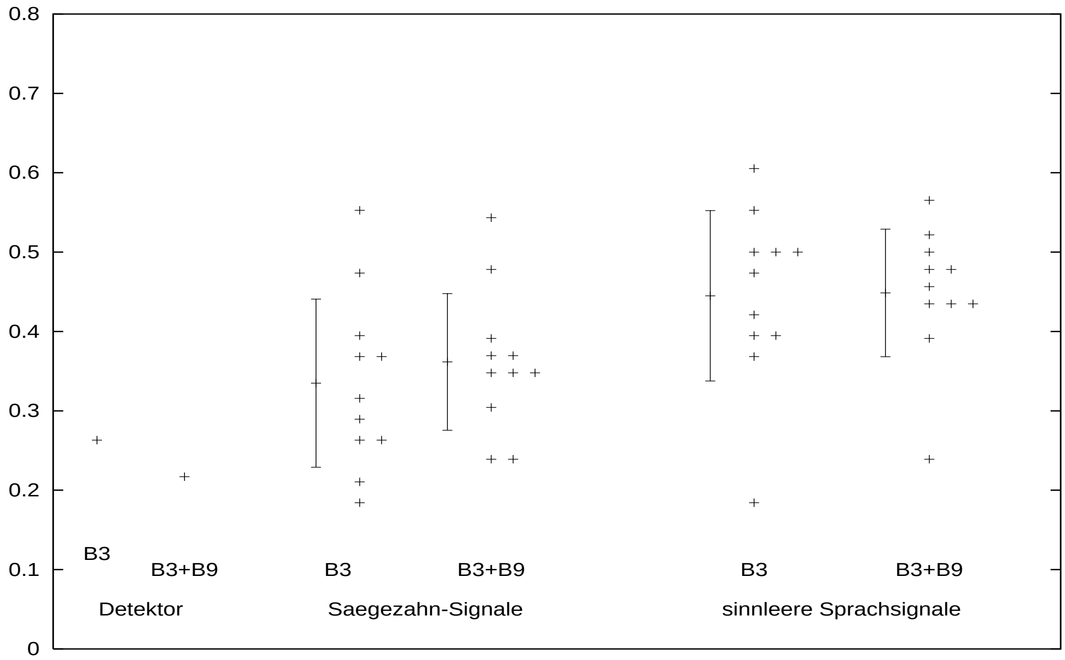

Abbildung 8.1: *Akkuratheit bei der Re-Etikettierung von Phrasengrenzen in delexikalisierter Sprache durch 11 Hörer. Mittelwerte und Standardabweichungen bei den Hörerdaten sind als Fehlerbalken dargestellt. Zum Vergleich links die Akkuratheit des Detektors.*

Es wurden die gleichen 20 Turns wir in Anschnitt 8.2 verwendet. Phrasengrenzen sollten in der Verschriftung durch senkrechte Striche zwischen den betreffenden Silben markiert werden.

Die Hörer erkannten im Durchschnitt 23.7 der 38 **B3**-Grenzen, fügten aber nur 6.1 **B3**-Grenzen ein. Dies führte zu einer gegenüber den Sägezahnsignalen höheren Akkuratheit von 44%. Werden bei der Auswertung **B3**- und **B9**-Grenzen zusammengefaßt, ändert das Ergebnis nur unwesentlich. In Abbildung 8.1 ist die Akkuratheit der einzelnen Hörer und ihr Durchschnitt bei diesem Perzeptionsexperiment dargestellt, zum Vergleich auch das Ergebnis des Perzeptionstest mit den Sägezahnsignalen und die Akkuratheit des Detektors.

8.3.2 Akzente

Es wurden die gleichen 22 Phrasen wir in Anschnitt 8.2 verwendet. Akzente sollten in der Verschriftung durch Unterstreichen der betreffenden Silben markiert werden.

Im Durchschnitt wurden von den 51 Akzenten 26.9 erkannt und 15.2 eingefügt, was eine Akkuratheit von 23% ergibt. Abbildung 8.2 zeigt die Einzelergebnisse und stellt sie denen für die Sägezahnsignale und für den Detektor gegenüber.

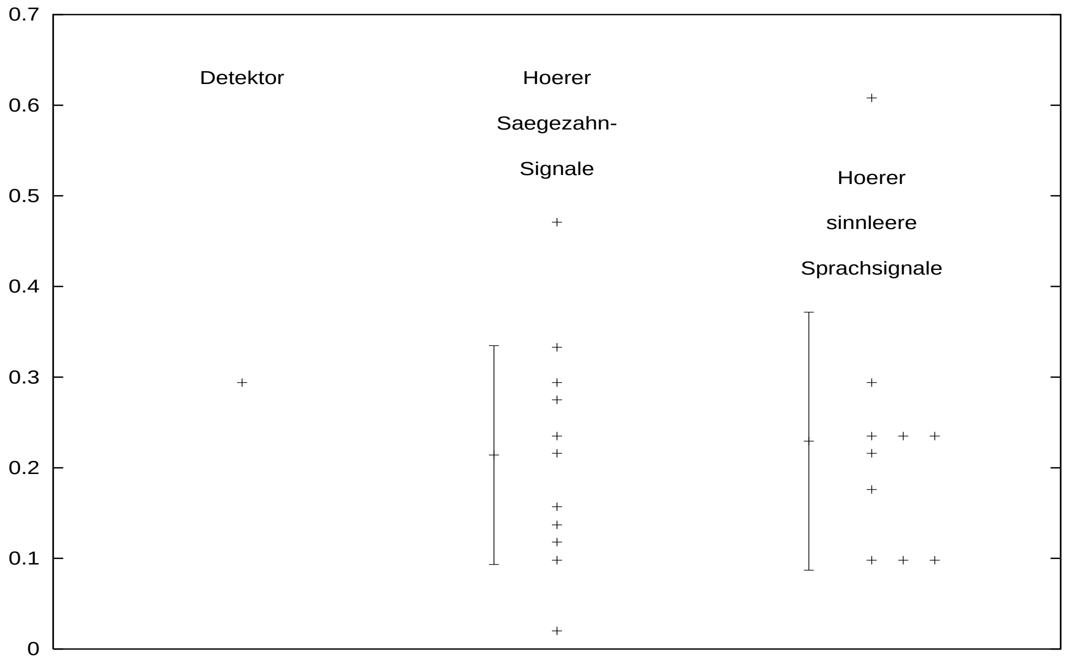

Abbildung 8.2: *Akkuratheit bei der Re-Etikettierung von Akzenten in delexikalisierter Sprache durch 11 Hörer. Mittelwerte und Standardabweichungen bei den Hörerdaten sind als Fehlerbalken dargestellt. Zum Vergleich links die Akkuratheit des Detektors.*

8.4 Folgerungen

Beim Vergleich des Detektors mit den Hörern als auch bei der Streuung innerhalb der Hörer muß der mögliche Wertebereich der Akkuratheit berücksichtigt werden, aufgrund dessen die Abbildungen 8.1 und 8.2 Unterschiede stärker hervorheben als es Erkennungsraten tun würden.

Aber inwieweit sind diese Zahlen überhaupt aussagekräftig? Inwieweit ist die Akkuratheit des Detektors mit der der Hörer vergleichbar?

Der Detektor wurde mit einer größeren Menge sehr ähnlichen Materials trainiert, während den Hörern nur einige Beispiele vorgeführt wurden. Ihnen wurde nicht näher erläutert, wie die Begriffe *Akzent* und *Phrasengrenzen* in diesem Experiment genau zu verstehen sind, während das Trainingsmaterial des Detektors prosodische Etiketten umfaßte, die zumindest gewissenhaft erstellt wurden. Allerdings waren die Versuchsteilnehmer sowohl mit der CD1 der Verbmobil-Stichprobe (und der badischen Färbung einiger Turns) als auch mit den Begriffen *Akzent* und *Phrasengrenze* im allgemeinen vertraut. Darüber hinaus basiert die menschliche Sprachkompetenz auf einer weit größeren „Stichprobe“ als die jedes ASV-Systems. Sie enthält z.B. die Erwartung, daß nach einer gewissen Redezeit eine Phrasengrenze kommen muß, während dies im Detektor mit seinem Kontext von nur 4 Silben nicht modelliert ist.

Nimmt man an, daß die Akzentuiertheit einer Silbe graduell ist, könnte man annehmen, daß der Detektor dem Hörer gegenüber im Vorteil ist, weil seine Grenze zwischen den beiden Akzentklassen durch sein Training besser an diese spezielle Stichprobe angepaßt ist. Die Akkuratheit als Maß nivelliert allerdings die „Neigung“, sich eher für AKZENTUIERT oder für NICHT-NICHT-AKZENTUIERT zu entscheiden, auch bei den Hörern untereinander (in diesem Punkt war die Streuung viel größer als bei den Akkuratheiten selbst).

Die geringe Stichprobengröße in dieser Versuchsreihe schränkt die Aussagekraft ebenfalls ein; gemessen am Aufwand ist die Stichprobe allerdings groß.

Die Akkuratheit des Detektors gemessen an der gesamten Teststichprobe, nämlich 33.6% für Phrasengrenzen und -2.3% für Akzente, weicht zwar deutlich von den Abbildungen 8.1 und 8.2 gezeigten Werten ab, dies zeigt aber nur, daß die hier gewählte Stichprobe nicht repräsentativ im Sinne der „Schwierigkeit“ ist.

Unter diesen Einschränkungen kann man sagen, daß der Detektor ungefähr gleich gut wie der durchschnittliche Hörer abschneidet, etwas besser bei der Akzenterkennung, etwas schlechter bei der Phrasengrenzenerkennung (was an der strengeren Wertung der Einfügungsfehler beim Phrasengrenzendetektor liegen mag).

Das Überraschende daran ist nicht das relativ gute Abschneiden des Detektors, sondern das relativ schlechte Abschneiden der Hörer. Es ist erstaunlich

```
<A> < Klicken> also am dritten Mai     um viertel vor drei
                                   B3                     B3
                              |     |    |
                      1       1     2    4       3   1

kommen Sie zu mir ins B"uro     <P> <Schmatzen> alles klar
                            B3                             B3
                             |
          1                 11                       1     11
```

Abbildung 8.3: *Vergleich der automatisch detektierten Phrasengrenzen mit den Hörerurteilen am Beispiel des Turns 001k/nbs1k020a: In der 1. Zeile die Orthographie, in der 2. die Handetiketten (**B2** und **B9** kamen hier nicht vor), in der 3. die vereinfachten Detektorausgaben (dargestellt als „|“), in der 4. die Anzahl der Höreretiketten für „Phrasengrenze“ (maximal 11).*

schwierig, in delexikalisierter Sprache Akzente und Phrasengrenzen zu hören, zumindest die gleichen, die man in der Sprache ohne Delexikalisierung hören würde.

In Abbildung 8.3 sind die Hörer-Etiketten für Phrasengrenzen (hier in Sägezahnsignalen markiert) den Ausgaben des Detektors an einem Beispiel gegenübergestellt, an dem man sehen kann, daß die „Fehler“ der Hörer denen des Detektors ähnlich sind: Die turnfinale Grenze [3] und die mit der Pause (von 400 ms) nach *Büro* wurde von allen Hörern wahrgenommen. Die Grenze nach *Mai* wurde nur von zwei, die nach *drei* von keinem Hörer erkannt. Der Detektor fügte nach *um* eine Grenze ein (mit geringerer Konfidenz), wo auch vier der elf Hörer eine Grenze wahrgenommen hatten.

Es liegt der Schluß nahe, daß die Wahrnehmung prosodischer Kategorien wie *Akzent* und *Phrasengrenze* stark von linguistischen Vorerwartungen geprägt sind.

Dies deckt sich auch mit der Tatsache, daß die in [Kom95, Tabelle 5] genannte „Erkennungsrate“[4] für **B3**-Grenzen eines n-Gramms über Wortkategorien und prosodischen Etiketten (siehe Abschnitt 7.2) allein schon höher ist als die Erkennungsrate, die dort nur mit akustisch-prosodischen Merkmalen erzielt wurde (siehe Abschnitt 6.4): Das Beispiel in Abbildung 8.3 zeigt, daß zwischen *um* und *viertel vor drei* offensichtlich ein prosodisch markierter Einschnitt vorliegt, der aus linguistischer Sicht wenig sinnvoll ist. Deshalb haben die Etikettierer in Braunschweig die Grenze nach *Mai* wahrgenommen. Dieses Wissen läßt sich schon mit

[3] Der Detektor fand die turnfinale Grenze nicht, weil der Silbenkern in *klar* nicht detektiert wurde.

[4] Da es sich um Vorhersage handelt, wäre „Trefferquote“ passender.

einem n-Gramm modellieren, wenn genügend Trainingsmaterial vorhanden ist.

Zu Beginn dieses Kapitels wurde die Frage gestellt, wie gut sich die durch die Handetiketten gegebenen prosodischen Klassen trennen lassen, wenn nur prosodische Information zur Verfügung steht, aber keine Wortinformation und andere linguistische Information.

Unter prosodischen Informationen wurden die aus dem akustischen Signal extrahierbaren Größen Grundfrequenz und Energie verstanden, aus denen sich Pausen und Silbenkerne bestimmen lassen (mit den in Abschnitt 5.3.1 genannten Einschränkungen), und daraus wiederum Dauermerkmale (absolute Werte) und die Sprechgeschwindigkeit. Zu den prosodischen Informationen zählen auch Wissensquellen, die sich mit diesen Parametern darstellen lassen, hier in Form von Verteilungsannahmen der Klassifikatoren.

Liegt hingegen auch Wortinformation in Form einer Wortsegmentierung oder eines Worthypothesengraphen vor, können absolute Dauerwerte auf intrinsische Lautdauern normiert werden, die Sprechgeschwindigkeit kann zuverlässiger bestimmt werden, und außerdem können prosodische Ereignisse allein aufgrund des oben erwähnten n-Gramms als zusätzlicher Wissensquelle vorhergesagt werden.

Auch wenn sich die Frage der Trennbarkeit prosodischen Klassen mit rein akustisch-prosodischen Merkmalen hier nicht abschließend beantwortet läßt, geben die Ergebnisse der Perzeptionsexperimente Anlaß zur Vermutung, daß mit dem Ansatz, der in dieser Arbeit verfolgt wurde, die Grenze des Möglichen nahezu erreicht ist.

Zwar können die Einbeziehung von mehr Kontext und die Verwendung globaler Merkmale[5] die Erkennungsleistung durchaus noch verbessern. Die ohne Wortinformation schwierig zu bestimmende, aber perzeptiv selbst in delexikalisierter Sprache relativ gut einzuschätzende lokale Sprechgeschwindigkeit könnte beispielsweise eine nützliche Normierungsgröße ergeben. Aber substantielle Verbesserungen lassen sich wahrscheinlich nur dann erzielen, wenn auch Wortinformation mitverwendet wird.

[5] Die Vorgabe der inkrementellen Arbeitsweise im Verbmobil-Architektur-Teilprojekt bedingte, möglichst wenig rechten Kontext und damit keine globalen Merkmale zu verwenden.

Kapitel 9

Ausblick

Künftige Verbmobil-Systeme sollen wie das INTARC-System inkrementell arbeiten. Dazu ist die prosodische Segmentierung eines Dialogbeitrags noch vor der Worterkennung in sog. Turnsegmente vorgesehen. Bei den meisten Sprechern wäre diese Segmentierung aufgrund einer Pausendetektion möglich. Zur Segmentierung längerer Turns, die keine Sprechpausen enthalten, bietet sich das hier vorgestellte Verfahren zur Phrasengrenzendetektion an.

Die nächstliegende Anwendung für Prosodieerkennung ohne Wortinformation ist jedoch die Worterkennung selbst. Es gibt noch keine Untersuchungen darüber, wie sich prosodisch detektierte Akzente in der Worterkennungsphase ausnutzen lassen. Die diesbezüglichen Arbeiten im Verbmobil-Architektur-Teilprojekt sind leider eingestellt worden, obwohl der Ansatz von der Zielsetzung her interessant war. Das Thema wird in dem mittlerweile bewilligten DFG-Projekt „Prosodiegetriebene Erkennung spontaner Sprache mit unbegrenztem Wortschatz" am IKP wieder aufgegriffen werden.

In diesem Projekt soll ein Worterkenner entwickelt werden, der Wörter aus Wortteilen zusammensetzt, die etwa Morpheme größer oder gleich einer Silbe sind, wie *offen-bar* oder *April-woche*, um die Beschränkung auf ein festes Inventar von Wortvollformen aufzuheben. Die Wortteile sollen so beschaffen sein, daß sie zum einen akustisch gut detektierbar sind, zum anderen sollen statistische Modelle für Folgen von Wortteilen möglichst geringe Perplexitäten aufweisen. Dabei wird die Akzenterkennung eine große Rolle spielen. Ausgehend von den Akzentetiketten zum Verbmobil-Korpus sollen in einem Bootstrapping-Verfahren größere Datenmengen automatisch so etikettiert werden, daß die entstehenden Akzentklassen einerseits akustisch evident sind, andererseits die Korrelation mit den Wortteilen auch statistisch gut modellierbar ist. Ausschlaggebend soll aber nicht die akustische oder linguistische Relevanz der Akzentklassen sein, sondern der Beitrag der Akzenterkennung zur Disambiguierung und zur Einschränkung des Suchraums bei der Worterkennung.

Wird prosodische Information erst nach der Worterkennung, also für die syntaktische und semantische Analyse, eingesetzt, sollte die Wortinformation zur Steigerung der Erkennungsrate auch genutzt werden. Die Anwendungsmöglichkeiten der Prosodieerkennung können durch bessere Modellierung der Prosodie auf der linguistischen Ebene noch erheblich vergrößert werden. In einem Übersetzungssystem wie Verbmobil schließlich sollte die Prosodie genutzt werden, um die synthetisierte Übersetzung der Eingabesprache möglichst ähnlich klingen zu lassen.

Die Integration prosodischer Information in ASV-Systeme steht erst am Anfang, die paralinguistischen und indexikalischen Funktionen der Prosodie wurden dabei bisher noch gar nicht genutzt. Hier bieten sich reichlich Möglichkeiten für innovative Forschungsaktivitäten.

Kapitel 10

Zusammenfassung

Schon seit den 50er Jahren ist die **a***utomatische* **S***prach***v***erarbeitung* (ASV) Gegenstand intensiver Forschung. Ihr Ziel ist die Vereinfachung der Mensch-Maschine-Kommunikation. Die Fortschritte auf diesem Gebiet konnten mittlerweile in einige kommerzielle Produkte umgesetzt werden wie Kommandosysteme zur Gerätesteuerung oder Diktiersysteme, die mit gewissen Einschränkungen eine Schreibkraft ersetzen können. Für sprachverstehende Systeme, wie Dialogsysteme zur Flugbuchung oder Hotelreservierung, gibt es mittlerweile Forschungsprototypen. Neben der Erkennung der Wörter leisten sie auch eine grammatikalische und inhaltliche Analyse des Gesprochenen.

Ein Phänomen, das in der ASV bisher kaum berücksichtigt wurde, in der sprachlichen Kommunikation zwischen Menschen aber eine große Rolle spielt, ist die Prosodie. Zur Prosodie gehören die Betonung inhaltlich wichtiger Wörter (Akzentuierung), die Markierung von Einschnitten im Redefluß (Phrasierung) und des Satzmodus. Phrasierung und Satzmodus können in der Schriftsprache durch Satzzeichen markiert werden, Akzentuierung durch Unterstreichen oder Wechsel der Schriftart. In gesprochener Sprache werden diese Funktionen realisiert u.a. durch Änderung der Tonhöhe, der Lautheit und der Sprechgeschwindigkeit sowie durch Dehnung von Silben und durch Pausensetzung. Nach einer Klärung der Terminologie wurden die wichtigsten Funktionen der Prosodie erläutert und Beispiele für Fälle gegeben, in denen die Prosodie das Verständnis einer gesprochenen Äußerung erleichtert oder zwischen Mehrdeutigkeiten auflösen kann.

Verglichen mit der Worterkennung steht die Prosodieerkennung noch am Anfang der Entwicklung. Ein wesentlicher Grund dafür ist, daß es für eine bestimmte prosodische Funktionsdomäne kein vorgegebenes Inventar von Klassen gibt, das den Wörtern in einem Lexikon entspricht. Anzahl und Art der Klassen, z.B. für die Funktionsdomäne „Akzentuierung", müssen abhängig von der geplanten Weiterverarbeitung in anderen Modulen bestimmt werden. Eine weitere Schwierigkeit besteht darin, daß es keine eindeutige Beziehung gibt zwischen den funktionalen Klassen und den formalen Klassen, die durch prosodische Ausdrucksmittel (wie Tonhöhe und Dauer) bestimmt werden. Zwar können auch Wörter leicht unterschiedlich ausgesprochen werden, im Bereich der Prosodie ist die Beziehung zwischen Funktion und Form jedoch weit stärker geprägt durch sprecherspezifische Eigenheiten, durch gegenseitige Beeinflussung, Mehrdeutigkeit und Fakultativität der Ausdrucksmittel. Aus diesen Gründen ist es auch schwierig und sehr zeitaufwendig, große und konsistente Mengen prosodischer Etiketten manuell zu erstellen, wie sie für überwachte automatische Lernverfahren notwendig sind.

In früheren Studien wurde anhand kleiner, ausgewählter oder extra für diesen Zweck hergestellten Stichproben untersucht, welche prosodischen Phänomene auftreten und durch welche Mittel die verschiedenen prosodischen Funktionen ausgedrückt werden. Erst in den letzten Jahren geht der Trend hin zur statistischen Modellierung größerer, realistischerer Datenmengen. Diese Untersuchungen zeigen, daß die Prosodieerkennung zur Verbesserung von ASV-Systemen beitragen kann.

Das erste System, in das eine Prosodiekomponente integriert wurde, ist das Zugauskunftssystem EVAR [MKE+94], das eine Satzmodusklassifikation zur Dialogsteuerung einsetzt. Mit dem innerhalb des Verbmobil-Projekts entstandene Forschungsprototyp [Kro96] und dem experimentellen INTARC-System liegen die ersten sprachverstehenden Systeme vor, in denen die Prosodieerkennung zur linguistischen Analyse beiträgt. Die hier vorgestellte Arbeit befaßt sich mit der Entwicklung von Erkennern für Akzente, Phrasengrenzen und Satzmodus, sowie deren Integration in das INTARC-System. Ein wesentlicher Unterschied des INTARC-Systems gegenüber dem Forschungsprototyp besteht in der inkrementellen Verarbeitung, d.h. die Analyse erfolgt schritthaltend mit dem Eintreffen des Sprachsignals, ohne das Äußerungsende abzuwarten.

Das Problem der Prosodieerkennung ist ein Spezialfall eines Mustererkennungsproblems. Die Erkennung besteht aus den Schritten Merkmalgewinnung und Klassifikation. Die Merkmalgewinnung besteht in der Transformation des Musters in einen Merkmalvektor. Dabei soll zum einen die Datenmenge reduziert werden, zum anderen soll die hinsichtlich der Klassifikation relevante Information möglichst erhalten bleiben. Die Merkmale sind daher so zu wählen, daß sie die Klassen möglichst gut trennen, d.h. Muster aus gleichen Klassen sollen Häufungsgebiete im Merkmalsraum bilden. Da es kaum systematische Methoden gibt, in diesem Sinn gute Merkmale zu finden, werden meist mithilfe von Expertenwissen über den Problemkreis heuristische Methoden angewendet. Im Fall der Prosodieerkennung beziehen sich diese Merkmale auf die Dauer (von Silben, Pausen etc., allgemein: auf die zeitliche Struktur), auf die Grundfrequenzkontur als akustisches Korrelat des Intonationsverlaufs und auf die Energiekontur als akustisches Korrelat des Lautheitsverlaufs.

Die Klassifikation ist die Abbildung des Merkmalvektors auf einen Klassennamen. Die optimale Entscheidungsregel, die die Fehlerwahrscheinlichkeit minimiert, wählt die Klasse mit der maximalen A-Posteriori-Wahrscheinlichkeit. Der statistische Klassifikator schätzt diese A-Posteriori-Wahrscheinlichkeiten mithilfe einer Trainingsstichprobe aus vorklassifizierten Merkmalvektoren, die auf manuell erstellten Etiketten basiert. Dem Normalverteilungsklassifikator liegt die Annahme zugrunde, daß die Merkmale klassenweise normalverteilt sind; das Training besteht in der Schätzung der klassenbedingten Mittelwertvektoren und Kovarianzmatrizen.

Den Prosodieerkennern in dieser Arbeit liegt der Normalverteilungsklassifikator zugrunde. Er hat gegenüber anderen Klassifikatortypen den Vorteil einer kurzen und numerisch relativ problemlosen Trainingsphase; darüber hinaus berechnet er tatsächlich Wahrscheinlichkeitsmaße, so daß bei der Verknüpfung mit Wahrscheinlichkeitsmaßen aus anderen Modulen in einem ASV-System Probleme der Metrik vermieden werden. Der möglichen Inadäquatheit der Normalverteilungsannahme kann durch explizite oder implizite Clusterung sowie durch Vorauswahl der am besten geeigneten Merkmale begegnet werden.

Zum Training der Klassifikatoren sind handetikettierte Sprachdaten in ausreichender Menge nötig. Für Voruntersuchungen zur Akzent- und Satzmoduserkennung wurden die Sätze eines Sprechers aus dem gelesenen Phondat-II-Zugauskunftskorpus verwendet. Für 134 Sätze wurde der Satzmodus aus der Orthographie abgeleitet. Zu 60 Sätzen, für die eine manuelle Phonemsegmentierung vorhanden war, wurden sowohl am IKP als auch an der TU Braunschweig Akzentetiketten erstellt.

In 27 Dialogen bzw. 716 Redebeiträgen des spontansprachlichen Verbmobil-Korpus wurden an der TU Braunschweig Etiketten für Akzente, Phrasengrenzen und Satzmodus erstellt. Neben diesen funktionalen Etiketten wurde auch die Intonation nach dem ToBI-System etikettiert. Diese prosodischen Handetiketten beziehen sich auf Wörter und Wortgrenzen, die zum Training verwendeten vorklassifizierten Merkmalvektoren aber auf Silbengrenzen und Frames (Sprachsignalabschnitte fester Länge, ein Phonem ist in der Regel mehrere Frames lang). Zur Abbildung der Handetiketten auf diese kleineren Einheiten war eine Phonemsegmentierung notwendig. Manuelle Phonemsegmentierung kam aus Zeitgründen nicht in Frage, deshalb wurden mehrere schrittweise verbesserte automatische Phonemsegmentierungen erzeugt.

Die oben skizzierte Vorgehensweise eines Mustererkennungssystems geht von der Klassifikation einfacher Muster aus, d.h. jedes Muster wird als Ganzes klassifiziert, nachdem seine Merkmale bestimmt wurden. Falls es sich bei dem Muster um ein Sprachsignal handelt, das genau einen Satz darstellt (wie beim Phondat-II-Zugauskunftskorpus), kann das ganze Muster in eine Satzmodusklasse abgebildet werden. Ein Satz enthält i.a. jedoch mehrere akzentuierte Silben, und die Dialogbeiträge im Verbmobil-Korpus bestehen i.a. aus mehreren Sätzen, die sich wiederum aus mehreren Intonationsphrasen zusammensetzen können. Zur Detektion von Akzenten und Phrasengrenzen müssen vor der Klassifikation die zu klassifizierenden Einheiten festgelegt werden. Da die hier vorgestellten Detektoren keine Hypothesen des Worterkenners als Eingabe verwenden, liegen die betreffenden Einheiten, nämlich Silben bzw. Wortgrenzen, nicht von vornherein fest. Zur Akzent- und zur Phrasengrenzendetektion wurde dieses Problem auf zwei unterschiedliche Arten gelöst: Zur Akzentdetektion werden zunächst Frames klassifiziert, wobei Nichtvokale von Vokalen unterschieden werden. Bei Vokalen werden weiterhinh die vier möglichen Akzentstufen der betreffenden Silbe unterschieden, nämlich nicht-akzentuiert, Nebenakzent, Hauptakzent und Emphase/Kontrast. Nach der Klassifikation werden Frames, die zum Vokal in einer akzentuierten Silbe gehören, zusammengefaßt. Zur Phrasengrenzendetektion dagegen wurden die zu klassifizierenden Einheiten, die Silbengrenzen (die eine Obermenge der Wortgrenzen sind), vorab durch eine Silbenkerndetektion grob bestimmt; mit grob ist hier gemeint, daß der Bereich zwischen zwei Silbenkernen (bei korrekter Silbenkerndetektion) genau eine Silbengrenze enthält.

Damit die Akzentdetektion im Zeitbereich von Frames möglich wird, muß der Merkmalvektor für einen Frame den Energie- und Grundfrequenzverlauf in dessen näherer zeitlicher Umgebung beschreiben. Zur Merkmalgewinnung wird nach der Grundfrequenzanalyse die Grundfrequenzkontur in stimmlosen Bereichen interpoliert. Der dafür entwickelte, auf Digitalfiltern basierende Interpolierer arbeitet inkrementell und ist auf die nachfolgende Dekomposition zugeschnitten, die ebenfalls auf Digitalfiltern beruht. Durch die Dekomposition wird die interpolierte Grundfrequenzkontur in einen lokalen und einen globalen Anteil zerlegt, wobei der globale Anteil die Tonhöhenbewegung stark geglättet wiedergibt, der lokale Anteil dagegen nur Tonhöhenbewegungen relativ zum globalen Anteil. Die Schwankungen des lokalen Anteils liegen etwa im Zeitbereich von Silben; die Überlagerung beider Komponenten ergibt (fast) wieder die ursprüngliche Grundfrequenzkontur. Zuvor wurde zum Zweck der Interpolation und Dekomposition eine inkrementell arbeitende Implementierung des Intonationsmodells von Fujisaki erstellt, die sich jedoch als nicht robust genug erwies. Um zu jedem Zeitpunkt nicht nur Information über absolute und relative Tonhöhe zu erhalten, sondern auch über die Änderung derselben, werden die Grundfrequenzkomponenten noch zeitlich abgeleitet.

Neben diesen Grundfrequenzmerkmalen werden für jeden Frame drei Energiemerkmale berechnet: In drei Frequenzbändern werden die Koeffizienten der Kurzzeit-DFT aufsummiert und zeitlich median-geglättet. Diese Merkmale wurden in einer anderen Arbeit [Nöt91] zur Silbenkerndetektion verwendet; sie sind nicht nur ein Maß für die momentane Lautheit, sondern eignen sich auch zur Unterscheidung von Vokalen und Nichtvokalen. Grundfrequenz- und Energiemerkmale ergeben zusammen die prosodischen Basismerkmale. Sie beschreiben die spektrale Energieverteilung und den Grundfrequenzverlauf in der näheren zeitlichen Umgebung des Frames und ändern sich daher von Frame zu Frame nur langsam.

Die Energiemerkmale dienen gleichzeitig der Silbenkerndetektion. Zur Phrasengrenzendetektion wird jeweils der Bereich zwischen zwei Silbenkernen detektiert. Dazu wird ein Analysefenster aus — soweit möglich — vier Silben betrachtet. Die Basismerkmale an den vier Silbenkernmitten werden zu einem komplexen Merkmalvektor zusammengefaßt, erweitert um die Länge der Silbenkerne und ihre Abstände als Dauermerkmale. An den Rändern einer Äußerung werden entsprechend kleinere Analysefenster verwendet.

Zur Akzenterkennung wurde ein Normalverteilungsklassifikator darauf trainiert, die Basismerkmale in 5 Klassen abzubilden. Die Klassifikatorentscheidung für jeden Frame wird bei der Nachverarbeitung auf die beiden Oberklassen „Vokal in einer akzentuierten Silbe ja/nein“ abgebildet, anschließend werden „akzentuierte Bereiche“ aus weniger als 9 Frames weggeglättet. Bei silbenweiser Auswertung ergibt sich für die Verbmobil-Stichprobe eine Erkennungsrate von 74%.

Die Phrasengrenzendetektion erfolgt silbenweise, die Merkmalvektoren beschreiben Grundfrequenz, Energie und Dauerverhältnisse in den vier umgebenden Silben. Es werden vier Phrasengrenzenklassen unterschieden: starke Grenzen (**B3**), schwache Grenzen (**B2**), irreguläre Grenzen (**B9**) und normale Silbengrenzen (**B0**). An **B3**-Grenzen wird zusätzlich der Satzmodus klassifiziert, wobei zwischen Fragen, Aussagen und Weiterführungen unterschieden wird. Zusammen mit den Tonetiketten gibt es für jede Silbengrenze 13 verschiedene Kombinationen aus Phrasengrenzentyp, Satzmodus und Tonetikett. Ein Normalverteilungsklassifikator bildet den Merkmalvektor in eine der 13 Klassen ab, diese wird anschließend der Phrasengrenzen- und der Satzmodusklasse zugeordnet.

Die Auswertung erfolgte wortweise und SILBEN-weise (siehe Abschnitt 6.3): Die SILBEN-weise Auswertung bezieht sich auf die detektierten Silben (-grenzen), die Fehler der Silbenkerndetektion werden hier nicht berücksichtigt, während sich die wortweise Auswertung an den tatsächlichen Wortgrenzen orientiert. Die Erkennungsrate für Phrasengrenzen und Satzmodus im INTARC-1.3-System bei SILBEN-weiser Auswertung beträgt 81% bzw. 86% (jeweils 4 Klassen), bei wortweiser Auswertung ist die Erkennungsrate für den Satzmodus 82%. Umgerechnet auf das Zweiklassenproblem **B3**/Nicht-**B3** ergibt sich eine Erkennungsrate von 84.0%.

Das verbesserte INTARC-2-System hat bei SILBEN-weiser Auswertung eine Erkennungsrate für Phrasengrenzen und Satzmodus von 86% bzw 91%; für das Zweiklassenproblem **B3**/Nicht-**B3** 91%.

Der Verbmobil-Prototyp kann aufgrund der englischen Schlüsselworterkennung und des Dialoggedächtnisses eine grobe, dialogaktbasierte Übersetzung ins Deutsche erstellen. Dabei entsteht ein Problem, wenn ein Dialogbeitrag aus mehreren Dialogschritten besteht, wie eine Terminablehnung gefolgt von einem alternativen Terminvorschlag. Daher wurde untersucht, ob sich die **B3**-Grenzen zur Dialogaktsegmentierung eignen. Diese Voruntersuchung wurde an deutschen Sprachdaten durchgeführt. Die Erkennungsrate für Dialogaktgrenzen betrug (wortweise ausgewertet) 87%. Die Schwelle zwischen **B3** und Nicht-**B3** konnte verschoben werden, um Einfügungs- und Auslassungsfehler gegeneinander auszubalancieren. Zum Vergleich wurden auch die **B3**-Grenzen des Erlanger Prosodiemoduls, das Bestandteil des Verbmobil-Prototyps ist, ausgewertet. Obwohl dieses Prosodiemodul einen vollständigen Worthypothesengraph als Eingabe benötigt, zeigte es ein ungünstigeres Verhältnis von Einfügungs- zu Auslassungsfehlern (Akkuratheit von 10% gegenüber 32%).

Die Phrasengrenzen- und Satzmodus-Erkennung wurde in das INTARC-System integriert. Die prosodisch detektierten Phrasengrenzen dienten zur Suchraumbeschränkung im Syntaxparser, **B3**-Grenzen und der Satzmodus zur Reduktion der Lesarten im Semantikparser.

Zur Koppelung des Phrasengrenzenerkenners mit dem Syntaxparser wurden die a posteriori Wahrscheinlichkeiten der Phrasengrenzenhypothesen kombiniert mit den Wahrscheinlichkeiten, die sich aus einem statistischen Sprachmodell für Folgen von Wortkategorien und Phrasengrenzen ergeben. Dieser sog. Prosodiefaktor steuerte zusammen mit der akustischen Bewertung der Worthypothesen, der Wort-Bigramm-Bewertung und der Grammatik-Bewertung die Präferenzen bei der Syntaxanalyse. Dadurch konnte zwar nicht eine bestimmte von mehreren möglichen syntaktischen Analysen für dieselbe Wortfolge prosodisch bevorzugt werden, da die Grammatik selbst keine prosodischen Phrasengrenzen modelliert, aber der Suchraum konnte dadurch beschränkt werden, indem Worthypothesenfolgen schlechter bewertet werden, die nicht kompatibel zu den detektierten Phrasengrenzen und den Vorhersagen des statistischen Sprachmodells für Wortkategorien und Phrasengrenzen sind. Dieser Effekt verstärkt sich, wenn das Sprachmodell mit den detektierten statt den handetikettierten Phrasengrenzen trainiert wird.

Da die Syntaxanalyse als Strahlsuche implementiert wurde, können die prosodisch detektierten Phrasengrenzen neben der Suchraumbeschränkung auch die Worterkennungsrate verbessern, die sich hier auf die Anzahl der von Beginn der

Äußerung an syntaktisch analysierbaren Wörter bezieht.

Bei einer Teststichprobe, die aus Dialogbeiträgen mit einfacher syntaktischer und semantischer Struktur bestand, verminderte der Prosodiefaktor die Anzahl der Kantenpaare (als Maß für Laufzeit) um 40% bei gleichzeitiger Erhöhung der Worterkennungsrate von 84% auf 86%. Bei einer schwierigeren Teststichprobe, vier Dialogen aus dem Verbmobil-Korpus, erhöhte die Prosodieerkennung bei gleicher Anzahl von Kantenpaaren die Worterkennungsrate von 48.2% auf 53.2%.

Die gliedernde Funktion der prosodischen Phrasengrenzen wird im INTARC-System erst bei der semantischen Analyse ausgenutzt. Die hier verwendete Grammatik, die nicht Sätze, sondern ganze Dialogbeiträge beschreibt, wurde so erweitert, daß satzverbindende bzw. segmentverbindende Regeln eine Phrasengrenze erfordern, während segmentinterne Regeln Phrasengrenzen ausschließen. Zusammen mit Restriktionen, die den Satzmodus betreffen, konnte die Prosodieerkennung in den Dialogen aus dem Verbmobil-Korpus die mittlere Anzahl der Lesarten um 25% verringern.

Ziel der Arbeit war es, die prosodische Ausprägung der linguistischen Konzepte Akzent, Phrasengrenze und Satzmodus so zu modellieren, daß eine automatische Klassifikation möglich wird. Das Problem besteht zum einen darin, daß die linguistische Funktion, z.B. die Phrasierung, mit der prosodischen Realisierung in keiner eindeutigen Beziehung steht. Zum anderen sind die prosodischen Handetiketten, die zum Training der Klassifikatoren dienten, zwar an der perzeptiven Wahrnehmung von Intonation, Lautheit, Lautdehnung, etc. orientiert; bis zu einem gewissen Grad fließen in sie jedoch auch syntaktische und semantische Vorerwartungen der Etikettierer ein. Zudem setzt die Wahrnehmung von Lautdehnung die Kenntnis von intrinsischen Lautdauern und die Erkennung der gesprochenen Wörter voraus. Es stellt sich die Frage, wie gut sich die durch die Handetiketten gegebenen prosodischen Klassen auch ohne ohne Wortinformation, allein mit akustischen Merkmalen, trennen lassen.

Dazu wurde in einer abschließenden Untersuchungsreihe ein kleiner Teil des Verbmobil-Korpus delexikalisiert, d.h. so verändert, daß die prosodischen Charakteristika wie Intonation, Lautheit und Sprachrhythmus erhalten bleiben, die Verständlichkeit aber verloren geht. 11 Hörer wurden dann gebeten, Akzente und Phrasengrenzen zu etikettieren. Es wurden zwei unterschiedliche Verfahren zur Delexikalisierung entwickelt. Erst wurden alle stimmhaften Bereiche durch ein Sägezahnsignal gleicher Grundfrequenz und Energie ersetzt, stimmlose Bereiche durch Stille. Das Problem für die Etikettierer bestand darin, daß die zu klassifizierenden Einheiten, Silben und Silbengrenzen, zwar akustisch meist noch wahrnehmbar, aber im Signal nicht mehr vorhanden waren. Das Plazieren und Überprüfen der Etiketten erfolgte daher rein akustisch und per Tastendruck.

Bei der zweiten Methode werden nach einer automatischen Phonemsegmentierung die Phoneme zufällig, aber unter Berücksichtigung der deutsche Phonotaktik durch ähnliche Phoneme ersetzt. Diese Phonemfolge wird durch einen Sprachsynthesizer gesprochen, wobei die Grundfrequenz- und Dauerwerte dem ursprünglichen Sprachsignal entnommen werden. Das so entstehende Sprachsignal wurde verschriftet und konnte auf herkömmliche Art etikettiert werden.

Die 11 Hörer markierten in 20 Dialogbeiträgen die Phrasengrenzen und in 22 Phrasen Akzente. Sie erreichten bei den Sägezahnsignalen eine Akkuratheit von 33% bzw. 21%, bei der resynthetisierten Sprache, die noch phonotaktische Information enthält, 44% bzw. 23%. Die Akkuratheit des Phrasengrenzen- und Akzenterkenners ist für diese Stichprobe 26% bzw. 29% und liegt damit für Phrasengrenzen etwas niedriger, für Akzente etwas höher. Dies läßt vermuten, daß sich die Erkennungsleistung ohne Verwendung von Wortinformation nicht mehr wesentlich steigern läßt.

Auch wenn sich bei der Prosodieerkennung ohne Wortinformation die Grenze des Möglichen abzeichnet, sind noch Verbesserungen denkbar. Wenn prosodische Information erst nach der Worterkennung eingesetzt wird, z.B. in der Syntaxanalyse, ist es natürlich besser, die Wortinformation mitzuverwenden; der Verbmobil-Prototyp hat z.B. gezeigt, daß die Syntaxanalyse erheblich verbessert werden kann, wenn prosodische Grenzen auch in der Grammatik modelliert werden. Dennoch eröffnen die hier vorgestellten Methoden neue Anwendungsmöglichkeiten in Situationen, in denen keine Wortinformation vorliegt: Künftige Verbmobil-Systeme sollen, wie das INTARC-System, inkrementell arbeiten. Dazu ist eine Vorsegmentierung des Sprachsignals nötig, die gegenwärtig nur durch Pausendetektion erfolgt. Auch gibt es noch keine Untersuchungen darüber, wie sich prosodisch detektierte Akzente in der Worterkennungsphase ausnutzen lassen. Dies wird ein Forschungsgegenstand in den nächsten Jahren werden.

Die Integration prosodischer Information in ASV-Systeme steht erst am Anfang; erste Erfolge geben Anlaß zur Vermutung, daß das Potential der Prosodieerkennung für die automatische Sprachverarbeitung bei weitem noch nicht ausgeschöpft ist.

Anhang A

Die SAMPA-Notation

Die SAMPA-Notation (Speech Assessment Methods Phonetic Alphabet) ist ein maschinenlesbares phonetisches Alphabet und besteht im wesentlichen aus einer Abbildung des Internationalen Phonetischen Alphabets (IPA) auf ASCII-Zeichen [SAM96]; zu den Modifikationen für das Verbmobil-Projekt siehe [Gib95].

Plosive:

Symbol	Wort	Transkription
p	Pein	paIn
b	Bein	baIn
t	Teich	taIC
d	Deich	daIC
k	Kunst	kUnst
g	Gunst	gUnst

Der Glottalverschluß:

?	Verein	fE6?aIn

Frikative:

f	fast	fast
v	was	vas
s	Tasse	tas@
z	Hase	ha:z@
S	waschen	vaS=n
Z	Genie	Zeni:
C	sicher	zIC6
j	Jahr	ja:6
x	Buch	bu:x
h	Hand	hant

Die Sonoranten bestehen aus den drei Nasalen /m/, /n/ und /N/ sowie den Liquiden /l/ und /R/. Falls das orthographische *r* gerollt ist, wird es mit /r/ transkribiert; zum vokalischen *r* siehe unten.

m	mein	maIn
n	nein	naIn
N	Ding	dIN
l	Leim	laIm
r	Reim	raIm

kurze Vokale:

I	Sitz	zIts
E	Gesetz	g@zEts
a	Satz	zats
O	Trotz	trOts
U	Schutz	SUts
Y	hübsch	hYpS
9	plötzlich	pl9tslIC

lange Vokale:

i:	Lied	li:t
e:	Beet	be:t
E:	spät	SpE:t
a:	Tat	ta:t
o:	rot	ro:t
u:	Blut	blu:t
y:	süß	zy:s
2:	blöd	bl2:t

Diphthonge

aI	Eis	aIs
aU	Haus	haUs
OY	Kreuz	krOYts

Schwache Vokale

@	bitte	bIt@
6	besser	bEs6

Literaturverzeichnis

[ACBD+95] F. Althoff, J. Carson-Berndsen, G. Drexel, D. Gibbon, K. Hübener, U. Jost, K. Kirchhoff, M. Pampel, A. Petzold, and V. Strom. Linguistische Worterkennung unter Berücksichtigung der Prosodie. Verbmobil Technisches Dokument Nr. 22, Universität Bielefeld, Universität Bonn, Universität Hamburg, 1995.

[ADD92] M. Adda-Decker and G. Decker. Experiments on stress-dependent phone modeling for continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, **1**, 561–564, San Francisco, 1992.

[Alt93] H. Altmann. Satzmodus. In J. Jacobs, A. v. Stechoow, W. Sternefeld, and T. Vennenmann, (Hrsg.), *Syntax – Ein Internationales Handbuch Zeitgenössischer Forschung – An International Handbook of Contemporary Research*, **1**, 1006–1029. Walter de Gruyter, Berlin, 1993.

[Aul84] A. M. Aull. Lexical stress and its application to large vocabulary speech recognition, 1984.

[Bat89a] A. Batliner. Fokus, Modus und die große Zahl. zur intonatorischen Indizierung des Fokus im Deutschen. In H. Altmann, A. Batliner, and W. Oppenrieder (Hrsg.), *Zur Intonation von Modus und Fokus im Deutschen*, 21–70. Niemeyer, Tübingen, 1989.

[Bat89b] A. Batliner. Zur klassifikation von fragen und nicht-fragen anhand intonatorischer merkmale. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, 335–338, Duisburg, 1989.

[Bat94] A. Batliner. Persönliche Mitteilung, Mai 1994.

[BBK95] J. Bos, A. Batliner, and R. Kompe. On the use of prosody for semantic disambiguation in Verbmobil. Verbmobil Memo Nr. 82, 1995.

[BE88] M. E. Beckman and J. Edwards. Articulatory timing and the prosodic interpretation of syllable duration. In *Phonetica*, **45**, 156–174, 1988.

[Bec86] M. E. Beckman. *Stress and Non-Stress Accent*. Foris Publication, Dordrecht, Holland/Riverton,USA, 1986.

[Bis92] K. Bishop. Modeling sentential stress in the context of a large vocabulary continuous speech recognizer. In *Proc. Int. Conf. on Spoken Language Processing*, **1**, 437–440, Banff, 1992.

[BKK+93] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. *The prosodic marking of accents and phrase boundaries: expectations and results*. In A. J. Rubino (Hrsg.), *New Advances and Trends in Speech Recognition and Coding*, **2** of *NATO ASI*, 89–92, Bubion (Granada), 1993.

[BOPSH90] J. Butzberger, M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. Isolated word intonation recognition using hidden markov models. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, **2**, 773–776, Albuquerque, 1990.

[BP90] J. Bear and P. J. Proce. Prosody, syntax, and parsing. In *Proc. Conf. of the Association for Computational Linguistics*, 17–22, Banff, 1990.

[BR94] A. Batliner and M. Reyelt. Ein Inventar prosodischer Etiketten für VERBMOBIL. Verbmobil Memo Nr. 33, TU Braunschweig, Ludwig-Maximilian-Universität München, 1994.

[Buß90] H. Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 1990.

[Cam92] N. Campbell. Prosodic encoding of English speech. In *Proc. Int. Conf. on Spoken Language Processing*, **1**, 663–666, Banff, 1992.

[Cam94] N. Campbell. Combining the use of duration and F0 in an automatic analysis of dialogue prosody. In *Proc. Int. Conf. on Spoken Language Processing*, 1111–1114, Yokohama, 1994.

[Cam95] N. Campbell. Prosodic influence on segmental quality. In *Proc. European Conf. on Speech Communication and Technology*, **2**, 1011–1014, Madrid, 1995.

[CL83] A. Cutler and D. Ladd. *Prosody: Models and Measurements.* Springer Verlag, Berlin, 1983.

[DKK95] A. K. Diagne, W. Kasper, and H. U. Krieger. Distributed parsing with HPSG grammars. In *Proceedings of the 4th International Workshop on Parsing Technologies, IWPT-95*, 79–86, 1995.

[dPS94] J.R. de Pijper and A.A. Sandermann. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. In *J. of the Acoustic Society of America*, **96**, 2037–2047, 1994.

[DPW96] F. Dellaert, T. Polzin, and A. Waibel. Detecting emotions in speech. In *Proc. Int. Conf. on Spoken Language Processing*, **3**, Philadelphia, 1996.

[DZ90] N. Daly and V. Zue. Acoustic, perceptual and linguistic analyses on intonation contours in human/machine dialogues. In *Proc. Int. Conf. on Spoken Language Processing*, 497–500, Kobe, 1990.

[EFK+92] W. Eckert, G. Fink, A. Kießling, R. Kompe, T. Kuhn, F. Kummert, M. Mast, H. Niemann, E. Nöth, R. Prechtel, S. Rieck, G. Sagerer, A. Scheuer, E. G. Schukat-Talamazzini, and B. Seestaedt. Evar: Ein sprachverstehendes Dialogsystem. In G. Görz (Hrsg.), *KONVENS 92*, Informatik aktuell, 49–58. Springer-Verlag, Berlin, 1992.

[EK96] A. Elsner and A. Klein. Erkennung des prosodischen Fokus und die Anwendung im dialogaktbasierten Transfer. Verbmobil Memo Nr. 107, Univ. Bonn, Univ. Hamburg, 1996.

[FHO79] H. Fujisaki, K. Hirose, and K. Otha. Acoustic features of the fundamental frequency contours of declarative sentences in japanese. In *Annual Bulletin of the Researche Institute for Logopedics and Phoniatrics*, **13**, 163–172. University of Tokyo, 1979.

[FN69] H. Fujisaki and S. Nagashima. A model for the synthesis of pitch contours of connected speech. In *Annual Bulletin of the Engineering Research Institute, Faculty of Engineering*, **28**, 53–60, Tokyo, 1969.

[Fry58] D. B. Fry. Experiments in the perception of stress. *Language and Speech*, 1:126–152, 1958.

[Fuj88] H. Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (Hrsg.), *Vocal physiology: voice production, mechanisms and funtions*, 347–355. Raven, New York, 1988.

[GE95] D. Gibbon and U. Ehrlich. Spezifikation für ein VERBMOBIL Lexikondatenbankkonzept. Verbmobil Memo Nr. 69, Universität Bielefeld, Daimler Benz AG, 1995.

[Geo93] E. Geoffrois. A pitch contour analysis guided by prosodic event detection. *Proc. European Conf. on Speech Communication and Technology*, II:793–796, 1993.

[Gib95] D. Gibbon. SAMPA-D-VMlex (deutsches SAMPA für die Verbmobil Lexikondatenbank. Universität Bielefeld, http://coral.lili.uni-bielefeld.de/Documents/sampa-d-vmlex.html, 1995.

[Gö88] G. Görz. *Strukturanalyse gesprochener Sprache — Ein Verarbeitungsmodell.* Addison-Wesley, Bonn, 1988.

[GRB+97] M. Grice, M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner. Consistency in transcription and labelling of German intonation with GToBI. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, 1997.

[HE94] D. Hirst and R. Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. In *Travaux de l'Institut de Phonétique d'Aix*, **15**, 75–85. Laboratoire Parole et Language URA CNRS 261, 1993–1994.

[Hes83] W. Hess. *Pitch Determination of Speech Signals*, **3** of *Springer Series of Information Sciences*. Springer Verlag, Berlin, 1983.

[HJH96] K. Hübener, U. Jost, and H. Heine. Speech recognition for spontaneously spoken german dialogs. In *ICSLP96*, 1996.

[Hü94] K. Hübener. Persönliche Mitteilung, Juli 1994.

[Hub88] D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation.* Dissertation, Universität Göteborg/Lund, 1988.

[HW94a] A. Hauenstein and H. Weber. An investigation of tightly coupled time synchronous speech language interfaces. In *Proceedings of the KONVENS 94*, Vienna, Austria, September 1994.

[HW94b] A. Hauenstein and H. Weber. An investigation of tightly coupled time synchronous speech language interfaces using a unification grammar. In Paul McKevitt (Hrsg.), *Proceedings of the Workshop on Integration of Natural Language and Speech Processing at AAAI 94*, 42–49, Seattle, August 1994.

[Jel89] F. Jelinek. Self organized language modeling for speech recognition. In A. Waibel and K.-F. Lee (Hrsg.), *Readings in Speech Recognition*. Morgan Kaufmann Publishers Inc., 1989.

[JKM+95] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogakte in Verbmobil. Verbmobil Technisches Dokument Nr. 26, Universität Hamburg, 1995.

[Jos96] U. Jost. Persönliche Mitteilung, März 1996.

[Kat87] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **35**, 400–401, 1987.

[Kie93] A. Kießling. Persönliche Mitteilung, Juni 1993.

[Kie97] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker Verlag, Aachen, 1997.

[Kir95] K. Kirchhoff. Two-level modelling of speech variant rules. Verbmobil Report Nr. 82, Universität Bielefeld, 1995.

[KK96a] W. Kasper and H. U. Krieger. Integration of prosodic and grammatical information in the analysis of dialogs. In *Proceedings of the 20th German Annual Conference on Artificial Intelligence, KI-96*, 1996. Springer: Lecture Notes in Computer Science, Berlin.

[KK96b] W. Kasper and H. U. Krieger. Modularizing codescriptive grammars for efficient parsing. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING-96*, 628–633, 1996.

[KKK+93] R. Kompe, A. Kießling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, and A. Batliner. Prosody takes over: A prosodically guided dialog system. In *Proc. European Conf. on Speech Communication and Technology*, **3**, 2003–2006, Berlin, 1993.

[KKN+92] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-based determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, II–17–II–20, 1992.

[KKSW96] W. Kasper, H. U. Krieger, J. Spilker, and H. Weber. From word hypotheses to logical form: An efficient interleaved approach. In D. Gibbon (Hrsg.), *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference*, 77–88. Mouton de Gruyter, Berlin, 1996.

[Kla76] D. H. Klatt. Linguistic use of segmental duration in English. In *J. of the Acoustic Society of America*, **59**, 1208–1221, 1976.

[KLP+94] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon. Handbuch zur Datenerhebung und Transliteration in TP14 von Verbmobil - 3-0. Verbmobil Technisches Dokument Nr. 11, Universität Kiel, 1994.

[Koh77] K. Kohler. *Einführung in die Phonetik des Deutschen*. Erich Schmidt Verlag, Berlin, 1977.

[Kom95] R. Kompe. Prosodic scoring of word hypotheses graphs. In *Proc. European Conf. on Speech Communication and Technology*, **2**, 1333–1336, Madrid, 1995.

[Kom97] R. Kompe. *Prosody in Speech Understanding Systems*, **1307** of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, New York, 1997.

[Kre82] J. Kreimann. Perception sentence and paragraph boundaries in natural conversation. In *Journal of Phonetics*, **10**, 163–175, 1982.

[Kro96] H.J. Kroner. Verbmobil Forschungsprototyp 0.3, Benutzerhandbuch. Verbmobil Technisches Dokument Nr. 69, DFKI GmbH Kaiserslautern, 1996.

[Lea73] W. Lea. Evidence that stressed syllables are the most readily decoded portions of continous speech. *J. of the Acoustic Society of America*, 55:436(A), 1973.

[Leh70] I. Lehiste. *Suprasegmentals.* MIT Press, Cambridge, MA, 1970.

[Leh79] I. Lehiste. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Ömann (Hrsg.), *Frontiers of speech communication research*, 191–201. Academic Press, New York, 1979.

[Leh94] M. Lehning. Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, 1257–1260, Dresden, 1994.

[Leh96] M. Lehning. *Statistische Verfahren zur Unterstützung der prosodischen Segmentierung und Etikettierung deutscher Spontansprache.* Shaker Verlag, Aachen, 1996.

[Lin63] B. Lindblom. Spectrographic study of vowel reduction. In *J. of the Acoustic Society of America*, **35**, 1773–1781, 1963.

[LKJ+85] P. Lieberman, W. Katz, A. Jongman, R. Zimmermann, and R. Miller. Measures if the sentence intonation of read and spontaneous speech in American English. In *J. of the Acoustic Society of America*, **77**, 648–657, 1985.

[LMS75] W. Lea, M. Medress, and T. Skinner. A prosodically guided speech understanding strategy. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **23**, 30–38, 1975.

[LW76] I. Lehiste and W. S-Y Wang. Perception of sentence boundaries with and without semantic information. In W. Dressler and O. Pfeiffer (Hrsg.), *Phonologica*, **19**, 277–283. Innsbruck, 1976.

[Mas95] M. Mast. Schlüsselwörter zur Detektion von Diskontinuitäten und Sprechhandlungen. Verbmobil Memo Nr. 57, Universität Erlangen, 1995.

[MG76] J.D. Markel and A.H. Gray. *Linear Prediction of Speech.* Springer Verlag, Berlin, 1976.

[MKE+94] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, and G. Sagerer. A speech understanding and dialog system with a homogeneous linguistic knowledge base. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **16**, 2, 179–194, 1994.

[MKH+96] M. Mast, R. Kompe, St. Harbeck, A. Kießling, H. Niemann, and E. Nöth. Dialog act classification with the help of prosody. In *Proc. Int. Conf. on Spoken Language Processing*, **3**, Philadelphia, 1996.

[Möb93] B. Möbius. *Ein quantitatives Modell der deutschen Intonation*. Max Niemeyer Verlag, Tübingen, 1993.

[NDK+94] H. Niemann, J. Denzler, B. Kahles, R. Kompe, A. Kießling, E. Nöth, and V. Strom. Pitch Determination Considering Laryngealization Effects In Spoken Dialogs. In *Proc. Int. Conf. on Neuronal Networks*, **7**, 4457–4461, Orlando, 1994.

[Nie83] H. Niemann. *Klassifikation von Mustern*. Springer Verlag, Berlin, 1983.

[Nie90] H. Niemann. *Pattern Analysis and Understanding*, **4** of *Series in Information Sciences*. Springer Verlag, Heidelberg, 1990.

[Nie89] H. Niemann. Vorlesung Mustererkennung 1, Wintersemester 1988/89.

[NK89] E. Nöth and R. Kompe. Verbesserung der Worterkennung mit prosodischer Information. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, 343–346, Duisburg, 1989.

[NNK+97] H. Niemann, E. Nöth, A. Kießling, R. Kompe, and A. Batliner. Prosodic Processing and its use in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1997. (to appear).

[Nöt91] E. Nöth. *Prosodische Information in der automatischen Spracherkennung*. Max Niemeyer Verlag, Tübingen, 1991.

[NS63] L. H. Nakatani and J. A. Schaffer. Prosodic cues for word perception. In *J. of the Acoustic Society of America*, **63**, 234–245, Philadelphia, 1963.

[Öhm67] S.E.G Öhman. Word and sentence intonation: a quantitative model. In *STL-QPSR*, **2–3**, 20–54, Stockholm, 1967. Royal Institute of Technology.

[ÖL66] S.E.G Öhman and J. Lindqvist. Analysis–by–synthesis of prosodic pith contours. In *STL-QPSR*, **4**, 1–6, Stockholm, 1966. Royal Institute of Technology.

[OPBW90] M. Ostendorf, P. J. Price, J. Bear, and C. W. Wightman. The use of relaive duration in syntactic disambiguation. In *Speech and Natural Language Workshop*, 26–31, Hidden Valley, Pensylvania, 1990. Morgan Kaufmann.

[O'S87] D. O'Shaughnessy. *Speech Communication.* Addison-Wesley Publishing Company, 1987.

[Ott93] K. Ott. Prosodisch basierte Dialogsteuerung in EVAR. Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, Januar 1993.

[Pät91] M. Pätzold. Nachbildung von Intonationskonturen mit dem Modell von Fujisaki — Implementierung des Algorithmus und erste Experimente mit ein– und zweiphrasigen Aussagesätzen. Magisterarbeit, Institut für Kommunikationsforschung und Phonetik, Universität Bonn, 1991.

[Pet95] A. Petzold. Strategies for focal accent detection in spontaneous speech. In *Proc. Int. Conf. on Phonetic Sciences*, **3**, 672 – 675, Stockholm, 1995.

[PH97] T. Portele and B. Heuft. Towards a prominence-based speech synthesis system. *Speech Communication*, 21:61–72, 1997.

[PHH97] T. Portele, F. Höfer, and W. Hess. A mixed inventory structure for German concatenative speech synthesis. In J.P. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Hrsg.), *Progress in speech synthesis*, 263–277. Springer, New York, 1997.

[Pie80] J. Pierrehumbert. The phonlogy and phonetics of English intonation. Dissertation, 1980.

[POSHF91] P. J. Price, M. Ostendorf, Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. In *J. of the Acoustic Society of America*, **90**, 2956–2970, 1991.

[PSZ91] J. Polifroni, S. Seneff, and V. W. Zue. Collection of spontaneous speech for the atis domain and comparative analyses of data collected at mit and ti. In *Speech and Natural Language Workshop*, San Mateo, California, 1991. Morgan Kaufmann.

[PWOB90] P. J. Price, C. W. Wightman, M. Ostendorf, and J. Bear. The use of relaive duration in syntactic disambiguation. In *Proc. Int. Conf. on Spoken Language Processing*, **1**, 13–18, Kobe, 1990.

[Rey94] M. Reyelt. Untersuchungen zur Konsistenz prosodischer Etikettierungen. In H. Trost (Hrsg.), *Proc. Konferenz Verarbeitung Natürlicher Sprache*, 290–299. Springer Verlag, 1994.

[Rey98] M. Reyelt. Experimentelle Untersuchungen zur Festlegung und Konsistenz suprasegmentaler Einheiten für die automatische Sprachverarbeitung. Dissertation, TU Braunschweig, 1998.

[Rin93] A. Rinscheid. Automatische Bestimmung von Periodenmarken mit dem emark-Algorithmus. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, 1048–1051, Frankfurt a. M., 1993. DPG–GmbH.

[Rip96] R. Ripley. *Pattern Recognition and Neuronal Networks.* Cambridge University Press, Cambridge, 1996.

[SAM96] SAMPA. speech assessment methods phonetic alphabet. University College London, Phonetics and Linguistics Home Page, http://www.phon.ucl.ac.uk/home/sampa/home.htm, 1987-1996.

[SBP+92] K. Silverman, M. Beckman, J. Pitrelli, M. Osterndorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi: A standard for labeling English prosody. In *Proc. Int. Conf. on Spoken Language Processing*, 867–870, 1992.

[Sch84] D. Schaffer. The role of intonation as a cue to topic management in conversation. In *Journal of Phonetics*, **12**, 327–344, 1984.

[Sch88] H.W. Schüssler. *Digitale Signalverarbeitung Band 1: Analyse diskreter Signale und Systeme.* Springer Verlag, Berlin, 1988.

[SD83] B. G. Secrest and G. R Doddington. An integrated pitch tracking algorithm for speech systems. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1352–1355, 1983.

[SEG+97] V. Strom, A. Elsner, G. Görz, W. Hess, W. Kasper, A. Klein, H.U. Krieger, J. Spilker, and H. Weber. On the use of prosody in a speech-to-speech translator. In *Proc. European Conf. on Speech Communication and Technology*, Rhodes, 1997.

[Sie85] E. Sievers. *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Leibzig, 1885.

[SK92] H. Shimodaira and M. Kimura. Accent phrase segmentation using pitch pattern clusering. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, **1**, 217–220, San Francisco, CA, 1992.

[SN94] H. Shimodaira and M. Nakai. Prosodic phrase segmentation by pitch pattern clusering. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, **2**, 185–188, Adelaide, 1994.

[ST91] E. G. Schukat-Talamazzini. Skriptum zur Vorlesung Sprachverstehen. Lehrstuhl für Informatik 5 (Mustererkennung), Wintersemester 1990/91.

[Str93a] V. Strom. Detektion von Laryngalisierungen in invers gefilteren Sprachsignalen. Diplomarbeit, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1993.

[Str93b] V. Strom. Verbesserung der Grundfrequenzbestimmung durch Optimierung der Stimmhaft/Stimmlos-Entscheidung und besondere Behandlung von Laryngealisierungen. Manuscript, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen–Nürnberg, 1993.

[Str94] V. Strom. Die Prosodiekomponente in INTARC I.2 — Satzmodusbestimmung aus der f_0. Verbmobil Technisches Dokument Nr. 6, Universität Bonn, 1994.

[Str95a] V. Strom. Akzent- und Phrasengrenzendetektion allein aus dem Grundfrequenz- und Energieverlauf. In *Proc. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik – DAGA*, **2**, 975–978, Saabrücken, 1995.

[Str95b] V. Strom. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Proc. European Conf. on Speech Communication and Technology*, **3**, 2039–2041, Madrid, 1995.

[Str95c] V. Strom. Die Prosodiekomponente in INTARC I.3. Verbmobil Technisches Dokument Nr. 33, Universität Bonn, 1995.

[Str96] V. Strom. What's in the pure prosody? In *Proc. Forum Acusticum*, 232, Antwerpen, 1996.

[SW96] V. Strom and C. Widera. What's in the "pure" prosody? In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.

[Swe97] M. Swerts. Prosodic features at discourse boundaries of different strength. In *J. of the Acoustic Society of America*, **101**, 1, 514–521, 1997.

[SZ90] M. Soclof and V. Zue. Collection and analysis of spontaneous and read corpora for spoken language system development. In *Proc. Int. Conf. on Spoken Language Processing*, **2**, 1105–1108, Kobe, 1990.

[Wah93] W. Wahlster. Verbmobil – translation of face-to-face dialogs. In *Proc. European Conf. on Speech Communication and Technology*, Opening and Plenary Sessions, 29–38, Berlin, 1993.

[Wai88] A. Waibel. *Prosody and Speech Recognition*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

[Web95] Hans H. Weber. *LR-inkrementelles probabilistisches Chartparsing von Worthypothesenmengen mit Unifikationsgrammatiken: Eine enge Kopplung von Suche und Analyse*. Dissertation, Universität Hamburg, FB Informatik, 1995.

[WH92] M. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries. In *Computer Speech and Language*, **6**, 175–196, 1992.

[Wig92] C. Wightman. Automatic detection of prosodic constituents. Dissertation, 1992.

[Win83] T. Winograd. *Language as a Cognitive Process. Volume I: Syntax*. Addison-Wesley Publishing Company, Reading, Mass., 1983.

[WO92] C. Wightman and M. Ostendorf. Automatic recognition of intonational features. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, **1**, 221–224, San Francisco, CA, 1992.

[WO94] C. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. In *IEEE Trans. on Speech and Audio Processing*, **2**, 3, 469–481, 1994.

[WS94] M.B. Wesenick and F. Schiel. Applying speech verification to a large data base of German to obtain a statistical survey about rules of pronunciation. In *Proc. Int. Conf. on Spoken Language Processing*, 279 – 282, Yokohama, 1994.

Lebenslauf

10. Mai 1966	Geburt in Erlangen
9/72 bis 7/76	Grundschule in Möhrendorf
9/72 bis 4/77	Albert-Schweitzer-Gymnasium in Erlangen
4/77 bis 6/85	Gymnasium Forchheim
Juni 1985	Abitur
10/85 bis 12/86	Grundwehrdienst
11/86 bis 1/93	Informatik-Studium an der Universität Erlangen–Nürnberg mit Schwerpunktfach Mustererkennung, Nebenfach Elektrotechnik mit Ausrichtung Nachrichtentechnik
Oktober 1988	Vordiplom
Januar 1993	Abschluß mit Diplom
4/93 bis 10/93	wissenschaftliche Hilfskraft am Lehrstuhl für Mustererkennung der Universität Erlangen–Nürnberg
seit 11/93	wissenschaftlicher Mitarbeiter am Institut für Kommunikationsforschung und Phonetik der Universität Bonn
11/93 bis 12/96	Mitarbeit im VERBMOBIL–Teilprojekt „Architektur“, Arbeitspaket 15.5 „Prosodiebezogene Interaktion“
1/97 bis 12/97	Mitarbeit im VERBMOBIL–Teilprojekt „Synthese“, Arbeitspaket 5.7 „Concept - to - speech: akustische Prosodie“