DER NUTZWERT PROSODISCHER MERKMALE FÜR DIE AUTOMATISCHE SPRACHERKENNUNG

E. Paulus, H.-D. Gerken, J. Reinecke und J. Veidt

Institut für Nachrichtentechnik, TU Braunschweig

1. Einführung

Gesprochene Sprache vermittelt dem Zuhörer im allgemeinen weitaus mehr Information als eine Niederschrift des gesprochenen Textes. Die über den reinen Wortlaut hinausgehende Information ist vielfältiger Natur und betrifft die auditive Analyse des Sprachsignals, die Erkennung des Worlauts und das Verstehen der Nachricht. Prosodische Information hat redundante, relevante und irrelevante Anteile. Der redundante Anteil erleichtert dem Zuhörer die Erkennung der Nachricht, vermittelt aber keine Inhalte, die nicht auch allein aus einer Niederschrift des gesprochenen Textes erschlossen werden können. Beispiele sind u.a. die richtige Betonung der Silben der gesprochenen Wörter und die Verdeutlichung der syntaktischen Struktur durch Aufgliederung der Äußerung in sogenannte prosodische Gruppen (Phrasen). Demgegenüber dient der relevante Anteil dazu, immer dann das richtige Verständnis sicherzustellen, wenn der Wortlaut mehr als eine Interpretation zuläßt. Bekannte Beispiele sind u.a. die Markierung des Satzfokus und die Unterscheidung zwischen Frage- und Aussagesätzen. Schließlich kann der Zuhörer aus der Prosodie auch Aufschlüsse über z.B. die Stimmungslage des Sprechers oder dessen Einstellung zum mitgeteilten Sachverhalt gewinnen, die für ihn durchaus wertvoll sein mögen, im engeren Sinn von Spracherkennung und -verstehen aber als irrelevant gelten sollen und hier nicht mehr weiter zu betrachten sind.

Um den vielfältigen prosodischen Merkmalen von Gesprochenem gerecht zu werden, muß zumindest zwischen Mikroprosodie, Wortprosodie, Phrasenprosodie und Satzprosodie unterschieden werden, je nachdem ob die Funktion eines Merkmals an einen Laut oder eine Lautfolge im Umfang von weniger als einer Silbe, an ein Wort, an eine prosodische Gruppe (Phrase) oder an einen Satz gebunden ist. Träger prosodischer Information im Sprachsignal sind die Verläufe der Intensität und der Sprachgrundfrequenz sowie Variationen der Silbendauer. Die Einflüsse von Mikro-, Wort-, Phrasen- und Satzprosodie überlagern sich dabei auf komplizierte und vielfach noch ungeklärte Weise. Während bei der Sprachsynthese die Umsetzung einer symbolischen Darstellung der prosodischen Information in die entsprechenden Merkmale des Sprachsignals bereits einigermaßen befriedigend gelingt, gibt es bei der Spracherkennung noch sehr viele offene Fragen über die Umsetzung von Signalmerkmalen in eine symbolische Darstellung der prosodischen Information oder allgemeiner über Art und Nutzen einer prosodischen Signalanalyse.

Für die symbolische Darstellung prosodischer Information gibt es - trotz mehrerer bei der Sprachsynthese erfolgreicher Vorgehensweisen - noch keinen einheitlichen und allgemein aner-

kannten Formalismus. Ein solcher, auf gesicherte Kenntnisse gestützter Formalismus wäre nicht nur für die Sprachsynthese, sondern vor allem auch für die Spracherkennung zu wünschen. Vorstellbar ist ein Formalismus, der als Kernstücke einerseits je ein Inventar elementarer Einheiten der Wort-, Phrasen- und Satzprosodie und andererseits Regeln für den Aufbau komplexer prosodischer Einheiten aus elementaren umfaßt. Durch die Regeln wären insbesondere auch die Wechselwirkungen zwischen Wort-, Phrasen- und Satzprosodie zu modellieren, die sehr stark von der grammatikalischen Gliederung der Sätze und dem semantisch-pragmatischen Gehalt der Äußerungen abhängen. Die Regeln sollten also nicht nur ein rein prosodisches Modell sondern auch ein syntaktisch-prosodisches und ein semantisch/pragmatisch-prosodisches Modell umfassen. Ergänzend zu einem derartigen Formalismus müßte geklärt werden, an welchen Signalmerkmalen die verschiedenen elementaren prosodischen Einheiten erkannt werden können, d.h. es müßte ein akustisch-prosodisches Modell erarbeitet werden. Es scheint zweckmäßig zu sein, die Mikroprosodie in diesem Modell zu berücksichtigen, so daß von einem akustisch/phonetisch-prosodischen Modell zu sprechen wäre.

Die Spracherkennung bzw. das Sprachverstehen betreffend gibt es für das Deutsche bisher erst sehr wenige Untersuchungen über die akustisch-prosodische Modellierung. Solche Untersuchungen sind jeweils nur auf einen Teilaspekt der prosodischen Analyse gerichtet und nicht aufeinander abgestimmt. Eine explizite Unterscheidung zwischen elementaren und komplexen prosodischen Einheiten wird im allgemeinen nicht eingeführt. Das Hauptinteresse galt dabei immer der Auswertung von Verläufen der Grundfrequenz F_0 [1] [2]. In [1] sind Ansätze zur Erkennung des Satzfokus unmittelbar aus einer sehr groben parametrischen Darstellung des F_0-Verlaufs beschrieben. [2] behandelt Versuche zur Zerlegung des F_0-Verlaufs in die von der Wort-, Phrasen- und Satzprosodie herrührenden Anteile. Während [1] die Erkennung relevanter prosodischer Information zum Zweck der richtigen Interpretation eines mehrdeutigen Wortlauts betrifft, gelten die eigenen Untersuchungen der Verwertbarkeit redundanter prosodischer Information bei der automatischen Spracherkennung. Langfristiges Ziel ist die Ausarbeitung eines möglichst umfassenden Formalismus zur symbolischen Darstellung prosodischer Information für die deutsche Sprache und die darauf abgestimmte akustisch-prosodische Modellierung.

2. Definition, Nutzwert und Merkmale prosodischer Einheiten

Auf der Grundlage der bereits an anderen Stellen [3] [4] zu Synthesezwecken eingesetzten Kenntnisse wurde ein sehr einfacher Formalismus zur silbenorientierten symbolischen Darstellung prosodischer Information zusammengestellt. Bei der Wortprosodie werden für die Betonung einer Silbe drei Stufen "unbetont", "schwach betont" und "betont" unterschieden. Auch bei der Phrasenprosodie wurden für die Betonung einer Silbe drei Stufen, nämlich "unbetont", "betont" und "Gruppenakzent", berücksichtigt. Bei der Satzprosodie sind drei Typen von Phrasen angenommen, nämlich "progredient", "terminal" und "interrogativ", und unabhängig vom Typ kann eine der prosodischen Gruppen eines Satzes das Merkmal "Fokus" tragen. Für das

Zusammenwirken der Wort-, Phrasen- und Satzprosodie untereinander und mit dem Lexikon und der Grammatik und Syntax wurden nur einige wenige, allgemein als gesichert geltende einfache Regeln berücksichtigt. Einige Einzelheiten sind in [5] beschrieben.

Die auf diesen Formalismus und die zugehörigen Regeln abgestimmte prosodische Analyse des Sprachsignals sollte als Ergebnis eine symbolische Darstellung der Phrasen- und der Satzprosodie liefern, die

1) für jede Silbe die zugehörige Betonungsstufe
2) die Lage der Phrasengrenzen
3) für jede Phrase den zugehörigen Typ
4) zu bestimmten Phrasen das Merkmal "Fokus"

angibt. Die symbolische Darstellung der Wortprosodie (d.h. die Betonungsstufen der einzelnen Silben) ist für jedes zugelassene Wort im Lexikon verzeichnet und ist somit nicht unmittelbar Gegenstand der Signalanalyse, sondern gehört zur Wissensbasis des Spracherkennungssystems. Allerdings sollte die Signalanalyse auch Merkmale der Wortprosodie feststellen, die bei der Erkennung insbesondere von mehrsilbigen Wörtern hilfreich sein können (siehe Abschnitt 3). Aber auch ohne die Bestimmung solcher Merkmale läßt sich die Wortprosodie bei der Erkennung von Wörtern dadurch nutzbringend berücksichtigen, daß die Verträglichkeit der im Lexikon verzeichneten Betonungsstufen mit der bei der Signalanalyse festgestellen Phrasenprosodie überprüft wird.

Bei der ersten experimentellen Erprobung im Rahmen des in [5] und [6] beschriebenen komplexen Spracherkennungssystems wurde die symbolische Darstellung der Phrasen- und der Satzprosodie nicht aus dem Sprachsignal gewonnen, sondern vom Experimentator vorgegeben. Dadurch wurde gleichsam die Wirksamkeit eines idealen Prosodieanalysators überprüft. Die Prosodieanalyse sollte die sonst nur durch Grammatik, Syntax und Semantik gesteuerte Suche nach dem wahrscheinlichsten Satz in einem aus der akustisch-phonetischen Signalanalyse resultierenden Netz aus teilweise konkurrierenden bewerteten Worthypothesen erleichtern. Die bereits in [5] erwähnten Ergebnisse bestätigen die Erwartungen: Insbesondere die Wort- und Phrasenprosodie führte sowohl zu einer Senkung der Fehlerwahrscheinlichkeit als auch zu einer Verminderung des Suchaufwands. Nicht ganz so wertvoll erwies sich die Satzprosodie, doch wurde vorerst nur die Lage der Gruppengrenzen und nicht auch der Gruppentyp verwertet.

Die neueren Untersuchungen gelten der Frage nach den für die Erkennung prosodischer Merkmale geeigneten Signalmerkmalen. Erste Beobachtungen führten zu der Annahme, daß - zumindest bei den beiden ausführlicher untersuchten Sprechern - die Wortprosodie vor allem aus dem Intensitätsverlauf und die Phrasen- und Satzprosodie eher aus dem F_0-Verlauf erkannt werden können. Zu beiden Themenkreisen wurden Ansätze entwickelt und erste experimentelle Ergebnisse erzielt. In den nachfolgenden Abschnitten werden die wichtigsten Überlegungen und die bisher durchgeführten Experimente grob und auszugsweise skizziert.

3. Untersuchungen an Silbenquotienten

Für mehrsilbige Wörter wurden sogenannte Silbenquotienten

$$Q(i) = \log [P(i-1)/P(i)] \qquad i = 2,3,\ldots,S$$

eingeführt. Dabei ist P(i) die maximale Intensität innerhalb der i-ten Silbe und S die Anzahl der Silben eines Wortes. In 72 gesprochenen Sätzen (18 verschiedene Sätze von 2 Sprechern je 2 mal geäußert) wurden die Silbenquotienten für 4 verschiedene Wörter in jeweils 3 verschiedenen Stellungen im Satz bestimmt, das sind je Wort und Sprecher 6 Ergebnisse. Im Mittel liegt die Standardabweichung der 6 für ein bestimmtes Q(i) vorliegenden Werte bei 0.15. Demgegenüber variieren die Zahlenwerte von Wort zu Wort weitaus deutlicher zwischen -0.7 und +0.7. Dieses Ergebnis stützt die Annahme, daß Silbenquotienten in erster Linie die Wort- und Mikroprosodie wiedergeben und von der Stellung des Wortes im Satz weitgehend unabhängig sind. Tatsächlich weichen auch die beiden Ergebnisse, die für die gleiche Stellung gelten, vergleichbar stark voneinander ab wie die drei Ergebnisse für verschiedene Stellungen.

Vom Standpunkt der Prosodieanalyse aus betrachtet, wäre es willkommen, wenn die Q(i) nur von der Wortprosodie, d. h. von der Abfolge der Silbenbetonungen, und nicht auch noch von den im Wort vorkommenden Sprachlauten abhingen. Tatsächlich wurde aber eine sehr deutliche Abhängigkeit von den Sprachlauten festgestellt, die so weit geht, daß Wörter allein anhand ihrer Q(i) ziemlich gut unterscheidbar sein können. Untersucht wurde die Unterscheidbarkeit von 10 verschiedenen viersilbigen Wörtern (unter denen es 4 verschiedene Wortprosodien gibt), die von einem Sprecher je 10 mal gesprochen wurden. Die Klassifizierung der Wörter anhand ihrer drei Silbenquotienten gemäß der Nächste-Nachbar-Regel nach der "leave-one-out"-Methode führte in 59 der 100 Fälle zur richtigen Erkennung des Wortes. Bei den 41 falsch erkannten Wörtern wird in 31 Fällen auch die Wortprosodie verwechselt, d. h.: die Q(i) sind zwar vermutlich von der Stellung des Wortes weitgehend unabhängig, sie kennzeichnen aber das Wort insgesamt und nicht nur allein dessen Wortprosodie. Bei allen bisher genannten Untersuchungen ist die Intensität P(i) rein physikalisch definiert gewesen. Wenn statt dessen ein Maß für die subjektiv empfundene Lautheit verwendet wird, prägen sich - wie Tabelle 1 zeigt - die zuvor beschriebenen Tendenzen sogar noch deutlicher aus.

		Wort: richtig	Wort: falsch	Wort: richtig	Wort: falsch
Betonungsmuster:	richtig	59	10	70	8
	falsch		31		22
Intensität P(i):		physikalisch		subjektiv	

Tabelle 1: Anzahl klassifizierter Wörter

Obwohl die Silbenquotienten die Wortprosodie nicht isoliert erfassen, können sie - unter bestimmten Bedingungen - dennoch in komplexen Sprachererkennungssystemen mit Vorteil ausgewertet werden. Zur Demonstration wurde das Lexikon des verfügbaren Spracherkennungssystems dadurch erweitert, daß für die 620 mehrsilbigen Wörter des zugelassenen Wortschatztes für jeden Silbenquotienten ein Meßwert eingetragen wurde. Diese zusätzlichen Einträge können bei der Suche nach dem wahrscheinlichsten Satz zur Bewertung von Worthypothesen mitherangezogen werden, vorausgesetzt, daß bei der Signalanalyse für jede Silbe ein Silbenquotient bestimmt wird. Bei 50 Testsätzen stieg dadurch z.B. die Anzahl derjenigen Sätze, bei denen alle wirklich gesprochenen Wörter durch die an der jeweiligen Stelle am besten bewertete Worthypothese angezeigt wurden, von 16 auf 21.

Die Eintragung von Q(i)-Werten in das Lexikon erfordert allerdings einigen Aufwand, der bei kleinem Wortschatz (z. B. weniger als 1000 Wörter) aber tragbar sein dürfte und sich durch eine Steigerung der Erkennungssicherheit oder eine Verminderung des Suchaufwandes bezahlt machen sollte.

4. Untersuchungen an F_0-Verläufen

Zu den Aufgaben der prosodischen Signalanalyse gehört die Bestimmung einer symbolischen Darstellung der Phrasen- und der Satzprosodie. Wie schon in Abschnitt 2 erwähnt, lassen sich dabei vier Teilaufgaben unterscheiden. Schwerpunkte der bisherigen Arbeiten waren die Teilaufgaben 1) und 2). Die Untersuchung von 3) wurde vorbereitet. Das Versuchsmaterial besteht aus 53 Sätzen, die bisher erst von einem einzigen Sprecher je einmal gesprochen wurden. Die 53 Sätze gliedern sich in 92 Phrasen und 574 Silben. Für jeden gesprochenen Satz liegen gemessene F_0-Verläufe vor, die für die Experimente etwas vorverarbeitet wurden, indem Lücken, die durch stimmlose Sprachlaute verursacht sind, durch glatte und stetige Interpolation geschlossen wurden. Für die Verwendung des Versuchsmaterials als Lern- und/oder Teststichprobe beim Entwurf bzw. der Erprobung des Prosodieanalysators wurde für jede Silbe die vom Experimentator festgestellte richtige Betonungsstufe verzeichnet. Im Hinblick auf die Teil-aufgabe 3) wurden zwei verschiedene Klassen unbetonter Silben unterschieden, je nachdem ob die betreffende Silbe innerhalb der Phrase vor oder nach der Silbe mit dem Gruppenakzent steht (siehe auch unten). Ebenfalls vom Experimentator festgestellt wurden die Phrasengrenzen und für jede Phrase der zugehörige Typ.

Um einen ersten Eindruck von der Brauchbarkeit von F_0-Verläufen für die Erkennung der Betonungsstufen der einzelnen Silben (Teilaufgabe 1)) zu gewinnen, wurden die Silben gemäß der Nächste-Nachbar-Regel nach der "leave-one-out"-Methode klassifiziert. Um eine Entfernung zwischen zwei Silben zu bestimmen, können die beiden zugehörigen F_0-Verläufe durch dynamische Programmierung in der als "dynamic-time-warping" geläufigen Form optimal aufeinander abgebildet werden.

Tabelle 2 zeigt das Ergebnis der Klassifizierung in Form einer Verwechselungsmatrix.

		erkannt				
		^	'	-	.	Summe
gespr.	^	50	7	21	14	92
	'	19	3	10	4	36
	-	33	6	147	51	237
	.	33	6	44	126	209
Summe		135	22	222	195	574

Tabelle 2: Anzahl klassifizierter Silben

Die Klassifizierung sollte sich dadurch verbessern lassen, daß ein einfaches Modell für die Abfolge von Betonungsstufen innerhalb einer Phrase berücksichtigt wird. Dieses Modell stützt sich auf einige bekannte Regeln (siehe z.B. [3] und [4]) und läßt sich durch das in Bild 1 gezeigte Übergangsnetzwerk darstellen.

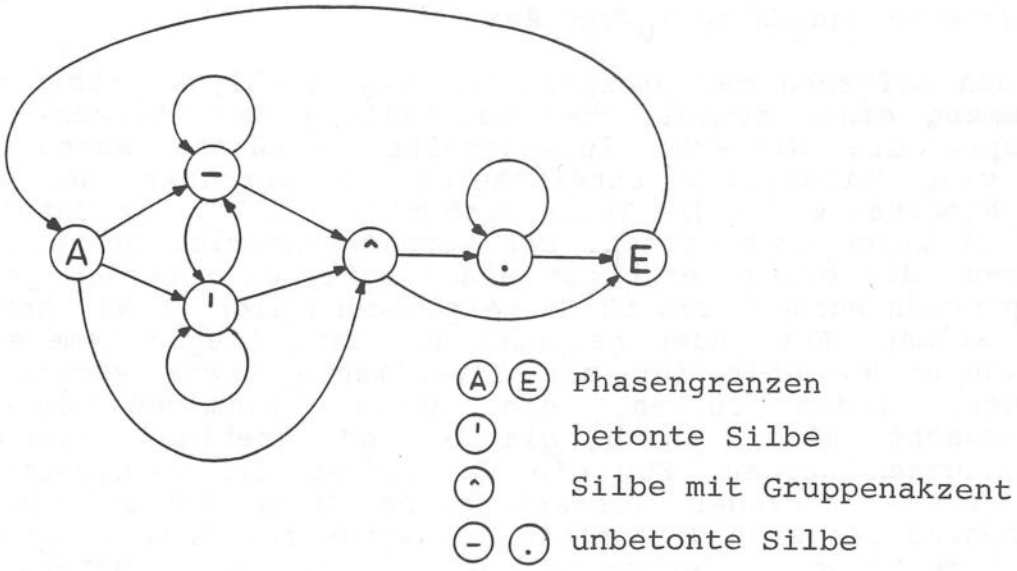

Bild 1: Übergangsnetzwerk als Modell für die Phrasenprosodie

Die Klassifizierung einer Silbe nach Betonungsstufen kann nun in zwei Schritten erfolgen. Im ersten Schritt wird eine Rangliste aller (hier vier) möglichen Klassifizierungsergebnisse aufgestellt, wobei jede Eintragung in dieser Liste eine Bewertung trägt, die im Idealfall die Wahrscheinlichkeit dafür ausdrückt, daß das betreffende Ergebnis richtig ist. In den ersten Versuchen wurde die Rangliste dadurch gewonnen, daß zu jeder Silbe getrennt nach den vier Betonungsstufen jeweils der nächste Nachbar bestimmt wurde, d.h. insgesamt also vier nächste Nachbarn zusammen mit den zugehörigen Entfernungen, aus denen sich die Rangfolge ergibt. (Kleinste Entfernung entspricht höchstem Rang!) Im zweiten Schritt wird unter den in den Ranglisten angebotenen bewerteten "Betonungshypothesen" die endgültige Auswahl dadurch getroffen, daß unter allen Abfolgen, die

1) den gesprochenen Satz lückenlos abdecken
2) nur aus Phrasen bestehen, die dem Modell entsprechen,

diejenige mit der besten Gesamtbewertung bestimmt wird. Dadurch werden nicht nur Betonungsstufen (Teilaufgabe 1)) sondern auch Phrasengrenzen (Teilaufgabe 2)) bestimmt. Als Gesamtbewertung wurde vorerst die Summe aller Entfernungen verwendet. Die Abfolge mit der kleinstmöglichen (d.h. besten) Bewertung kann auch hier durch dynamische Programmierung bestimmt werden. Tabelle 3 zeigt das Klassifizierungsergebnis wiederum in Form einer silbenbezogenen Verwechselungsmatrix.

		erkannt				
		^	'	-	.	Summe
gespr.	^	55	7	12	18	92
	'	19	7	7	3	36
	-	28	6	165	38	237
	.	24	2	50	133	209
Summe		126	22	234	192	574

Tabelle 3: Anzahl klassifizierter Silben bei modellgestützter Klassifizierung

Aus dem Vergleich der Tabellen 2 und 3 läßt sich eine gewisse Überlegenheit der modellgestützten Klassifizierung erkennen, obgleich der Umfang des verwendeten Versuchsmaterials für gesicherte Aussagen noch viel zu gering ist. Die Überlegenheit läßt sich aber auch dadurch belegen, daß es für eine n-silbige Phrase ohne Modell 3^n mögliche Abfolgen von Betonungsstufen gibt, während das Modell nur 2^n-1 verschiedene Abfolgen zuläßt. (Bei dieser Überlegung sind die beiden Unterklassen unbetonter Silben zu einer einzigen Klasse "unbetont" zusammengefaßt!)

Trotz des geringen Umfangs des verwendeten Versuchsmaterials läßt sich bereits behaupten, daß die bisher versuchte Auswertung des F_0-Verlaufs nur eine sehr unzuverlässige Unterscheidung der vier Betonungsstufen erlaubt, so daß Verbesserungsmöglichkeiten untersucht werden müssen. In den weiterführenden Untersuchungen soll daher neben dem F_0-Verlauf selbst auch dessen erste Ableitung betrachtet werden. Darüber hinaus ist auch an die Einbeziehung des bereits bei der Wortprosodie verwendeten Intensitätsverlaufs zu denken. Während bisher immer kurvengetreue Darstellungen zeitlicher Verläufe verwendet wurden, sollen künftig auch vereinfachte parametrische Darstellungen untersucht werden. Die grundsätzliche Vorgangsweise der modellgestützten Generierung einer symbolischen Darstellung der Phrasen- und Satzprosodie in zwei Schritten kann bei allen genannten Ausweitungen der Untersuchungen beibehalten werden.

Möglicherweise läßt sich die Zuverlässigkeit der Klassifizierung auch dadurch steigern, daß im ersten Schritt der modellgestützten Klassifizierung die Silben nicht einzeln und unabhän-

gig voneinander, sondern im Verbund von zwei Silben klassifiziert werden. Dabei ist für jede Silbe eine Rangliste aller im Modell vorkommenden Kanten aufzustellen. (Jede solche Kante repräsentiert im allgemeinen einen Übergang zwischen zwei Betonungsstufen!). Im zweiten Schritt ist wiederum unter den in diesen Ranglisten angebotenen bewerteten "Übergangshypothesen" die endgültige Auswahl dadurch zu treffen, daß unter allen mit dem Modell verträglichen Abfolgen diejenige mit der besten Gesamtbewertung bestimmt wird. Experimentelle Ergebnisse für diese Vorgangsweise liegen noch nicht vor; mehr noch als die zuvor beschriebenen Versuche mit einzelnen klassifizierten Silben erfordern sie ein sehr umfangreiches Versuchsmaterial, um verläßliche Aussagen machen zu können.

Ebenso noch ausständig sind Versuche zur Erkennung des Phrasentyps. Dazu ist geplant, das bisher verwendete, der Struktur nach vom Phrasentyp unabhängige Modell geeignet zu erweitern. Insbesondere sollen dabei ab dem Gruppenakzent mehr Knoten und/oder Kanten eingeführt werden, die nicht - wie bisher - auf nur einen (E)-Knoten, sondern auf drei verschiedene (E)-Knoten führen, die den einzelnen Phrasentypen entsprechen. Um den im ersten Schritt benötigten, auf dieses Modell abgestimmten Klassifikator (sei es für die Klassifikation einzelner Silben oder von Silbenübergängen) zu trainieren, muß eine sehr umfangreiche Lernstichprobe erfaßt werden. Daran wird noch gearbeitet.

5. Ausblick und Danksagung

Die bisherigen Ergebnisse sind zwar noch sehr unvollkommen, ermutigen aber zur Fortsetzung und Intensivierung der begonnenen Arbeiten. Dabei werden die laufenden Untersuchungen über die Phrasen- und Satzprosodie an Gewicht gewinnen.

Der DFG ist für die Förderung der hier beschriebenen Arbeiten im Rahmen des Schwerpunktprogramms "Modelle und Strukturanalyse bei der Auswertung Bild- und Sprachsignalen" zu danken.

Literatur

[1] BATLINER, A. & E. NÖTH: The Prediction of Focus. Proc. Eurospeech 89, Paris 1989, 210-213

[2] DEMENKO, G.; B. MÖBIUS & M. PÄTZOLD: Statistische Analyse zur Klassifizierung von Intonationsverläufen. Tagungsband DAGA'90, Wien 1990

[3] CLASS, F.; D. STALL & R. ZELINSKI: Probleme bei der automatischen Erkennung und Synthese von gebundener Sprache. GMD-Spiegel 1/83, 32-37

[4] WOLF, H.E.: Intonationsmotive zur Prosodiesteuerung nach Regeln für die Sprachsynthese. Tagungsband FASE/DAGA'82, Göttingen 1982, 1031-1034

[5] PAULUS, E.: Von der verständnisfreien zur sprachkundigen Erkennung gesprochener Prosatexte. Tagungsband DAGA '88, Braunschweig 1988, 23-41

[6] PAULUS, E. & J. MUDLER: Grundzüge eines sprachkundigen Systems zur automatischen Spracherkennung. it 5/89, 358-364