Der Einsatz prosodischer Information im Spracherkennungsyssystem EVAR

E. Nöth, R. Kompe

Lehrstuhl für Informatik 5 (Mustererkennung)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstraße 3
8520 Erlangen, F.R. Germany

Kurzfassung: Es werden einige Experimente vorgestellt zum Einsatz von prosodischem Wissen auf der lexikalischen und pragmatischen Ebene des Spracherkennungssystems EVAR. Die Qualität von Worthypothesen an pragmatisch wichtigen Stellen einer Äußerung und die Wirksamkeit von zwei prosodisch motivierten Worthypothesenfiltern wird betrachtet. Die Häufigkeitsverteilungen von Wörtern in Abhängigkeit von der Silbenanzahl werden für verschiedene Korpora und Hypothesenmengen untersucht. Ansätze für eine prosodische Verifikation von Worthypothesen werden beschrieben.

1. Einleitung

Die Bedeutung der prosodischen Information beim menschlichen Perzeptionsprozeß wird allgemein anerkannt. In den letzten Jahren ist auch das Interesse an prosodischer Information als Wissensquelle für die automatische Spracherkennung stark angestiegen. W. Lea schlug bereits 1975 ein prosodisch gesteuertes Spracherkennungssystem vor [LEA 75]. Trotzdem gibt es kaum ein System, in das prosodische Information tatsächlich integriert wurde, und in dem konkrete Ergebnisse über den Beitrag der prosodischen Information am Gesamterkennungsverhalten vorliegen. Ein Überblick über den Einsatz prosodischer Information in der automatischen Spracherkennung findet sich z.B. in [VAI 88]. Wir befassen uns in dieser Arbeit mit dem Einsatz prosodischer Information auf der lexikalischen und pragmatischen Ebene. Die im folgenden beschriebenen Experimente dienen als grundlegende Voruntersuchungen für ein Prosodie-Modul. Es werden drei Fragestellungen untersucht:

- Werden pragmatisch wichtige Stellen der Äußerung besser erkannt (dies ermöglicht eine prosodisch gesteuerte Kontrolle)?
- Läßt sich an pragmatisch wichtigen Stellen das Lexikon einschränken (dies ermöglicht eine prosodisch motivierte, erwartungsgesteuerte Worthypothesengenerierung)?
- Wieviele falsche Worthypothesen kann man aufgrund der Divergenz zwischen lexikalischem und automatisch berechnetem Betonungsmuster verwerfen (prosodische Wortverifikation)?

Wir definieren als **pragmatisch wichtige Stelle** den Nukleus der Konstituente, die den **Fokus** einer Äußerung realisiert. (Der Begriff Fokus bezeichnet das Informationszentrum eines Satzes [BUS 83].) Zwar sind die pragmatisch wichtigen Stellen eines Sprachsignals

nicht identisch mit den am stärksten betonten, es besteht jedoch ein enger Zusammenhang. Dies wurde für die in [NÖT 88] beschriebene Stichprobe von Zugauskunftsdialogen durch folgendes Experiment nachgewiesen:

Die Stichprobe wurde 15 Hörern vorgespielt, die die betonten Silben mit 1 und die unbetonten mit 0 markieren mußten. Durch Aufaddieren der Markierungen ergab sich für jede Silbe eine Bewertung zwischen 0 und 15. (Dies stellt selbstverständlich keine 16-stufige Betonungsbewertung dar.) Die Akzentsilben der pragmatisch wichtigen Wörter hatten eine durchschnittliche Bewertung von 12, der Durchschnitt für alle Silben betrug 2. Weiterhin handelte es sich bei allen Wörtern, die von mindestens der Hälfte der Testpersonen als betont markiert wurden, um pragmatisch wichtige Stellen.

Unsere Ergebnisse stehen teilweise in engem Zusammenhang mit dem Spracherkennungssystem, in welches das Prosodie-Modul integriert werden soll. In Kapitel 2 werden wir daher zunächst das System EVAR vorstellen, soweit dies für das Verständnis der durchgeführten Experimente notwendig ist. Die Ergebnisse werden in Kapitel 3 vorgestellt. In Kapitel 4 fassen wir die Einzelergebnisse zusammen und geben einen Ausblick auf weitere Untersuchungen.

2. Das Spracherkennungssystem EVAR

Am Lehrstuhl für Informatik 5 wird seit 1979 am System EVAR (Erkennen, Verstehen, Antworten, Rückfragen) gearbeitet. Die Systemstruktur lehnt sich an ein geschichtetes linguistisches Modell an. Das System ist in mehrere Module aufgeteilt, die die verschiedenen linguistischen Ebenen repräsentieren. Eine Übersicht des Systems findet sich in [NIE 85].

Das **Akustik-Phonetik**-Modul [REG 88] zerlegt das Sprachsignal in äquidistante Zeitscheiben von 12,8 Millisekunden (im folgenden als **Frame** bezeichnet). Diese werden nach Lautkomponenten klassifiziert und mit Hilfe eines syntaktischen Verfahrens zu **Segmenten** zusammengefaßt. Es werden keine alternativen Segmentgrenzen betrachtet. Einem Segment werden maximal fünf Lauthypothesen zugewiesen. Unterschieden werden 49 Lautkomponentenklassen und 36 Lautklassen.

Das **Worthypothesen**-Modul [KUN 88] vergleicht die Lauthypothesenfolge mit jedem Eintrag eines Lexikons. Es wird eine **Word-Spotting-Technik** verwendet, die auf einem statistischen Verfahren (**H**idden **M**arkov Modell) basiert. Eine Worthypothese besteht aus Anfangs- und Endframe des Bereichs, über dem die Hypothese generiert wurde, Wortnummer, Bewertung, sowie einem optimalen **Zuordnungspfad** der Standardaussprache zu der Lauthypothesenfolge. Das **Lexikon** [EHR 86] enthält für jedes Wort u.a. eine systemweit eindeutige Wortnummer, die Standardaussprache nach Duden und drei Bitvektoren. Diese repräsentieren die Zugehörigkeit des Wortes zu den möglichen syntaktischen Klassen sowie den semantischen und pragmatischen Konzepten. Momentan verwenden wir zwei Vollformen-Lexika: **GLEXS** mit ca. 4000 Einträgen und eine Untermenge **KLEXS** (ca. 500 Einträge).

Zeitlich aufeinanderfolgende Hypothesen werden auf syntaktische Konsistenz überprüft und zu **Konstituenten** zusammengefaßt [BRI 87]. Diese werden semantisch verifiziert und

pragmatisch interpretiert und, soweit möglich, zu komplexen Konstituenten zusammengefaßt [EHR 88]. Dabei können zeitliche Lücken auftreten. Semantisch verifizierte und komplexe Konstituenten bezeichnen wir als **Semantikhypothesen.** Pragmatische Interpretation in unserem Sinn ist der anwendungsabhängige Teil der semantischen Analyse.

Noch nicht integriert in das System ist das **Prosodie**-Modul. Aufgabe des Moduls ist die Erstellung einer symbolischen Betonungsbeschreibung. Als Träger der Betonungsinformation gilt die Silbe; die Veränderung der prosodischen Parameter Tonhöhe, Lautdauer und Lautstärke zum Zwecke der Betonung betrifft insbesondere den vokalischen Silbenkern. Grundlage des Moduls wird das in [NÖT 88] beschriebene Verfahren zur automatischen Silbenkerndetektion und Betonungsbewertung sein. Das Betonungsmaß errechnet sich aus der Veränderung von Merkmalen, welche die drei prosodischen Parameter repräsentieren.

3. Experimente

3.1. Qualität der Worthypothesen an pragmatisch wichtigen Stellen

Im Rahmen der Arbeiten am Pragmatik-Modul wurde eine Stichprobe von 62 Äußerungen erstellt, für die eine Pragmatikanalyse sinnvoll erschien. Diese Stichprobe ist disjunkt zu der oben erwähnten Dialogstichprobe, es handelt sich um eine Teilmenge der in [NIE 85] beschriebenen Sprachstichprobe. Alle pragmatisch wichtigen Wörter wurden markiert (z.B. "Wann fährt der **nächste** Zug nach **München**"). Im folgenden werden diese Wörter Fokus-Wörter genannt. Die Stichprobe umfaßt 354 Wörter, 84 Fokus-Wörter und 562 Silben. Jeder Silbenkern in der Handsegmentierung des Sprachsignals nach Lauten wurde auf das am besten passende, automatisch erzeugte Segment abgebildet. Dies simuliert eine ideale Silbenkerndetektion. Im folgenden soll gezeigt werden, daß die Qualität der Worterkennung an betonten Stellen einer Äußerung besser ist als an unbetonten Stellen.

Bild 1 zeigt die Erkennungsrate in Abhängigkeit von der Anzahl der generierten Worthypothesen. Die Kurven bedeuten z.B. für den Eintrag 17 auf der Abszisse, daß bei der Generierung von 17*n Hypothesen 64% **aller** Wörter und 71% der **Fokus**-Wörter erkannt wurden. Dabei ist **n** die Anzahl der Segmente. Ein Wort mit der Wortnummer wnr heißt **erkannt**, wenn in der Hypothesenmenge eine Worthypothese wnr existiert, die die Handsegmentierung des tatsächlich gesprochenen Wortes zu mindestens 50% zeitlich überlappt.

Die Erkennungsrate für die Fokus-Wörter ist um durchschnittlich 7 Prozentpunkte höher als die aller Wörter; in dem für eine weitere Verarbeitung interessanten Bereich von weniger als 20 Hypothesen/Segment ist die Erkennungsrate um durchschnittlich 10 Prozentpunkte besser. Unter der Voraussetzung einer robusten automatischen Silbenkerndetektion und Akzentbewertung kann das Prosodie-Modul Informationen über Stellen im Sprachsignal liefern, an denen die Worthypothesengenerierung deutlich besser ist. Dies ist insbesondere deshalb von Bedeutung, da diese Stellen für die weitere Analyse besonders wichtig sind.

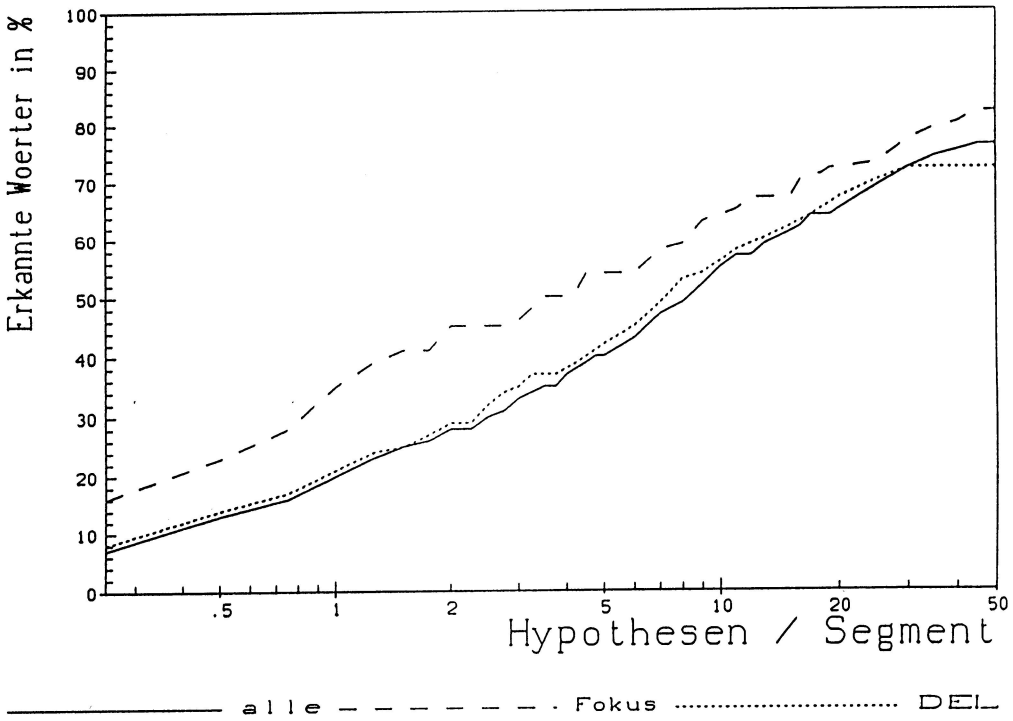

Bild 1: Erkennungsraten für **alle** Wörter, **Fokus**-Wörter und alle Wörter nach Anwendung des **DEL**-Filters (siehe 3.3.2.)

3.2. Prosodisch motiviertes Einschränken des Lexikons

Die Erkennung auf Worthypothesenebene läßt sich zusätzlich verbessern, wenn man an den pragmatisch wichtigen Stellen das Lexikon auf solche Wörter einschränkt, die ein pragmatisches Konzept realisieren können. Durch nachträgliches Herausfiltern der Worthypothesen ohne solch einen Pragmatikeintrag wurden insgesamt ca. 22% der Worthypothesen verworfen. Es ergab sich eine durchschnittliche Verbesserung der Erkennungsrate um ca. 3 Prozentpunkte, im Bereich unter 20 Hypothesen/Segment um ca. 5 Prozentpunkte. Dieses Ergebnis läßt sich automatisch nur erzielen, wenn alle betonten Silbenkerne korrekt detektiert werden. Deshalb stellt unser Ergebnis eine optimistische Abschätzung dar. Das Experiment ist vor allem unter dem Gesichtspunkt inselgetriebener bzw. erwartungsgesteuerter Worthypothesengenerierung zu sehen.

3.3. Prosodische Wortverifikation

Die im Lexikon eingetragene Umschrift eines Wortes schließt die Kennzeichnung der den Wortakzent tragenden Silbe ein. Mit einer automatisch berechneten Akzentuierung von Silbenkernen sind daher zwei Ansätze für eine Verifikation durch Konsistenzprüfung der Akzentstellen denkbar:

(i) In dem durch eine Worthypothese vorgegebenen Zeitbereich wird unabhängig von den Lauthypothesen und dem Zuordnungspfad eine automatische Silbenkerndetektion durchgeführt. Jedem Silbenkern wird ein Betonungsmaß zugewiesen. Zur Silbenkerndetektion und Betonungsberechnung wird das in [NÖT 88] beschriebene Verfahren benutzt. Das Betonungsmaß wird mit der Standardaussprache aus dem Lexikon verglichen.

(ii) Zu jedem Silbenkern im Zuordnungspfad einer Worthypothese wird ein Betonungsmaß berechnet, das mit dem Lexikoneintrag konsistent sein muß (siehe 3.3.2.).

Die beiden Methoden unterscheiden sich in der Art der Silbenkerndetektion, zur Betonungsberechnung dient jeweils der gleiche Algorithmus. Wir haben bisher Fall (ii) realisiert.

3.3.1. Optimistische Abschätzung

Über den Wortakzent können nur zwei- oder besser n-silbige (n>2) Wörter verifiziert werden. Deshalb haben wir zunächst einige Voruntersuchungen zur Häufigkeitsverteilung von 1-, 2- und n-silbigen Wörtern durchgeführt.

Zur Verfügung stand das Frequenzwörterbuch einer Teilmenge des FACID-Korpus [HIT 86]. Bei dieser Teilmenge handelt es sich um eine repräsentative Sammlung von Dialogen aus dem Bereich der Bundesbahn-Zugauskunft. Diese Wortmenge besteht aus 9421 Wörtern (1288 verschiedene). Tabelle 1 zeigt den prozentualen Anteil der drei Klassen an der jeweiligen Teilmenge der gesamten Wortmenge. Dabei bedeuten n% der Wortmenge: in dem nach Häufigkeit des Auftretens sortierten Frequenzwörterbuch wurde, vom häufigsten Wort beginnend, die relative Häufigkeit aufsummiert bis zu n% der Gesamtwortmenge. Die bis zu n% aufgetretenen verschiedenen Wörter wurden in die drei Klassen eingeteilt. Der Verlauf spiegelt die Tatsache wider, daß es einen Basiswortschatz mit relativ kurzen, häufig auftretenden Wörtern und einen anwendungsbezogenen Wortschatz mit eher längeren Wörtern gibt.

% Wortmenge	30	40	50	60	70	80	90	100
1-Silber	93	86	79	76	74	69	65	61
2-Silber	7	12	19	20	22	24	26	28
n-Silber	0	2	2	3	5	7	9	12

Tabelle 1: Relative Häufigkeit von ein- und mehrsilbigen Wörtern in den Bundesbahndialogen des FACID-Korpus

Das Lexikon GLEXS, die Äußerungen unserer Stichprobe (WOR) und die dazu generierten Worthypothesen (WHYP) und Semantikhypothesen (SHYP) wurden ebenfalls ausgezählt (Tabelle 2). Die Werte in der Spalte WOR beziehen sich auf 100% der Wortmenge. Sie stimmen gut überein mit den Ergebnissen für das FACID-Korpus. Die Angaben zu WHYP beziehen sich auf Hypothesen zum GLEXS und einer Menge von 50

Hypothesen pro Segment; wir konnten aber zeigen, daß die Häufigkeiten weitgehend unabhängig von der Größe der Hypothesenmenge sind. Die Semantikhypothesen wurden auf KLEXS-Basis und 150 Worthypothesen pro Äußerung generiert. (Semantikhypothesen mit einer zeitlichen Lücke wurden wie zwei Hypothesen behandelt.)

Die Auszählungen lassen eine prosodische Verifikation mit dem im folgenden beschriebenen Verfahren als lohnend erscheinen, da 65% aller Worthypothesen und über 80% aller Semantikhypothesen mindestens 2 Silben haben.

	GLEXS	WOR	WHYP	SHYP
1-Silber	16	57	35	19
2-Silber	39	32	48	22
n-Silber	45	12	17	59

Tabelle 2: Relative Häufigkeit von ein- und mehrsilbigen Wörtern im Lexikon, in den untersuchten Äußerungen und in den Hypothesenmengen

3.3.2. Ein prosodisches Wortfilter

Wir wollen uns hier darauf beschränken, ein prosodisches Wortfilter vorzustellen; ein Filter für Semantikhypothesen wird momentan implementiert [UNG 88]. Das Filter bildet bei den zwei- oder mehrsilbigen Worthypothesen jeden Silbenkern der Standardaussprache über den Zuordnungspfad auf ein Segment ab. Für jeden so ermittelten Silbenkernbereich wird ein Betonungsmaß berechnet. Anschließend wird die Differenz zwischen der Betonung der Wortakzentsilbe und der Betonung der anderen Silbe(n) gebildet.

Aus der Sicht der Prosodie nehmen die Silbenkerne eine besondere Stellung im Wort ein; sie sind Orte höherer Schallfülle [KOH 77] und werden beim Sprechen zumeist ausgeprägter realisiert als die Konsonanten. Aus diesem Grund fordern wir, daß im Zuordnungspfad einer Worthypothese kein Silbenkern gelöscht sein darf, was einem Delete-Übergang beim Mustervergleichsalgorithmus der Worthypothesengenerierung entspricht. Mit dem DEL-Filter werden deshalb alle Worthypothesen verworfen, bei denen ein Silbenkern gelöscht wurde. Die Erkennungsrate der so gefilterten Hypothesen ist in Bild 1 als Kurve **DEL** bezeichnet. Das Filter reduziert die Gesamtmenge von rund 80000 Hypothesen (dies entspricht 50 Hypothesen/Segment) um ca. 32% auf rund 55000 (ca. 35 Hypothesen/Segment). Die Qualität der Worthypothesen verbessert sich im Bereich unter 30 Hypothesen/Segment um durchschnittlich 2 Prozentpunkte. Oberhalb von 30 wird die Erkennungsrate schlechter, da bei einigen richtigen, aber schlecht bewerteten Hypothesen, ein Silbenkern gelöscht wurde. Es wurden vorwiegend schlecht bewertete Hypothesen verworfen, die nur bei mehr als 20 Hypothesen/Segment erzeugt wurden. Daher erhöhte sich die Erkennungsrate trotz der erheblichen Verkleinerung der Gesamthypothesenmenge nur um 2 Prozentpunkte.

Für die weitere Filterung mittels Betonung gingen wir von der Annahme aus, daß eine Wortakzentsilbe nicht weniger betont sein darf als eine andere Silbe. Wir haben zunächst die oben erwähnten Betonungsdifferenzen, getrennt nach richtigen und falschen

Hypothesen, gemessen. Die Verteilung der Differenzen ließ Unterschiede erkennen: der Mittelwert war bei den richtigen Worthypothesen höher als bei falschen Worthypothesen. Trotzdem war bei zuvielen richtigen Wörtern die Akzentsilbe schwächer bewertet als eine andere Silbe des Wortes. Mit der momentanen Version des Moduls zur Betonungsbewertung läßt sich somit keine weitere Verbesserung der Worterkennung erreichen, da zuviele richtige Hypothesen verworfen werden. Die Fehlbewertungen sind auf die in [NÖT 88] beschriebenen Schwachpunkte bei der Berechnung der prosodischen Merkmale zurückzuführen, insbesondere darauf, daß die intrinsische Energie der verschiedenen Vokale unberücksichtigt bleibt.

4. Zusammenfassung und Ausblick

Ziel dieser Arbeit war es, Möglichkeiten des Einsatzes prosodischer Information auf der Ebene der Worterkennung und der pragmatischen Analyse zu untersuchen. Daher lag der Schwerpunkt auf grundsätzlichen Voruntersuchungen.

Wir stellten fest, daß pragmatisch wichtige Wörter im Mittel um 7 Prozentpunkte besser erkannt werden. Unter der Voraussetzung einer optimalen Detektion betonter Silbenkerne könnte man an diesen Stellen durch Einschränken des Lexikons auf pragmatisch relevante Wörter die Gesamterkennungsrate zusätzlich verbessern. Bei dem von uns untersuchten Material bedeutete dies eine Reduktion der Worthypothesenmenge um 22% . Die Erkennungsrate stieg durchschnittlich um 3 Prozentpunkte.

In einem weiteren Experiment erhöhten wir die Erkennungsrate durch Verwerfen der Hypothesen, bei denen ein Silbenkern im optimalen Pfad des Mustervergleichsalgorithmus der Generierung gelöscht wurde. Dabei wurde die Hypothesenmenge um 32 % reduziert, die Erkennungsrate stieg bis zu 30 Hypothesen/Segment im Mittel um 2 Prozentpunkte. Die Verifikation von Worthypothesen durch Konsistenzprüfung eines automatisch berechneten Betonungsmusters brachte keine Verbesserung der Erkennungsrate. Ursache dafür waren Schwächen in der Akzentuierungsberechnung, insbesondere die Nichtberücksichtigung der intrinsischen Energie der verschiedenen Vokale. Dies wird in einer nächsten Version des Akzentmoduls korrigiert.

Die Verifikation erscheint aufgrund des hohen Anteils der zwei- und mehrsilbigen Wörter als lohnend. Ihr Anteil beträgt bei den Äußerungen unserer Stichprobe und denen des FACID-Korpus ca. 40% , bei den Worthypothesen 65% und bei den Semantikhypothesen 81 % .

5. Literaturverzeichnis

[BRI 87] Brietzmann, A.: "Stufenweise Syntaktische Analyse mit integrierter Bewertung für die kontinuierliche Spracherkennung", Arbeitsberichte des IMMD der FAU, Band 20, No. 9, Erlangen, 1987.

[BUS 83] Bussmann, H.: "Lexikon der Sprachwissenschaft", Alfred Kröner Verlag, Stuttgart, 1983.

[EHR 86] Ehrlich, U.: "Ein Lexikon für das natürlich-sprachliche System EVAR", Arbeitsberichte des IMMD der FAU, Band 19, No. 3, Erlangen, 1986.

[EHR 88] Ehrlich, U., Niemann, H.: "Using Semantic and Pragmatic Knowledge for the Interpretation of Syntactic Constituents", in H. Niemann, M. Lang, G. Sagerer: "Recent Advances in Speech Understanding and Dialog Systems", Springer Verlag, Berlin, 1988.

[HIT 86] Hitzenberger, L., Ulbrand, R., Kritzenberger, H., Wenzel, P.: "FACID Fachsprachlicher Corpus informationsabfragender Dialoge", Universität Regensburg, FG Linguistische Informationswissenschaft, 1986.

[KOH 77] Kohler, K.: "Einführung in die Phonetik des Deutschen", Erich Schmidt Verlag, Berlin, 1977.

[KUN 88] Kunzmann, S., Kuhn, T., Niemann, H.: "An Experimental Environment for Generating Word Hypotheses in Continuous Speech", in H. Niemann, M. Lang, G. Sagerer: "Recent Advances in Speech Understanding and Dialog Systems", Springer Verlag, Berlin, 1988.

[LEA 75] Lea, W., Medress, M., Skinner, T.: "A Prosodically Guided Speech Understanding Strategy", IEEE Trans. Vol. ASSP-23, 30-38, 1975.

[NIE 85] Niemann, H., Brietzmann, A., Mühlfeld, R., Regel, P., Schukat, G.: "The Speech Understanding and Dialog System EVAR", in De Mori, R., Suen, C.: "New Systems and Architectures for Automatic Speech Recognition and Synthesis", Springer Verlag, Berlin, S. 271-302, 1985.

[NÖT 88] Nöth, E., Schmölz, S., Niemann, H.: "Prosodic Features in German Speech: Stress Assignment by Man and Machine ", in H. Niemann, M. Lang, G. Sagerer: "Recent Advances in Speech Understanding and Dialog Systems", Springer Verlag, Berlin, 1988.

[REG 88] Regel, P.: "Akustisch-phonetische Transkription für die automatische Spracherkennung", VDI-Verlag, Düsseldorf, 1988.

[UNG 88] Unglaub, J.: "Prosodische Verifikation von Semantik-Hypothesen", Magisterarbeit, Lehrstuhl für Informatik 5 (Musterkennung), Universität Erlangen, (in Vorbereitung).

[VAI 88] Vaissiere, J.: "The Use of Prosodic Parameters in Automatic Speech Recognition" in H. Niemann, M. Lang, G. Sagerer: "Recent Advances in Speech Understanding and Dialog Systems", Springer Verlag, Berlin, 1988.