Elmar Nöth *)
Roswitha Lang

Lehrstuhl für Informatik 5
(Mustererkennung)
Friedrich-Alexander-Universität
Erlangen-Nürnberg
Martensstr. 3
8520 Erlangen

Anton Batliner **)
Wilhelm Oppenrieder

Institut für Deutsche Philologie
Ludwig-Maximilians-Universität
München
Schellingstr. 3
8000 München 40

AUTOMATISCHE GRUNDFREQUENZANALYSEN UND SATZMODUSDIFFERENZIERUNG

1. Grundfrequenzalgorithmen

Für die Intonationsforschung ist eine zuverlässige Bestimmung der Stimmgrundfrequenz von entscheidender Bedeutung. In den letzten Jahrzehnten wurden viele Verfahren zur Bestimmung des Grundfrequenzverlaufs in gesprochenen Äußerungen entwickelt. Dabei handelt es sich um Messungen mit Instrumenten (mechanische, optische, usw.) und um Algorithmen für den Einsatz am Rechner. Eine ausführliche Übersicht über gängige Verfahren findet sich in /Hess 83/.

Die Aufgabe eines jeden Grundfrequenz-Verfahrens besteht aus 2 Schritten, die nicht unbedingt in dieser Reihenfolge ablaufen müssen:

1) Unterteilung des Sprachsignals in stimmhafte und stimmlose Bereiche (im folgenden als SH/SL-Entscheidung bezeichnet).

2) Angabe eines Grundfrequenzwertes für jeden Zeitpunkt der stimmhaften Bereiche. Dazu kann entweder die Dauer einer jeden Periode angegeben, oder das Zeitsignal in gleichlange Zeitscheiben zerlegt und für jeden solchen Zeitbereich ein mittlerer Grundfrequenzwert berechnet werden.

Man unterscheidet daher bei Fehlern von Grundfrequenz-Verfahren auch zwischen SH/SL-Entscheidungsfehlern und fehlerhaften Grundfrequenzwerten in den stimmhaften Bereichen. Hierbei unterscheidet man weiter in Grobfehler (z.B. Oktavsprünge) und Feinfehler (z.B. Fehler aufgrund des Auflösungsvermögens des Verfahrens).

Vergleicht man verschiedene Verfahren, so hängen die Kriterien, anhand derer ein Verfahren als besser oder schlechter bezeichnet wird, sehr stark von der Anwendung ab. Im Rahmen des Spracherkennungssystems EVAR /Niemann 85/ wurden für ein zu erstellendes Prosodie-Modul einige Grundfrequenzalgorithmen implementiert und verglichen. Unter den im System EVAR gestellten Rahmenbedingungen lieferte ein Verfahren nach /Sennef 78/ die besten Ergebnisse. Wie sehr die Anwendung bei der Beurteilung eine Rolle spielt, sieht man z.B. daran, daß im Rahmen von EVAR der Grundfrequenzalgorith-

mus auch bei Beschränkung auf Telefon-Bandbreite (300-3400 Hz) noch vernünftige Werte liefern muß. Eine solche Bedingung ist für einen Phonetiker, der mit Aufnahmen in Tonstudio-Qualität arbeitet, absolut unerheblich.

Bei dem Verfahren nach /Sennef 78/ handelt es sich um ein Frequenzbereichsverfahren. Nach der SH/SL-Entscheidung wird in den stimmhaften Bereichen alle 12.5 msek. für einen 37.5 msek. langen Ausschnitt des Sprachsignals das Fourier-Spektrum erzeugt (damit sind auch bei einer Grundfrequenz von 55 Hz noch 2 volle Perioden in dem betrachteten Bereich). Die Spitzen im Spektrum entsprechen der Grundfrequenz und den höheren Harmonischen. Im Bereich zwischen 100 und 1100 Hz wird die Differenz zwischen den Spitzen im Spektrum ermittelt. Aus der Beziehung

$$\text{Harmonische}_i - \text{Harmonische}_{(i-1)} = \text{Grundfrequenz}$$

ergibt sich für die 12.5 msek. ein Grundfrequenz-Schätzwert. Die Schätzwerte werden mit einem Medianfilter geglättet.

Für die SH/SL-Entscheidung sind momentan drei Varianten implementiert:

1) Da das zu entwickelnde Prosodie-Modul im Rahmen des Sprachverarbeitungssystems EVAR implementiert wird, kann für die SH/SL-Entscheidung der Ausgang des Akustik-Phonetik-Moduls herangezogen werden.

2) Um ein eigenständiges Prosodie-Modul zu ermöglichen, wurden zwei zusätzliche SH/SL-Entscheidungs-Algorithmen implementiert, die auf der Energie in einem Frequenz-Bereich (300-2300 Hz) basieren. In der einen Version wird die Grundfrequenz dort berechnet, wo der Bandpaß eine gewisse Schwelle überschreitet.

3) Bei der zweiten Variante werden in dem Ausgang des Bandpass-Filters signifikante Maxima gesucht, die den Silbenkernen zugeordnet werden. Nur in diesen Silbenkernen wird die Grundfrequenz berechnet.

Der Grundfrequenzalgorithmus soll nun anhand einiger Beispiele erläutert werden; die Sprachdaten stammen aus dem weiter unten in Punkt 2 skizzierten Korpus.

Bildfolge 1 verdeutlicht das Prinzip des Grundfrequenz-Algorithmus. **Bild 1a** zeigt einen 25 msek. langen Ausschnitt aus einer Realisierung des Diphtongs [ɔʏ] (Zeitsignal, weibl. Stimme), **Bild 1b** zeigt das dazugehörige logarithmische Fourier-Spektrum. Man erkennt deutlich die Grundfrequenz und die ersten drei Harmonischen als Spitzen im Spektrum. Der Abstand zwischen den Harmonischen beträgt ca. 250 Hz, ebenso wie die im Zeitsignal gemessene Periodendauer von 4 msek. einer Grundfrequenz von 250 Hz entspricht.

Bild 1a

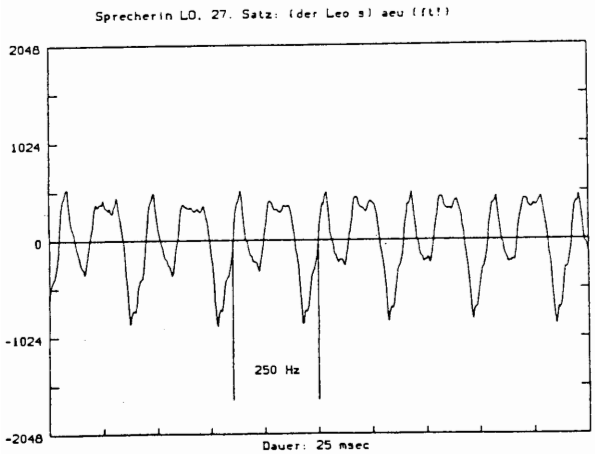
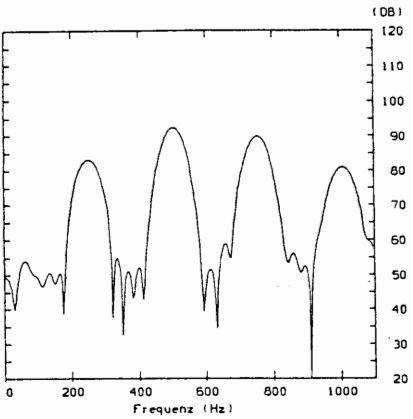


Bild 1b



Im folgenden werden typische Fehler des Algorithmus erläutert. **Bild 2a** zeigt einen 25 msek. langen Ausschnitt aus einer stimmlosen Realisierung des Lautes /g/, der fälschlicherweise als stimmhaft klassifiziert wurde, **Bild 2b** das dazugehörige Spektrum. Bildet man die Differenzen der Maxima, so ergibt sich kein Häufungspunkt. Das Verfahren entscheidet sich für die Differenz zwischen den beiden größten Maxima (450 Hz und 650 Hz), also für eine Grundfrequenz von 200 Hz.

Bild 2a

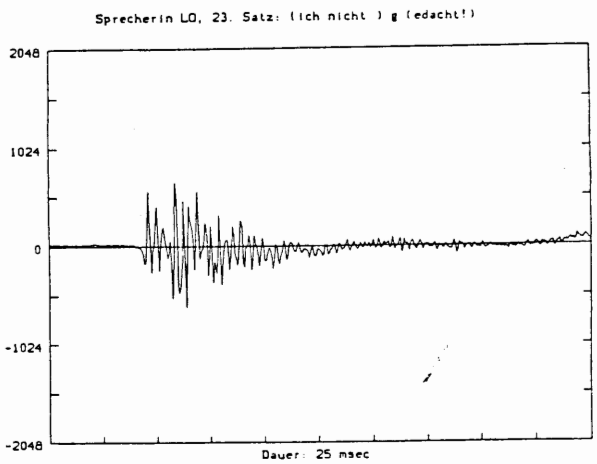
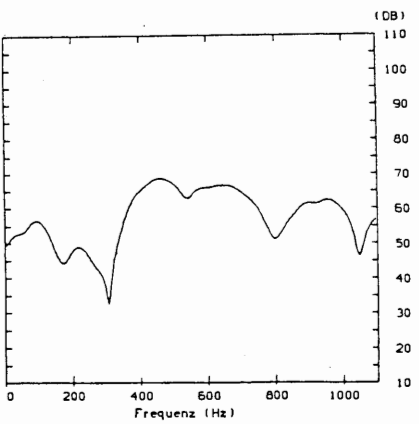


Bild 2b



Bildfolge 3 zeigt einen Oktav-Fehler. **Bild 3a** zeigt 25 msek. des Lautes [m], **Bild 3c** 25 msek. des Lautes [l]. Beide Sprachausschnitte stammen aus einer Realisierung der Wörter *dem Leo* (männl. Stimme); zwischen den Ausschnitten liegen 50 msek. Die Bilder 3b und 3d zeigen die jeweiligen Spektren. In **Bild 3b** erkennt man, daß

die geradzahligen Harmonischen soweit nach unten gedrückt sind, daß sie nicht mehr als Spitzen im Spektrum erscheinen. Der Algorithmus entscheidet sich für das Doppelte der Grundfrequenz (180 statt 90 Hz). Im Zeitsignal dagegen ist die richtige Periodendauer gut zu erkennen. In Bild 3d sind auch die meisten der geradzahligen Harmonischen als Maxima zu sehen. Der Algorithmus entscheidet sich für die richtige Grundfrequenz.

Bild 3a

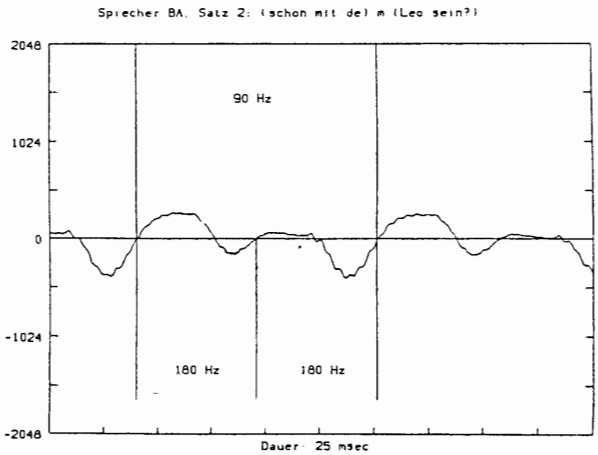


Bild 3b

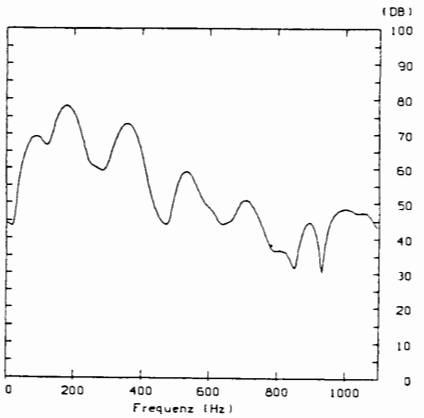


Bild 3c

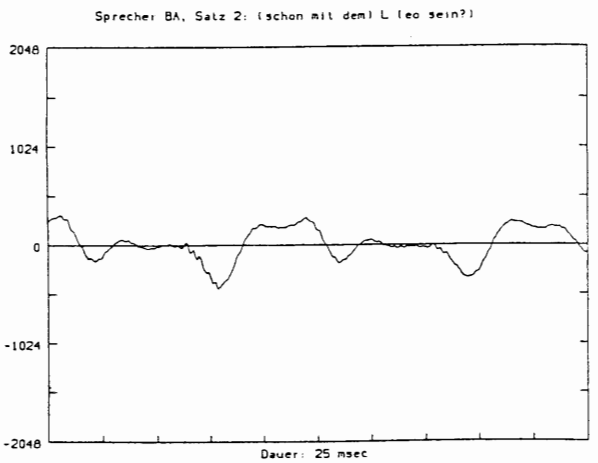
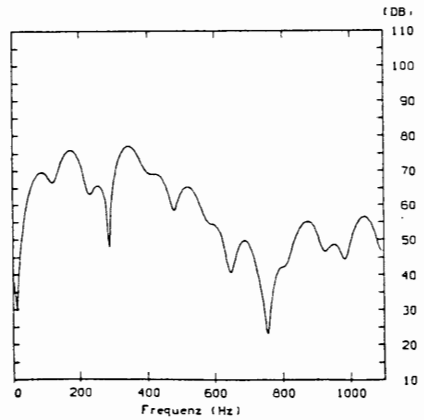


Bild 3d



Oktavsprünge unterscheiden sich von schnellen, vom Sprecher beabsichtigten Grundfrequenzbewegungen dadurch, daß die Bewegungen sehr abrupt sind, während bei einem tatsächlichen Heben bzw. Senken der Stimme die Bewegung relativ langsam ist. Daher ist es möglich, einen großen Teil der Oktavsprünge zu korrigieren, ohne re-

levante Grundfrequenzveränderungen wegzuglätten. Ein solcher Korrekturalgorithmus ist für die nächste Version geplant.

2. Untersuchte Satzmoduskonstellationen

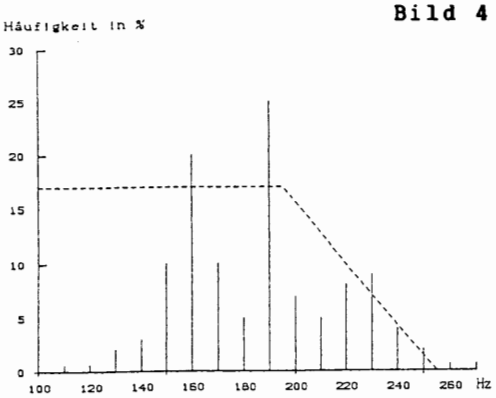
Der soeben dargestellte Grundfrequenzalgorithmus wurde nun bei der Bestimmung der intonatorisch markierten linguistischen Kategorie des Satzmodus eingesetzt. Für eine solche Fragestellung bietet es sich an, den Grundfrequenzalgorithmus und den Bestimmungsalgorithmus zunächst an systematisch erstellten Korpora zu überprüfen, bei denen andere relevante Faktoren wie segmentaler Einfluß, Länge der Äußerung, Redestil etc. nicht variieren. Wir wählten deshalb ein Teilkorpus aus dem DFG-Projekt "Modus-Fokus-Intonation", die sog. "Leo-Sätze". (Das in München zugrundegelegte Satzmodussystem ist in diesem Band an anderer Stelle, in /Batliner 87/, skizziert.) Die Sätze *Der Leo säuft* und *Säuft der Leo* indizieren dabei folgende Modus-Konstellationen: normaler **Aussagesatz**, **Aussagesatz mit Kontrastakzent**, **assertiver Fragesatz**, **Verb-Zweit-Exklamativsatz**; **Verb-Erst-Fragesatz**, **Verb-Erst-Exklamativsatz**. Bei jeder dieser Konstellationen kann der Fokus entweder die NP (mit einem Akzent auf *Leo*) oder das Verb (mit einem Akzent auf *säuft*) umfassen. Sechs Sprecher (3 weibl., 3 männl.) produzierten jede der möglichen Modus-Fokus-Kombinationen mindestens zweimal, wobei die intendierte Lesart nicht explizit vorgegeben, sondern durch einen passenden Vorgängersatz sichergestellt wurde. (Bei der Kombination *Der Theo säuft nicht, der Leo säuft* etwa indiziert der Vorgängersatz einen Aussagesatz mit Kontrastfokus auf *der Leo*, bei *Der Leo trinkt nicht, der Leo säuft* einen Aussagesatz mit Kontrastfokus auf *säuft*.) Für die folgenden Analysen wurden Fälle, bei denen der Grundfrequenzalgorithmus eindeutig fehlerhafte Werte lieferte, per Hand ausgesondert.

3. Bestimmung des Satzmodus

Als Klassifikationsgrundlage wurden die folgenden Parameter gewählt: (Fo-)**Onset**, (Fo-)**Offset**, (Fo-)**Maximum**, (Fo-)**Minimum**, (Fo-)**Mittelwert**, **Steigung** des Fo-Verlaufs am Ende der Äußerung, **Energiemaximum** (über 75 msek. integriert). Einmal gingen alle Fälle in Analyse und Klassifikation ein, um damit eine obere Grenze der Prädiktionsgüte bestimmen zu können; das andere Mal wurde nach dem "leave-one-out"-Verfahren eine Sprecherunabhängigkeit simuliert; dabei werden reihum nach n-1 Sprecher analysiert und der jeweils ausgelassene klassifiziert. Zwei verschiedene Klassifikationsverfahren kamen zur Anwendung:

(i) Bei den in Erlangen durchgeführten Untersuchungen wurde für die Entscheidung Frage/Nicht-Frage ein Schwellwertverfahren benutzt, das neben der harten Entscheidung für eine Klasse noch ein Maß für die Zugehörigkeit zu jeder der beiden Klassen liefert, z.B. für einen Satz i: $\text{Sicherheit}_{\text{Aussage}}$ (Satz i) = 0.9
$\text{Sicherheit}_{\text{Frage}}$ (Satz i) = 0.4.
Die Sicherheitsfunktion nimmt Werte zwischen 0 und 1 an und wird folgendermaßen berechnet: Für einen Parameter (z.B. Offset) wird für jede Klasse ein Histogramm der auftretenden Werte erstellt. Damit wird eine Trapezfunktion konstruiert (im Falle eines Zweiklassenproblems eine einseitig offene Trapezfunktion). Der Beginn der abfallenden Flanke wird so gewählt, daß X Prozent der beobachteten Fälle die maximale Bewertung erhalten. Bild 4 zeigt ein Histogramm von beobachteten Offset-Werten für Aussagen und die daraus erzeugte Sicherheitsfunktion (gestrichelte Linie, x=65). Bei dieser Bewertungsfunktion handelt es sich um eine Fuzzy-Membership-Function /de Mori 83/. Im Augenblick ist es nur möglich, Bewertungsfunktionen für je einen Parameter zu berechnen; es ist aber geplant, mehrere dieser Funktionen zu verknüpfen.

Bild 4



(ii) In München wurde die Diskriminanzanalyse eingesetzt; vgl. zu diesem Verfahren /Batliner 87/ in diesem Band.

4. Ergebnisse

Die Tabelle gibt für verschiedene Satzmoduskonstellationen und Kombinationen von Prädiktorvariablen die richtig klassifizierten Fälle in Prozent wieder. Die erste Zeile, (a)1, zeigt das Ergebnis für die Fuzzy-Membership-Function, alle anderen Zeilen zeigen die Ergebnisse der Diskriminanzanalysen. Die erste Spalte führt die drei verschiedenen Moduskonstellationen auf: (a) alle Fälle (157), Frage vs. Nicht-Frage, Erwartungswert: 50%; (b) Verb-Zweit-Sätze (94 Fälle), Frage vs. Aussage ohne vs. Aussage mit Kontrastakzent vs. Exklamativ, Erwartungswert 25%; (c) Verb-Zweit-Sätze (94 Fälle), Frage vs. Aussage ohne und mit Kontrastakzent vs. Exklamativ, Erwartungswert 33.33%. Die nächste Spalte zeigt die verwendeten Kombinationen der Prädiktorvariablen: Off.: Offset allein; Alle: Offset, Onset, Maximum, Minimum; Off., Energ.: Offset und Energiemaximum; Alle, Energ.: Onset, Offset, Maximum, Mi-

nimum, Energiemaximum. Die dritte Spalte zeigt an, ob nach der leave-one-out-Methode klassifiziert wurde (n-1), oder ob alle Fälle in Analyse und Klassifikation eingingen (n). Für die Konstellationen (b) und (c) wurde die leave-one-out-Methode nicht angewandt, da hier jeweils nur sehr wenig Fälle klassifiziert werden könnten. Die folgenden Spalten geben die verschiedenen Transformationen der Hz-Werte an: (a) keine Transformation, Hz-Rohwerte: Hz; (b) Transformation der Hz-Werte in Halbtonwerte zur Basis 1: Ht; (c) Hz-Werte transformiert zum sprecherspezifischen Basiswert, der sich aus dem vom jeweiligen Sprecher tiefsten erreichten Offset-Wert ergibt: Hz_b; (d) Halbtonwerte transformiert zum sprecherspezifischen Basiswert: Ht_b; (e) Hz-Werte transformiert zum Mittelwert der jeweiligen Äußerung in Hz: Hz_m; (f) Halbtonwerte transformiert zum Mittelwert in Halbtönen: Ht_m.

Tabelle

Konst.	Präd. var.	An.	Hz	Ht	Hz_b	Ht_b	Hz_m	Ht_m
(a)1	Off.	n-1	71.34	67.51	91.72	88.54	91.09	89.17
(a)2	Off.	n-1	70.18	66.25	92.80	90.05	91.50	90.62
	Off.	n	75.16	75.16	92.99	91.72	91.08	90.45
	Alle	n-1	93.80	94.22	92.65	95.52	92.07	92.50
	Alle	n	96.18	94.90	93.63	95.54	93.63	94.27
(b)	Off.	n	34.04	34.04	47.87	47.87	56.38	54.26
	Off., Energ.	n	44.68	44.68	54.26	58.51	55.32	54.26
	Alle	n	56.38	59.57	64.89	56.38	67.02	64.89
	Alle, Energ.	n	67.77	67.02	61.70	63.83	69.15	67.02
(c)	Off.	n	44.68	43.62	64.89	59.57	69.15	68.09
	Off., Energ.	n	56.38	55.32	70.21	67.02	69.15	67.02
	Alle	n	69.15	70.21	71.28	72.34	74.47	73.40
	Alle, Energ.	n	70.21	72.34	70.21	72.34	75.53	76.60

Auf eine eingehende Diskussion der Ergebnisse müssen wir verzichten; wir beschränken uns auf eine stichpunktartige Zusammenfassung:

1) Die beiden verschiedenen Bestimmungsalgorithmen erzielen in etwa die gleichen Ergebnisse, vgl. (a)1 und (a)2 in der Tabelle.

2) Grundsätzlich werden mit den transformierten Werten bessere Ergebnisse erzielt, wobei kaum ein Unterschied zwischen der Basiswert- und der Mittelwert-Transformation besteht. Der Algorithmus braucht also keine Information, die, wie der Basiswert, in der jeweiligen Äußerung nicht vorhanden ist.

3) Die Steigung der Grundfrequenz am Ende der Äußerung als Prädiktorvariable führt zu einer richtigen Klassifikation von 77.07% (Erlangen) bzw. 80.47% (München). Der Offset als einzige Prädiktorvariable erzielt ein um ca 10% besseres Ergebnis. Das

erklärt sich aus der Tatsache, daß die Fragen fast immer einen hohen Offset, aber doch manchmal bei einem hohen Offset einen leichten Abfall am Ende und damit eine negative Steigung aufweisen.

4) Alle vier Fo-Variablen als Prädiktoren ergeben im Verhältnis zum Offset allein eine bis zu 5% bessere Prädiktion; darin dürfte sich die Tatsache widerspiegeln, daß neben dem Offset auch der übrige Fo-Verlauf eine zwar untergeordnete, aber doch relevante Rolle bei der Distinktion Frage vs. Nicht-Frage spielt.

5) Die Erkennungsrate für die Verb-Erst-Sätze (Frage vs. Exklamativ) ist in der Tabelle nicht aufgeführt; sie ist genauso hoch wie die für das ganze Korpus, mit einem Maximum bei 95.52%. Bei den Verb-Zweit-Sätzen werden die weniger trennscharfen Klassen, vgl. (b) und (c) in der Tabelle, immerhin noch mit signifikanter Sicherheit unterschieden.

Zu diesen Punkten sind weitere Untersuchungen geplant. Insbesondere soll geprüft werden, ob durch eine Bestimmung des fokussierten Elements in der Äußerung und durch die Berücksichtigung von Dauerverhältnissen eine noch bessere Trennschärfe zwischen den Nicht-Fragen erzielt werden kann.

*) Die Arbeiten an der FAU Erlangen entstanden im Rahmen des BMFT Verbundvorhabens "Sprachverarbeitung" in Zusammenarbeit mit der Siemens AG, München.
**) Die Arbeiten an der LMU München entstanden im Rahmen des DFG-Projekts "Modus-Fokus-Intonation".

Literaturverzeichnis:

/Batliner 87/ A. Batliner: Der Einsatz der Diskriminanzanalyse zur Prädiktion des Modus. In: H.G. Tillmann, G. Willée: Analyse und Synthese gesprochener Sprache. Vorträge im Rahmen der Jahrestagung 1987 der Gesellschaft für Linguistische Datenverarbeitung e.V., Bonn, 4.-6. März 1987. (Dieser Band)

/Hess 83/ W. Hess: Pitch Determination of Speech Signals. Springer Verlag, Berlin, 1983.

/de Mori 83/ R. de Mori: Computer Models of Speech Using Fuzzy Algorithms. Plenum Press, New York, 1983

/Niemann 85/ H. Niemann, et al.: The Speech Understanding and Dialog System EVAR. In: R. DeMori, C. Y. Suen: New Systems and Architectures for Automatic Speech Recognition and Synthesis. NATO ASI Series F, Springer Verlag, Berlin, 271-302, 1985.

/Sennef 78/ S. Sennef: Real-time harmonic pitch detector. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, 358-364, 1978.