ZU ERGEBNISSEN DER INTONATIONSFORSCHUNG AN DER HUMBOLDT-UNIVERSITÄT ZU BERLIN

Dieter Mehnert

Sektion Rehabilitationspädagogik und Kommunikationswissenschaft, Bereich Phonetik/Technik der Humboldt-Universität zu Berlin

1. Einleitung

Stellt man die Kommunikationsfähigkeiten des Menschen denen moderner Rechner gegenüber, so wird das Problem der Kommunikationsgeschwindigkeit im Dialog mit dem Automaten deutlich. Die Kommunikationsgeschwindigkeit ist im Verhältnis zu Verarbeitungszeiten der Information sehr langsam. Sie kann nur erhöht werden durch die Anpassung des Rechners an die Kommunikationsfähigkeiten des Menschen. Möglich ist das über Technologien, die eine Kommunikation mit dem Rechner über die natürliche Sprache realisieren. Die mündliche Sprache ist dabei von besonderer Bedeutung, weil sie das direkteste Kommunikationsmittel darstellt und ohne Verwendung von Zwischendatenträgern verwendet werden kann. Die Erfahrungen haben gezeigt, daß durch die Sprachkommunikation mit dem Automaten herkömmliche Ein- und Ausgabegeräte keineswegs ersetzt werden können. Es hat sich aber herausgestellt, daß sich neue Anwendungsmöglichkeiten anbieten, die aus den Besonderheiten der Sprache resultieren.

Der enorme technologische Fortschritt auf den Gebieten der Mikroelektronik hat die erforderlichen Voraussetzungen für die Sprachkommunikation mit Automaten geschaffen. Dies gilt gleichermaßen für die Spracherkennung und die Sprachsynthese. Da für die Verwirklichung des Dialogs beide Komponenten benötigt werden, steigt mit den Anwendungsmöglichkeiten der Spracherkennung auch der Bedarf an leistungsfähigen Synthesesystemen.

Effektiv wird die Mensch-Maschine-Kommunikation nur dann sein, wenn die Spracherkennung sicher arbeitet und die Sprachqualität bei der Sprachausgabe einem Standard entspricht, der weitgehend die Natürlichkeit des vom humanen Sprachgenerator erzeugten Sprachsignals erreicht und damit eine hohe Verständlichkeit hat, wenn sich beide Kommunikationspartner also problemlos verstehen. Intonationsstrukturen spielen bei der Realisierung dieser Forderung, wie überhaupt im formalen linguistischen Kommunikationscode, eine wesentliche Rolle.

2. Wirksamkeit prosodischer Parameter

Intonation wird als ein Komplex verstanden, in dem die 3 prosodischen Parameter Tonhöhe, Lautstärke und Dauer kompliziert zusammenwirken. Diese Auffassungen vom Komplexcharakter der Intonation sind experimentell bestätigt worden. Es erhebt sich die Frage, welche Wirksamkeit die prosodischen Parameter einzeln oder kombiniert auf das Sprachsignal ausüben, wie groß der Einfluß auf die Verständlichkeit und Natürlichkeit ist und ob es eine Hierarchie unter den Parametern gibt. Daraus sollte abgeleitet werden, welchem oder welcher Gruppe von Parametern man sich bei der Sprachsynthese besonders widmen muß.

2.1. Grundfrequenz-Mikrokontur

Zunächst richtete sich die Untersuchung auf die Feinstruktur des Anregungssignals. Es war das Ziel, charakteristische, lautspezifische Grundfrequenzverläufe an Lautübergängen von natürlichem Sprachmaterial festzustellen und den Versuch zu unternehmen, das Charakteristische zu schematisieren.

Des weiteren sollten die gefundenen Grundfrequenzmuster auf synthetische Lautverbindungen angewendet werden, um Kenntnisse über ihre perzeptive Relevanz unter dem besonderen Aspekt des Verständlichkeits- und Natürlichkeitsgewinns zu erlangen.

Zur Auswahl des Untersuchungsmaterials wurde von der Häufigkeitsstatistik deutscher Sprachlaute ausgegangen. Die Zusammenstellung der Beispielwörter wurde so vorgenommen, daß die jeweils interessierende Lautverbindung in unterschiedlichen Positionen enthalten war. Es sollte u.a. die Frage beantwortet werden, ob es für lautübergangsspezifische f_G-Konturen positionelle Varianten gibt.

Bei der Auswertung des analysierten Sprachmaterials wurden Daten über f_G-Verläufe an Konsonant-Vokal-Lautübergängen bei isoliert gesprochenen Wörtern und Kontextwörtern ermittelt. Zugleich wurden auch Werte über f_G-Hübe in Silben, Wörtern und Sätzen gewonnen, die eine Grundlage für den späteren Aufbau von suprasegmentalen Makro-f_G-Konturen darstellen.

In Abb. 1 ist zu erkennen,

- daß die mittleren Frequenzhübe Δf_{Gm} am Lautübergang in isoliert gesprochenen Wörtern annähernd gleich denen in Kontextwörtern sind, d.h., daß die Mikro-f_G-Kontur unabhängig von einer bewußt gesteuerten Makro-f_G-Kontur ist und
- daß die mittlere Frequenzänderung in der Lautübergangsumgebung von Konsonant-langer Vokal etwas größer ist, als die in der Lautübergangsumgebung Konsonant-kurzer Vokal. Das zeigt, daß die deutlichere Ausformung aller gedehnt gesprochenen Vokale einen größeren Frequenzhub zur Folge hat.

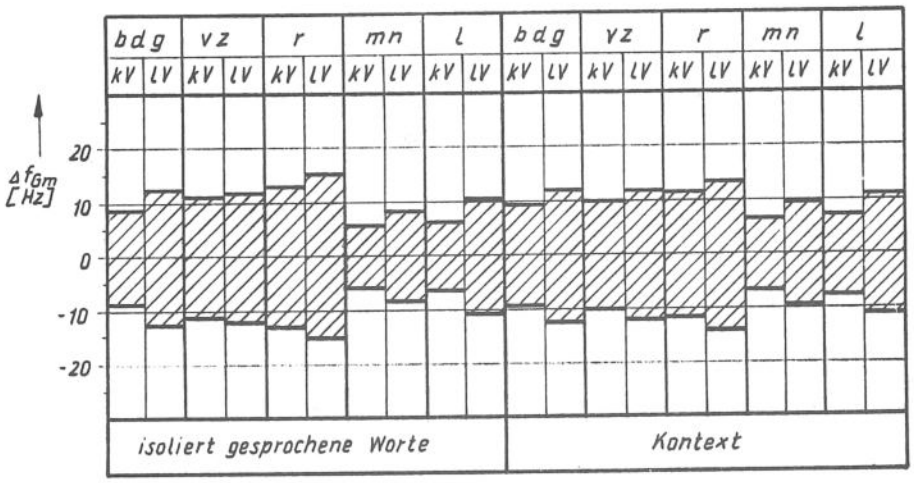


Abb. 1: Mittl. Frequenzänderung an LÜ-Umgebung (bez. auf 120 Hz)

Abb. 2 zeigt die gefundenen f_G-Konturen in der Lautübergangsumgebung der ausgewählten KV-Gruppen. Es ist zu erkennen, daß Grundfrequenzbewegungen an Lautübergängen lautspezifisch sind.

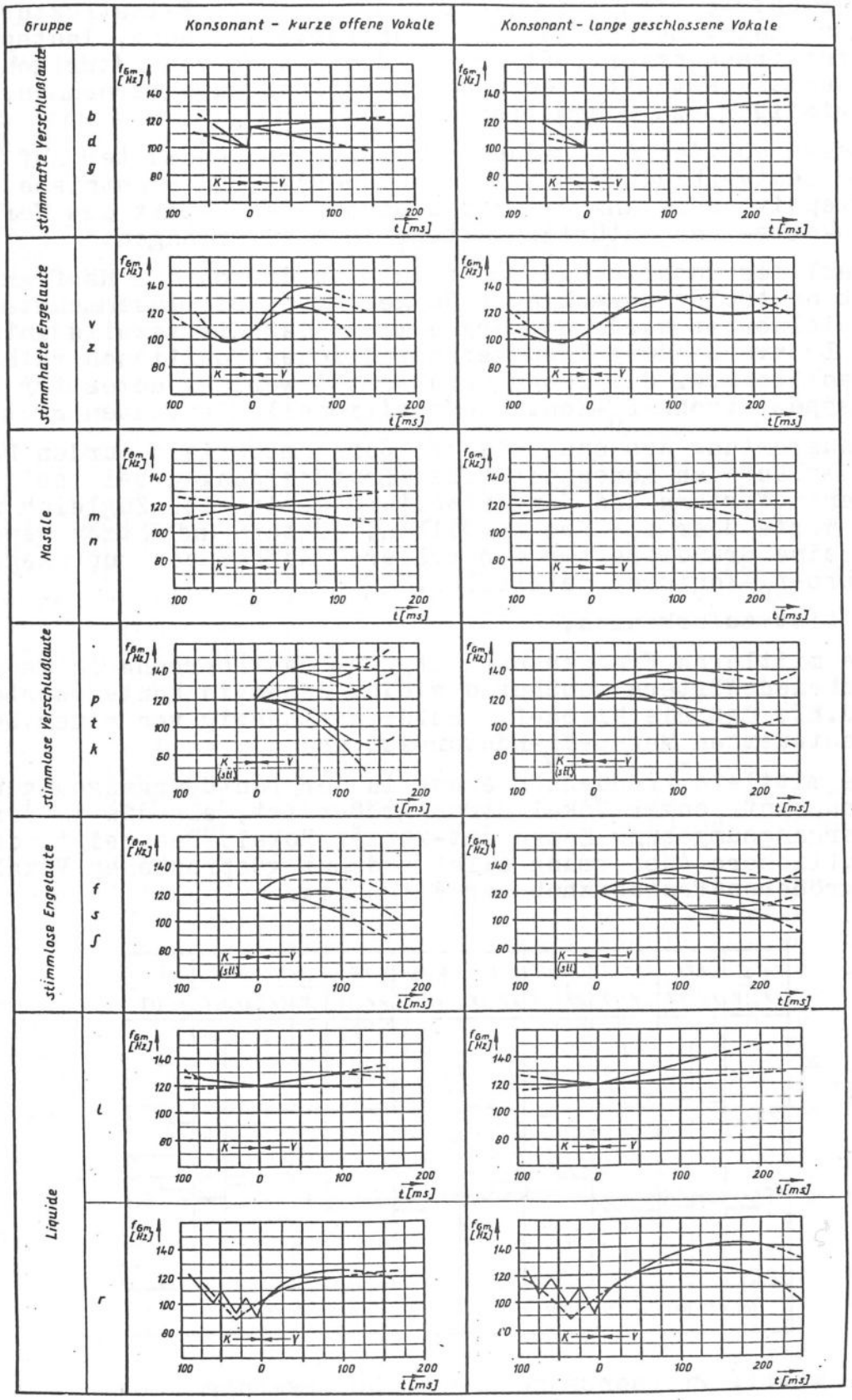


Abb. 2: Umgebung vom Lautübergang, männl. Sprecher (gemittelte Werte, bez. auf f_{Gm} = 120 Hz)

Sie treten als

- Sprung (stimmhafte Konsonanten-Vokale),
- verschliffen (stimmhafte Engelaute und r-Vokale),
- fließend (Nasale und l-Vokale) und
- abrupt (stimmlose Verschlußlaute und stimmlose Engelaute-Vokale)

auf, dabei ist die lautspezifische f_G-Kontur unabhängig von der Position und Länge der Vokale.

Entsprechend diesen Ergebnissen sind 5 lautübergangsspezifische f_G-Konturen für die technische Sprachsynthese vorgeschlagen und an einem zeichengesteuerten Sprachsynthetisator mit einem speziell dazu entwickelten f_G-Konturengenerator aufgebaut und erprobt worden /1, S. 175/.

Hörer		ungeübt							
Kons.-Vokal Gruppen		einzeln				gemischt			
f_G-Muster		r	f	m	nv	r	f	m	nv
Gruppen der Kons.-Vokal Verbindungen	b,d,g	87,4	71,4	68,7	98,3	89,8	72,2	62,0	86,0
	v,z	98,4	63,3	75,0	96,6	90,4	49,3	53,6	94,6
	m,n	98,0	92,0	96,0	100	91,5	92,8	92,8	100
	p,t,k	89,8	-	64,0	99,2	90,8	-	67,4	94,2
	f,s,ʃ	100	-	87,4	96,1	98,1	-	89,7	97,6
	l,r	98,2	-	73,4	99,1	80,6	-	51,7	88,7

r - lautspez.
f - nicht lautspez.
m - monoton
nv- natürlich, über BP verzerrt

Abb. 3: Ergebnisse der Hörversuche (Gruppenverständlichkeit in %)

In Abb. 3 sind auszugsweise die Resultate der angestellten Perzeptionsversuche zusammengetragen. Es ist zu ersehen,

- daß synthetisch aufgebaute Beispiele (Logatome) mit lautspezifischen f_G-Konturen von ungeübten Hörern sicherer erkannt werden als Beispiele mit monotoner oder nicht lautspezifischer Kontur und
- daß Unterschiede innerhalb der einzelnen KV-Gruppen bestehen, die im Zusammenhang mit der Kompliziertheit der Signalstruktur zu erklären sind.

Alle mit lautübergangsspezifischen f_G-Konturen versehenen Strukturen ergaben gegenüber monotonen und nicht lautspezifischen f_G-Konturen höhere Verständlichkeitswerte.

Es kann für diesen Abschnitt zusammenfassend festgestellt werden, daß die Grundfrequenz-Feinstruktur nicht in allen Einzelheiten perzeptiv relevant zu sein scheint, daß aber charakteristische Mikro-f_G-Verläufe, auch Periodizitätsschwankungen < 4-5 Perioden für die Natürlichkeit und somit auch für die Verständlichkeit verantwortlich sind und eine katalytische Wirkung auf den Sprachinformationsfluß ausüben.

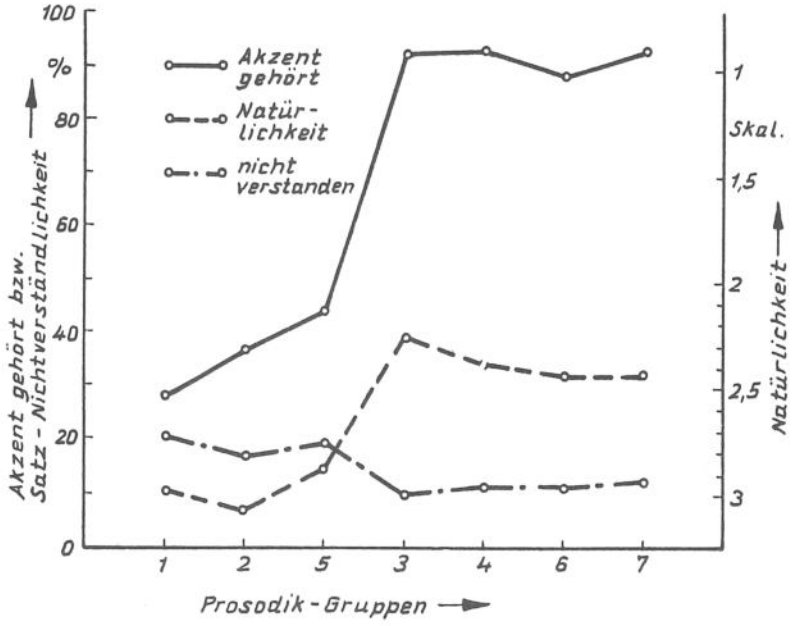
2.2. Grundfrequenz-Makrokontur

In einem weiteren Untersuchungsabschnitt wurden Sätze synthetisch, zunächst monoton, aufgebaut und mit den prosodischen Merkmalen Grundfrequenz, Intensität und Zeit, entsprechend Abb. 4 ergänzt.

Gruppe 1	f_G J (t)	Gruppe 4	(f_G) J (t)
Gruppe 2	f_G (J) t	Gruppe 5	f_G (J) (t)
Gruppe 3	(f_G) J t	Gruppe 6	(f_G) (J) t
		Gruppe 7	(f_G) (J) (t)

○ - variable
übrige - const.

Abb. 4: Perzeptionstest Strategie



Gr.1 < Gr.2 < Gr.5 << Gr.3 < Gr.4 ≈ Gr.6 ≈ Gr.7

$t_{(m)} < J < J+t << f_G < f_G+t \approx f_G+J \approx f_G+J+t$

Abb. 5: Wirksamkeit prosodischer Parameter

Hörergruppen hatten die Aufgabe, das synthetische Sprachmaterial hinsichtlich Verständlichkeit, Akzentwahrnehmung und Natürlichkeit zu beurteilen. Die Zusammenfassung aller auditiv ermittelten Ergebnisse der Perzeptionsversuche ist in Abb. 5 dargestellt. Es ist zu erkennen, daß, sobald der Grundfrequenzparameter allein oder mitvertreten ist, die Werte für Verständlichkeit, Natürlichkeit und Akzentwahrnehmung erheblich zunehmen. Der Parameter Grundfrequenz ist in der Lage, beim Hörer mehr akustisches Referenzwissen zu aktivieren, als es die Parameter Intensität oder Zeit oder beide zusammen vermögen. Der Hörer wird bei Anwesenheit des Parameters Grundfrequenz deutlich zu größerer Perzeptionsleistung veranlaßt.

Das bedeutet für die Anwendung in der Sprachsynthese, daß es sicher genügt, sich vorerst dem Parameter Grundfrequenz und seinen Variationen zuzuwenden. Der Gewinn bei zusätzlicher Berücksichtigung der anderen beiden Parameter wird, im Verhältnis zum Aufwand gesehen, gering sein. Letztlich ist, zumindest in den europäischen Sprachen, die Funktion der Intonation hinsichtlich ihrer kommunikativen Aufgabe im wesentlichen an Grundfrequenzbewegungen gebunden. Der Parameter Grundfrequenz und dessen zeitlicher Verlauf ist dominierend für die Akzentwahrnehmung, die Natürlichkeit und somit auch für die Verständlichkeit.

3. Intonationsmodell im Deutschen

Betrachtet man bisherige Untersuchungen zur Intonation, so kann man feststellen, daß neben einigen Ausnahmen fast alle sprachwissenschaftlich orientierten Methoden das Ziel verfolgt haben, die Tonhöhenphänomene in der Sprache hinsichtlich ihrer distinktiven Funktion zu beschreiben. Danach war zu erwarten, daß die Analyse der deutschen Intonation von einer anderen Methode profitieren könnte, die z.B. alle Verbindungen zur sprachwissenschaftlichen Funktion vermeidet und Fragen zu beantworten versucht, wie:

- Ist es überhaupt notwendig, die Intonation als eine kontinuierliche Grundfrequenzbewegung zu beschreiben oder könnte eine Approximation durch diskrete Grundfrequenzänderungen als eine Funktion der Zeit erfolgen?
- Welches ist die kleinste Einheit für eine Wahrnehmungsbeschreibung und wieviele Einheiten gibt es?
- Welches sind ihre akustisch-phonetischen Eigenschaften und wie groß ist der akustische Toleranzbereich?

Alle diese Fragen befassen sich mit den melodischen Aspekten der Äußerung und münden in eine Darstellung der wahrnehmungsrelevanten Tonhöheneigenschaften der Sprache, die in enger Beziehung zu den physiologischen Tätigkeiten des Sprechers stehen. Das heißt, daß eine Beschreibung der Tonhöhenbewegung zu den sprecherischen Aktivitäten führt und zu Modellen, die in der Kommunikation zwischen Sprecher und Hörer repräsentiert werden.

Verfolgt man auditiv den Intonationsverlauf einer Äußerung, so bemerkt man, daß sich Tonhöhenanstiege und -abfälle und Abschnitte einer fast gleichbleibenden Tonhöhe abwechseln. Betrachtet man Mehrkanal-Registrierungen derartiger Äußerungen, findet man den Gehörseindruck bestätigt. Demnach ist also für eine mögliche Approximation der Grundfrequenz von Interesse, wieviele unterschiedliche Anstiege, Abfälle und quasistationäre Abschnitte sich unterscheiden lassen und ob diese segmentalen Tonhöhenbewegungen eine gemeinsame Bezugslinie aufweisen.

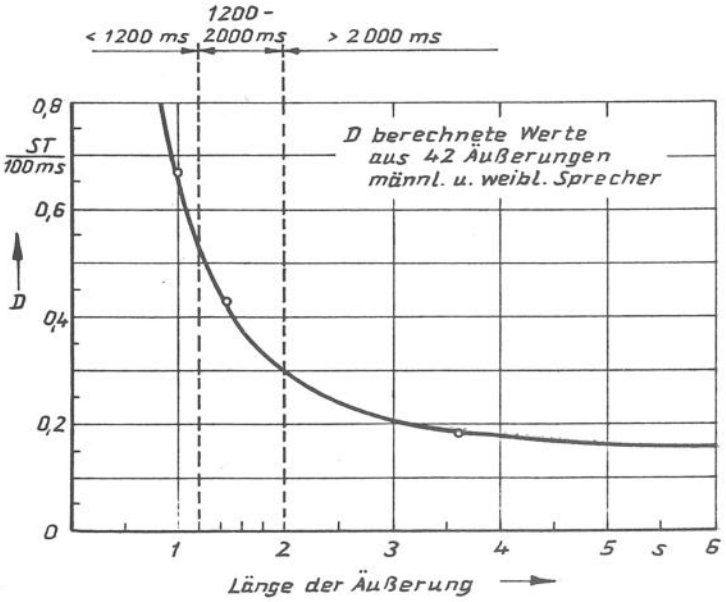


Abb. 6: Deklination als Funktion der Länge der Äußerung

Das analytische Zuhören, ebenso wie die apparative Analyse eines repräsentativen Sprachmaterials ergaben, daß in den meisten Äusserungen, wo keine bemerkenswerten Anstiege oder Abfälle vorkommen, die Tonhöhe nicht monoton bleibt, sondern allmählich nach unten abweicht. Dieses Phänomen wird als Deklination bezeichnet/2/.

Die Auswertung des gesamten Analysematerials ergab für die Deklination eine Abhängigkeit nach Abb. 6.

Danach wurden für die Sprachsynthese 3 Deklinationstypen, abhängig von der Länge der Äußerung vorgeschlagen (Abb. 7). Die nach unten abweichende Linie kann nun als Grundlinie (0) angesehen werden, der die übrigen Tonhöhenbewegungen überlagert sind. Die Deklinationslinie verbindet ebenfalls Anstiege und nachfolgende Abfälle, d.h., sie kann auch auf einem höheren Niveau (Ø) weitergeführt werden. Für den Abstand beider Linien, den Frequenzhub Δf_G, sind entsprechende Werte gefunden und in Abb. 8 gemittelt dargestellt.

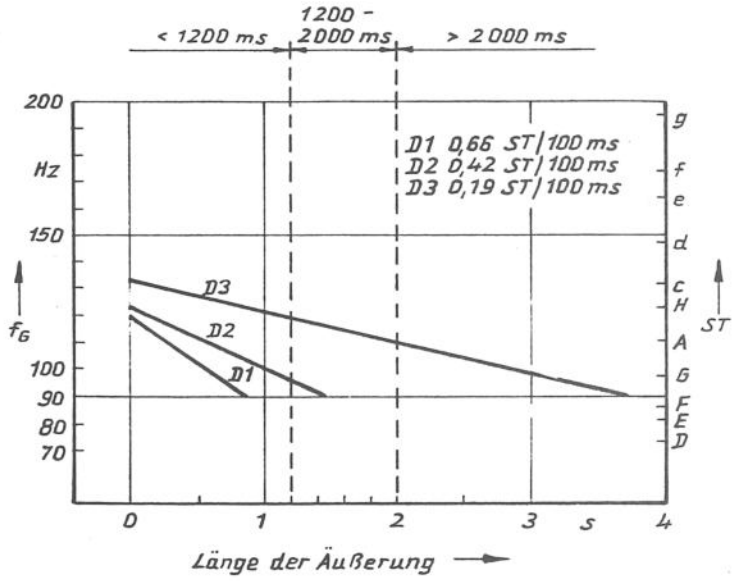


Abb. 7: Deklinationstypen in Relation zum f_G-Bereich

Die Deklination hat für den synthetischen Aufbau der Grundfrequenzkontur eine große Bedeutung. Wird die Deklination durch eine monotone Grundlinie ersetzt, wird der Eindruck der Natürlichkeit sogleich erheblich herabgesetzt.

Zu hören und im Registriermaterial zu beobachten sind desweiteren Tonhöhenanstiege und -abfälle. Sie werden als Übergänge von einer niedrigen zu einer höheren bzw. umgekehrt von einer hohen zu einer niedrigen Tonhöhe wahrgenommen. Diese repräsentieren sich in verschiedenen Formen als steile und flache Übergänge.

In dem gesamten Material konnten nun verschiedene Anstiegsformen und Formen des Abfalls gefunden und hinsichtlich Position und Funktion in der Silbe klassifiziert werden. Schließlich wurden 4 Anstiegsformen (1, 2, 3, 4) und 4 Formen des Abfalls der Grundfrequenz (A, B, C, D) unterschieden. Aus den experimentell gesammelten Daten sind so für die Wahrnehmung relevante segmentale Tonhöhenkonturen (Minimaleinheiten) herausgefunden worden, die die Grundlage für schematisierte größere Intonationseinheiten, für suprasegmentale Grundfrequenzkonturen bilden (Abb. 8).

Die Feststellung eines Grundmodells und einiger Varianten führt nun zu der Frage, ob ein derartiges Modell zur Beschreibung der Intonation überhaupt ausreicht. Vom Standpunkt der Kommunikation gäbe es keinen Grund, wonach es weitere Modelle geben sollte, denn das Grundmodell ist in der Lage, eine Äußerung durch die Intonationskontur zu komplettieren, es beinhaltet Tonhöhenbewegungen, die Silben zu betonten Silben

Tonhöhenbewegung		zur Akzentstelle beitragend	Position in der Silbe	$\Delta f_{gem.}$ [ms]	$\Delta f_{gem.}$ [ST]	Symbol
Anstiege	1	ja	Anfang	100 - 150	4 - 5	Silbe 1 ϕ
	2	ja	Ende	50	2 - 3	ϕ 2A
	3	nein	Ende	100	4 - 5	0 3 ϕ
	4	ja oder nein	geht über variierende Anzahl von Silben	offen	4 - 5	0 4 ϕ
Abfälle	A	ja	Ende	75-100	4 - 5	ϕ A
	B	ja	Ende	50	2 - 3	ϕ B ϕ
	C	nein	Anfang bis Mitte	75 - 100	4 - 5	ϕ C 0
	D	ja oder nein	geht über variierende Anzahl von Silben	offen	4 - 5	1 D 0

Abb. 8: Daten für wahrnehmungsrelevante segmentale Tonhöhenbewegungen

machen können, und es kann syntaktische Strukturen innerhalb einer Äußerung verbinden usw. Es gibt jedoch Hinweise, daß weitere Modelle existieren müssen, die sich vom Grundmodell unterscheiden, letztlich auch aus der Feststellung, daß die Alltagssprache nicht so lebendig klingen würde, wenn sie nur aus der stereotypen Verkettung von Grundmodellen bestände. Alle diese Modelle sind synthetisch mittels eines speziellen f_G-Konturengenerators aufgebaut und danach perzeptiv überprüft worden. Die Aussagen der Hörer waren erwartungsgemäß eindeutig, die Intonationskontur 'Deklination + Akzent' wurde höher bewertet als nur 'Deklination' und beide zusammengenommen wesentlich höher als 'monoton'. Die Auswertungen der Registrierungen und die Hörergebnisse lassen die Schlußfolgerung zu, daß sich die Intonation des Deutschen durch eine Reihe von Grundfrequenz-Konfigurationen beschreiben läßt /3, S. 134/.

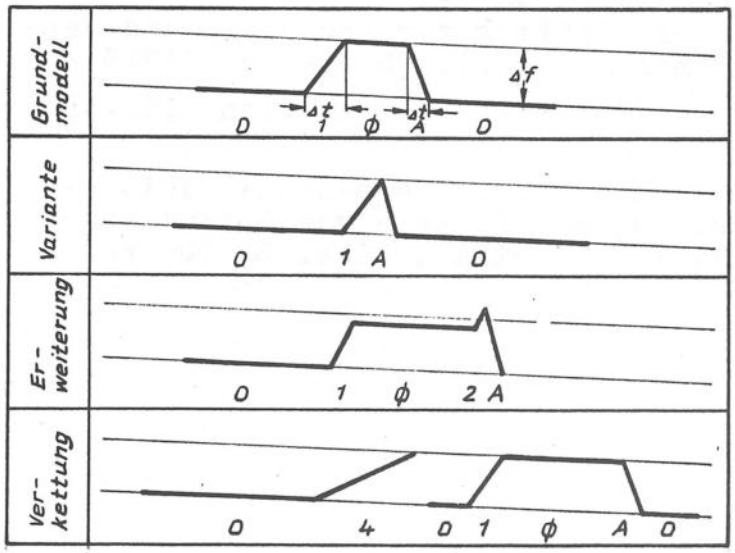


Abb. 9: Modelle suprasegmentaler f_G-Konturen

4. Anwendung für Sprachkommunikationsgeräte

Wie nachgewiesen werden konnte, leisten die beschriebenen Grundfrequenzmodelle einen Beitrag zur Qualitätsverbesserung synthetischer Sprache.

Realisiert wurde eine zunächst halbautomatische Intonationssteuerung des Sprachsynthesegerätes ROSY 4201 (TU Dresden), d.h., daß durch gesetzte Zusatzzeichen in der Phonemkette an den zu akzentuierenden Stellen der Rechner die gewünschte suprasegmentale f_G-Kontur selbständig aufbaut. (Später werden diese Zusatzzeichen bereits bei der Graphen-Phonem-Umsetzung mitbehandelt.) Der Rechner ermittelt aus den weiteren Satzzeichen die Länge des Satzes, bestimmt die entsprechende Deklinationslinie und berechnet, unter Berücksichtigung der Position der Akzentstellung, eine der Hochlautung angenäherte suprasegmentale Grundfrequenzkontur, die dann über den Synthetisator realisiert wird. Das Programm erlaubt die terminale Intonationskontur mit einer oder mehreren Akzentstellen und auch die progrediente Intonationskontur, wenn die Satzteile durch 'Komma' oder 'und' voneinander getrennt (gekennzeichnet) sind. Der zweite Satzteil wird dann hinsichtlich der Akzentstellen wie ein autonomer Satz behandelt.

Die gleichen Grundfrequenzmodelle lassen sich auch zur Tonhöhensteuerung eines Elektrolarynx verwenden, was zu einer erheblichen Verbesserung der Verständlichkeit der Elektrolarynx-Sprache im Vergleich zur monotonen beiträgt. Aus den Erfahrungen, die mit der Grundfrequenzsteuerung bei der Sprachsynthese gemacht worden sind, wurde ein Steuerteil für den Elektrolarynx entwickelt, der eine Reihe von 'normierten' f_G-Bewegungen allein auf Abruf erzeugen kann.

Mit diesen Elementen ist die Erzeugung des Grundmodells und seiner Varianten möglich. Das bedeutet zwar eine Einschränkung, sie erscheint aber aus dem Grunde erlaubt, da sichetwa 60 - 70 % aller Äußerungen im Deutschen mit dem Grundmodell und seinen Varianten realisieren lassen. Ein Grundmodell kann auch ohne wesentliche Natürlichkeitseinbuße eine kompliziertere f_G-Kontur ersetzen.

Literatur

/1/ Mehnert, D.: Grundfrequenzanalyse und -synthese der stimmhaften Anregungsfunktion, ein Beitrag zur Erzeugung und Verarbeitung sprachlicher Signale, Diss.A, TU Dresden (1975)

/2/ Cohen, A., H.'t Hart: On the anatomy of intonation, Lingua 19 (1967), 177 - 192

/3/ Mehnert, D.: Analyse und Synthese suprasegmentaler Intonationsstrukturen des Deutschen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme, Diss. B, TU Dresden (1985)