MODELLIERUNG VON INTONATIONSKONTUREN DES DEUTSCHEN - ANWENDUNGEN FÜR SPRACHKOMMUNIKATIONSGERÄTE

DIETER MEHNERT

Sektion Rehabilitationspädagogik und
Kommunikationswissenschaft, Bereich Technik,
Humboldt-Universität zu Berlin, DDR

ZUSAMMENFASSUNG

Aus einer umfangreichen Untersuchung repräsentativen Sprachmaterials konnten für die Wahrnehmung relevante segmentale und suprasegmentale Grundfrequenzkonturen ermittelt werden, die die Grundlage für schematisierte Grundfrequenzmodelle bilden. Es wird an Beispielen gezeigt, welchen Einfluß die Grundfrequenzkonturen auf die Wahrnehmung künstlich erzeugter Sprachsignale ausüben und inwieweit Grundfrequenzmodelle geeignet sind, Sprachkommunikationsgeräte zu qualifizieren.

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 Vernältnis 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 eine Sprachkommunikation mit Automaten geschaffen. Dies gilt gleichermaßen für die Spracherkennung und 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.

WIRKSAMKEIT PROSODISCHER PARAMETER

Intonation wird als ein Komplex verstanden, in dem die 3 prosodischen Parameter Tonhöhe, Lautstärke und Dauer kompliziert zusammenwirken. Es erhebt sich die Frage, welche Wirksamkeit die prosodischen Parameter einzeln oder kombiniert auf das synthetische 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. Dazu wurden in einer größeren Untersuchung /1/ Sätze synthetisch, zunächst monoton, aufgebaut und mit den prosodischen Merkmalen Grundfrequenz, Intensität und Zeit entsprechend Abb. 1 ergänzt.
Hörergruppen hatten die Aufgabe, das synthetische Sprachsignal hinsichtlich Verständlichkeit, Akzentwahrnehmung und Natürlichkeit zu beurteilen. Die Zusammenfassung aller auditiv ermittelten Ergebnisse der Perzeptionsversuche ist in Abb. 2 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.

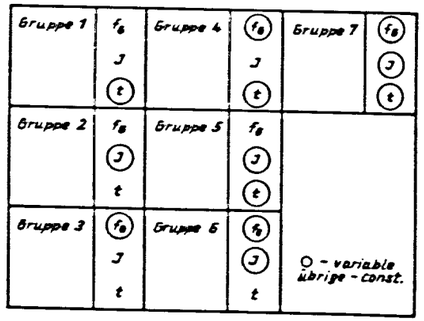


Abb. 1 Perzeptionstest - Strategie

INTONATIONSMODELLE 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 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 Äußerungen 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 Tonhöhenbewegungen zu den sprecherischen Aktivitäten führt und zu Modellen, die in der Kommunikation zwischen Sprecher und Hörer repräsentiert werden.

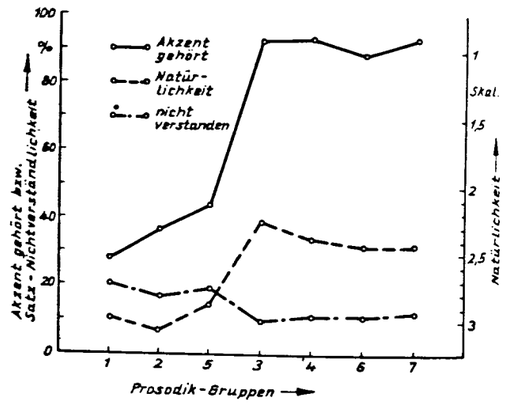


Abb. 2 Wirksamkeit prosodischer Parameter

Verfolgt man auditiv den Intonationsverlauf einer Äußerung, so bemerkt man, daß sich Tonhöhenanstiege und -abfälle und Abschnitte einer fast gleichbleibenden Tonhöhe abwechseln. Betrachtet man Mehrkanalregistrierungen 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.
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. 3.
Danach wurden für die Sprachsynthese 3 Deklinationstypen, abhängig von der Länge der Äußerung, vorgeschlagen (Abb.4). Diese gefundenen Werte für deutsche Satzrealisierungen sind mit denen in der Literatur veröffentlichten Daten anderer europäischer Sprachen vergleichbar/1/.

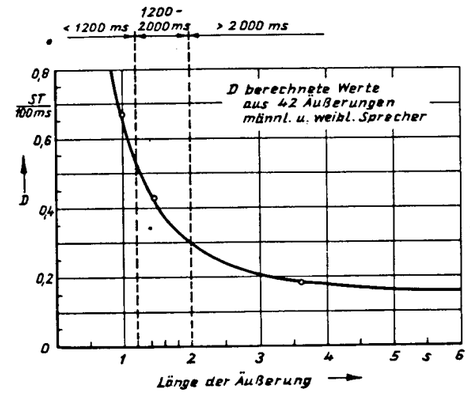


Abb. 3 Deklination als Funktion der Länge der Äußerung

Die nach unten abweichende Linie kann nun als Grundlinie (O) angesehen werden, der die übrigen Tonhöhenbewegungen überlagert sind.

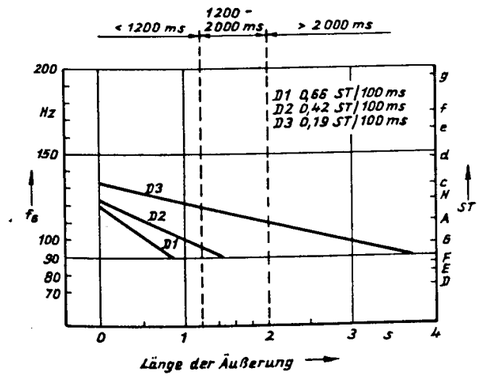


Abb. 4 Deklinationstypen in Relation zum f_G-Bereich

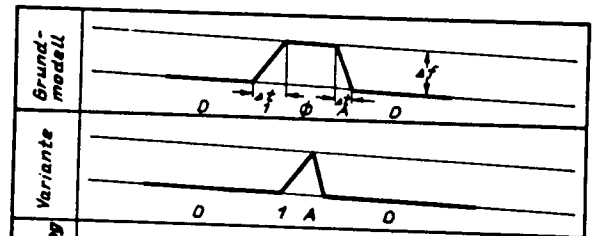
Die Deklinationslinie verbindet ebenfalls Anstiege und nachfolgende Abfälle, d.h. sie kann auch auf einem höheren Niveau (Ø) weitergeführt werden. Für den Abstand beider Linien, den Frequenzhub Δf, sind entsprechende Werte gefunden und in Abb. 5 gemittelt dargestellt.
Die Deklination hat für den synthetischen 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. Sie repräsentieren sich in verschiedenen Formen als steile und flache Übergänge.

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

Abb. 5 Daten für wahrnehmungsrelevante segmentale Tonhöhenbewegungen

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.
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 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 All-

tagssprache 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.



Abh. 6 Modelle suprasegmentaler f_G-Konturen

Die Aussagen der Hörer waren erwartungsgemäß eindeutig, die Intonationskontur 'Deklination + Akzent' wurde höher bewertet als nur 'Deklination' und beide zusammengenommen wesentlich höher als 'monoton' /1, S. 134/.
Die Auswertungen der Registrierungen und die Hörergebnisse lassen die Schlußfolgerung zu, daß sich die Intonation des Deutschen durch eine Reihe von Grundfrequenzkonfigurationen beschreiben läßt.

ANWENDUNGEN 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 Graphem-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 Akzentstellen, 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. Bisher war mit derartigen Anregungsgeneratoren für Laryngektomierte nur monotone Sprache möglich. (Frühere Versuche mit steuerbaren Geräten brachten nicht den gewünschten Erfolg /1, S. 158/).
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 /3/.
Mit diesen Elementen ist die Erzeugung des Grundmodells und seiner Varianten möglich. Das bedeutet zwar eine Einschränkung, sie scheint aber aus dem Grund erlaubt, da sich etwa 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.
Zusammenfassend kann festgestellt werden, daß bei einer Gegenüberstellung der intonationsgesteuerten zur monotonen Elektrolaryfx-Sprache die Natürlichkeit der erstgenannten höher eingeschätzt wird. Davon profitiert indirekt auch die Verständlichkeit, das Hören von richtig intonierter Elektrolarynx-Sprache ist für den Kommunikationspartner angenehmer, er kann sich besser auf das erzeugte Sprachsignal konzentrieren.

/1/ D. Mehnert, Analyse und Synthese suprasegmentaler Intonationsstrukturen des Deutschen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme
Diss. B, Technische Universität Dresden (1985)
/2/ A. Cohen, H't Hart, On the anatomy of intonation, Lingua 19 (1967), 177-192
/3/ D. Mehnert, Anwendung suprasegmentaler Intonationskonturen zur Verbesserung von Elektrolarynx-Sprache, Studientexte zur Sprachkommunikation 2 TU Dresden (1986), 110 - 118

Doz. Dr.sc.techn. D. Mehnert, Humboldt-Universität zu Berlin, DDR 1086 Berlin, Unter den Linden 9 - 11