BMBF

Verbmobil

Verbundvorhaben

Ein System zur Modifikation prosodischer Eigenschaften fließend gesprochener Sprache

Jörg Reinecke

TU Braunschweig

Vm

Report 74
Juli 1995

Juli 1995

Jörg Reinecke

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstraße 22
38092 Braunschweig

Tel.: (0531) 391 - 2479
Fax: (0531) 391 - 8218
e-mail: reinecke@ifn.ing.tu-bs.de

Gehört zum Antragsabschnitt: 14.3 Werkzeuge zur prosodischen Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Überblick

In diesem Artikel wird ein System vorgstellt, das in eine Arbeitsstation zur Segmentierung und Etikettierung prosodscher Einheiten integriert ist. Es ermöglicht die Resynthese einer gesprochenen Äußerung bei gleichzeitiger Modifikation der prosodischen Eigenschaften. Ausgehend von einer abstrakten prosodischen Etikettenfolge werden unter Berücksichtigung eines Intonationsmodells die Synthesesteuerparameter generiert. Das verwendete Intonationsmodell beschreibt den zu synthetisierenden Grundfrequenzverlauf als stückweise lineare Änderungen im Halbtonbereich relativ zu einer Deklinationsgeraden. Die eigentliche Synthese erfolgt nach dem PSOLA-Verfahren. Die Detektion der hierbei erforderlichen Anregungszeitpunkte im Sprachsignal geschieht automatisch. Die Güte der automatischen Detektion wird anhand eines Vergleichs mit einer manuell erstellten nachgewiesen.

1 Einleitung

Neuere Konzepte zur Erkennung fließend gesprochener Sprache oder zur Erkennung von Spontansprache beinhalten auch die Auswertung der im Sprachsignal enthaltenen prosodischen Information. Zum Training derartiger Spracherkennungssysteme muß prosodisch etikettiertes Sprachmaterial im großen Umfang bereitgestellt werden. Die Erstellung dieses Materials erfolgt derzeit noch manuell durch geeignet geschulte Transkribenten, die für jede von ihnen auditiv wahrgenommene prosodische Kategorie ein Etikett aus einem festen Etikteninventar vergeben. Als Hilfsmittel stehen ihnen hierbei Rechnerarbeitsplätze zur Verfügung, die es gestatten, das gesamte Sprachsignal oder beliebige Ausschnitte daraus anzuhören. Ferner besteht meist die Möglichkeit der Bildschirmanzeige des Signalverlaufs bzw. der Zeitverläufe daraus abgeleiteter Signalparameter wie z.B. der Sprachgrundfrequenz oder der Lautheit. Der Transkribent hat jedoch keine Möglichkeit seine Etikettenvergabe zu kontrollieren. Aus diesem Grund kann es insbesondere bei umfangreicheren Inventaren mitunter zu falschen Etikettenzuordnungen kommen, die dann die Gesamtkonsistenz der prosodischen Transkription mindern.
Hier setzt das Konzept einer Arbeitsstation zur prosodischen Etikettierung an, das am Institut für Nachrichtentechnik erarbeitet wurde [1]. Es soll den Transkribenten ermöglichen, seinen natürlichen "Eingabekanal" für prosodische Ereignisse - die auditive Wahrnehmung - auch zur Kontrolle der Etiketten einsetzen zu können. Erreicht wird dies durch die Integration einer Synthesekomponente mit flexibler Prosodiesteuerung in eine Arbeitsstation, wie sie bereits oben beschrie-

ben wurde. Die Steuerung der prosodischen Parameter während des Syntheseprozesses geschieht dabei durch die bereits vergebenen Etiketten. Die synthetische Äußerung soll jedoch keine akustisch exakte Kopie der Originaläußerung sein, sondern eine linguistisch gleichwertige. Der Transkribent wird so in die Lage versetzt, durch auditiven Vergleich der gesprochenen Äußerung mit der synthetischen, seine Entscheidung für ein bestimmtes Etikett gezielt zu überprüfen und ggf. zu korrigieren. Das angestrebte Ziel ist eine Erhöhung der Konsistenz prosodischer Etikettierungen und eine Reduzierung des Zeitaufwands bei deren Erstellung.
Technisch realisert wurde ein System zur Resynthese der gesprochenen Äußerung. Dies ist möglich, da nur die prosodischen Eigenschaften der Äußerung modifiziert werden sollen, der Wortlaut aber beibehalten wird. Gegenüber einer vollständigen Text-nach-Sprache Umsetzung ist die Resynthese weniger aufwendig und bietet zudem den Vorteil, daß charakteristische Sprechereigenschaften erhalten bleiben. Die synthetische Äußerung klingt, als wäre sie vom gleichen Sprecher gesprochen worden. Für den Transkribenten bedeutet dies, daß er sich beim auditiven Vergleich stärker auf die prosodischen Unterschiede konzentrieren kann und nicht noch zusätzlich durch die Synthese verursachte Sprecherunterschiede "überhören" muß.
Konkret wurde eine PSOLA-Synthese mit modifizierbarer Prosodiesteuerung implementiert. Gegenüber vergleichbaren Realisierungen bietet das PSOLA-Verfahren einige deutliche Vorteile. Es läßt sich bei vergleichsweise geringem Aufwand bereits eine sehr gute Sprachqualität der synthetischen Äußerung erzielen. Für die hier angestrebte Resynthese ist es vorteilhaft, daß als einzige Syntheseparameter lediglich die Anregungszeitpunkte (vgl. Abschnitt 3) einmal für jede zu untersuchende Äußerung bestimmt werden müssen. Hierfür wurde ein sehr effizientes automatisches Detektionsverfahren entwickelt, das im Abschnitt 4 beschrieben wird. All diese Punkte führen dazu, daß die Vorverarbeitung und die eigentliche Generierung der synthetischen Äußerung auf einem 486er PC sehr schnell durchgeführt werden können. Dies ist ein nicht zu vernachlässigender Gesichtspunkt, da dieses System den Transkribenten unterstützen und nicht zusätzlich behindern soll.

2 Intonationsmodell

Das wichtigste prosodische Gestaltungsmittel für einen Sprecher ist die Intonation. Über die Variation der Sprachgrundfrequenz hat er die Möglichkeit, seine Äußerung zu gliedern (Phrasierung) bzw. bestimmte Äußerungsteile hervorzuheben (Akzentuierung). Ein Intonationsmodell versucht mit Hilfe eines mehr oder

weniger umfangreichen Regelwerks, die für eine Sprache gültigen und möglichen Intonationsmuster zu beschreiben. Abhängig vom Grad der Abstraktion kann es sich hierbei um Anweisungen handeln, die lediglich den zeitlichen Verlauf der Grundfrequenz beschreiben oder die angeben, wie der Grundfrequenzverlauf unter bestimmten äußeren linguistischen Bedingungen (Phrasengrenzen, Akzente) zu verändern ist. Hierbei spielen nicht nur die linguistischen Ereignisse an sich eine Rolle, sondern auch deren Abfolge. In der Literatur sind mehrere unterschiedliche Ansätze derartiger Intonationsmodelle fürs Deutsche beschrieben. Da es bisher kein allgemeingültiges Modell gibt und geklärt werden muß, welches das geeignetste Modell für den hier skizzierten Anwendungsfall ist, bietet das hier beschriebene System die Möglichkeit zur einfachen Modellmodifikation bzw. zum Austausch des Modells.
Das gegenwärtig implementierte Modell beschreibt in Anlehnung an [4] und [5] den zu synthetisierenden Grundfrequenzverlauf als Geradenabschnitte im Halbtonbereich. Abbildung 1 zeigt eine synthetische Intonationskontur, wie sie von diesem Modell generiert werden kann.

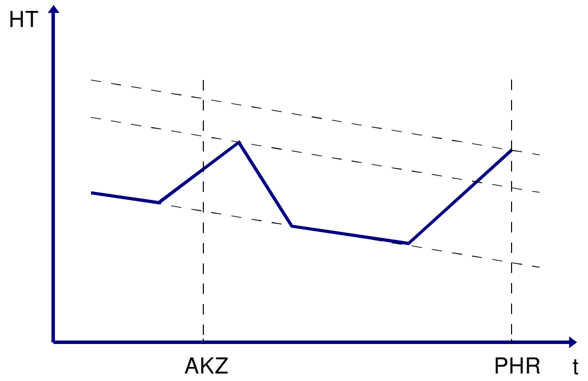

Abbildung 1: Beispiel einer synthetischen Intonationskontur

Dargestellt ist der Grundfrequenzverlauf im Bereich einer Akzentposition (AKZ) und einer Phrasengrenze (PHR) mit "Fragekontur". Die Grundfrequenzbewegungen beginnen bzw. enden jeweils auf der Deklinationsgeraden oder auf parallel verlaufenden Hilfslinien. Das Modellinventar beschreibt eine feste Anzahl derartiger Bewegungen, die sich jeweils in der zeitlichen Lage des Beginns der Bewegung relativ zu einem Bezugspunkt (AKZ oder PHR) bzw. im Frequenzhub unterscheiden. In [5] sind außerdem Regeln zur Aneinanderreihung der einzelnen Grundfrequenzbewegungen angegeben. Die dort ausgearbeiteten Modellparameter, wie die

Steigung der Deklinationsgeraden, der Abstand der Hilfslinien oder die Startzeitpunkte der einzelnen Grundfrequenzbewegungen, gelten für gelesene Sprache. Da die eigenen Arbeitsschwerpunkte aber im Bereich der Spontansprache angesiedelt sind und erste informelle Untersuchungen gezeigt haben, daß die Modellparameter für diesen Bereich nicht unverändert übernommen werden können, muß eine entsprechende Anpassung erfolgen. Hierzu sind bisher keine verläßlichen Aussagen möglich, weshalb an dieser Stelle noch keine konkreten Parameter genannt werden können.

3 Prinzip der PSOLA-Synthese

Der Begriff PSOLA steht für Pitch Synchronous Overlap and Add, was mit Grundperiodensynchroner Überlagerung und Addition übersetzt werden kann. Beschrieben wird hiermit ein zeitbereichsbasiertes Sprachsyntheseverfahren [2][3]. Kennzeichnendes Merkmal dieses Verfahrens ist die grundperiodensynchrone Verarbeitungsweise. Es werden im Abstand der aktuellen Grundperiode zwei Grundperioden lange Abschnitte aus dem Originalsprachsignal ausgeschnitten, mit einem Hammingfenster gewichtet und im Abstand der Synthesezeitpunkte überlagert. Die Abstände der Synthesezeitpunkte bestimmen die lokale Periodendauer und somit auch den Grundfrequenzverlauf der synthetischen Äußerung. Bei der PSOLA-Synthese können die prosodischen Parameter Grundfrequenz, Lautdauer und Energie des synthetischen Sprachsignals sehr einfach beeinflußt werden. Die Modifizierung des Grundfrequenzverlaufs erfolgt durch entsprechende Variation der Abstände der Synthesezeitpunkte. Die Lautdauer wird durch Weglassen einzelner Signalabschnitte (Stauchung) oder mehrfacher Aneinanderreihung des gleichen Signalabschnitts (Dehnung) modifiziert. Der Energieverlauf kann durch die Gewichtung jedes Signalabschnitts mit einem Amplitudenfaktor variiert werden. Die Einstellung dieses Faktors ist dabei völlig unabhängig von den beiden zuvor genannten Parametern. Die Variation des Grundfrequenzverlaufs und der Lautdauer dagegen beeinflussen sich gegenseitig, d.h. die Erhöhung der Grundfrequenz (Verkürzung der Periodendauer) geht mit einer Verringerung der Lautdauer einher und umgekehrt. Wenn dieser Effekt nicht erwünscht ist, was meistens der Fall sein dürfte, muß an dieser Stelle eine entsprechende Dauerkorrektur vorgenommen werden. Im hier vorgestellten Resynthesesystem erfolgt diese Korrektur implizit, so daß die von außen einstellbaren Parameter Lautdauer und Grundfrequenz quasi unabhängig voneinander arbeiten.
Besondere Bedeutung kommt bei der PSOLA-Synthese der Bestimmung der Anregungszeitpunkte im Sprachsignal zu. Sie markieren in stimmhaften Lauten je-

weils den Beginn einer Grundperiode. Bei der hier vorgestellten PSOLA-Synthese werden auch in stimmlosen Signalabschnitten Anregungszeitpunkte markiert. Diese werden im festen Abstand zueinander gesetzt und beschreiben keinerlei Signaleigenschaften. Sie dienen lediglich der vereinfachten Berechnung des synthetischen Signals, da während der Synthese keine stimmhaft/stimmlos Unterscheidung erforderlich ist. Die korrekte Bestimmung der Anregungszeitpunkte als Startpunkte der Grundperioden beeinflußt in hohem Maße die Sprachqualität der synthetischen Äußerung, da sie sowohl bei der Festlegung der aktuellen Grundperiodendauer als auch bei der Umsetzung der prosodischen Modifikationen als Ankerpunkte fungieren.

4 Systembeschreibung

Die Konzeption des im folgenden beschriebenen Systems zur Modifikation prosodischer Eigenschaften gesprochener Sprache erfolgte als Experimentalsystem. Mit ihm sollte und soll untersucht werden, ob der einleitend skizzierte Anwendungsfall der auditiven Verifikation prosodischer Etiketten überhaupt möglich ist. Die dort aufgestellte Forderung nach linguistischer Gleichwertigkeit der Originaläußerung und der synthetischen Äußerung bedeutet, daß es möglich sein muß eine abstrakte prosodische Kategorie auf die Modifikation einer Anzahl hierfür relevanter Signalparameter abzubilden. Mit Hilfe des hier beschriebenen Systems sollen diese Abbildungsvorschriften erarbeitet und getestet werden.
Das Kernstück des Systems bildet ein PSOLA-Synthese-Modul. Die darum herum gruppierten Module zur ANREGUNGSZEITPUNKTBESTIMMUNG, STEUERPARAMETERGENERIERUNG und MODIFIKATIONSGENERIERUNG dienen der Festlegung der Syntheseparameter bzw. der Umsetzung der abstrakten Etikettenfolge in konkrete Synthesesteueranweisungen. Die Implementierung der einzelnen Module erfolgt als eigenständige Programme, die miteinander über ASCII-Dateien kommunizieren. Diese Konzeption ermöglicht die Portierung des Systems auf verschiedene Rechnerplattformen bei gleichzeitiger Beibehaltung der Steuerdateien. In Abbildung 2 ist die Struktur des Gesamtsystems dargestellt und die Kommunikation der Module untereinander skizziert. Es ist zu erkennen, daß die Umsetzung der Etiketten in konkrete PSOLA-Steueranweisungen über die Zwischenebene der prosodischen Modifikationen geschieht. Diese Aufteilung geschah im Hinblick auf eine einfache Austauschbarkeit des Intonationsmodells und zur einfachen manuellen Korrektur bzw. Einflußnahme auf den Syntheseprozeß.

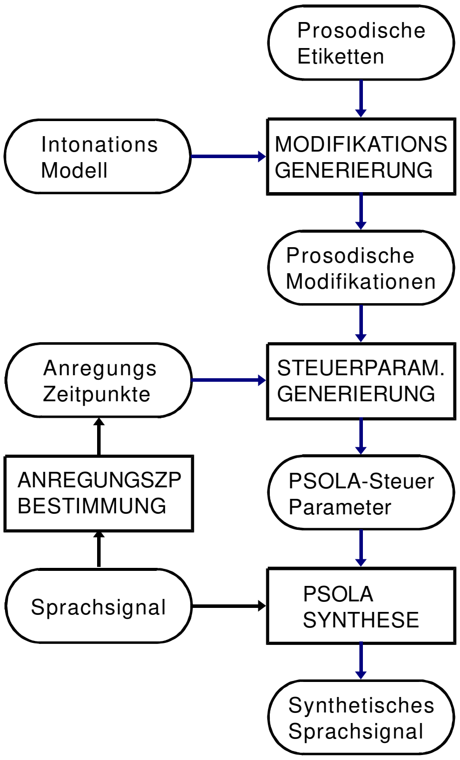

Abbildung 2: Struktur des Systems zur Modifikation prosodischer Eigenschaften

In einer Initialisierungsphase werden zunächst die Anregungszeitpunkte im Sprachsignal automatisch bestimmt. Dieser Arbeitsschritt ist für jede zu untersuchende Äußerung nur einmal nötig.
Der eigentliche Resynthesevorgang beginnt mit der Umsetzung der abstrakten prosodischen Etiketten in Modifikationsanweisungen für die PSOLA-Synthese. Dies geschieht im Modul MODIFIKATIONSGENERIERUNG unter Berücksichtigung des verwendeten Intonationsmodells. Die STEUERPARAMETERGENERIERUNG erzeugt aus diesen Modifikationsanweisungen die Steuerparameter für die PSOLA-SYNTHESE, die letztendlich das synthetische Sprachsignal generiert.

4.1 Modifikationsgenerierung

In diesem Modul erfolgt die Umsetzung der prosodischen Etiketten in Modifikationsanweisungen für die PSOLA-Synthese unter Berücksichtigung des Intonationsmodells. Die gegenwärtig verwendeten Etiketten sind lediglich abstrakte Beschreibungen der vom Intonationsmodell zugelassenen Grundfrequenzbewe-

gungen (vgl. Abschnitt 2). Im Endausbau des Systems sollen an dieser Stelle die gleichen Etiketten benutzt werden, die auch zur Erstellung der prosodischen Transkription herangezogen werden. Diese beschreiben, losgelöst von den prosodischen Signaleigenschaften, prosodische Kategorien, wie z.B. Akzentuierung oder unterschiedliche Formen der Phrasierung (terminal, progredient).
Die resultierenden prosodischen Modifikationen beziehen sich jeweils auf einen Abschnitt des zu synthetisierenden Signals. Angegeben werden der Anfangs- und Endzeitpunkt des Abschnitts sowie die zugehörigen Grundfrequenzwerte in Halbtonschritten. Der Grundfrequenzverlauf innerhalb des Bereichs verläuft zwischen beiden Punkten linear im Halbtonbereich. Optional können für einen Modifikationsbereich noch Skalierungsfaktoren für die Grundfrequenz, die Dauer sowie für die Signalamplitude angegeben werden.

4.2 Steuerparametergenerierung

Hier werden die prosodischen Modifikationsanweisungen in konkrete Steuerkommandos für den Syntheseprozeß umgesetzt. Zunächst werden unter Berücksichtigung der Anregungszeitpunkte sämtliche Synthesezeitpunkte gemäß der Modifikationsanweisungen bestimmt. Diese werden dann zusammen mit den Grenzen der zugehörigen Sprachsignalabschnitte und den bereits in den Modifikationsanweisungen verzeichneten Amplitudenfaktoren als
PSOLA-Steuerparameter abgelegt.

4.3 PSOLA-Synthese

Diese Systemkomponente erzeugt aus dem Originalsprachsignal gemäß der Steueranweisungen die synthetische Äußerung. Die jeweils zwei Grundperioden langen Abschnitte werden aus dem Sprachsignal ausgeschnitten, mit einem Hammingfenster gewichtet und mit dem in der Steueranweisung für diesen Abschnitt angegebenen Amplitudenfaktor skaliert. Die derart aufbereiteten Signalabschnitte werden im Abstand ihrer zugehörigen Synthesezeitpunkte zum synthetischen Ausgangssignal überlagert.

4.4 Bestimmung der Anregungszeitpunkte

Wie bereits in Abschnitt 3 erläutert wurde, kommt der Bestimmung der Anregungszeitpunkte im Sprachsignal für den hier skizzierten Anwendungsfall besondere Bedeutung zu. Das Kennzeichen stimmhafter Laute im Sprachsignal ist deren periodische bzw. quasiperiodische Struktur. Der Anfang einer solchen Grundperiode wird als Anregungszeitpunkt bezeichnet. Das dem Anregungszeitpunkt

folgende Signalextremum trägt die Bezeichnung Führungsamplitude. Wie später noch besprochen wird, kann es sich hierbei sowohl um ein lokales Maximum als auch um ein lokales Minimum handeln.
Da eine manuelle Markierung der Anregungszeitpunkte im vorliegenden Anwendungsfall der Resynthese wechselnder Äußerungen aus Aufwandsgründen nicht praktikabel ist, stellt die automatische Detektion die einzig vertretbare Alternative dar. Im folgenden werden ein relativ schnelles Verfahren zur automatischen Detektion der Anregungszeitpunkte vorgestellt und die damit erzielten Ergebnisse mit einer manuell erstellten Referenzmarkierung verglichen.
Das Verfahren zur Detektion der Anregungszeitpunkte basiert auf der Markierung der Führungsamplituden der jeweiligen Grundperioden. Wie bereits oben erwähnt, stehen diese in direktem Zusammenhang mit den eigentlich gesuchten Anregungszeitpunkten. Ist erst einmal eine Führungsamplitude bestimmt, ergibt sich der zugehörige Anregungszeitpunkt aus dem vorangehenden Signalnulldurchgang, der ohne Schwierigkeit bestimmt werden kann. Im vorliegenden Fall erfolgt die Suche nach relevanten lokalen Extrema allerdings nicht direkt im Sprachsignal, sondern in einem daraus abgeleiteten, stark tiefpaßgefilterten Signal. Konkret wird diese Filterung durch Faltung des Sprachsignals mit einem Hammingfenster realisiert. Die Länge des Fensters wird hierbei so gewählt, daß dessen Spektrum bei 800 Hz die erste Nullstelle aufweist.
In diesem tiefpaßgefilterten Signal werden nun gemäß der unten angegebenen Regeln sämtliche Führungsamplituden bestimmt. Ob hierbei ein Maximum oder ein Minimum gesucht wird, richtet sich nach den Signaleigenschaften. Aufgrund der nicht bekannten Aufnahme- und Vorverarbeitungsbedingungen kann mitunter eine Invertierung des Signals aufgetreten sein. Daher wird zuvor das Vorzeichen des betragsmäßig größten Extremwertes bestimmt und als Auswahlkriterium für eine Maximum- oder Minimumsuche herangezogen.

(1) Suche nächstes lokales Max/Min

(2) Betrag größer SCHWELLE?
ja: weiter(3)   nein: weiter(1)

(3) Ist Max/Min absolutes Max/Min in ±3ms Umgebung?
ja: weiter(4)   nein: weiter(1)

(4) Max/Min ist Führungsamplitude
weiter(1)

Nach Beendigung der Extremwertsuche werden in einem Korrekturdurchlauf noch einmal sämtliche gefundenen lokalen Extrema in Relation zu ihren Vorgängern und Nachfolgern gesetzt und entschieden, ob der jeweilige Wert tatsächlich eine

Führungsamplitude ist.
Um eine Aussage über die Qualität der automatischen Detektion machen zu können, wurden für 64 Äußerungen eines Sprechers die Führungsamplituden im Signalverlauf von Hand markiert. Insgesamt ergaben sich so 18387 Zeitpunkte im Referenzmaterial. Das Testmaterial wurde mit Hilfe des oben beschriebenen Detektionsverfahren automatisch generiert, wobei insgesamt 19035 Zeitpunkte markiert wurden.
Die Markierungen der Führungsamplituden im Referenz- und Testmaterial wurden nun miteinander verglichen. Hierbei zeigten sich bei einer ersten visuellen Kontrolle der automatisch und der manuell generierten Daten Unterschiede hinsichtlich der Lage der jeweiligen Markierungen. Dies ist jedoch auch nicht verwunderlich, da zwei unterschiedliche Signale zur Detektion herangezogen wurden und sich in dem tiefpaßgefilterten Signal, aufgrund des Fehlens der höherfrequenten Signalanteile, die Extrema gegenüber dem Originalsignal verschoben haben. Die Auswertung der Daten erfolgte daher nicht nur für die exakten Zeitpunkte sondern auch innerhalb kleiner Intervalle um die Testmarkierung herum.
Es wurden insgesamt 3% mehr Führungsamplituden automatisch als manuell detektiert. Bei Berücksichtigung eines 2 ms großen Intervalls (1 ms links und rechts der Markierung) wurden bereits 94% der manuell gesetzten Markierungen auch automatisch detektiert. Bei Vergrößerung des Intervalls auf 8 ms bestand in 98% der Fälle Übereinstimmung.
Dieser hohe Grad der Übereinstimmung zeigt, daß eine automatische Detektion der Anregungszeitpunkte mit dem angegebenen Verfahren möglich ist. Die gute Qualität der automatischen Detektion hat sich auch bei der Resynthese gesprochener Äußerungen nach dem PSOLA-Verfahren gezeigt. Hierbei waren perzeptiv keine Unterschiede gegenüber der Verwendung der von Hand markierten Anregungszeitpunkte feststellbar.

5 Bewertung und Ausblick

Die bisherigen Arbeiten haben gezeigt, daß die automatische Detektion der Anregungszeitpunkte im Sprachsignal prinzipiell möglich ist und die damit erzielbaren Ergebnisse durchaus die Qualität einer manuell durchgeführten Markierung aufweisen. Auch bei der Bearbeitung spontansprachlicher Signaldaten hat sich das Verfahren bereits bewährt.
Die Implementierung des Systems zur Modifikation prosodischer Eigenschaften auf einem handelsüblichen 486er PC zeigt aufgrund der kurzen Antwortzeiten bei der Umsetzung prosodischer Etiketten in synthetische Sprache, daß der angestrebte Einsatz zur Unterstützung von Transkribenten prinzipiell möglich ist. Messun-

gen haben ergeben, daß für eine 7,5 Sekunden lange Äußerung der vollständige Resyntheselauf von der Umsetzung der prosodischen Etiketten bis zum synthetischen Sprachsignal nur 6 Sekunden dauert.
Laufende Untersuchungen sollen klären, ob die prosodischen Etiketten auf charakteristische, prototypische prosodische Signalparameter abgebildet werden können. Hiermit einher geht auch die Anpassung der Parameter des Intonationsmodells an das spontansprachliche Umfeld.

6 Literatur

[**1**] Reinecke, J.: *Konzept einer Arbeitsstation zur Segmentierung und Etikettierung prosodischer Einheiten.* Proc. DAGA '93, Frankfurt a.M. 1993, 960-963

[**2**] Charpentier, F.; Moulines, E.: *Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones.* Proc. EUROSPEECH '89, Paris 1989, Vol. 2, 13-19

[**3**] Valbret, H.; Moulines, E.; Tubach, J.P.: *Voice Transformation using PSOLA technique.* Speech Communication 11, North-Holland, 1992, 175-187

[**4**] Meyer, P. et al: *PHRITTS - A Text-To-Speech Synthesizer for the German Language.* Proc. EUROSPEECH '93, Berlin 1993, 877-880

[**5**] Adriaens, L.M.H.: *Ein Modell deutscher Intonation.* Dissertation, TU Eindhoven 1993