Gefördert vom

bmb+f

Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil

Verbundvorhaben

# Resynthese als Hilfsmittel bei der prosodischen Etikettierung

Jörg Reinecke

TU Braunschweig

Vm

**Report 162**
August 1996

August 1996

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.

# Kurzfassung

Der vorliegende Beitrag beschreibt ein Modul einer Arbeitsstation zum prosodischen Etikettieren, das es dem Transkribenten erlaubt, seine Etikettenvergabe im Kontext der Gesamtäußerung auditiv zu verifizieren. Hierdurch soll die Fehleranfälligkeit prosodischer Etikettierungen gesenkt und deren Konsistenz erhöht werden. Erreicht wird dies durch die Resynthese der Originaläußerung unter Berücksichtigung der prosodischen Etikettierung. Dabei werden die prosodischen Etiketten auf prototypische Realisierungsformen abgebildet. Ziel der Resynthese ist eine linguistisch gleichwertige Kopie der Originaläußerung.

# 1 Prosodische Etikettierung

Unter dem Begriff *Prosodie* wird die Gesamtheit aller linguistisch relevanten suprasegmentalen Eigenschaften gesprochener Sprache subsumiert. Hierunter fallen sowohl übergeordnete linguistische Funktionen, wie beispielsweise Phrasierung oder Akzentuierung, als auch die sie kennzeichnenden prosodischen Empfindungsgrößen. Konkret handelt es sich dabei um die Intonationskontur sowie die Rhythmus- und Intensitätsstruktur der Äußerung. *Prosodische Etikettierung* umfaßt das Auffinden prosodischer Kategorien im Sprachsignal und deren Kennzeichnung durch die jeweilige Vergabe eines entsprechenden Etiketts aus einem festen Etiketteninventar. Im Rahmen des Verbundprojekts VERBMOBIL hat sich ein am amerikanischen Tobi-System angelehntes Etikettiersystem etabliert, bei dem eine explizite Charakterisierung der Phrasen-, Akzent- und Intonationsstruktur der gesprochenen Äußerung erfolgt [4]. Ausgehend von der Markierung prosodischer Phrasen als mehr oder minder tiefe Einschnitte im Redefluß (starke bzw. schwache Grenzen), erfolgt innerhalb jeder Phrase die Kennzeichnung besonders hervorgehobener Wörter gemäß deren Prominenz. Die zeitliche Fixierung der Phrasengrenzetiketten erfolgt zum Ende jeder Phrase. Die Akzentetiketten hingegen werden dem Silbenkern der akzenttragenden Silbe des hervorgehobenen Wortes zugeordnet.
Für die hier beschriebenen Arbeiten von Bedeutung ist die Etikettierung des Intonationsverlaufs der Äußerung. Dabei werden charakteristische Intonationsmuster im Bereich der Akzentmarkierungen und vor Phrasengrenzen durch die Angabe von Tonwerten **H** (hoch) und **L** (tief), entsprechend der jeweiligen Tonhöhenempfindung, gekennzeichnet. Neben diesen statischen Intonationsempfindungen können auch dynamische Übergänge von L nach H (steigend) oder umgekehrt (fallend) charakterisiert werden. Abhängig vom Auftreten der markierten Intonationskontur erfolgt die zusätzliche Kennzeichnung durch ein * für Akzenttöne, ein – für Phrasentöne an *schwachen* Grenzen und ein % für Grenztöne an *star-*

*ken* Grenzen. Die jeweiligen Tonwertangaben werden dem entsprechenden Akzent- bzw. Grenzetikett zugeordnet und erhalten somit auch dessen zeitliche Zuordnung im Sprachsignal.

Im Bereich der Akzentposition erfolgt die Charakterisierung durch Angabe der folgenden Etiketten:

| | |
|---|---|
| H* | Kennzeichnet eine hohe Tonhöhenempfindung in der akzentuierten Silbe. |
| L+H* | Beschreibt einen starken Anstieg der Intonationskontur in der akzentuierten Silbe mit einer tiefen Tonhöhenempfindung in der vorangehenden Silbe. |
| L*+H | Markiert einen Anstieg der Intonation zum Ende der akzentuierten Silbe mit nachfolgender hoher Tonhöhenempfindung in der folgenden Silbe. |
| H+L* | Beschreibt einen Abfall der Intonation von einer hohen Tonhöhenempfindung in der vorangehenden Silbe hin zu einer tiefen Tonhöhenempfindung in der akzentuierten Silbe. |
| L* | Charakterisiert eine tiefe Tonhöhenempfindung in der akzentuierten Silbe. |

Zusätzlich zu den verzeichneten Etiketten kann noch *Downstepping* markiert werden. Hierunter wird das Absinken der Tonhöhe in einer Folge *hoher* Akzente verstanden. Die Kennzeichnung geschieht durch Hinzufügen eines ! zu den oben genannten Akzentetiketten (!H*, L+!H*, L*+!H, !H+L*).

Die Kennzeichnung der Intonation im Bereich schwacher Phrasengrenzen geschieht durch die Vergabe der Etiketten:

| | |
|---|---|
| L- | für eine tiefe Tonhöhenempfindung und |
| H- | für eine hohe Tonhöhenempfindung |

an der betreffenden Position.

Der Intonationsverlauf vor starken Phrasengrenzen wird bitonal durch einen Phrasen- und einen Grenzton markiert. Beide zusammen charakterisieren die Intonationskontur zwischen der letzten Akzentposition innerhalb der Phrase und der Grenze selbst.

| | |
|---|---|
| L-L% | Kennzeichnet einen tiefen, zur Phrasengrenze hin fallenden Intonationsverlauf. |
| H-H% | Markiert einen hohen, zur Grenze hin steigenden Tonverlauf. |
| L-H% | Beschreibt einen Abfall der Intonation mit anschließendem Anstieg zur Phrasengrenze hin. |
| H-L% | Charakterisiert eine mittlere ebene bzw. hohe leicht fallende Intonationskontur. |

# 2 Konzept der Resynthese

Die Erstellung prosodischer Etikettierungen geschieht durch geeignet geschulte Transkribenten, die für jede von ihnen wahrgenommene prosodische Kategorie ein entsprechendes Etikett vergeben. Hierbei kann es mitunter vorkommen, daß ein Transkribent ein Etikett vertauscht, er also die wahrgenommene Kategorie falsch markiert. Geschieht dies vereinzelt, resultieren hieraus statistische Fehler in der Transkription. Setzt sich eine derartige Fehlzuweisung allerdings in der Erinnerung des Transkribenten fest, wird er jedesmal die betreffende Kategorie falsch markieren und die Transkription weist einen systematischen Fehler auf.
Die beschriebene Problematik beruht auf der Tatsache, daß der Transkribent bisher keine Möglichkeit besaß, seine Etikettierung zu überprüfen. An dieser Stelle setzt das eigene Konzept einer Arbeitsstation zum prosodischen Etikettieren an [2]. Es sieht vor, dem Transkribenten neben den üblichen visuellen und akustischen Hilfsmitteln auch die Möglichkeit der unmittelbaren Kontrolle seiner Etikettenvergabe an die Hand zu geben. Erreicht wird dies durch die Resynthese der Originaläußerung unter Berücksichtigung der prosodischen Etikettierung, wobei die prosodischen Etiketten auf prototypische Realisierungsformen abgebildet werden. Angestrebt wird dabei keine akustisch identische Kopie des Originalsignals, sondern eine linguistisch gleichwertige. Durch den auditiven Vergleich des synthetischen Signals mit dem zugrundeliegenden Originalsignal wird der Transkribent in die Lage versetzt, seine Etikettenvergabe im Kontext der Gesamtäußerung zu verifizieren. Das Ziel dieser Vorgehensweise ist es, die Fehleranfälligkeit prosodischer Transkriptionen zu reduzieren und deren Konsistenz zu erhöhen. Realisiert wurde eine PSOLA-Synthese, die um die flexible Steuerung der prosodischen Signalparameter $F_0$, Dauer und Intensität erweitert wurde (Abbildung 1). Das PSOLA-Verfahren gewährleistet eine gute Synthesequalität bei vergleichsweise geringem Aufwand [5]. Als einzige Syntheseparameter sind die Grundperioden im Sprachsignal zu detektieren. Hierzu wurde ein automatisches Verfahren entwickelt und implementiert, das in etwa die Güte einer manuellen Detektion erreicht [3].
Da hier lediglich intonatorische Merkmale etikettiert werden, erfolgt keine expli-

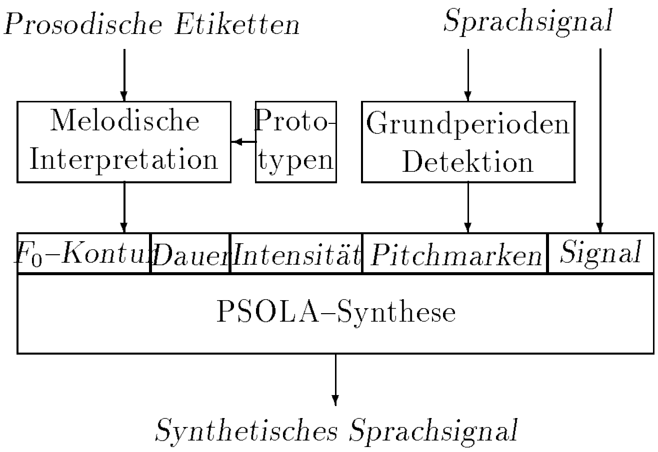


Abbildung 1: Architektur des Resynthesesystems

zite Steuerung der Dauer- und Intensitätsstruktur des synthetischen Signals. Die entsprechenden Werte werden unverändert vom Originalsignal übernommen.

# 3 Melodische Interpretation

Die melodische Interpretation betrifft die Umsetzung der abstrakten prosodischen Intonationsetiketten in prototypische Grundfrequenzbewegungen. Diese sollen nach der Resynthese vom Transkribenten auf die der Etikettierung zugrundeliegenden prosodischen Kategorien bezogen werden können. Die Grundidee der melodischen Interpretation beruht auf der stückweisen linearen Modellierung der synthetischen $F_0$–Kontur im Halbtonbereich. Motiviert wird diese Vorgehensweise durch Untersuchungen [1], die zeigen daß es möglich ist, eine reale Grundfrequenzkontur derart durch Geradensegmente anzunähern, daß die daraus resultierende synthetische Äußerung und die Originaläußerung perzeptiv gleichwertig sind.
Den Rahmen der hier beschriebenen prototypischen Realisierungen bilden drei zueinander parallele Geraden, die den Tonwerten L, !H und H zugeordnet sind (Abbildung 2). Der Abstand der Geraden zueinander ist fest, während die Stimmlage der synthetischen Äußerung durch die absolute Höhe des Basisniveaus L festgelegt wird. Deklination, also das allmähliche Absinken der $F_0$–Kontur über der Äußerung, kann durch Neigen der Hilfslinien eingestellt werden. Sämtliche Grundfrequenzbewegungen beginnen und enden auf den Niveaulinien bzw. laufen auf ihnen entlang. In Testreihen wurden die Abstände der Niveaus zu:

$$\Delta_H = 6\ \mathrm{HT} \qquad \text{und} \qquad \Delta_{!H} = 3\ \mathrm{HT}$$

ermittelt. Für die Deklination konnte kein einheitlicher Wert bestimmt werden, so daß sie für die prototypischen Realisierungen zu

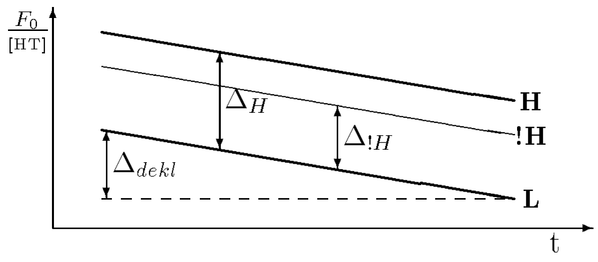


Abbildung 2: Rahmen zur melodischen Interpretation prosodischer Etiketten

$$\Delta_{dekl} \equiv 0$$

gewählt wird.

# 4 Prototypische Realisierungen

Zur Bestimmung der prototypischen Realisierungsformen wurden zunächst aus realen Sprachsignalen besonders typisch erscheinende Realisierungen der jeweiligen prosodischen Kategorien ausgewählt und mit den gemäß der prosodischen Etikettierung resynthetisierten Äußerungsteilen verglichen. In einem iterativen Prozeß wurden die Modellparameter solange variiert, bis synthetische Realisierung und Original auditiv gut übereinstimmten. Hierbei wurden möglichst standardisierte und sprecherunabhängige Prototypen angestrebt.
Die folgenden Abbildungen zeigen die für jedes Etikett gefundenen Prototypen. Dargestellt sind jeweils die beiden Niveaulinien L und H sowie die charakteristischen Realisierungen. Die angegebenen Zeitpunkte $t_{akz}$ und $t_{grenz}$ markieren dabei die Signalposition des zugehörigen Akzent- bzw. Grenzetiketts. Gestrichelt angedeutet sind mögliche Fortsetzungen der $F_0$–Kontur.
Die beiden Akzenttöne L* und H* werden durch einen *Zielpunkt* auf der zugehörigen Niveaulinie an der angegebenen Signalposition ($t_{akz}$) realisiert (Abbildung 3). Bei den bitonalen Akzentetiketten (L+H*, L*+H, H+L*) wird zunächst auch der durch * gekennzeichnete Tonwert an der markierten Position realisiert. Zusätzlich wird aber noch ein zweiter Tonwert, 150 ms davor oder dahinter festgelegt (Abbildung 4). Hierdurch kann ein Anstieg (L+H*) oder Abfall (H+L*) der Intonation im Silbenanfang bzw. ein Anstieg (L*+H) zum Silbenende hin modelliert werden. Die prototypische Realisierung der *Downstep*–Etiketten (Kennzeichen !) geschieht in der gleichen Weise, nur daß die zugehörigen Konturen auf dem !H–Niveau beginnen bzw. enden.
Die Modellierung der Intonation im Bereich schwacher Grenzen (L-, H-) erfolgt

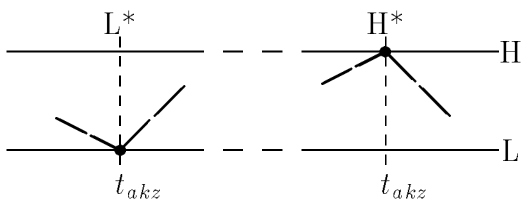


Abbildung 3: Melodische Interpretation der Etiketten L* und H*

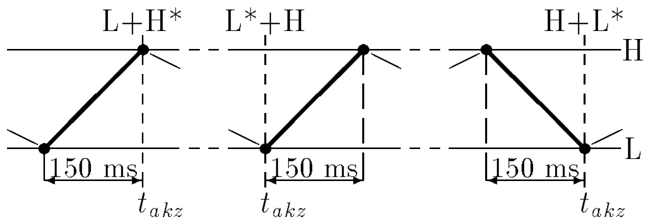


Abbildung 4: Melodische Interpretation der Etiketten L+H*, L*+H und H+L*

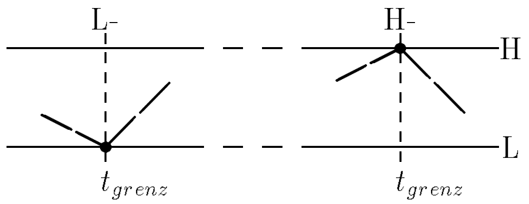


Abbildung 5: Melodische Interpretation der Etiketten L- und H-

wiederum durch Angabe eines zugehörigen Zielpunktes für den Grundfrequenzverlauf auf der entsprechenden Niveaulinie (Abbildung 5).
Schwieriger gestaltet sich die Realisierung der bitonalen Markierung starker Grenzen (L-L%, H-H%, L-H%, H-L%). Hierbei wird das Ende der zugehörigen Grundfrequenzkontur an der markierten Grenzposition durch den Grenzton (L%, H%) festgelegt (Abbildung 6). Der Grundfrequenzverlauf zwischen der letzten Akzentrealisierung und der Grenze selbst wird durch den Phrasenton (L-, H-) beeinflußt. Eine gute Näherung realer $F_0$–Verläufe konnte hier durch Wahl eines zeitlichen Abstandes von 150 ms zum Ende der vorangehenden Akzentrealisierung, in der Abbildung fett dargestellt, erzielt werden.
Bei der Umsetzung der prosodischen Etikettierung werden zunächst sämtliche Prototypen an den zugehörigen Signalpositionen realisert und dann durch Geradensegmente zur synthetischen $F_0$–Kontur verbunden. Diese wird bei der anschlie-

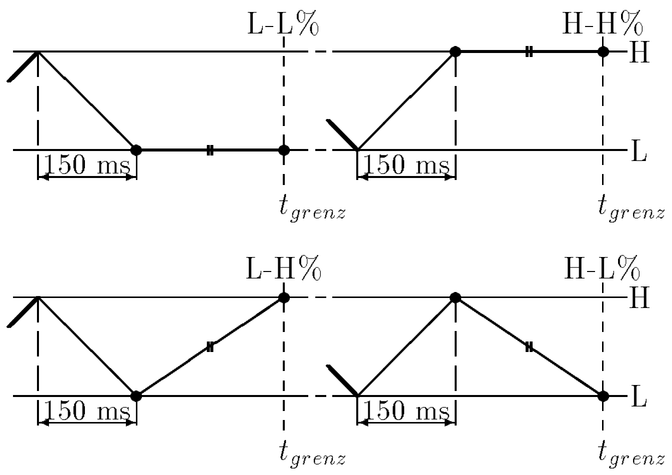


Abbildung 6: Melodische Interpretation der Etiketten L-L%, H-H%, L-H% und H-L%

ßenden PSOLA–Synthese dem Originalsignal gleichsam aufgeprägt.
Abbildung 7 zeigt das Resultat der prototypischen Umsetzung einer prosodischen Etikettierung im Vergleich mit der realen $F_0$–Kontur der Äußerung. Es sind rein visuell Unterschiede feststellbar, die auch beim akustischen Vergleich der synthetischen Äußerung mit dem Original zutage treten. Trotzdem sind Original und Synthese linguistisch gleichwertig, da sie vom Zuhörer in der gleichen Weise interpretiert werden.

# 5 Experimentelle Ergebnisse

Der direkte Nachweis des Nutzens der hier vorgestellten Resynthese beim prosodischen Etikettieren steht noch aus. Allerdings konnte in einem Experiment mit 12 Testhörern ein erster Eindruck vom Potential der Resynthese gewonnen werden. Hierbei wurden jeweils zwei unterschiedliche prosodische Etikettierungen der gleichen Äußerung resynthetisiert, und die Testpersonen hatten die Aufgabe zu entscheiden, welche der beiden Realisierungen der zugrundeliegenden Originaläußerung ähnlicher sei. In der überwiegenden Zahl der Fälle präferierten die Testhörer eine der beiden Etikettierungen. Lediglich bei zwei von 51 untersuchten Testphrasen entschieden sich gleich viele Testpersonen (6) für beide Realisierungen. Dieses Testergebnis kann als Indiz dafür gewertet werden, daß die angestrebte linguistische Gleichwertigkeit prinzipiell erzielbar ist und daß die Transkribenten beim prosodischen Etikettieren hiervon profitieren können.

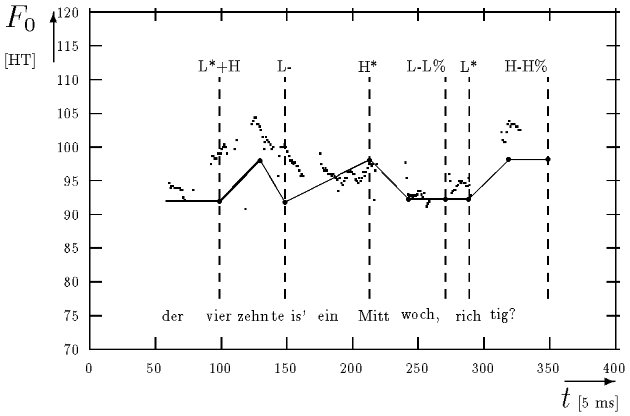


Abbildung 7: Gemessener Grundfrequenzverlauf (punktiert) und prototypische Grundfrequenzkontur

# Literatur


[1] Adriaens, L.M.H.: *Ein Modell deutscher Intonation.* Dissertation, Technische Universität Eindhoven 1991.

[2] Reinecke, J.: *Konzept einer Arbeitsstation zur Segmentierung und Etikettierung prosodischer Einheiten.* Fortschritte der Akustik, Tagungsband DAGA 93, Frankfurt am Main 1993, 960 – 963.

[3] Reinecke, J.: *Ein System zur Modifikation prosodischer Eigenschaften fließend gesprochener Sprache.* Studientexte zur Sprachkommunikation, Heft 11, Tagungsband Elektronische Sprachsignalverarbeitung, Berlin 1994, 213 – 220.

[4] Reyelt, M.: *Ein System Prosodischer Etiketten zur Transkription von Spontansprache.* Studientexte zur Sprachkommunikation, Heft 12, Hrsg.: R. Hoffmann, R. Ose Tagungsband Elektronische Sprachsignalverarbeitung, Wolfenbüttel, September 1995, 167 – 174.

[5] Valbret, H.; Moulines, E.; Tubach, J.P.: *Voice transformation using PSOLA technique.* Speech Communication 11 (1992), 175 – 187.