Konzept einer Arbeitsstation zur Segmentierung und Etikettierung prosodischer Einheiten

Jörg Reinecke

TU Braunschweig

Report 73
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.

Einleitung

Der sprachliche Kommunikationsprozeß zwischen Menschen wird wesentlich durch prosodische Informationen im Sprachsignal beeinflußt und gesteuert. Ein Sprecher benutzt prosodische Gestaltungsmerkmale, um seine Äußerung zu gliedern und so dem Zuhörer das Verständnis des Gesprochenen zu erleichtern. Des weiteren gibt es Situationen, in denen die korrekte Interpretation einer Äußerung seitens des Zuhörers überhaupt erst durch die Auswertung prosodischer Informationen möglich wird [1].
Es ist naheliegend, die überaus wichtigen prosodischen Informationen auch in automatischen Spracherkennungssystemen auszuwerten, insbesondere wenn fließend gesprochene Sprache erkannt werden soll. Doch der Einbeziehung prosodischer Merkmale in den Erkennungsprozeß automatischer Spracherkennungssysteme steht das teilweise noch recht geringe Wissen um Art, Funktion und Ausprägung dieser Merkmale im Sprachsignal entgegen.
An dieser Stelle setzt nun die Idee der im folgenden beschriebenen Arbeitsstation an. Dem Benutzer sollen komfortable Werkzeuge zur Verfügung gestellt werden, mit deren Hilfe er Sprachsignale prosodisch segmentieren und etikettieren kann. Darüber hinaus wird erwartet, daß durch den Umgang mit der Arbeitsstation und den prosodischen Einheiten auch neue Einsichten in die Wirkungsweise prosodischer Merkmale gewonnen werden können.

Konzeption

Grundlage für den Einsatz der geplanten Arbeitsstation ist zunächst einmal die Festlegung eines geeigneten Transkriptionssystems. Tendenziell können zwei verschiedene Wege beschritten werden. Zum einen kann eine signalnahe Transkription erfolgen, bei der bestimmte Signalformen oder aus dem Signal abgeleitete Parameter möglichst genau beschrieben werden sollen. Zum anderen besteht die Möglichkeit, eine abstrakte Kategorisierung der sprachlich relevanten prosodischen Ereignisse vorzunehmen. Die erstgenannte Variante bedingt einen relativ umfangreichen Symbolvorrat, da prinzipiell alle auftretenden Realisierungsvarianten eines prosodischen Merkmals erfaßt werden müssen.
Wir haben uns bei unseren Arbeiten, in Anlehnung an [2], für die zweite Variante der rein symbolischen Repräsentation abstrakter prosodischer Kategorien entschieden. Die Einführung dieser symbolischen Transkriptionsebene beinhaltet den Vorteil, daß pro Kategorie nur ein Symbol zur Verfügung gestellt werden muß. Daraus resultiert ein relativ kleiner Symbolvorrat, was der Konsistenz der Transkriptionsergebnisse sicherlich zugute kommt. Wie ein Sprecher die entspre-

chende Kategorie, zum Beispiel den Wortakzent, realisert hat, bleibt zunächst im verborgenen. Eine Zuordnung der Kategorie zu dieser Realisierungsform kann im nachhinein, ggf. sogar automatisch, erfolgen.
Der vorläufige Symbolvorrat an prosodischen Kategorien umfaßt Wort-, Phrasen- und Satzakzent sowie Phrasengrenzen und Pausen. Diese Auflistung ist insofern als vorläufig anzusehen, als daß im Verlauf der Arbeiten sicherlich noch die eine oder andere Modifikation am prosodischen Inventar vorgenommen werden muß. Diese möglichen Variationen der Symbolliste müssen allerdings beim Entwurf der Arbeitsstation und insbesondere der Dateiformate bereits berücksichtigt werden.
Der Leitgedanke bei der Konzeption der Arbeitsstation war, den Segmentierer in die Lage zu versetzen, sämtliche von ihm getroffenen Entscheidungen auch akustisch verifizieren zu können. Hierzu ist geplant, die zu etikettierende Äußerung durch eine nach dem PSOLA-Verfahren [3] arbeitende Synthese zu resynthetisieren. Dabei sollen die prosodischen Parameter der Äußerung gemäß der vom Benutzer gesetzten Etiketten modifiziert werden. Die Resynthese bietet an dieser Stelle gegenüber einer Vollsynthese den Vorteil, daß die Sprachcharakteristika des Sprechers erhalten bleiben und sich der Segmentierer einzig und allein auf die prosodischen Modifikationen konzentrieren kann. Die Verwendung des PSOLA Verfahrens ermöglicht eine relativ einfache und schnelle Verarbeitung bei gleichzeitiger hoher Sprachqualität.
Auch bei der Wahl der Etiketten ist an eine Unterstützung des Segmentierers durch das System gedacht. Dies soll durch eine hypothetische Etikettenfolge erreicht werden, die aus einer linguistischen Analyse der orthographischen Niederschrift des Gesprochenen extrahiert wird. Der Benutzer kann dann die prototypischen Etiketten übernehmen, die mit den vom Sprecher realiserten Kategorien übereinstimmen und muß nur die falschen Hypothesen ersetzen.
Die genannten Maßnahmen haben das Ziel, die Konsistenz der Transkriptionsergebnisse zu erhöhen und eröffnen die Möglichkeit, die für die prosodische Segmentierung und Etikettierung notwendige Einarbeitungszeit deutlich zu verringern.

Realisierung

Aufgrund der komplexen Gesamtstruktur der Arbeitsstation ist ein mehrstufiger Ausbau vorgesehen. Die erste, bereits abgeschlossene, Ausbaustufe umfaßt eine einfache Arbeitsstation zum manuellen Segmentieren und Etikettieren. Im Bild 1 ist die Struktur dieser Ausbaustufe dargestellt. Es können sowohl der Signalverlauf als auch zuvor bereits berechnete, zugehörige Parameterverläufe, wie beispielsweise Grundfrequenzverlauf oder Lautheitsverlauf graphisch dargestellt werden. Der Benutzer kann über die **Manuelle Eingabe** Segmentgrenzen setzen und zughörige Etiketten vergeben.

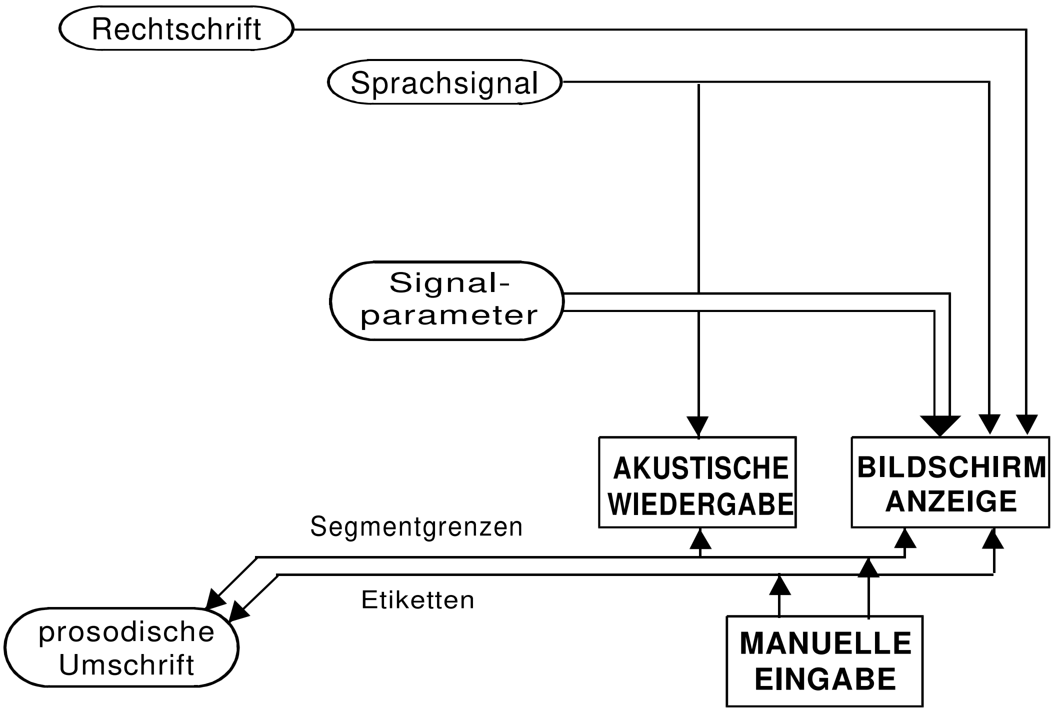

Bild 1: Grundaufbau der Arbeitsstation

Die **Akustische Wiedergabe** ist bereits in diesem Ausbaustadium vorhanden und erlaubt dem Benutzer, sowohl die gesamte Äußerung als auch einzelne Segmente oder beliebige Äußerungsteile anzuhören. Diese Ausbaustufe ist nicht allein der prosodischen Transkription von Sprachsignalen vorbehalten, vielmehr können im Prinzip beliebige Einheiten segmentiert und etikettiert werden. So wurden bereits Wortsegmentierungsaufgaben oder die Markierung von Anregungszeitpunkten zur PSOLA-Synthese mit der Arbeitsstation vorgenommen.
Der geplante Endausbau der Arbeitsstation ist in Bild 2 wiedergegeben. Gegenüber der Basisversion aus Bild 1 sind zunächst einmal die beiden Analysemodule **Linguistische Analyse** und **Signalanalyse** hinzugekommen. Ihre Aufgabe ist die in Kapitel 2 beschriebene Bereitstellung hypothetischer Etikettenfolgen sowie die Berechnung der Signalparameter, die teilweise auch vom Synthesemodul benötigt werden.
Über das **Auswahlmodul** hat der Benutzer nun die Möglichkeit, die hypothetische Etikettenfolge der gesprochen Äußerung entsprechend anzupassen und

schließlich als prosodische Umschrift des Gesprochenen abzuspeichern.
Das **Synthesemodul** führt die in Kapitel 2 beschriebene Resynthese der Äußerung durch. Maßgeblich sind hierbei die aktuellen Änderungen der hypothetischen Etikettenfolge, so daß der Benutzer seine Entscheidungen jederzeit auditiv verifizieren kann.

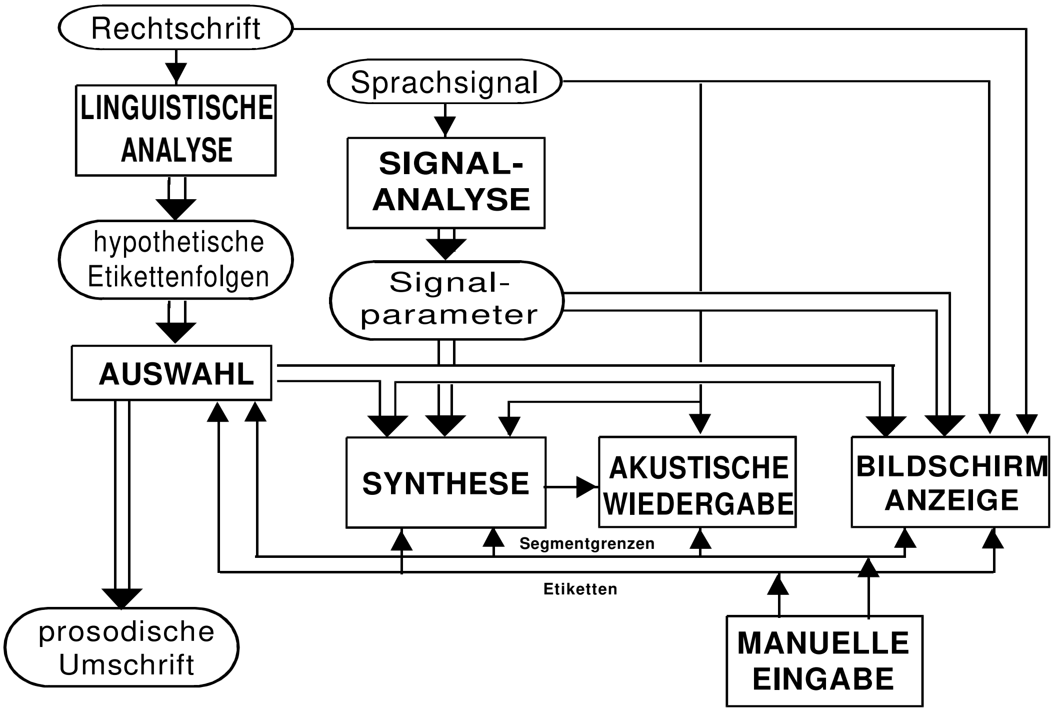

Bild 2: Endausbau der Arbeitsstation

Rechnerplattform

Die Arbeitsstation ist als PC-Basiertes System unter WINDOWS konzipiert. Diese Entscheidung gründet sich auf die große Verbreitung derartiger Systeme und, verbunden mit der einheitlichen Betriebssoftware (DOS und WINDOWS), einfachen Installation der Arbeitsstation. Für den Benutzer ergeben sich aufgrund der bekannten Programmoberfläche unter WINDOWS sehr geringe Einarbeitungszeiten. Da moderne 486er Systeme über akzeptable Rechenleistungen verfügen steht auch von dieser Seite her der Entscheidung nichts entgegen.

Für den institutsinternen Gebrauch wird die Station auch noch für das Betriebssystem VMS entwickelt, da hierfür eine sehr gut ausgebaute Rechnerinfrastruktur vorhanden ist.

Literatur

[1] Paulus, E.; Gerken, H.-D.; Reinecke, J; Veidt, J.: *Der Nutzwert prosodischer Merkmale für die automatische Spracherkennung.* Tagungsband Elektronische Sprachsignalverarbeitung, Berlin 1990

[2] Kohler, K.J.: *Prosodisches Transkriptionssystem für die Etikettierung von Sprachsignalen.* Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK) 26, 239 - 252

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