BMBF

Verbmobil

Verbundvorhaben

Untersuchungen zur Etikettierung prosodischer Einheiten

Matthias Reyelt

TU Braunschweig

Vm

Report 75
Juni 1995

Juni 1995

Matthias Reyelt

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstraße 22
38092 Braunschweig

Tel.: (0531) 391 - 2479
Fax: (0531) 391 - 8218
e-mail: m.reyelt@tu-bs.de

Gehört zum Antragsabschnitt: 14.6 Prosodische Etikettierung

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

Einleitung

Automatische Spracherkennungssysteme beschränken sich zur Zeit meist auf die Analyse der segmentalen Information des Sprachsignals, während die prosodische Information nicht ausgewertet wird. Daß für den menschlichen Hörer die Prosodie einer Äußerung sehr wohl von großer Bedeutung ist, zeigt nicht zuletzt die Tatsache, daß bei Sprachsynthesesystemen die Qualität der Prosodiesteuerung die Verständlichkeit der Sprache sowie die Akzeptanz bei den Hörern stark beeinflußt.

Bei der Sprachgenerierung wird vom Sprecher ausgehend von Wortlaut, lexikalischer Wortbetonung und syntaktischer Struktur die Prosodie der Äußerung generiert[2]. Die *prosodischen Kategorien* sind dann im Sprachsignal durch die prosodischen Parameter *Sprachgrundfrequenz*, *Intensität* und *Silbendauer* realisiert. Zusätzlich werden die Parameter noch durch mikroprosodische Effekte und die Atmung beeinflußt.

Sprachsynthesesysteme können sich sowohl bei der Generierung von Akzent und Phrase als auch bei der Steuerung der Signalparameter auf eine spezielle Realisierung beschränken, ohne daß die synthetische Sprache unnatürlich klingt. Bei der Prosodie menschlicher Sprecher überlagern sich jedoch eine Vielfalt von Informationen und Einflüssen, was eine automatische Extraktion für die Automatische Spracherkennung relevanter Information stark erschwert. Eine Verbesserung der Erkennungsleistung von Spracherkennungssystemen (etwa eine Erkennung von Satzende, Satzmodus oder Wortbetonung) läßt sich zwar theoretisch zeigen [1], praktisch ist die Ermittlung prosodischer Kategorien aus den Signalparametern jedoch fehlerbehaftet[5].

Untersuchungen zur Konsistenz prosodischer Etikettierungen

Vorbedingung für die Integration der Prosodieanalyse in die automatische Spracherkennung ist die Bereitstellung prosodisch etikettierten Sprachmaterials als Referenz. Schon früher wurde darauf hingewiesen, daß die Konsistenz dieser prosodischen Etikettierungen problematisch ist [4]. Um erste Erkenntnisse darüber zu erhalten, welche Abweichungen zwischen verschiedenen Transkribenten zu erwarten sind, wurden Untersuchungen durchgeführt, in denen mehreren Testpersonen dasselbe Material zur Etikettierung vorgelegt wurde. Als Sprachmaterial dienten 46 Sätze aus dem PHONDAT90-Korpus[6], die auf Silbenebene segmentiert wurden. Als Inventar prosodischer Etiketten wurden für die Akzentuierung vier Kategorien verwendet: unbetont, Nebenakzent, Hauptakzent, Emphase. Weiterhin wurden Phrasengrenzen markiert. Als Transkribenten wurden vier Studenten eingesetzt, die phonetisch nicht geschult waren. Es wurden zwei Gruppen gebildet. Im einzelnen wurden die folgenden Versuche durchgeführt:

1. Gruppe (Transkribent 1 und 2):

V1: Nur die Akzentuierung wurde etikettiert. Die Transkribenten wurden angewiesen, vier *Betonungsstufen* durch Zahlen 1(unbetont) bis 4(Emphase) zu markieren.

V2: Die Akzente wurden wie im ersten Versuch etikettiert. Allerdings wurden die Anweisungen geändert. Die Transkribenten wurden angewiesen, Haupt- bzw. Nebenakzente zu etikettieren. Unbetonte Silben wurden nicht mehr markiert.

V3: Hier wurden Phrasengrenzen markiert. Die Transkribenten wurden angewiesen, Einschnitte im Sprachsignal zu markieren.

V4: V3 wurde wiederholt.

2. Gruppe (Transkribent 3 und 4):

V5: Hier wurden Akzente und Phrasengrenzen in einem Durchgang etikettiert. In den Anweisungen wurde das Etikettieren von Phrasen und Betonung folgendermaßen verkoppelt: Die Transkribenten sollten in jeder Äußerung zunächst die Phrasengrenzen markieren. In jeder Phrase sollte dann die Silbe etikettiert werden, auf der der *Hauptakzent* bzw. *Fokus* dieser Phrase liegt. Zusätzlich konnten *Nebenakzente* etikettiert werden. Die Transkribenten wurden also dahingehend eingeschränkt, daß für jede Phrase nur ein Hauptakzent etikettiert werden durfte.

Zwischen den einzelnen Versuchen lagen jeweils einige Tage. Als Vergleich stand eine dritte prosodische Etikettierung zur Verfügung. Diese war schon früher vom Autor im Rahmen von Untersuchungen zur automatischen Prosodieanalyse durchgeführt worden.

Die Sprachdaten wurden den Transkribenten auf der Braunschweiger Sprach–Arbeitsstation dargestellt [3]. Angezeigt wurde der Signalverlauf mit den Silbengrenzen und dem dazugehörigen Text. Die Transkribenten konnten sich sowohl den gesamten Satz als auch einzelne Silben anhören, und zwar so oft wie gewünscht. Für die Dauer der einzelnen Transkriptionssitzungen gab es keine zeitliche Einschränkung.

Ergebnisse

Bei V1 führte die Anweisung, alle Silben – auch die unbetonten – zu etikettieren, sowie die Verwendung von Zahlen als Etiketten zu einer Transkription, die in etwa einer Quantisierung des Grundfrequenzverlaufs entsprach. Die Versuche zielten eigentlich jedoch auf eine Etikettierung abstrakterer prosodischer Kategorien ab. Bei den Anweisungen in den folgenden Versuchen wurde auf den Begriff "Betonungsstufe" bewußt verzichtet und statt dessen der Begriff "Akzent" verwendet. Der Begriff *Betonungsstufe* wird hauptsächlich bei der Sprachsynthese verwendet, z.B. zur Angabe der Höhe eines Grundfrequenzpeaks. Durch seine implizierte Signalnähe scheint er für die Etikettierung prosodischer Kategorien weniger geeignet zu sein als der eher abstrakte Begriff *Akzent*. Auf eine weitere Auswertung dieses Versuches wurde verzichtet.

V2 wurde mit diesen geänderten Anweisungen durchgeführt. Die Entscheidungen der beiden Testpersonen wurden untereinander und mit der Referenzetikettierung des Autors (Ref.) verglichen. Die Übereinstimmungen sind in Tabelle 1 zusammengestellt. Durchschnittlich lag die Übereinstimmung bei knapp 80 %. Die Werte sind für die Nebenakzente deutlich schlechter als für die Hauptakzente.

Für 70 % der Silben vergaben alle drei Transkribenten (Testpersonenen und Autor) dasselbe Etikett; bei weiteren 27 % waren sich zwei der Transkribenten einig und bei 3 % etikettierten alle drei verschieden.

Bei den Phrasengrenzen lassen sich zwischen V3 und V4 auch intrapersonelle Übereinstimmungen angeben. Hier ist das Ergebnis für beide Studenten sehr unterschiedlich. Während der eine sehr konstant etikettierte (55/53 Phrasengrenzen, von diesen stimmten 48 überein), unterschieden sich die Etikettierungen des zweiten stark (18/70 Phrasengrenzen, 18 übereinstimmende).

Die interpersonellen Übereinstimmungen sind wieder in Tabelle 1 angegeben, allerdings nur für V4. Hier waren sich alle drei Transkribenten bei 87 % der Fälle einig.

Für V5 waren die Anweisungen nach den Erfahrungen der ersten Gruppe modifiziert worden. Nach diesen Anweisungen etikettierten nun die beiden anderen Testpersonen. Die neuen Anweisungen hatten den Erfolg, daß weniger Rückfragen auftraten und die Etikettierungen sofort ausgewertet werden konnten. Im einzelnen sind die Übereinstimmungen der beiden Testpersonen (Transkribenten 3 und 4) untereinander und mit der Referenzetikettierung in Tabelle 1 angegeben. Alle drei sind sich bei 82 % aller Silben einig, bei weiteren 15 % weicht ein Transkribent ab und nur bei 3 % hat jeder ein anderes Etikett vergeben. Bei den Phrasengrenzen sind sich alle drei in 94 % der Fälle einig.

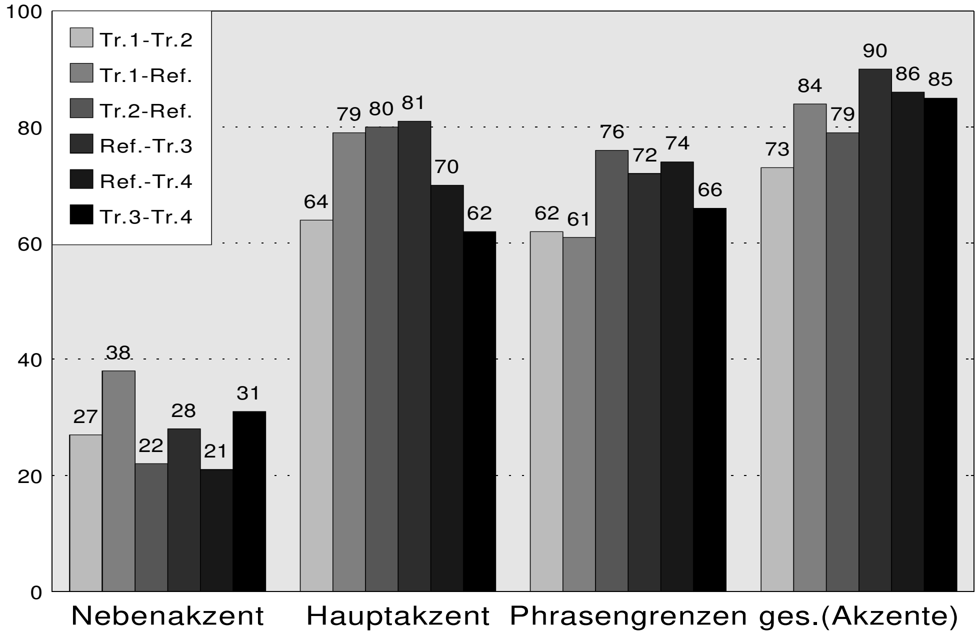

Abbildung 1: Übereinstimmungen zwischen den Transkribenten

Vergleich der Versuche V2/V4 und V5

Der Vergleich der beiden Transkribentengruppen zeigt keine besonderen Unterschiede. Insgesamt ist die zweite Gruppe etwas besser, vor allem, wenn die unterschiedlichen Voraussetzungen der Transkribenten berücksichtigt werden. Die erste Gruppe hatte zu diesem Zeitpunkt das Material zum dritten bzw. vierten Mal etikettiert, während die zweite Gruppe weder das Material kannte noch Transkriptionserfahrung hatte.

Bei beiden Gruppen sind die Abweichungen untereinander größer als die zum Autor. Dies spricht dafür, daß diese auf Unsicherheiten beruhen und nicht auf prinzipiellen Abweichungen. Die Anzahl vergebener Etiketten zeigt Abb. 2. Es zeigt sich, daß die Transkribenten in der ersten Gruppe teilweise deutlich mehr Etiketten vergaben als die in der zweiten. Die Schwankungen sind besonders bei den Nebenakzenten sehr groß.

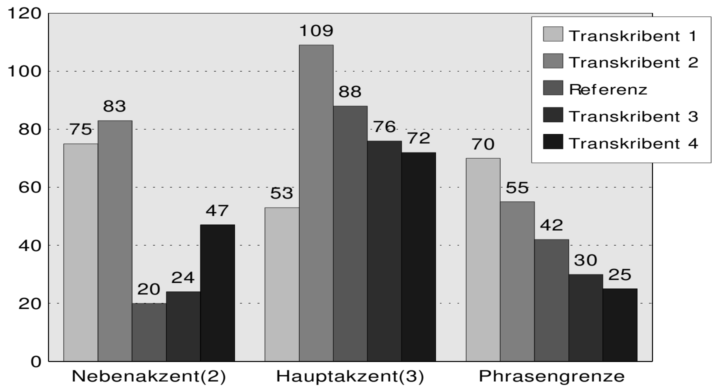

Abbildung 2: Anzahl der von den Transkribenten vergebenen Etiketten

Zusammenfassende Beurteilung und Ausblick

Insgesamt sind die Übereinstimmungen zwischen den Transkribenten mit ca. 80% schon recht hoch, vor allem, wenn man bedenkt, daß sie keinerlei Transkriptionserfahrung hatten. Durch Training läßt sich die Konsistenz vermutlich noch weiter verbessern. Das Inventar prosodischer Etiketten und Anweisungen können bei den folgenden Versuchen beibehalten werden. Die Versuche sollen für weiteres Material und mit mehr Transkribenten wiederholt werden. Insbesondere muß noch untersucht werden, wie aus mehreren parallelen Etikettierungen eine verläßliche prosodische Referenzetikettierung erzeugt werden kann.

Literatur

[1] E. Paulus et.al. Der Nutzwert prosodischer Merkmale für die automatische Spracherkennung. In K. Fellbaum, Hrsg., *Tagungsband Elektronische Sprachsignalverarbeitung*, 1990.

[2] Carsten Günther. Das prosodische System des Deutschen aus Sicht der Sprachproduktion. In R. Hoffmann, Hrsg., *Elektronische Sprachsignalverarbeitung*, 1992.

[3] J. Reinecke. Konzept einer Arbeitsstation zur Segmentierung und Etikettierung prosodischer Einheiten. In *Fortschritte der Akustik – DAGA 93*, Bad Honnef, 1993. DPG-GmbH.

[4] M. Reyelt. Automatische Extraktion prosodischer Merkmale aus den Verläufen von Sprachgrundfrequenz und Lautheit. In G. Görz, Hrsg., *KONVENS 92*, S. 385 – 389, Berlin, 1992. Springer.

[5] M. Reyelt. Einfluß der Prosodie auf die Verläufe von Sprachgrundfrequenz und Lautheit. In *Fortschritte der Akustik – DAGA 92*, Bad Honnef, 1992. DPG-GmbH.

[6] Werner Thon & Wim A. van Dommelen. PHONDAT 90: Rechnerverarbeitbare Sprachaufnahmen eines umfangreichen Korpus des Deutschen. In K.J. Kohler, Hrsg., *AIPUK*. Inst. für Phonetik, Universität Kiel, 1992.