BMBF

Verbmobil

Verbundvorhaben

Ein System zur prosodischen Etikettierung von Spontansprache

Matthias Reyelt

TU Braunschweig

Vm Report 86

Juli 1995

Juli 1995

Matthias Reyelt

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 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 (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

1 Zusammenfassung

Zur prosodischen Etikettierung von Spontansprache im Rahmen des Projektes VERBMOBIL wurde ein Etiketteninventar entworfen, das sich einerseits an den Anforderungen unterschiedlicher Projektpartner orientiert, andererseits aber auch die Bedürfnisse der Transkribenten berücksichtigt. Die Etikettierung findet auf mehreren Ebenen statt, auf denen unterschiedliche prosodische Information markiert wird. Zusammen mit dem Inventar wurde auch eine Arbeitsumgebung für die Etikettierung entwickelt, sowie eine Evaluierung der Transkriptionen durchgeführt. Bisher wurden so Sprachdaten im Umfang von einer Stunde etikettiert und von Projektpartnern zur automatischen Erkennung eingesetzt.

2 Einleitung

Bei der automatischen Spracherkennung ist auf die Dauer die Auswertung prosodischer Information unverzichtbar, besonders wenn, wie in [37] zur Erkennung von Spontansprache übergegangen wird. Vor allem bei der Dialoganalyse hat die Prosodie wichtige Disambiguierungsfunktion [13, 4].

Es gibt entsprechend viele Untersuchungen zur automatischen Prosodieanalyse [5, 39, 38, 3, 8]. Eine Schwierigkeit der häufig eingesetzten statistischen Verfahren ist allerdings die Bereitstellung entsprechend etikettierten Sprachmaterials in ausreichendem Umfang. Während es für die segmentale Etikettierung Standard–Beschreibungssysteme (z.B. SAMPA oder IPA) gibt, ist die prosodische Etikettierung selbst noch Gegenstand der Forschung [29, 30, 11, 14, 2].

Die Entwicklung eines Etikettensystems zur Transkription großer Korpora stellt besondere Anforderungen an Art und Definition der Etiketten. Zum einen soll das Material meist von mehreren unterschiedlichen Anwendern benutzt werden und deren Anforderungen erfüllen. Zum anderen ist es notwendig, die Transkriptionsarbeit auf mehrere, speziell trainierte Transkribenten aufzuteilen. Dabei stellt sich das Problem, wie eine ausreichende intersubjektive Konsistenz der Transkriptionen gewährleistet werden kann. Letzteres muß schon in der Designphase in die Definition der Etiketten mit einfließen, um das Training möglichst zu erleichtern.

3 Das Etikettensystem

Das für die prosodische Etikettierung im Rahmen des Projekts VERBMOBIL entwickelte Inventar wurde aus den oben genannten Gründen in enger Zusammenarbeit mit den Projektpartnern, aber auch unter Einbeziehung der Ergebnisse von Perzeptionsexperimenten mit Transkribenten [24, 26] entwickelt. Neben der Konsistenz der Transkriptionen stand auch die unbedingte maschinelle Verarbeitbarkeit im Vordergrund.

Die Etikettierung wird bei diesem System auf vier Ebenen in zwei Stufen durchgeführt. Zunächst werden die Äußerungen auf Wortebene segmentiert (Orthographische Ebene). Die orthographische Transliteration kann dazu von Projektpartnern übernommen werden. In einem zweiten Schritt wird dann die eigentliche prosodische Etikettierung des Materials durchgeführt. Hier werden drei Ebenen parallel transkribiert, die unterschiedliche Informationen über die Prosodie bereitstellen:

1. Funktionale Ebene

2. Phrasengrenzen
3. Intonation

Alle drei Ebenen können später zusammen oder auch einzeln verwendet werden. Die Projektpartner können sich die für sie relevante prosodische Information aus den Etikettendateien heraussuchen. Im folgenden sollen die drei Ebenen noch einmal einzeln beschrieben werden:

4 Funktionale Ebene

Auf dieser Ebene werden eher funktionale Aspekte der Prosodie beschrieben (zur Unterscheidung funktional/formal vgl. [9]). Es werden unterschiedlich starke Prominenz und Satzmodus etikettiert. Im einzelnen gibt es die folgenden Etiketten:

Hauptakzent (PA) Das prominenteste Wort innerhalb jeder Intonationsphrase (B3, s.u.) erhält den Hauptakzent der Phrase. Er ist im Grunde ein perzeptives Äquivalent zum Fokus. Im Prinzip kann pro Phrase nur ein Hauptakzent auftreten; in Fällen, wo keine eindeutige Entscheidung möglich ist (bei Spontansprache nicht selten), können aber auch mehrere Wörter den Hauptakzent erhalten. Jedoch ist der Hauptakzent in seiner Definition auf die jeweilige Phrase begrenzt; er stellt keine Akzentuierungs*stufe* dar.

Nebenakzent (NA) Weitere, schwächer akzentuierte Wörter erhalten als Etikett den Nebenakzent.

Emphase/Kontrast (EK) Ist das Wort besonders stark hervorgehoben, wird statt des Hauptakzents ein emphatischer Akzent etikettiert.

Satzmodus (?) Bei Spontandialogen ist die Unterscheidung Frage/Nichtfrage mitunter problematisch und läßt sich nur durch Anhören des Sprachsignals durchführen. An Phrasengrenzen werden daher Fragen etikettiert, und zwar vier genau definierte Fragetypen. Durch die Information auf den anderen Ebenen läßt sich so angeben, ob eine Frage prosodisch oder anders (z.B. durch Wortstellung) markiert wurde.

5 Phrasengrenzen

Auf dieser Ebene werden unterschiedliche Arten von Phrasengrenzen etikettiert:

Intonationsphrasengrenze (B3) Dies ist eine “starke” Phrasengrenze, die eine vollständige Intonationseinheit beendet. Diese Phrasengrenze wird durch eine starke intonatorische Markierung und/oder Dehnung oder Pause gekennzeichnet. Alternativ kann auch ein Wechsel des Sprechtempos den Beginn einer neuen Phrase kennzeichnen. Intonationsphrasengrenzen werden auf der Intonationsebene (s.u.) immer bitonal markiert.

Intermediäre Phrasengrenze (B2) Diese Phrasengrenzen haben strukturierende Funktion innerhalb der Intonationsphrase [21]. Sie sind schwächer als B3 und sind meist nur durch eine schwache tonale Markierung gekennzeichnet.

Irreguläre Phrasengrenze (B9) Als irregulär werden Phrasengrenzen markiert, die nicht intentional strukturierend sind, also Performanzgrenzen wie Häsitationen, Abbrüche etc. An diesen Stellen treten z.B. Pausen auf, nach denen die Intonationskontur unverändert fortgeführt wird. Diese Grenzen erhalten zwar auch eine tonale Markierung, meist ist diese jedoch nicht entscheidbar.

6 Intonation – Das ToBI-System

Das ToBI–System (Tone and Break Indices) zur prosodischen Etikettierung wurde Anfang der 90er Jahre in den USA entworfen und hat sich mittlerweile zum Standardsystem für die Beschreibung der Prosodie des Englischen entwickelt [32, 31].

Das System basiert auf dem Tonsequenzansatz [22, 17, 16], der für Englisch sehr verbreitet ist und auch schon für die prosodische Phonologie des Deutschen verwendet wurde [10, 35, 36]. Für das ToBI–System wurde der Tonsequenzansatz speziell für die Etikettierung auf eine gute Konsistenz hin optimiert und auch evaluiert. Weiterhin wurde dafür Trainingsmaterial erarbeitet, um das Training an unterschiedlichen Orten zu ermöglichen [6, 7].

Dies System wurde für das Deutsche adaptiert. Dies wurde dadurch erleichtert, daß das ToBI–System auch in Deutschland an mehreren Stellen verwendet wird [20, 12, 28]. Die Zusammenarbeit mit diesen Stellen soll zu einem ToBI–Standard auch für das Deutsche führen.

Auch das ToBI–System besteht aus mehreren Ebenen; besonders wurde für das hier beschriebene System die tonale Beschreibung übernommen. Merkmal dieses Systems ist die Intonationsbeschreibung durch aufeinanderfolgende hohe (H) und tiefe (L) Töne. Diese Tonfolge wird der segmentalen Silbenfolge zugeordnet. Die Beschränkung auf nur zwei Töne erleichtert die Beschreibung erheblich, und durch Kombinationen von zwei Tönen lassen sich auch Tonbewegungen elegant ausdrücken. Bei den Tönen gibt es Akzenttöne, die durch ein nachgestelltes '*' gekennzeichnet werden (also z.B. H*,L*), Phrasentöne (durch '–' gekennzeichnet) und Grenztöne (durch '%' gekennzeichnet). Bitonale Akzente werden durch einen vor- bzw. nachgestellten Ton markiert.

Eine Äußerung wird als Folge von Intonationsphrasen beschrieben, die jeweils aus einer Folge von Tonakzenten bestehen. Der letzte Akzent (Nuklearakzent) hat dabei eine besondere Bedeutung. Der Verlauf zwischen Nuklearakzent und folgender Phrasengrenze wird durch den Phrasenakzent markiert; zusätzlich erhalten Intonationsphrasen am Ende noch einen Grenzton.

Für das Deutsche wurden im hier beschriebenen System die folgenden Etiketten verwendet:

6.1 Akzente

H* Gipfelakzent, hohe F_0 in akzentuierter Silbe

L+H* Gipfelakzent mit starkem Anstieg innerhalb der akzentuierten Silbe, tiefe F_0 in Vorgängersilbe.

L*+H "verzögerter" Gipfel. Anstieg in der akzentuierten Silbe, Gipfel erst in darauf folgender Silbe.

H+!H* “früher” Gipfel. Fallender F_0–Verlauf beginnend auf einer der vorhergehenden Silben, mittlere bis tiefe F_0 in der akzentuierten Silbe.

L* Talakzent

Zusätzlich kann noch *Downstepping* etikettiert werden. Als Downstepping wird eine Folge von Gipfelakzenten bezeichnet, bei der jeweils die folgenden Gipfel bei einer niedrigeren F_0 liegen, also eine absteigende Treppe bilden. Downstepping wird diakritisch mit einem **!** markiert. Es tritt nur bei Gipfelakzenten auf (also !H*, L+!H*, L*+!H).

6.2 Phrasengrenzen

Intonationsphrasen (B3) werden grundsätzlich bitonal etikettiert; sie bestehen aus einem Phrasenakzent und einem Grenzton. Durch Kombination der beiden ergeben sich vier verschiedene Möglichkeiten:

L–L% tiefe, fallende Phrasengrenze, “terminal”.

H–H% hohe, steigende Phrasengrenze, “Frageintonation”.

L–H% alternative Frageintonation nach Gipfelakzent, also z.B. H* L–H% (Hamburger Dialekt).

H–L% mittlere ebene bzw. hohe leicht fallende Kontur. “progredient” (continuation rise).

B2 und B9 erhalten nur einen Phrasenakzent als tonale Markierung, also **L–** bzw. **H–**.

7 Beispiel

Dies Inventar soll nun an einem Beispiel näher erläutert werden. In der folgenden Äußerung wurde die prosodische Etikettierung dem Text zugeordnet. Die drei Ebenen sind dabei untereinander dargestellt.

```
ja prima   dann lassen Sie uns doch noch einen Termin ausmachen
    PA           NA                                PA
        B3                    B2                                B3
  L+H*  H-H%    H*            L-                  H*            L-L%

wann w"ar´s Ihnen denn recht
            NA          PA  ?
                            B3
            H*          !H* L-L%
```

In dieser Äußerung wurden drei Intonationsphrasen etikettiert. In allen drei Phrasen ist der letzte Akzent vor der Phrasengrenze der PA; dies ist der Defaultfall für das Deutsche, also “normale” Akzentuierung. In der zweiten Intonationsphrase befindet sich zwischen `uns` und `doch` eine B2–Grenze, die an dieser Stelle ganz erheblich zur Disambiguierung beitragen kann. Die letzte Intonationsphrase ist eine Frage, die durch ein Fragepartikel, aber nicht durch Frageintonation markiert ist. Dies ist durchaus der Normalfall, daß der Satzmodus vom Sprecher nicht doppelt markiert wird.

8 Transkriptionstools

Die Etikettierung großer Korpora erfordert eine Arbeitsumgebung, die speziell an die jeweilige Aufgabe angepaßt ist. Zusätzlich sollten automatisierbare Vorgänge auch automatisch bearbeitet werden, um die Transkribenten von derartigen Tätigkeiten zu entlasten. Z.B. werden die Sprachdaten automatisch auf Wortebene vorsegmentiert [19], diese Wortgrenzen müssen nur noch korrigiert werden, was die Segmentierung der Wortgrenzen stark vereinfacht. Als Arbeitsumgebung dient das Programm *fish*, welches durch seinen modularen Aufbau eine flexible Anpassung an die Transkriptionsaufgabe ermöglicht [25]. Die resultierenden Transkriptionen werden beim Speichern auf formale Konsistenz überprüft (ob z.B. alle markierten Akzente auch eine tonale Etikettierung erhalten haben).

Für die Transkriptionsdateien wird das SAM–Format verwendet [1]. Bei diesem Dateiformat können unterschiedliche Informationen (z.B. orthographische, phonetische, prosodische) in einer einzigen Datei gehalten werden, ohne daß diese unübersichtlich wird. Die einzelnen Informationen lassen sich sehr schnell automatisch aus dieser Datei extrahieren.

Ein Teil des Materials wird zur Evaluierung der Transkriptionen verwendet. Dazu werden die Etikettierungen unterschiedlicher Transkribenten verglichen und die Konsistenz untersucht [24, 26, 27]. Diese Untersuchungen lassen quantitative Aussagen über die Qualität der Etikettierung zu. Dabei lassen sich auch Verbesserungen durch Training messen [33, 34].

9 Ausblick

Mit dem hier beschriebenen Transkriptionssystem wurde bisher Sprachmaterial im Umfang von einer Stunde etikettiert. Erste Ergebnisse zeigen, daß dies Material für automatische Prosodieanalyse als Trainingsmaterial nur zur Unterscheidung weniger Klassen (wie akzentuiert/unakzentuiert) ausreicht. Zur Klassifikation der Tonakzente ist das Material noch nicht umfangreich genug [15].

Weiterhin soll auch die prosodische Etikettierung teilweise automatisiert werden, so daß die Transkribenten nur noch die Aufgabe der Verifikation der Etikettierung haben [18]. Dabei sollen sie durch eine Resynthese der hypothetischen Etikettierung unterstützt werden [23].

[1] *User guide to ETR tools*. SAM-UCL-G007, pp. 15–19, 1992.

[2] L.M.H. Adriaens. *Ein Modell deutscher Intonation. Eine experimentell- phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzänderungen in vorgelesenem Text*. Diss., University of Eindhoven, 1991.

[3] Paul C. Bagshaw. An investigation of acoustic events related to sentential stress and pitch accents, in English. *Speech Communication*, 13:333–342, 1993.

[4] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, & U. Kilian. The prosodic marking of accents and phrase boundaries: Expectations and results. In A.J. Rubio, Hrsg., *NATO ASI: New Advances and Trends in Speech Recognition and Coding*, S. 89–92, Bubion (Granada), June-July 1993.

[5] A. Batliner, C. Weiand, A. Kießling, & E. Nöth. Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody. In *Working papers*, Vol. 41, S. 112–115, Lund University, 1993. Dept. of Linguistics.

[6] Mary E. Beckman & Gayle M. Ayers. Guidelines for ToBI labelling. Ms., Linguistics Dept., Ohio State University, Feb. 1994.

[7] Mary E. Beckman & Julia Hirschberg. TheToBI annotations conventions. Ms., Linguistics Dept., Ohio State University, 1994.

[8] Nick Campbell. Automatic detection of prosodic boundaries in speech. *Speech Communication*, 13:343–354, 1993.

[9] Anne Cutler & D. Robert Ladd. *Prosody: Models and Measurements*. Springer, Berlin/New York, 1983.

[10] Caroline Féry. *German Intonational Patterns*. Linguistische Arbeiten 285. Niemeyer, Tübingen, 1993.

[11] Dafydd Gibbon & Margret Selting. Intonation und die Strukturierung eines Diskurses. *Zeitschrift für Literaturwissenschaft und Linguistik*, 49:53–73, 1983.

[12] Martine Grice & Ralf Benzmüller. Transcription of German using tobi–tones: The Saarbrücken System. Ms., Institut für Phonetik, Universtität Saarbrücken, Saarbrücken, 1995.

[13] A. Kiešling, R. Kompe, H. Niemann, E. Nöth, & A. Batliner. “roger”, “sorry”, “i’m still listening”: Dialog guiding signals in information retrieval dialogs. In *Working papers*, Vol. 41, S. 112–115, Lund University, 1993. Dept. of Linguistics.

[14] K.J. Kohler. PROLAB – the Kiel system of prosodic labelling. Erscheint in: Proc. International Congress of Phonetic Sciences, Stockholm, August 1995.

[15] R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, & A. Batliner. Prosodic scoring of word hypotheses graphs. Erscheint in: Proc. EUROSPEECH, Madrid 1995, September 1995.

[16] D. Robert Ladd. Phonological features of intonational peaks. *Language*, 59(4), 1983.

[17] D. Robert Ladd, Kim E.A. Silverman, & Klaus R. Scherer. Parametrische und kategoriale Ansätze bei der Erforschung intonatorischer Funktion. *Zeitschrift für Literaturwissenschaft und Linguistik*, 49:124–133, 1983.

[18] Michael Lehning. Statistical methods for the automatic labelling of German prosody. Erscheint in: Proc. EUROSPEECH, Madrid 1995, September 1995.

[19] Michael Lehning & Rainer Grünheid. Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen. In *Fortschritte der Akustik – DAGA 94*, Bad Honnef, 1994. DPG-GmbH.

[20] Jörg Mayer. Transcription of German intonation: The Stuttgart System. Ms., Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, 1995.

[21] Marina Nespor & Irene Vogel. Prosodic structure above the word. In Anne Cutler & D. Robert Ladd, Hrsgg., *Prosody: Models and Measurements*, S. 123–140. Springer, Berlin, 1983.

[22] Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation*. Diss., M.I.T., 1980.

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

[24] Matthias Reyelt. Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Working papers*, Vol. 41, S. 238–241, Lund University, 1993. Dept. of Linguistics.

[25] Matthias Reyelt. Ein flexibles Programmpaket zur Visualisierung von Sprachdaten. In K. Fellbaum, Hrsg., *Tagungsband Elektronische Sprachsignalverarbeitung*, S. 358–365, Berlin, 1994.

[26] Matthias Reyelt. Untersuchungen zur Konsistenz prosodischer Etikettierungen. In H. Trost, Hrsg., *KONVENS 94*, S. 290–299, Berlin, 1994. Springer.

[27] Matthias Reyelt. Consistency of prosodic transcriptions. labelling experiments with trained and untrained transcribers. Erscheint in: Proc. International Congress of Phonetic Sciences, Stockholm 1995, August 1995.

[28] Matthias Reyelt & Anton Batliner. Ein Inventar prosodischer Etiketten für VERBMOBIL. Verbmobil–Memo 33/94, 1994.

[29] Margret Selting. Descriptive categories for the auditive analysis of intonation in conversation. *Journal of Pragmatics*, 11:777–791, 1987.

[30] Margret Selting. The role of intonation in the organization of repair and problem handling sequences in conversation. *Journal of Pragmatics*, 12:293–322, 1988.

[31] K. Silverman, E. Blaauw, J. Spitz, & J. Pitrelli. Prosodic comparison of spontaneous speech and read speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, S. 1299–1302, 1992.

[32] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, & Julia Hirschberg. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, S. 867–870, 1992.

[33] Eva Strangert & Mattias Heldner. Labelling of boundaries and prominences by phonetically experienced and non experienced transcribers. In *PHONUM*, Vol. 3, S. 85–109, UmeåUniversity, 1995. Dept. of Phonetics.

[34] Eva Strangert & Mattias Heldner. The labelling of prominence in Swedish by phonetically experienced transcribers. Erscheint in: Proc. International Congress of Phonetic Sciences, Stockholm, August 1995.

[35] Susanne Uhmann. Akzenttöne, Grenztöne und Fokussilben. Zum Aufbau eines phonologischen Intonationssystems für das Deutsche. In Hans Altmann, Hrsg., *Intonationsforschungen*, Linguistische Arbeiten 200, S. 65–88. Niemeyer, Tübingen, 1988.

[36] Susanne Uhmann. *Fokusphonologie: eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*. Linguistische Arbeiten 252. Niemeyer, Tübingen, 1991.

[37] W. Wahlster. Verbmobil: Translation of face-to-face dialogs. In *Proceedings Eurospeech 93*, 1993.

[38] Michelle Q. Wang & Julia Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196, 1992.

[39] Colin W. Wightman & Mari Ostendorf. Automatic recognition of intonational features. In *Proceedings ICASSP 1992 ?*, S. 221–224, 1992.

Dieser Report erscheint als Beitrag im Tagungsband:
Elektronische Sprachsignalverarbeitung, Wolfenbüttel, 4.–6. September 1995.