BMBF

Verbmobil
Verbundvorhaben

Tempo und Tempowechsel in Verbmobil–Dialogen

A. Batliner, A. Kießling,
R. Kompe

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Vm **Memo 110**
August 1996

August 1996

A. Batliner, A. Kießling,
R. Kompe

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (089) 2180 - 2916
e-mail: Anton.Batliner@phonetik.uni-muenchen.d400.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 F/4 und 01 IV 102 H/0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Tempo und Tempowechsel in VERBMOBIL–Dialogen

Anton Batliner, Andreas Kießling, Ralf Kompe

Inhaltsverzeichnis

1 Einleitung

Es gibt Sprecher, die grundsätzlich schnell, und solche, die grundsätzlich langsam reden, wobei im Augenblick dahingestellt bleibt, wie viele Stufen man annehmen soll und ob diese Stufen klar getrennt werden können. Ein solches Sprechtempo ist sprecherspezifisch und charakterisiert einen Sprecher/eine Sprecherin. Der Mensch kann sich als Hörer sehr schnell auf das Sprechtempo seines Gegenübers einstellen, er kalibriert also seine Perzeption am jeweiligen Dialogpartner. Bei extremen Schnellsprechern kann es aber auch für ihn – zumindest in einer Anfangsphase – schwierig werden. Es gibt genügend Hinweise darauf, daß es in der automatischen Spracherkennung ähnlich ist: Üblicherweise sind die Worterkennungssysteme nicht auf unterschiedliche Sprechtempos hin trainiert; in die Lernstichprobe gehen einfach alle möglichen Arten von Sprechern, Langsamsprecher und Schnellsprecher, mit ein. Versuche, etwa mit zwei Worterkennern, einem für Langsamsprecher und einem für Schnellsprecher, zu arbeiten, haben aber gezeigt, daß damit die Worterkennung verbessert werden kann, vgl. [Mir96].

Ein und derselbe Sprecher kann aber auch das Sprechtempo wechseln. Das kann zum einen von seinem affektiven Zustand abhängen: Im Zustand der Erregung (Ärger, partiell auch Freude) tendiert man eher zu einem schnelleren Sprechtempo, im Zustand der Resignation (z.B. Trauer) tendiert man eher zu einem langsameren Sprechtempo; vgl. den Literaturüberblick in [Tis93]. Es gibt aber auch einen Tempowechsel bei gleichbleibendem affektiven Zustand, so wie er etwa bei 'normalem', nicht mißglücktem Gesprächsverlauf im VERBMOBIL–Szenario anzunehmen ist[1]. Hier liegen andere Auslöser vor, die eher im Planungsverhalten des Sprechers zu suchen sind, analog zu Zögerungsphänomenen, vgl. etwa [Bat94] sowie [Lev89]: Die Dehnung eines einzelnen Wortes oder eine gefüllte Pause können die Funktion haben, die Redeübernahme des Dialogpartners (turn taking) zu verhindern, solange man seinen Beitrag noch plant (paralinguistische Funktion). Aus demselben Grund kann ein Sprecher eine ganze Phrase oder einen ganzen Satz langsamer produzieren, als es sonst für ihn typisch ist. Allerdings kann auch eine temporäre Beschleunigung des Sprechtempos genau diesem Zweck dienen; vgl. dazu die Akzentuierung, bei der es im Grunde auf die *Abweichung* von der Norm ankommt, sei das nun eine deutliche Anhebung des Tonverlaufs oder eine Absenkung des Tonverlaufs, auch wenn die Anhebung (Rise–Fall, Gipfelakzent, H*) zumindest im Standarddeutschen eher üblich ist. Eine Beschleunigung des Tempos kann aber auch damit zu tun haben, daß der Sprecher schon Bekanntes quasi der Vollständigkeit halber nur noch schnell erwähnt, ohne es zu sehr hervorheben zu wollen. Wenn das regulär der Fall ist, kann man auch versuchen, solche Tempoänderungen für die semantische Analyse einzusetzen (schneller = thematische Passagen, langsamer = rhematische Passagen).

2 Stand der Forschung

Die bisherigen Aussagen waren eher tentativ, weil es zwar relativ viele Arbeiten zum Sprechtempo gibt, der Stand der Forschung zur Änderung des Sprechtempos aber gleichermaßen tentativ ist. [Tis93] erwähnt 28 Arbeiten innerhalb der Emotionsforschung, in denen das Sprechtempo entweder durch subjektives Rating oder durch eine Maßzahl wie Silben pro Sekunde über die ganze Äußerung untersucht wurde. Die Bestimmung des Sprechtempos in der automatischen Spracherkennung, vgl. [Wig92], wird – abgesehen vom Einsatz in der Worterkennung, vgl. [Mir96] – eher zur Lautdauernormierung und damit zur besseren Modellierung der Akzentuierung und der Phrasierung verwendet als zur Modellierung der spezifischen Funktionalität des Tempowechsels. Das Tempo wird daher pro Sprecher, pro Satz oder pro Turn bestimmt, nicht aber für kleinere Phrasierungseinheiten.

Innerhalb der Phonetik gibt es fast nur Arbeiten, die kontrolliertes und elizitiertes Material zugrunde legen, vgl. etwa den Beitrag von [Not91]; eher singulär ist [Mil84], wo lokale Tempoänderungen bei 30 Sprechern in einer Interview–Situation untersucht wurden. Auch in der psycholinguistischen Literatur finden sich meist entweder Aussagen zu kontrolliertem, nicht spontansprachlichem Material oder Eindrücke, die eher nicht quantifizierbar bzw. auf Spontansprache übertragbar sind.

[1] Mißglückt ist etwa eine Dialogsequenz, in der das System eine Äußerung wiederholt falsch versteht; der Ärger des Benutzers kann sich dann z.B. in einem schnelleren Sprechtempo niederschlagen.

Voraussetzung für die automatische Bestimmung prosodischer Ereignisse ist eine etikettierte Referenzstichprobe, in der die zeitlichen Positionen der Ereignisse geeignet annotiert sind. Für die prosodische Etikettierung werden üblicherweise manuell perzeptive Etiketten erstellt; es hat sich aber gezeigt, daß für bestimmte Anwendungsgebiete, wie etwa syntaktische Phrasierung, Etiketten auch erstellt werden können, ohne daß man abhören muß, vgl. [Bat96]. Eine solche Vorgehensweise ist für die Etikettierung von Tempo bzw. Tempowechsel nicht möglich, das heißt, eine auditive Inspektion des Sprachmaterials ist unabdingbar. Bei der segmentalen Transkription findet sich pro Buchstabe/Laut grosso modo mindestens ein Etikett, dito bei der Transliteration einer Äußerung pro Wort ein transliteriertes Wort. Auf der einen Seite geht es natürlich schneller, ein großes Material nach Tempowechsel durchzugehen und entsprechend zu etikettieren, weil es um Größenordnungen weniger Tokens dafür gibt als etwa für Laute oder Wörter. Auf der anderen Seite müssen für seltene Phänomene wie Tempowechsel umfangreichere Korpora etikettiert werden, damit für ein Training genügend Tokens zur Verfügung stehen. Darüberhinaus ist es eine notwendige Voraussetzung, daß bereits Laute, Silben und/oder Wörter transliteriert und/oder etikettiert sind, da das Auftreten eines Tempowechsels sinnvollerweise Silben– bzw. Wortgrenzen zugeordnet wird. Für die automatische Berechnung prosodischer Merkmale zur Detektion von Tempowechsel ist dann auch die zeitliche Position eines Tempowechsels im Sprachsignal festlegbar (z.B. durch 'automatic time alignment').

Eine automatische Bestimmung des Sprechtempos und der Änderung des Sprechtempos ist prinzipiell zwar auch basierend auf automatisch detektierten Silbenkernen möglich (in Anzahl Silbenkerne pro Zeiteinheit), muß aber auf jeden Fall an einer manuell etikettierten Referenzstichprobe evaluiert werden. Untersuchungen, wie sie in diesem Beitrag dargestellt werden, sind also letztlich erst durch die gründliche Aufbereitung großer Korpora möglich.

Mit den im folgenden dargestellten Arbeiten soll versucht werden, einen ersten Beitrag zur Untersuchung des Sprechtempos und des Sprechtempowechsels in spontaner Sprache und ihrer automatischen Klassifikation anhand eines großen Korpus zu leisten.

3 Sprechtempo

Die akustisch-prosodische Realisierung von Akzenten und Grenzen hängt in starkem Maße von der Sprechgeschwindigkeit ab. Insbesondere hat die Sprechgeschwindigkeit einen Einfluß auf die Laut– bzw. die Silbendauer, also auf wichtige Merkmale für die Erkennung von Akzentuierung (prominenzbezogene Dehnung) und Grenzen (phrasenfinale Dehnung). Die Berücksichtigung der Sprechgeschwindigkeit bei der Merkmalberechnung ist deshalb von besonderer Bedeutung. Im folgenden beschäftigen wir uns mit der Frage, inwieweit die implizite oder explizite Normierung mit Hilfe der Sprechgeschwindigkeit zu einer Verbesserung der Erkennungsrate führt.

3.1 Berechnung

In [Cry88] konnte experimentell gezeigt werden, daß der Mittelwert und die Standardabweichung der Lautdauer in etwa gleich starkem Maße linear von der Sprechgeschwindigkeit abhängen; eine Feststellung, die konsistent ist mit der Lautdauermodellierung durch Gammaverteilungen [Cry82]. Eine höhere Sprechgeschwindigkeit führt somit zu einer Reduktion der mittleren Laut– bzw. Silbendauer sowie deren Standardabweichungen. In [Wig92] wird als geeigneter Skalierungsfaktor die *mittlere normierte Sprechdauer* τ eingeführt, die z.B. über die gesamte Äußerung (oder einen bestimmten, lokal begrenzten Sprachsignalabschnitt) ermittelt werden kann:

$$(1) \quad \tau = \frac{1}{k} \sum_{i=1}^{k} \frac{d(i)}{\mu_{laut(i)}}$$

wobei

k	Anzahl der Laute in der Äußerung, ohne Berücksichtigung von Pausen
$d(i)$	Dauer des i-ten Lautsegments mit dem zugehörigen Laut $laut(i)$
$\mu_{laut(i)}$	Mittelwert der Dauer des Lautes $laut$, der dem Segment i zugrunde liegt

Mit Hilfe der mittleren normierten Sprechdauer läßt sich die *mittlere normierte Lautdauer* DAUER_{norm} von Lauten, Silbenkernen, Silben, Wörtern etc. auf einfache Weise berechnen mittels:

$$(2) \quad Dauer_{norm} = \frac{1}{l}\sum_{i=1}^{l}\frac{d(i)-\tau\cdot\mu_{laut(i)}}{\tau\cdot\sigma_{laut(i)}}$$

wobei

l	Anzahl der Laute in der Silbe (bzw. 1 bei Silbenkernen)
$d(i)$	Dauer des i-ten Lautsegments mit dem zugehörigen Laut $laut(i)$
$\mu_{laut(i)}$	Mittelwert der Dauer des Lautes $laut$, der dem Segment i zugrunde liegt
$\sigma_{laut(i)}$	Standardabweichung der Dauer des Lautes $laut$ des i-ten Lautsegments

Die lautintrinsischen Werte μ_{laut} und σ_{laut} müssen vorab anhand einer Trainingsstichprobe geschätzt werden. Wir betrachten dabei unterschiedliche alternative Kontexte eines Lautes:

- Bei DAUER_{norm} werden alle Auftreten des Lautes $laut$ der gesamten Stichprobe zur Schätzung von μ_{laut} und σ_{laut} verwendet.
- Bei $\text{DAUER}^{WorAkz}_{norm}$ werden unterschiedliche μ_{laut}, σ_{laut} geschätzt abhängig davon, ob die Silbe, in welcher sich der Laut befindet, den lexikalischen Wortakzent trägt oder nicht; für jede Lautklasse sind also zwei weitere (μ,σ)–Paare aus der Trainingsstichprobe zu schätzen, mit denen die Laute normiert werden.
- Bei $\text{DAUER}^{SilPos}_{norm}$ werden unterschiedliche μ_{laut}, σ_{laut} geschätzt abhängig davon, ob der Laut sich in einem einsilbigen Wort oder in der wortinitialen, wortfinalen oder einer wortinternen Silbe eines mehrsilbigen Wortes befindet. Da dies getrennt für Silben mit lexikalischem und ohne lexikalischem Wortakzent berechnet wird, ist also die Schätzung von acht weiteren (μ,σ)–Paaren pro Lautklasse nötig, mit denen die Laute normiert werden.

Diese Arten der Normierung werden im folgenden als explizite Normierung bezeichnet. In analoger Weise lassen sich auch auf verschiedene Weise normierte Sprechdauern (s.u.) gemäß Gleichung 1 ermitteln, indem die jeweils entsprechenden $\mu_{laut(i)}$ verwendet werden.

3.2 Klassifikation von Akzenten und Grenzen

Erste Untersuchungen, die sich mit der Frage beschäftigen, inwieweit die Berücksichtigung der Sprechgeschwindigkeit zu einer Verbesserung bei der Klassifikation von Grenzen und Akzenten führt, wurden in [Wig92] an von professionellen Radiosprechern gelesenem Sprachmaterial durchgeführt. Für spontansprachliches Material sind unseres Wissens bislang noch keine umfangreichen Untersuchungen dieser Fragestellung bekannt.

Die im folgenden beschriebenen Experimente wurden an dem von der TU Braunschweig prosodisch etikettierten VERBMOBIL-Material durchgeführt (cf. [Rey94, Rey95]): Für das Training wurden 30 Dialoge (797 Turns; 100 Minuten Sprache) von 53 männlichen und 7 weiblichen Sprechern verwendet. Die Teststichprobe wurde von der TU Braunschweig ausgewählt und umfaßt 3 Dialoge (64 Turns; 12 Minuten Sprache) von 3 männlichen und 3 weiblichen Sprechern. Die Berechnung der akustisch-prosodischen Merkmale basiert auf der automatischen Zeitzuordnung der gesprochenen Wortkette. Für jede Wortendesilbe wird hier ein Vektor mit 276 prosodischen Merkmalen über einem maximalen Kontext von ± 2 Silben beziehungsweise von ± 2 Wörtern um die aktuelle Silbe berechnet, der die prosodischen Eigenschaften (Grundfrequenzbewegungen,

Merkmalmengen (SET)	Anzahl Merk-male	SET alleine ¬A | A $\mathcal{ER}$	SET alleine ¬A | A $\mathcal{ER}_{\overline{K}}$	SET alleine B3 | B[029] $\mathcal{ER}$	SET alleine B3 | B[029] $\mathcal{ER}_{\overline{K}}$	ALLE \ SET ¬A | A $\mathcal{ER}$	ALLE \ SET ¬A | A $\mathcal{ER}_{\overline{K}}$	ALLE \ SET B3 | B[029] $\mathcal{ER}$	ALLE \ SET B3 | B[029] $\mathcal{ER}_{\overline{K}}$
ALLE	276/0	**82.6**	(**82.2**)	**88.3**	(**86.8**)	—		—	
DAUER_{alle}	60/216	74.9	(74.7)	78.7	(77.7)	81.7	(81.4)	83.9	(85.1)
SPRECHDAUER	3/273	50.4	(51.3)	48.6	(54.9)	82.0	(81.5)	87.7	(86.2)
		SET alleine				ALLE \ DAUER_{alle} ∪ SET			
$\text{DAUER}_{unnormiert}$	15/231	67.0	(67.0)	74.4	(75.0)	82.2	(81.8)	85.6	(84.8)
DAUER_{norm}	15/231	69.5	(69.2)	72.3	(74.1)	81.8	(81.4)	87.2	(85.1)
$\text{DAUER}_{norm}^{WorAkz}$	15/231	66.7	(66.2)	72.9	(73.6)	82.4	(82.0)	86.4	(85.2)
$\text{DAUER}_{norm}^{SilPos}$	15/231	68.5	(67.7)	71.9	(73.3)	82.0	(81.6)	85.4	(84.6)

Tabelle 1: Erkennungsraten zur Klassifikation von Akzenten (¬A | A) und prosodischen Grenzen (B3 | B[029]) mit unterschiedlichen Merkmalsätzen in VERBMOBIL. Differenziert wird dabei die Klassifikation mit verschiedenen Merkmalmengen (SET alleine) sowie die Klassifikation mit allen Merkmalen ohne diese Menge (ALLE \ SET). In der Spalte 'ALLE \ DAUER_{alle} ∪ SET' werden neben den 216 'Nichtdauer-Merkmalen' nur die in der ersten Spalte spezifizierten Dauermerkmale verwendet. Neben der mittleren Erkennungsrate $\mathcal{ER}$ ist in Klammern auch der Mittelwert der klassenweisen Erkennungsraten $\mathcal{ER}_{\overline{K}}$ angegeben; weitere Erläuterungen im Text.

Dauer– und Energieverhältnisse, Pausensetzung) des Wortes und seiner Umgebung charakterisiert. Für die Klassifikation der Grenzen und Akzente werden mit den annotierten Merkmalvektoren der Trainingsstichprobe Mehrschichtenperzeptren trainiert und an der Teststichprobe ausgewertet. Detailliertere Beschreibungen zur Berechnung der prosodischen Merkmale und zur Klassifikation finden sich z.B. in [Kie96a, Kie96b] sowie in [Kom95, Bat96].

Basierend auf den bis dato besten Erkennungsraten für akzentuierte Wörter und prosodische Grenzen [Kie96a, Kie96b], haben wir zum einen die Einflüsse der unterschiedlichen Normierungsarten (explizite Normierung) und zum anderen die direkte Verwendung der mittleren normierten Sprechdauer als Merkmal (implizite Normierung) untersucht. Die Ergebnisse dieser Experimente sind in Tabelle 1 zusammengefaßt, und zwar sowohl für die Klassifikation der prosodischen Grenzen (B3 | B[029]) als auch für die Akzentklassifikation (¬A | A). Für jede Merkmalmenge ist in Tabelle 1 das jeweils beste damit erzielte Resultat angegeben. Dabei bezeichnet die Spalte 'SET alleine' die Ergebnisse, die sich mit der in der ersten Spalte spezifizierten Merkmalmenge ergeben, die Spalte 'ALLE \ SET' bezeichnet das Komplementärereignis dazu, also die Erkennungsraten unter Verwendung aller Merkmale *ohne* die in der ersten Spalte spezifizierten. In der Spalte 'ALLE \ DAUER_{alle} ∪ SET' werden neben den 216 'Nichtdauermerkmalen' nur die 15 in der ersten Spalte spezifizierten Dauermerkmale verwendet.

In den Experimenten wurden insgesamt drei verschiedene Sprechdauern (Zeile 'SPRECHDAUER') verwendet, die jeweils über den gesamten Turn berechnet werden: die mittlere normierte Sprechdauer, die normierte Sprechdauer unter Berücksichtigung der Akzentposition sowie die normierte Sprechdauer unter Berücksichtigung der Silbenposition im Wort (vgl. oben). Ohne die Berücksichtigung weiterer prosodischer Merkmale (Zeile 'SPRECHDAUER', Spalte 'SET alleine') sind diese Sprechgeschwindigkeitsmaße natürlich nicht für eine direkte Klassifikation von Grenzen und Akzenten geeignet. Ein Vergleich von Zeile 'ALLE' mit 'ALLE \ SPRECHDAUER' läßt aber erkennen, daß die implizite Normierung der Merkmale durch die Sprechdauer zu einer deutlichen Verbesserung der Erkennungsraten sowohl bei den Grenzen als auch bei den Akzenten führt.

Werden bei der Grenzklassifikation (B3 | B[029]) nur Dauermerkmale verwendet, so zeigen sich für die unnormierten Dauermerkmale geringfügig bessere Werte (Zeile '$\text{DAUER}_{unnormiert}$', Spalte 'SET alleine'); in Kombination mit den übrigen 216 'Nichtdauermerkmalen' zeigt sich DAUER_{norm} als etwas geeigneter. Die gemeinsame Verwendung aller 60 Dauermerkmale (Zeile 'DAUER_{alle}', Spalte 'SET alleine') anstelle von nur einer Normierungsart ergibt allerdings durchwegs bessere Resultate. So wird beispielsweise gegenüber den 15 unnormierten Merkmalen (Zeile '$\text{DAUER}_{unnormiert}$') die Fehlerrate bei Verwendung aller 60 Dauermerkmale (Zeile 'DAUER_{alle}') um ca. 17 % reduziert.

In Bezug auf die Akzentklassifikation ($\neg$A | A) zeigen die normierten Dauermerkmale (DAUER$_{norm}$) bei 'SET alleine' zum Teil deutlich bessere Resultate als die anderen Dauermerkmale; in Kombination mit den übrigen 216 'Nichtdauermerkmalen' ist allerdings die Normierung unter Berücksichtigung der Wortakzentpositionen (DAUER$^{WorAkz}_{norm}$) am geeignetsten. Die gemeinsame Verwendung aller 60 Dauermerkmale (Zeile 'DAUER$_{alle}$', Spalte 'SET alleine') anstelle von nur einer Normierungsart ergibt auch hier das beste Resultat.

Zusammenfassend läßt sich also feststellen: Die implizite Normierung der akustisch-prosodischen Merkmale durch die Sprechdauer führt sowohl bei den Grenzen als auch bei den Akzenten zu einer deutlichen Verbesserung der Erkennungsraten. Die Berücksichtigung von Dauerinformation ergibt eine bessere Erkennungsleistung von ca. 1% bei der Akzentklassifikation und von ca. 4% bei der Grenzklassifikation. Ein Vergleich der verschiedenen Dauernormierungen zeigt nur geringfügige Unterschiede in den Erkennungsraten. In jedem Fall ist aber die gemeinsame Verwendung von normierten und unnormierten Dauermerkmalen deutlich besser als die Verwendung von nur einer Normierungsart alleine.

4 Wechsel des Sprechtempos

4.1 Etikettierung des Tempowechsels

Es wurden die Turns aus den VM-CDs 1-5 zugrundegelegt, für die auch M-Labels vergeben wurden, also die beiden prosodischen Trainings- und Testkorpora (hauptsächlich von CD1), sowie die M-Trainingskorpora (fast alle Turns von CD2-5); im einzelnen dazu vgl. [Bat96]. Die Turns wurden von einem Phonetiker einzeln abgehört, wobei aus Aufwandsgründen mit einem Programm gearbeitet wurde, das jeweils den ganzen Turn vorspielt. Nach einigen Testdurchgängen und dem Abwägen von Pro und Kontra entschieden wir uns für das folgende Vorgehen:

An den Wortgrenzen, an denen eine deutliche Veränderung der Sprechgeschwindigkeit perzipiert werden konnte, wurde entweder " aa " für eine Beschleunigung des Tempos ("Allegro") oder " ll " für eine Verlangsamung ("Lento") etikettiert. Das Etikett wurde im Prinzip direkt rechts an die entsprechende Wortgrenze gesetzt, links und rechts jeweils mit einem Blank getrennt[2]. Aus Aufwandsgründen wurde auch darauf verzichtet, das *Ende* einer Passage mit geändertem Tempo zu bestimmen, es sei denn, es folgt eindeutig wieder eine Passage mit geändertem Tempo, vgl. die Beispiele 1 und 2. Defaultmäßig kann man auch annehmen, daß sich der Tempowechsel (insbesondere) über die nächste größere syntaktische Phrasierungseinheit erstreckt, vgl. dazu im einzelnen weiter unten.

Im weiteren Verlauf werden die folgenden Akronyme verwendet:

TCA Tempo Change Allegro, Beschleunigung des Sprechtempos

TCL Tempo Change Lento, Verlangsamung des Sprechtempos

TC Tempo Change, Änderung des Sprechtempos, also entweder TCA oder TCL

T0 jede andere Wortgrenze ohne Änderung des Sprechtempos

Das Analysefenster, über das sich der Tempowechsel bemerkbar machen mußte, betrug mehr als ein Wort. Das heißt, daß kürzere Tempoänderungen, wie sie z.B. bei Häsitationen im Wort beobachtet werden können, *nicht* etikettiert wurden. Diese Phänomene sind schon in der Basistransliteration verzeichnet und müssen daher nicht nochmals gelabelt werden. Es handelt sich dabei auch um eine eher kurzfristige Arhythmik, quasi ein 'Außer-Tritt- geraten', ohne daß damit eine echte Änderung des Sprechtempos verbunden wäre;

[2] Allerdings wurde diese Strategie aus – prima vista auch verständlichen – Gründen nicht immer durchgehalten, wenn etwa zwischen der rechten Wortgrenze und dem folgenden Wort, bei dem sich die Änderung bemerkbar machte, 'non-linguistische', in spitzen Klammern gesetzte Ereignisse wie '<Pause>, <Atmen>' o.ä. stehen. Sinngemäß gehört das Label für einen Tempowechsel ja an den Beginn des Wortes, bei dem er anfängt. Bei einer automatischen Weiterverarbeitung müßten also gegebenenfalls die Positionen vereinheitlicht werden.

	#
Turn	7286
Sprecher	362
TCA	208
TCL	114
TC	322
Turn mit TC	201
Sprecher mit TC	80
Wörter	149643
max. Turnlänge	158
max. TCA/Turn	3
max. TCL/Turn	2
max. TC/Turn	5

Tabelle 2: Kenndaten der Etikettierung

ihr Geltungsbereich erstreckt sich normalerweise nur über das jeweils betroffene Wort. Wenn aber eine Häsitation einen Tempowechsel einleitet, wird TC gelabelt.

Der intuitive Eindruck des Etikettierers war, daß die meisten Sprecher immer ein konstantes Sprechtempo haben, andere dagegen Tempowechsel einsetzen. Seine weiteren Eindrücke:

> *"Einfluß auf die Geschwindigkeitsempfindung scheinen mir auch pitch und Lautstärke zu nehmen. Leise und/oder tief scheint mir langsamer zu klingen als laut und/oder hoch, wofür ich allerdings keine objektiven Anhaltspunkte habe. Je komplexer die syntaktische Struktur, desto eher findet sich eine Änderung der Geschwindigkeit und zwar in der Weise, daß bestimmte Satzteile, die schon Bekanntes enthalten, meist etwas schneller gesprochen werden. Die meisten Variationen lassen sich wohl auf 'rhetorische' Gründe zurückführen."*

4.2 Ergebnisse und Diskussion

Tabelle 2 gibt einen ersten Überblick über die Ergebnisse der Etikettierung. In 2.2 Prozent der Turns kommt mindestens ein TC vor, auf 2.8 Promille der Wörter folgt ein TC. Bei 78 Prozent der Sprecher findet sich kein TC; das Phänomen ist also stark sprecherspezifisch. Bei den Sprechern mit TC finden sich im Schnitt pro Sprecher 2,5 Turns mit TC sowie 4 TC's. Die Verteilung auf die einzelnen CDs ist stark ungleichgewichtig: in CD2 finden sich 136 TCAs und 75 TCLs, in den anderen CDs 72 bzw. 39. Das Szenario und die Instruktion der Sprecher haben sich bei CD1 bis CD5 nicht grundsätzlich geändert, allerdings war nach der CD2 die Auswahl der Dialoge etwas 'verschärft', da einige Dialoge, die zu lange dauerten oder zu starke Sprecheridiosynkrasien aufwiesen, verworfen wurden.

Tabelle 3 zeigt die durchschnittliche Anzahl der Wörter vom Turnanfang bis zum ersten TC (Phase 1), vom ersten TC bis zum zweiten TC bzw. zum Turnende (Phase 2) usw. Diese Mittelwerte sind natürlich mit Vorsicht zu interpretieren, da die Standardabweichung relativ hoch ist und die Zahl der Fälle gegen Ende zu recht niedrig ist.

4.2.1 Merkmalausprägungen

Tabelle 4 zeigt die Mittelwerte der prosodischen Merkmale Dauer, Energie und F0 sowie die Mittelwerte der Regressionskoeffizienten von Energie und F0. Die Mittelwerte wurden jeweils an den TC-Tokens der Trainingsstichprobe (s.u.) ermittelt. Da sich mit dem größten zeitlichen Kontext die beste Erkennung ergibt, vgl. unten, werden die Werte dieses Kontexts dargestellt: zuerst der Wert jeweils über die drei Wörter vor

Phase	1	2	3	4	5	6
# Token	201	201	87	23	9	2
Mittelwert	16.8	10.8	10.8	7.1	10.4	11.5
Std.Abw.	12.1	10.3	8.1	5.2	6.2	12.0

Tabelle 3: Durchschnittliche Zahl der Wörter in den TC-Passagen

Merkmale	Kontext TCA (# = 128)			Kontext TCL (# = 69)		
	[-3,-1]	0	[1,3]	[-3,-1]	0	[1,3]
normierte Dauer	0.30	0.76	-0.74	-0.59	0.34	0.68
absolute Dauer	0.24	0.55	-0.52	-0.36	0.13	0.35
Energie Mittelwert	-0.07	-0.10	0.37	0.13	-0.16	0.04
Energie Regressionskoeffizient	-0.05	-0.02	0.73	-0.13	-0.19	0.39
F0 Mittelwert	-0.15	-0.10	0.30	-0.16	-0.13	0.05
F0 Regressionskoeffizient	-0.06	0.11	0.27	-0.11	0.03	0.00

Tabelle 4: Mittelwerte/Regressionskoeffizienten relevanter Merkmale für TA und TL

dem Wort, an dessen rechter Wortgrenze TC gelabelt wurde [-3,-1], dann der Wert dieses Wortes selbst (0), und zuletzt der Wert der drei nachfolgenden Wörter [1,3]. Die entsprechenden Werte für T0 (n=3816) liegen sämtlich erwartungsgemäß nahe Null und sind deshalb in der Tabelle nicht dargestellt. Die Tendenzen, die sich in diesen Werten widerspiegeln, lassen sich wie folgt zusammenfassen: Bei einem TCA verlangsamt sich das Tempo in Richtung auf den TC hin, insbesondere das des Wortes direkt vor dem TC (langsamer als der Durchschnitt bei T0); dann wird das Tempo deutlich schneller. Bei einem TCL ist es umgekehrt: Die letzten drei Wörter davor sind deutlich schneller als der Durchschnitt, dann wird das Tempo langsamer. Energie und F0 verhalten sich bei TCA und bei TCL im linken Kontext ähnlich: Die Werte vor dem TC sind etwas tiefer als der Durchschnitt oder fast gleich; nach dem TCA sind die Werte höher, nach dem TCL nur schwach höher oder gleich dem Durchschnitt. Das gleiche gilt für die Regressionskoeffizienten.

Zum einen zeigt sich damit die Korrelation der unterschiedlichen Merkmale untereinander, die allerdings sicher nicht nur eine Abhängigkeit redundanter von distinktiven Merkmalen ist, sondern auch vom Sprecher gesteuert wird. Zum anderen wird deutlich, daß die Sprecher quasi 'ausholen', also sich *vor* dem Wechsel antagonistisch verhalten. Sie gehen also nicht von einem neutralen Sprechtempo aus, sondern verlangsamen vor einer Beschleunigung oder werden vor einer Verlangsamung zuerst einmal schneller. Das bedeutet natürlich nicht, daß es keine Tempoänderung aus dem neutralen Tempo heraus geben kann, es bedeutet nur, daß das Etikettierkriterium "deutliche Tempoänderung" hauptsächlich eben von solchen Tempoänderungen erfüllt wird.

4.2.2 Auftreten an syntaktischen Grenzen

Tabelle 5 zeigt das Auftreten von TCA, TCL und TC an syntaktisch–prosodischen M–Grenzen, vgl. zu diesen Grenzen im einzelnen [Bat96] sowie Tabelle 6. M2I wurde nur für die 3 Testdialoge etikettiert, bei den restlichen Dialogen müssen sie daher unter M0 aufgeführt werden. Man beachte, daß diese Korrespondenzen automatisch ermittelt wurden. Ein manueller Durchgang ergab gewisse Abweichungen, die auf die unterschiedliche Abfolge der diversen Etiketten (M–Labels, irreguläre Grenzen, non–linguistische Ereignisse im Labelfile) zurückzuführen sind. Ein weiteres Caveat für die Interpretation ergibt sich aus der relativ geringen Zahl der Tokens von TC. Das grundsätzliche Bild dürfte sich aber nicht ändern und ist sinnvoll zu interpretieren:

TC's korrespondieren mit unterschiedlichen syntaktischen Grenzen; es ist daher nicht sehr wahrscheinlich,

M–Label	M–Anzahl	TCA	TCL	TC	% TC/M–label
M3S	11716	94	65	159	1.35
M3P	4551	27	8	35	0.76
M3E	1409	7	4	11	0.78
M3I	369	2	3	5	1.35
M3T	325	3	3	6	1.84
M3D	5148	9	0	9	0.17
M3A	733	4	0	4	0.54
M2I	132	1	1	2	1.51
M2I/M0	117974	61	31	92	0.07

Tabelle 5: Korrespondenzen zwischen Tempoänderungen und syntaktisch–prosodischen Grenzen

daß sie zur Klassifikation der einzelnen Grenztypen einen wesentlichen Beitrag leisten. Das war auch nicht zu erwarten, wenn man sie nicht als syntaktisches Phänomen, sondern als Phänomen der Sprechplanung versteht, analog zu den gefüllten Pausen; vgl. zur ähnlichen Korrespondenz der gefüllten Pausen mit syntaktischen Grenzen [Bat95]; dort fanden wir 9/10 der gefüllten Pausen an syntaktischen Grenzen. Wie bei den gefüllten Pausen ist also das Auftreten von Tempoänderungen nicht willkürlich und hauptsächlich an syntaktischen Grenzen unterschiedlicher Stärke. In der letzten Spalte von Tabelle 5 zeigt sich, daß bei vier M–Labels mehr als 1% TC's vorkommen (M3S, M3I, M3T, M2I), bei den anderen weniger, am wenigsten bei M3D. M3S ist die syntaktisch stärkste Grenze. M3I steht bei eingebetteten Sätzen (Parenthesen, Relativsätze, etc.); insb. Parenthesen werden oft als prototypische Kandidaten für TC angesehen. M3T steht bei Satzauftaktspartikeln, die durch eine Pause und/oder Atmen vom folgenden Satz getrennt sind; sie bieten sich also für eine Planungspause an, nach der die Planung entweder perseveriert (TCL) oder durch eine Beschleunigung abgelöst wird. Die manuelle Überprüfung zeigt, daß die meisten TC's der letzten Zeile in Tabelle 5 bei M2I, also bei einer Konstituentengrenze, stehen. Es finden sich aber auch Fälle wie " ... *um schon irgendwelche TCA Projekte vorzubereiten*", wo die TC *innerhalb* einer Konstituente steht.

Wenn man wie in Tabelle 6 alle syntaktischen Grenzen (M3S, M3P, M3E, M3I, M3T) auf 'Grenze' (M3, 18370 Tokens) abbildet, so findet sich hier in 1.17% der Fälle ein TC. Bei den ambigen Grenzen MU (M3D, M3A, 5881 Tokens) sind es 0.22%, beim Rest M0 (Konstituentengrenzen M2I und alle anderen Wortgrenzen M0, 117974 Tokens) 0.07. Damit wird deutlich, daß der Tempowechsel fast ausschließlich an einer syntaktischen Grenze stattfindet. An syntaktischen Grenzen tritt 17-mal häufiger eine TC auf als an einer normalen Wortgrenze.

4.2.3 Funktionen der TCs im Dialogverlauf

Beispiel 1 zeigt drei TC's in einem Turn, zwei an einer M3S–Grenze und der letzte an einer Konstituentengrenze. Beispiel 2 zeigt drei TC's an unterschiedlichen Grenzen, und Beispiel 3 zeigt einen Fall, wo *nur* der TC zwischen verschiedenen Lesarten disambiguiert. In Beispiel 4 findet sich quasi ein explizit performativer Akt der Verhinderung der Turnübernahme durch den Dialogpartner, da der Sprecher sowohl das Tempo beschleunigt als auch darauf hinweist, daß er noch nicht fertig ist. In Beispiel 5 findet sich eine Parenthese mit TC, also ein 'klassischer' Fall der Tempoänderung. Parenthesen gibt es allerdings relativ selten im VERBMOBIL–Material, und daher ist diese Parenthese auch eher singulär[3]. Grundsätzlich läßt sich fürs erste keine klare Tendenz feststellen, daß TCs etwa oft bei rhematischen oder oft bei thematischen Passagen bzw. oft bei dialogrelevanten oder redundanten Passagen auftreten. Eine genaue Analyse steht allerdings noch aus.

[3]Man beachte, daß manchmal unterschiedliche Möglichkeiten der syntaktischen Interpretation und damit der M–Etikettierung bestehen, insbesondere in der Umgebung von außergrammatischen Passagen. So wäre eine Labelung mit M3I als Parenthese in Beispiel 4 ebenso möglich.

Kurzbeschreibung	M3-Grenzen	Oberklasse
Haupt-/Nebensatz	M3S	M3
freie Phrase, elliptischer Satz	M3P	M3
extraponierte Phrase	M3E	M3
eingebettete(r) Phrase/Satz	M3I	M3
Satzauftakts-/Satzabtaktspartikel mit <P>/<A>	M3T	M3
Satzauftakts-/Satzabtaktspartikel ohne <P>/<A>	M3D	MU
syntaktisch ambig	M3A	MU
Konstituente, möglicherweise prosodisch markiert	M2I	M0
Konstituente, wahrscheinlich prosodisch nicht markiert	M1I	M0
jedes andere Wort (default)	M0I	M0

Tabelle 6: Vollständige Liste aller in VERBMOBIL vergebenen syntaktisch-prosodischen M3-Grenzetiketten. Neben einer Kurzbeschreibung ist auch eine Abbildung der M3-Klassen auf drei Oberklassen angegeben. M3 bezeichnet dabei eine in der Regel prosodisch stark markierte syntaktische Grenze, M0 ist im allgemeinen prosodisch unmarkiert und die MU treten sowohl prosodisch markiert als auch prosodisch unmarkiert auf.

Ex. 1 *<Atmung> ja M3D ich denke schon M3S <Atmung> das klappt wohl alles wunderbar TCA M3S dann nehmen wir den Montag TCL M3S <Pause> und treffen uns dann morgens TCA in der Halle A <Atmung>*

Ex. 2 *ee <Pause> da ist bei mir schlecht M3A vom IZE <Pause> siebten IZE bis zum IZE sechzehnten Februar TCA M3A ist bei mir schlecht TCL M3S <Schmatzen> <Atmung> aber IZE <Pause> <Räuspern> dann IZE <ähm> <Pause> siebzehnter IZE bis dreiundzwanzigster TCA paßt bei mir immer gut <Klicken>*

Ex. 3 *gut M3D <Pause> dann treffen wir uns Dienstag M3A morgen M3A vielleicht M3A um neun TCA M3A gleich M3A zum Frühstücken*

Ex. 4 *wer hat deinen Terminkalender ausgearbeitet M3S <Atmung> am dri IWN TCA ich bin noch nicht fertig M3E am dritten ab fünfzehn Uhr*

Ex. 5 *<ähm> ich seh' grade M3S ich hätte noch 'ne andere Möglichkeit M3E und zwar <äh> zwischen Sonntag den dritten M3I das ist Ostersonntag TCA M3I das ist häßlich aber machbar TCL M3I <Atmung> bis Sonntag den zehnten IZE <Klicken>*

4.2.4 Automatische Klassifikation

Tabelle 7 zeigt das Ergebnis einer automatischen Klassifikation mit 421 akustisch-prosodischen Merkmalen, womit das beste Ergebnis beim Vergleich unterschiedlich großer Merkmalsätze erzielt werden konnte, vgl. Tabelle 8. Aus Aufwandsgründen wurden bei Training und Test nur die Turns zugrundegelegt, die mindestens einen TC aufweisen. Die Turns der CD2 dienten zum Trainieren (insg. 123 Turns), alle anderen Turns zum Testen (insg. 71 Turns). Man muß annehmen, daß es einen relativ großen Übergangsbereich zwischen TC und T0 gibt: T0 kann zur einen Seite hin also mit TCA und zur anderen Seite hin mit TCL verwechselt werden, dito TCA mit T0 bzw. TCL mit T0. Eine Verwechslung von TCA mit TCL – 'über T0 hinweg' – sollte eher nicht vorkommen; hier gibt es genau 9 Fälle (0.37% aller Wortgrenzen), die sich eher nicht erklären lassen und die sicher auch Kandidaten für Fehllabelungen sind. Der Mittelwert der klassenweisen Erkennungsraten von TCA und TCL ist 76.1%, die mittlere Fehlerrate für die weniger kritischen Verwechslungen von TCA/TCL mit T0 und vice versa 34.2%.

In Tabelle 8 sind mittlere klassenbedingte Erkennungsraten dargestellt, wobei für die Klassifikation jeweils unterschiedlich große Merkmalsätze über unterschiedlich lange zeitliche Kontexte berechnet wurden. Es

		TCA	TCL	T0
TCA	72	78	7	15
TCL	39	10	75	15
T0	2332	19	16	65

Tabelle 7: Erkennungsraten für Tempowechsel in Prozent

Kontext: # Silben	0	1	2	1	2	3	3	4	5	5	6
Kontext: # Wörter	0	0	0	1	1	1	2	2	2	3	3
# Merkmale	45	83	121	127	165	203	251	289	327	383	421
Erkennungsrate (%)	57	61	66	62	67	69	69	66	68	70	72

Tabelle 8: klassenweise ermittelte mittlere Erkennungsraten in Prozent in Abhängigkeit von der Zahl der Merkmale

zeigt sich deutlich, daß die Erkennungsrate im Schnitt mit der Zahl der Merkmale und mit der Größe des linken und rechten Kontexts wächst. Damit widerspiegelt sich in einem gewissen Ausmaß das Kriterium des Etikettierers, nur TC's zu etikettieren, die sich über mehrere Wörter erstrecken.

Eine genaue Analyse der Beiträge einzelner Merkmale bzw. einzelner Merkmalgruppen steht noch aus. Die Dauerwerte sollten z.B. relevanter für die Erkennung von TC vs. T0 sein, als dies bei der Klassifikation von Grenze vs. Nicht–Grenze der Fall ist.

5 Schlußbemerkungen

Unterschiedliches Sprechtempo sollte in VERBMOBIL modelliert werden, da sich damit die Worterkennung verbessern läßt. In diesem Fall muß für jeden Sprecher am Anfang eines Dialogs das Sprechtempo und damit die Zuordnung (z.B. bei zwei unterschiedlichen Worterkennern für langsames oder für schnelles Sprechtempo) ermittelt werden. Es dürfte also kein großer Aufwand sein, auch im weiteren Verlauf des Dialogs diese Sprechtempoermittlung mitlaufen zu lassen, um zum Beispiel auf geeignetere akustische Modelle umschalten oder Parametergewichte dynamisch anpassen zu können.

Eine andere Anwendung für die höheren Module scheint sich aber im Augenblick nicht anzubieten, da es sich um ein Phänomen handelt, das psycholinguistische Prozesse widerspiegelt, die kurz- und mittelfristig nicht Gegenstand einer automatischen Verarbeitung sein werden. Hinzu kommt die starke Sprecherspezifität und die geringe Zahl der Tokens über alle Sprecher im VERBMOBIL-Material.

Es muß für's erste dahingestellt bleiben, auf welche Gründe im einzelnen die geringe Zahl von TC's zurückzuführen ist. Die Etikettierung wurde nur teilweise intersubjektiv überprüft; es ist also möglich, daß der Etikettierer dazu tendiert, nur (über–) deutliche TC's wahrzunehmen. Vorteilhaft ist dabei allerdings, daß eine Etikettierung im unscharfen Übergangsbereich vermieden wird. Es ist ebenfalls möglich, daß eine andere Strategie, etwa der Einbezug von kürzeren TC–Passagen, sinnvoller ist. Es ist aber genauso möglich, daß das spezielle experimentelle Setting TC's nicht gerade begünstigt. Und schlußendlich kann dieses Ergebnis auch einfach richtig sein: Die untersuchte Art des Tempowechsels ist *nur eine* unter vielen unterschiedlichen – prosodischen und nicht–prosodischen – Möglichkeiten, bestimmte Funktionen zu erfüllen – und es muß nicht die häufigste sein.

Literatur

[Bat94] A. Batliner, S. Burger, and A. Kießling. Außergrammatische Phänomene in der Spontansprache:

Gegenstandsbereich, Beschreibung, Merkmalinventar. Verbmobil Report 57, 1994.

[Bat95] A. Batliner, A. Kießling, S. Burger, and E. Nöth. Filled Pauses in Spontaneous Speech. In *Proc. of the 13th Int. Congress of Phonetic Sciences*, volume 3, pages 472–475, Stockholm, 1995.

[Bat96] A. Batliner, R. Kompe, A. Kießling, M. Mast, and E. Nöth. All about Ms and Is, not to forget As, and a comparison with Bs and Ss and Ds. Towards a syntactic–prosodic labeling system for large spontaneous speech data bases. Verbmobil Memo 102, 1996.

[Cry82] T.H. Crystal and A.S. House. Segmental durations in connected–speech signal: Preliminary results. *Journal of the Acoustic Society of America*, 72:705–716, 1982.

[Cry88] T.H. Crystal and A.S. House. Segmental durations in connected–speech signal: Current results. *Journal of the Acoustic Society of America*, 83:1553–1573, 1988.

[Kie96a] A. Kießling. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Dissertation. Technische Fakultät der Universität Erlangen–Nürnberg, 1996. (erscheint).

[Kie96b] A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Classification of Boundaries and Accents in Spontaneous Speech. In R. Kuhn, editor, *Proc. of the CRIM / FORWISS Workshop*, Montreal, 1996. (erscheint).

[Kom95] R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic scoring of word hypotheses graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.

[Lev89] W. Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

[Mil84] J.L. Miller, F. Grosjean, and C. Lomanto. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41:215–225, 1984.

[Mir96] Nikki Mirghafori, Eric Fosler, and Nelson Morgan. Towards Robustness to Fast Speech in ASR. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 335–338, Atlanta, 1996.

[Not91] S.G. Noteboom. Some observations on the temporal organization and rhythm of speech. In *Proc. of the 12th Int. Congress of Phonetic Sciences*, volume 1, pages 228–237, Aix–en–Provence, 1991. Université de Provence.

[Rey94] M. Reyelt and A. Batliner. Ein Inventar prosodischer Etiketten für Verbmobil. Verbmobil Memo 33, 1994.

[Rey95] M. Reyelt. Ein System zur prosodischen Etikettierung von Spontansprache. In R. Hoffmann and R. Ose, editors, *Elektronische Sprachsignalverarbeitung*, volume 12 of *Studientexte zur Sprachkommunikation*, pages 167–174. TU Dresden, Wolfenbüttel, 1995.

[Tis93] Bernd Tischer. *Die vokale Kommunikation von Gefühlen*, volume 18 of *Fortschritte der psychologischen Forschung*. Psychologie Verlags Union, Weinheim, 1993.

[Wig92] C.W. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University, 1992.