Parametrische Modellierung von Dauer und Energie prosodischer Einheiten

Viktor Zeißler and **Elmar Nöth** and **Georg Stemmer**
Lehrstuhl für Mustererkennung, Friedrich-Alexander Universität, 91058 Erlangen
zeissler@immd5.informatik.uni-erlangen.de

Abstract

In dieser Arbeit wird untersucht, ob die empirische Dauer- bzw. Energieverteilung von Wörtern und ihren Untereinheiten mit einer parametrischen Verteilungsfunktion modelliert werden kann. Als mögliche Kandidaten für den Vergleich dienen Normal-, Log-Normal- und die Gamma-Verteilungen. Im Gegensatz zu der bisher verwendeten heuristischen Methode wird ein statistischer Ansatz zur Berechnung der satzbezogenen Werte des Sprechtempos und der mittleren Lautheit angewendet. Die Experimente, die auf der umfangreichen VERBMOBIL-Stichprobe durchgeführt werden, zeigen, dass der neue Ansatz in Bezug auf die *Likelihood*-Werte annäherend doppelt so gut ist, wie die bisherige Vorgehensweise.

1 Einleitung

Die Erkennung *prosodischer* Phänomene, wie z.B. Akzente, Phrasengrenzen, Tonverlauf usw. ist ein wichtiger Teilbereich der automatischen Sprachverarbeitung (Kießling, 1997). Im Rahmen des VERBMOBIL-Projekts wurde zur Bewältigung dieser Aufgabe ein *Prosodiemodul* erstellt (Batliner et al., 2000), das zur Zeit im Rahmen des SMARTKOM-Projekts[1] weiterentwickelt wird. Die Prosodieerkennung in diesem Modul basiert auf einem heterogenen Merkmalssatz, der mehrere prosodische Indikatoren wie Tonhöhe, Dehnung und Klangstärke berücksichtigt. In dieser Arbeit wird ein Ansatz untersucht, der die Berechnung der *Dauer*- und *Energiemerkmale* und damit auch die prosodische Klassifikation verbessern soll.

Wie in mehreren Untersuchungen betont wurde, werden diese Merkmale von verschiedenen Faktoren geprägt. Dazu gehören neben der für die Erkennung nützlichen satzinternen prosodischen Information auch solche globale sprecherabhängige Eigenschaften wie mittleres Sprechtempo und mittlere Lautheit (Anastasakos et al., 1995). Zusätzlich spielt der Einfluss der segmentalen Information (Mikroprosodie) eine große Rolle, wobei die Abhängigkeit von der Identität der zugrunde liegenden prosodischen Einheit an erster Stelle steht. Die genannten Faktoren behindern die Erkennung der satzinternen Prosodie und sollen nach Möglichkeit reduziert werden.

Um dieses Problem zu lösen, werden in (Kießling, 1997; Wightman, 1992) die *normierten* Dauer- und Energiemerkmale eingeführt, die zusammen mit den berechneten Mittelwerten für Sprechtempo und Lautheit zum Merkmalssatz hinzugefügt werden. Die zur Normierung benötigten satzbezogenen Sprechtempo- und Lautheitswerte werden heuristisch und ohne Verwendung statistischer Annahmen berechnet. In der vorliegenden Arbeit wird dagegen ein statistischer Ansatz verfolgt, der eine theoretisch fundierte Schätzung der unbekannten Parameter erlaubt. Dazu wird zunächst untersucht, mit welcher parametrischen Verteilungsfamilie die empirische Verteilung von Dauer- und Lautheitswerten am besten modelliert werden kann. Anschließend werden verschiedene Varianten der Parameterschätzung miteinander verglichen, wobei die bestpassende parametrische Verteilung als Grundlage für die eingesetzte Dauer- bzw. Energiemodellierung verwendet wird.

Die Berechnung des Sprechtempos und der

[1]Das dieser Arbeit zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (*BMBF*) im Rahmen des SMARTKOM-Projekts unter dem Förderkennzeichen 01 IL 905 K7 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

mittleren Lautheit kann wahlweise auf der Wort-, Silben- oder Lautebene erfolgen. Bei den in Abschnitt 4 beschriebenen Experimenten wird sie ausschließlich auf der Wortebene durchgeführt. Bei der Untersuchung von empirischen Verteilungen werden dagegen auch die Wortuntereinheiten einbezogen, da die entsprechenden Ergebnisse für andere Ansätze relevant sein können, wie beispielsweise für die explizite Dauermodellierung von HMM-Zuständen (Burshtein, 1995).

2 Merkmalsberechnung

Zu den in dieser Arbeit behandelten Merkmalen gehören absolute und normierte Dauer- bzw. Energiemerkmale sowie die satzbezogenen Sprechtempo- und Lautheitswerte. Da die Vorgehensweise bei der Dauer- und Energienormierung identisch ist, werden im weiteren nur die Ausdrücke für die Dauermerkmale angegeben. Eine Ausnahme stellen die absoluten Merkmale dar, auf deren Berechnung in Abschnitt 2.1 eingegangen wird.

Die Bestimmung der normierten Dauermerkmale d_i^{norm} anhand der absoluten Werte d_i erfolgt nach einer einfachen Normierungsvorschrift (Wightman, 1992):

$$d_i = \tau_d d_i^{\text{norm}} \quad \Rightarrow \quad d_i^{\text{norm}} = \frac{d_i}{\tau_d} \,, \qquad (1)$$

wobei τ_d das mittlere Sprechtempo bezeichnet. Man kann den Wert d_i^{norm} als unverzerrte Dauer verstehen, die die i-te Phraseneinheit ohne den Einfluss der geänderten Sprechgeschwindigkeit hätte.

2.1 Berechnung der absoluten Dauer- und Energiemerkmale

Die Berechnung der absoluten Merkmale erfordert zunächst eine Segmentierung des Sprachsignals mit einem Worterkenner. Die absoluten Dauerwerte werden anschließend extrahiert und in Millisekunden umgerechnet. Wegen der festeingestellten Fortschaltzeit beim verwendeten Spracherkenner sind sie auf 10 ms gerastert. Die Energiewerte für jeden 10 ms-langen Frame werden nach der in (Kießling, 1997) beschriebenen Methode bestimmt. Für die i-te Phraseneinheit, die aus n Frames besteht, werden zwei Energiemerkmale berechnet:

$$\varepsilon_i^{\text{word}} = \sum_{j=1}^{n} \varepsilon_j \,, \qquad \varepsilon_i^{\text{frame}} = \frac{\varepsilon_i^{\text{word}}}{n} \,. \qquad (2)$$

ε_j steht dabei für den Energiewert in dem j-ten Frame der gegebenen Phraseneinheit. $\varepsilon_i^{\text{word}}$ und $\varepsilon_i^{\text{frame}}$ werden entsprechend als *wortweise* bzw. *frameweise normierte* Energiemerkmale bezeichnet.

2.2 Statistisches Modell

Um die Schätzungsformeln für die τ_d-Werte abzuleiten, müssen hier einige Annahmen bezüglich der normierten Merkmale d_i^{norm} getroffen werden. Für jede Phraseneinheit mit der Identität i (beispielsweise für das i-te Wörterbucheintrag) soll folgendes gelten:

- Die normierten Merkmale gehören einer bestimmten parametrischen Verteilungsfamilie an, die mit der Dichtefunktion $p(d_i^{\text{norm}}) = p(d_i^{\text{norm}} \mid \theta)$ und dem Parametersatz θ eindeutig beschrieben wird.
- Die Verteilungsparameter θ hängen nur von der Identität der betrachteten prosodischen Einheit i ab: $\theta = \theta_i$.
- Alle anderen beeinflussenden Faktoren, unter anderem prosodische Information, werden ausschließlich durch die Varianz des Merkmals abgedeckt.

2.3 Bestimmung des Sprechtempos und der mittleren Lautheit

Zur Schätzung aller unbekannten Werte, also des Parametersatzes θ_i und der τ-Werte, wird der *Maximum-Likelihood* (ML) Ansatz angewendet. Wenn die Wahrscheinlichkeit der gesamten Folge aller N Merkmalswerte in der Stichprobe maximiert wird:

$$\mathcal{L} = \prod_{j=1}^{N} p\,(\frac{d_j}{\tau_{d_j}} \mid \theta_j) \;\rightarrow\; \max \,, \qquad (3)$$

bekommt man die optimalen Parameterschätzungen $\tilde{\theta}_j$ und $\tilde{\tau}_{d_j}$. Die geschätzten Werte von θ_i für alle Einheiten i werden in speziellen Statistiktabellen abgespeichert, die in einem *Trainingsprozess* erstellt werden. Bei einem Testlauf dagegen wird anhand dieser Tabellen nur der τ_d-Wert ermittelt. Für eine

Äußerung, die aus n Wörtern besteht, gilt dafür der folgende Ausdruck:

$$\tilde{\tau}_d = \underset{\tau_d}{\operatorname{argmax}} \sum_{j=1}^{n} \log p\left(\frac{d_j}{\tau_d} \mid \theta_j\right) , \qquad (4)$$

der für die verwendeten Verteilungen $p(x \mid \theta)$ geschlossen gelöst werden kann.

3 Parametrische Modellierung der empirischen Verteilungen

Die beschriebene Parameterschätzung erfordert eine Modellierung der empirischen Dauer- und Energieverteilungen durch die parametrischen Dichtefunktionen. Um die bestpassendste Verteilungsfamilie zu bestimmen, wurde die *Kullback-Leibler* (KL) Distanz zwischen den empirischen Verteilung $p(x)$ und einer Reihe bekannter parametrischen Verteilungen $q(x)$ berechnet:

$$D_{KL}(p,q) = \int_x p(x) \ln \frac{p(x)}{q(x)}\, dx \; . \qquad (5)$$

$D_{KL}(p,q) = 0$ gilt nur dann, wenn $p(x)$ und $q(x)$ identisch sind, ansonsten ist die KL-Distanz immer positiv. Ihre Größe spiegelt dabei den Unterschied zwischen den Verteilungen $p(x)$ und $q(x)$ wider.

Da bereits im Vorfeld festgestellt wurde, dass sowohl Dauer- als auch Energieverteilungen meistens unsymmetrisch sind, wurden zu diesem Test neben der symmetrischen Normalverteilung auch zwei unsymmetrische, die *Log-Normal-* und die *Gamma*-Verteilung ausgewählt. Die Gamma-Verteilungsdichte

$$p(x \mid \alpha, p) = \frac{\alpha^p}{\Gamma(p)} \exp\{-\alpha x\}\, x^{p-1} \qquad (6)$$

wurde bereits in (Burshtein, 1995) getestet und hat dort die beste Übereinstimmung mit der empirischen Dauerverteilung sowohl von Lauten als auch von Wörtern gezeigt. Die Dichtefunktion der Log-Normalverteilung

$$p(x \mid \mu_l, \sigma_l) = \frac{1}{x\sqrt{2\pi\sigma_l^2}} \exp\left\{-\frac{(\ln x - \mu_l)^2}{2\sigma_l^2}\right\} \qquad (7)$$

hat einen steileren Anstieg und stärkere Assymmetrie als die Gamma-Dichte, was sie besonders geeignet für die Modellierung der kürzeren prosodischen Einheiten macht.

Die Berechnung der empirischen Verteilungen erfolgte auf der *VERBMOBIL*-Stichprobe, die ca. 46 Stunden spontaner Sprache umfasst. In der Stichprobe sind insgesamt 8807 unterschiedliche Wörter, 2971 unterschiedliche Silben und 71 Laute enthalten. Für die drei berechneten Merkmalsgruppen (Dauermerkmale, frame- und wortweise normierten Energiemerkmale) gab es jeweils zwei Varianten: der absolute und der nach (1) normierte Wert. Als Phraseneinheiten wurden Wörter, Silben und Laute getestet. Es wurden dabei nur die Einheiten ausgewählt, die oft genug (mindestens 50 Mal) in der Stichprobe vorkommen und somit statistisch zuverlässige Aussagen erlauben. Für jede solche Einheit wurde ihre eigene empirische Verteilungsdichte und die entsprechenden KL-Distanzen zu den bestpassendsten parametrischen Dichten berechnet. Diese Distanzen wurden anschließend über alle Einheiten eines Types (wie z.B. Wörter) gemittelt. Ein Beispiel der absoluten Dauerverteilung für das Wort *'das'* und die entsprechenden parametrischen Dichten sind in der Abbildung 1 zu sehen.

Die Ergebnisse für alle betrachteten Fälle sind in der Tabelle 1 angegeben. Demnach können sowohl die Dauerverteilungen als auch die Verteilungen von wortweise normierten Energiemerkmalen am besten durch die Log-Normalverteilung modelliert werden. Die Verteilung der frameweise normierten Energiemerkmale, die von der Dauer der Phraseneinheit nicht beeinflusst wird, entspricht am nächsten der Gamma-Verteilung, wobei die Unterschiede insbesondere bei den normierten Werten extrem gering ausfallen. Bei den Wörtern und Silben sind die Merkmale annäherend normalverteilt. Die Verteilung der absoluten Werte dieser Merkmale kann neben der Gamma-Verteilung ebenso gut der Log-Normalverteilung entsprechen.

3.1 Abhängigkeit der Verteilungen vom Silbenakzent

In (Anastasakos et al., 1995) wird berichtet, dass die Lautdauerverteilungen unter anderem vom lexikalischen Akzent beeinflusst werden. Um nachzuprüfen, ob diese Unterschiede rein quantitativer Natur sind oder sich auch in der stark geänderten Form der Verteilung manifestieren, wurden einige Vergleichstests zwi-

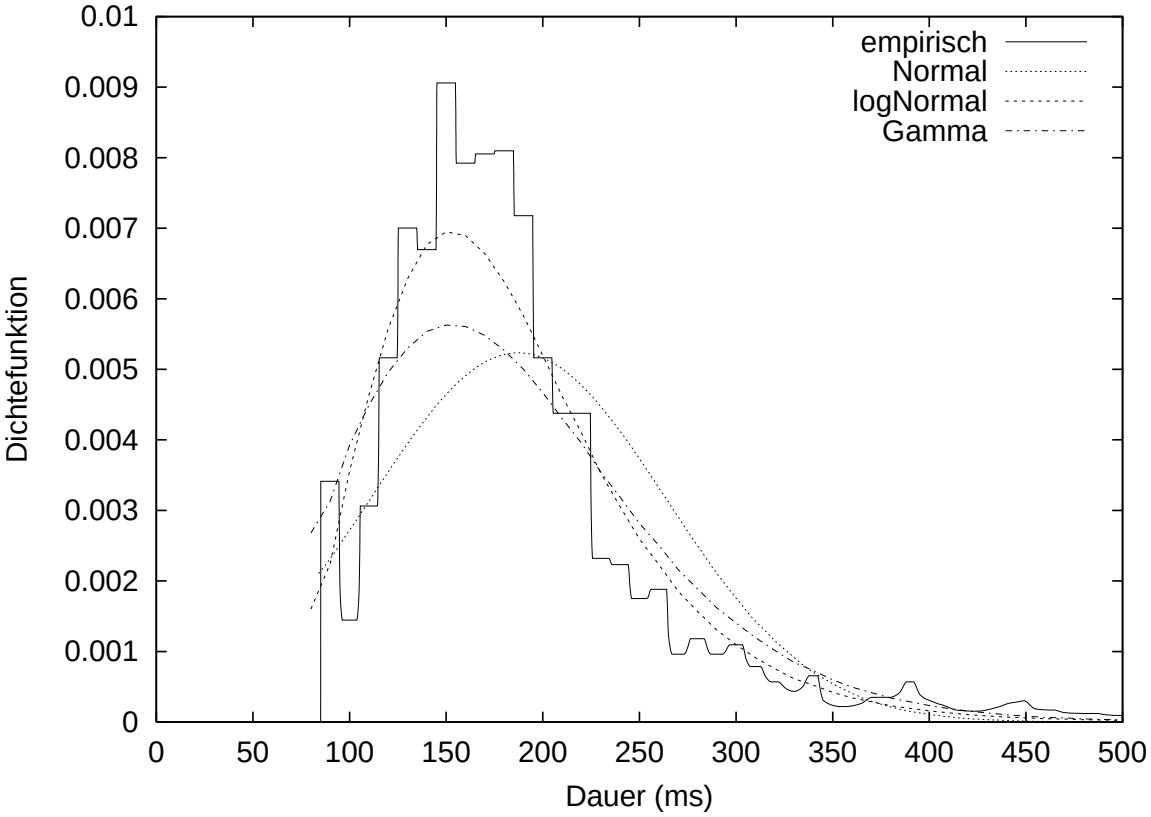

Abbildung 1: Empirische Dauerverteilungen für das Wort 'das' (durchgezogene Linie) und die angepassten parametrischen Verteilungsdichten: Normalverteilung (Punktlinie), LogNormalverteilung (geschtrichelte Linie), Gamma-Verteilung (strichpunktierte Linie)

schen akzentuierten und nicht akzentuierten Silben bzw. Vokalen durchgeführt. Eine prosodische Einheit wurde dabei als akzentuiert angenommen, falls sie den lexikalischen Wortakzent trägt, unabhängig davon, ob die entsprechende Wortausprägung in der betrachteten Äußerung prosodisch akzentuiert wurde. Bei diesen Experimenten wurde festgestellt, dass die Verteilungen aller getesteten Merkmale zwar zum Teil beträchtliche akzentbedingte Unterschiede im Mittelwert und der Varianz aufweisen, dennoch sind ihre Dichtefunktionen der Form nach ähnlich und gehören der gleichen parametrischen Verteilungsfamilie an. Diese Unterschiede können deutlich am Beispiel der Dauerverteilung vom Laut *'i:'* in Abbildung 2 beobachtet werden. Obwohl die dargestellten empirischen Verteilungen stark unterschiedlich sind, entsprechen sie in beiden Fällen der LogNormalverteilung.

4 Experimente zur Parameterschätzung

Wie bereits festgestellt, stimmt die Verteilung der Dauer- und der wortweise normierten Energiemerkmale $\varepsilon_i^{\text{word}}$ mit der LogNormalverteilung gut überein. Um zu prüfen, ob diese Annahme eine bessere Abschätzung der τ-Werte erlaubt, als die in (Kießling, 1997) beschriebene Vorgehensweise, wurden die logarithmierten Werte der *Likelihood*-Funktion nach dem Ausdruck (3) sowohl auf einer Trainingstichprobe (15647 Sätze) als auch auf einer separaten Teststichprobe (5000 Sätze) untersucht. Dabei wurden folgende vier Fälle unterschieden:

a. Keine τ-Schätzung. Das Sprechtempo und die mittlere Lautstärke werden auf den Wert 1.0 gesetzt.

b. Die τ_d-Werte werden nach der vorher verwendeten Methode als Mittelwert aller

Tabelle 1: KL-Distanzen für alle getesteten Merkmale, Verteilungsdichten und Phraseneinheiten

Dichte	Dauer		Energie ($\varepsilon^{\text{word}}$)		Energie ($\varepsilon^{\text{frame}}$)	
	abs.	norm.	abs.	norm.	abs.	norm.
	Wörter					
Normal	0.29	0.18	0.26	0.15	0.12	0.06
LogNormal	**0.12**	**0.08**	**0.06**	**0.07**	0.06	0.07
Gamma	0.17	0.10	0.09	0.07	**0.05**	**0.06**
	Silben					
Normal	0.31	0.19	0.27	0.16	0.11	0.04
LogNormal	**0.12**	**0.06**	**0.03**	**0.04**	0.04	0.05
Gamma	0.19	0.10	0.08	0.05	**0.03**	**0.03**
	Laute					
Normal	0.44	0.33	0.37	0.27	0.13	0.05
LogNormal	**0.17**	**0.10**	**0.03**	**0.04**	0.03	0.04
Gamma	0.26	0.17	0.10	0.08	**0.02**	**0.03**

wortbezogener Erwartungswerte τ_{d_j} in einem Satz berechnet (Kießling, 1997):

$$\tilde{\tau}_d = \frac{1}{n}\sum_{j=1}^{n}\mathcal{E}(\tau_{d_j}) = \frac{1}{n}\sum_{j=1}^{n}\frac{d_j}{\mu_j}\,, \qquad (8)$$

wobei n die Anzahl der Wörter in der Äußerung, d_j die Dauer des j-ten Wortes und μ_j die Dauermittelwert über alle Vorkommen des j-ten Wortes in der Trainingsstichprobe bezeichnet.

c. Die τ-Werte werden nach dem ML-Prinzip geschätzt (3):

$$\tilde{\tau}_d = \exp\left\{\frac{\sum_{j=1}^{n} 1 + \frac{\log d_j - \mu_{l_j}}{\sigma^2_{l_j}}}{\sum_{j=1}^{n}\frac{1}{\sigma^2_{l_j}}}\right\}\,, \qquad (9)$$

wobei μ_{l_j} und σ_{l_j} für jede j-te Wortinstatz auf den Trainingsdaten geschätzt werden. Für die Statistiken der oft vorkommenden Wörter wird dabei ebenfalls die ML-Schätzung verwendet. Bei den seltenen Wörtern werden μ und σ-Werte anhand der Silben- bzw. der Lautstatistik interpoliert.

d. Die τ-Werte werden wie in c. geschätzt. Bei der Berechnung von μ_l und σ_l Statistiken auf den Trainingsdaten werden die jeweiligen τ-Werte ebenfalls berücksichtigt. Die Optimierung der Parameter erfolgt mit einem iterativen EM^*-Algorithmus.

Die Hauptschwierigkeit im Umgang mit den absoluten Werten der Likelihood-Funktion besteht darin, dass keine vergleichbaren Referenzgrößen oder Extrema vorliegen, und die Ergebnisse deswegen nur in Relation zueinander interpretierbar sind. Um eine Aussage über den Nutzen der verwendeten Schätzungsverfahren zu erhalten, wird das Experiment *a.* als Baseline gewählt. Für alle anderen Fälle werden die Differenzen zur Baseline gebildet, die in der Tabelle 2 zu finden sind. Um den Vergleich zwischen den verschiedenen Stichproben zu ermöglichen, werden diese Werte zusätzlich mit der Anzahl der Wörter N in der Stichprobe normiert.

Die angeführten Ergebnisse zeigen, dass die ML-Schätzung von τ-Werten in *c.* etwa zwei- bis dreifach besser als die Standardmethode (in *b.*) im Bezug auf die log-Likelihood-Werte ist. Die iterative Optimierung auf der Trainingsstichprobe in *d.* bringt bei den Dauermerkmalen keinen weiteren Vorteil. Bei den Energiemerkmalen

Tabelle 2: Differenzen der gemittelten logarithmierten Likelihood-Werte für die getesteten Varianten

Merkmal	Training			Test		
	b.	c.	d.	b.	c.	d.
Energie	0.25	0.48	0.54	0.26	0.49	0.55
Dauer	0.06	0.17	0.18	0.05	0.18	0.18

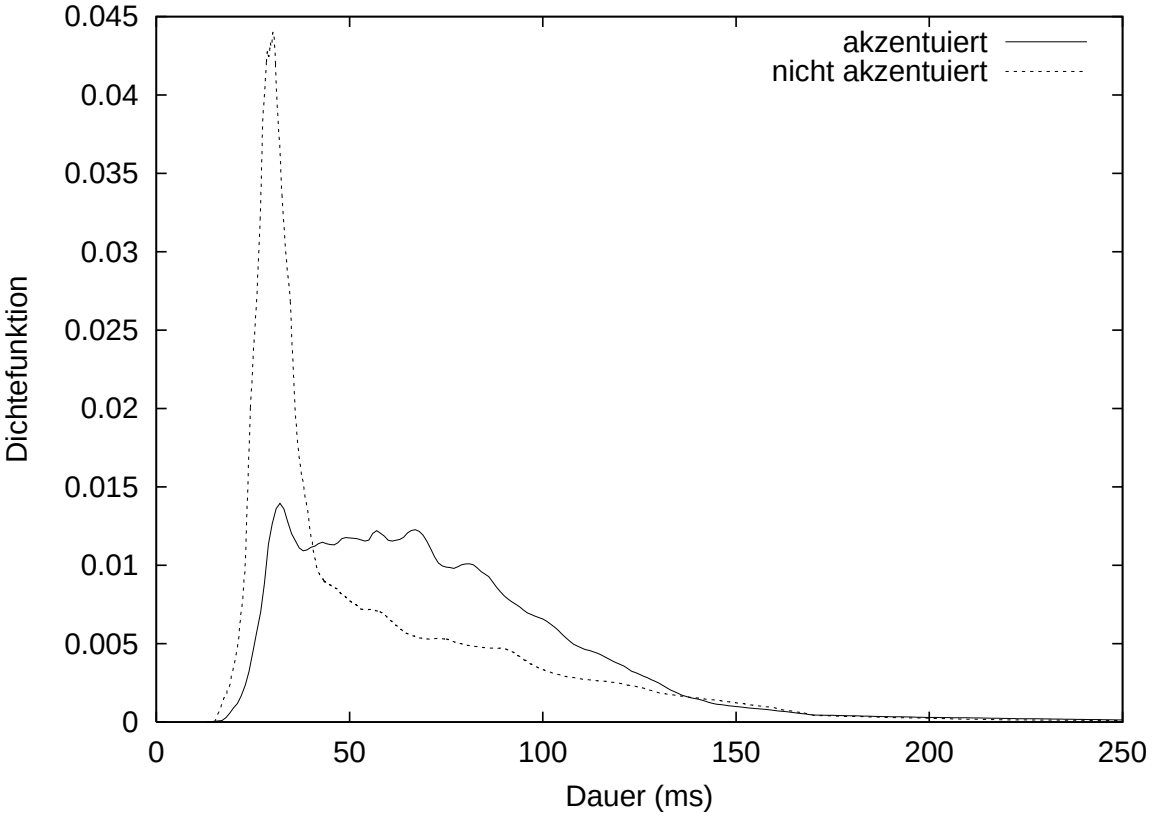

Abbildung 2: Dauerverteilungen vom Laut 'i:' mit Akzent (wie z.B. im Wort 'Termin') und ohne Akzent (wie in 'Juli')

erreicht man damit eine Verbesserung von log-Likelihood-Werten um weitere 10 %. Die Unterschiede zwischen der Test- und Trainingsstichprobe sind ebenfalls marginal, was auf eine gute Zuverlässigkeit der beim Training gewonnenen Statistiken deutet.

5 Ausblick

Die durchgeführten Experimente zeigen, dass die Bestimmung des Sprechtempos und der mittleren Lautheit durch die Annahme einer parametrischen Verteilungsform verbessert werden kann. Die Zuverlässigkeit dieser Schätzung übt einen wesentlichen Einfluss auf die Erkennung der prosodischen Phänomene aus. Die im Vorfeld durchgeführten Versuche haben bestätigt, dass der neue Ansatz zur Verbesserung der Prosodieerkennung beiträgt. Um jedoch eine quantitative Aussage zu treffen, müssen weitere Experimente zur Erkennung prosodischer Klassen, wie Phrasenakzente bzw. Phrasengrenzen durchgeführt werden.

Das in dieser Arbeit verwendete Modell für τ-Berechnung (1) stellt eine relativ grobe Annäherung an die Realität dar. Es wird darin implizit angenommen, dass alle Wörter vom Sprechtempo oder von der Satzlautstärke im gleichen Maß beeinflusst werden, was der empirischen Erfahrung widerspricht. Eine bessere Modellierung könnte durch die Verwendung komplexerer Annahmen für τ-Berechnung errreicht werden, z.B. wie in (Anastasakos et al., 1995) vorgeschlagen:

$$d_i = \tau_d^{\alpha_i} d_i^{\text{norm}} \ , \tag{10}$$

wobei α_i als Zusatzparameter anhand einer Trainingsstichprobe für jeden i-ten Worteintrag bestimmt werden kann.

Einen weiteren möglichen Untersuchungspunkt stellt die Bestimmung der optimalen Kontextbreite für die Berechnung der τ-Werte dar. Das ist besonders wichtig bei den langen Sätzen, wo sich die lokalen Werte des

Sprechtempos bzw. der Lautheit erheblich ändern können.

Nach der aktuellen Vorgehensweise werden die Statistiken für die selten vorkommenden Wörter nicht explizit berechnet, sondern anhand der Silben- bzw. Lautstatistiken interpoliert. Die Wahl der optimalen Interpolationsmethode stellt ebenfalls ein Problem dar, dessen Lösung eine erhöhte Zuverlässigkeit der Parameterschätzung und eine Reduzierung der benötigen Trainingsstichprobe verspricht.

References

A. Anastasakos, R. Schwartz, and H. Shu. 1995. Duration modeling in large vocabulary speech recognition. In *Proc. of the Int. Conf. on Acoustic, Speech, and Signal Processing, ICASSP'95*, pages 628–631.

A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. 2000. The Prosody Module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 106–121. Springer, New York, Berlin.

D. Burshtein. 1995. Robust parametric modelling of durations in hidden markov models. In *Proc. of the Int. Conf. on Acoustic, Speech, and Signal Processing, ICASSP'95*, pages 548–551.

A. Kießling. 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Shaker Verlag, Aachen.

C.W. Wightman. 1992. *Automatic Detection of Prosodic Constituents for Parsing.* Ph.D. thesis, Boston University Graduate School.