REGULIERBARE VERARBEITUNGSGESCHWINDIGKEIT AUTOMATISCHER SPRACHERKENNUNGSSYSTEME DURCH DYNAMISCHE SUCHRAUMEINSCHRÄNKUNG

Sascha Wendt, Gernot A. Fink, Franz Kummert

Universität Bielefeld, Technische Fakultät,
Postfach 100131, 33501 Bielefeld
{swendt,gernot,franz}@techfak.uni-bielefeld.de

Abstract: Bei der automatischen Spracherkennung müssen während des Erkennungsprozesses komplexe Zustandsräume durchsucht werden. Durch Einschränkung dieser Suchräume kann die Bearbeitungszeit deutlich reduziert werden, wobei jedoch i.A. auch die Wortakkuratheit (Erkennungsleistung) sinkt. Eine verstärkte Suchraumeinschränkung ist aber vor allem für zeitkritische Anwendungen, z.B. Interaktion mit agierenden Systemen oder mobilen Robotern, notwendig. Es gibt zwar eine Reihe von Verfahren zur Suchraumeinschränkung, die jedoch zumeist keine bestimmte Verarbeitungszeit garantieren können. Die meisten Verfahren werden über einen Parameter gesteuert, der i.d.R. heuristisch festgelegt werden muss und während des Erkennungsprozesses unverändert bleibt. Dadurch kann i.A. nur erreicht werden, dass die Spracherkennung in einem gewissen Zeitrahmen, nicht aber mit einer bestimmte Geschwindigkeit erfolgt. Außerdem ist der Zeitrahmen bei einem festgelegten Parameter stark ressourcenabhängig. Deshalb wurde ein Verfahren zur dynamischen Suchraumeinschränkung entwickelt, das während des Erkennungsprozesses die Parameter für verschiedene bekannte Verfahren durch Abschätzung der Verarbeitungsgeschwindigkeit und der Wortakkuratheit auf der Basis empirisch ermittelter Daten dynamisch optimiert. Dieses Verfahren ermöglicht es, Sprachsignale mit einer vorgegebenen Geschwindigkeit zu verarbeiten. Im Gegensatz zum Spracherkennungssystem ohne dynamische Suchraumeinschränkung, bei dem die Verarbeitungsgeschwindigkeit zwischen verschiedenen Äußerungen stark variiert, zeigt das erweiterte System eine annähernd konstante Verarbeitungsgeschwindigkeit, wobei die Wortakkuratheit bei gleicher Verarbeitungszeit nicht signifikant sinkt. Außerdem kann eine vorgegebenen Verarbeitungsgeschwindigkeit unabhängig von den verfügbaren Ressourcen eingehalten werden. Darüber hinaus kann das erweiterte System jederzeit so beschleunigt werden, dass eine automatische Spracherkennung in Echtzeit erfolgt, wodurch z.B. eine reflexartige Reaktion von agierenden Systemen auf sprachliche Interventionen ermöglicht wird.

1 Einleitung

Bei der Entwicklung von automatischen Spracherkennungssystemen wird das Hauptaugenmerk i.A. auf eine robuste Erkennung mit einer möglichst geringen Fehlerrate gelegt, während die Erkennungsgeschwindigkeit meist nur eine untergeordnete Rolle spielt. Für zeitkritische Anwendung, z.B. bei der Interaktion mit agierenden Systemen bzw. mobilen Robotern oder in einer

[1] Die vorliegende Arbeit wurde im Rahmen des Sonderforschungsbereichs 360 "Situierte künstliche Kommunikatoren" von der Deutschen Forschungsgemeinschaft (DFG) gefördert.

In R. Hoffmann, editor, Elektronische Sprachsignalverarbeitung, volume 24 of Studientexte zur Sprachkommunikation, pages 96-103, Dresden, 2002.

virtuellen Umgebung, kann die Geschwindigkeit jedoch sehr wichtig sein. Hier sollte grundsätzlich eine unnatürlich lange Verzögerung bei der Spracherkennung vermieden werden. Steht ein agierendes System darüber hinaus z.B. kurz davor, einen möglicherweise folgenschweren Fehler zu machen, sollten entsprechende sprachliche Interventionen von Seiten des Benutzers sogar schnellstmöglich, d.h. in Echtzeit, erkannt werden.

Eine entsprechend schnelle Spracherkennung kann durch verschiedene Methoden erreicht werden. Eine Möglichkeit besteht in der Entwicklung effizienter Algorithmen, die ein System deutlich beschleunigen können. Dadurch kann jedoch i.A. keine bestimmte Verarbeitungsgeschwindigkeit garantiert werden. Zum einen ist die Geschwindigkeit immer auch von den verfügbaren Ressourcen abhängig. Außerdem kann der Rechenaufwand bei der automatischen Spracherkennung stark schwanken, da sich die Größe der entstehenden Suchräume häufig ändert.

Eine andere Möglichkeit besteht darin, die beim Erkennungsprozess entstehenden Suchräume in ihrer Größe zu beschränken. In der Literatur sind dazu verschiedene Verfahren zu finden. Eine Einschränkung der Suchräume führt i.d.R. jedoch auch zu einer Zunahme der Wortfehlerrate. Außerdem können die meisten Verfahren keine bestimmte Verarbeitungszeit garantieren. Deshalb wurde ein Verfahren zur dynamischen Suchraumeinschränkung entwickelt. Dieses Verfahren passt die Parameter für verschiedene Verfahren zur Suchraumeinschränkung dynamisch an, um zum einen eine vorgegebene Verarbeitungsgeschwindigkeit einhalten zu können, und zum anderen den Verlust an Wortakkuratheit so gering wie möglich zu halten.

2 Allgemeine Verfahren zur Suchraumeinschränkung

In der Literatur gibt es eine Reihe von Verfahren zur Suchraumeinschränkung, von denen einige im Folgenden beschrieben werden.

Ein bekanntes und häufig eingesetztes Verfahren basiert auf dem sogenannte *Beamsearch* [1]: Innerhalb des Suchraums werden nur die Pfade weiter verfolgt, deren Bewertung nur um einen bestimmten Wert (dem Schwellwert) von der Bewertung des aktuell besten Pfades abweicht. Dieser Schwellwert wird über einen Parameter gesteuert, der heuristisch festgelegt werden muss und i.d.R. während des gesamten Erkennungsprozesses konstant bleibt. Dadurch kann dieses Verfahren keine bestimmte Verarbeitungsgeschwindigkeit garantieren, da im ungünstigsten Fall alle Bewertungen sehr dicht beieinander liegen und somit keine Einschränkung des Suchraums erfolgt. Das Verfahren kann an verschiedenen Stellen innerhalb eines Spracherkennungssystems eingesetzt werden, z.B. auf der Ebene des akustischen Modells und des Sprachmodells (siehe dazu z.B. [2]) oder auch der Mischverteilungen.

Einen ähnlichen Ansatz wie das vorherige Verfahren verfolgt das sogenannte *Histogramm-Pruning*, bei dem die Anzahl der weiter zu verfolgenden Pfade auf ein Maximum beschränkt wird [3]. Im Gegensatz zum Beamsearch kann dieses Verfahren durch die Festlegung einer maximalen Anzahl von Suchpfaden zumindest garantieren, dass eine bestimmte Verarbeitungszeit nicht überschritten wird. Allerdings darf dieses Maximum nicht zu klein gewählt werden, um eine robuste Erkennung zu gewährleisten. Dies kann wiederum zur Folge haben, dass die Anzahl der zu verfolgenden Pfade das Maximum zeitweise unterschreitet und damit der Suchraum kleiner wird als notwendig, also die verfügbare Rechenzeit nicht optimal genutzt wird. Somit stellt auch dieses Verfahren für sich alleine kein optimales Verfahren zur Berücksichtigung von Zeitrestriktionen dar.

Eine weitere Möglichkeit zur Suchraumeinschränkung besteht in der Zusammenfassung benachbarter Suchbäume (*merging of adjacent search trees*, siehe [4]): Suchbäume, die zu einem bestimmten Zeitpunkt Δt im selben Zustand enden und deren Startpunkt innerhalb eines kleinen Zeitfensters vor dem Zeitpunkt Δt liegt, werden zu einer Gruppe zusammengefasst. Innerhalb dieser Gruppe wird dann nur noch der am besten bewertete Pfad weiter verfolgt. Je größer das

Zeitfenster gewählt wird, desto mehr Suchbäume werden zusammengefasst und desto weniger Pfade müssen weiter verfolgt werden, aber desto größer ist auch die Wahrscheinlichkeit, den letztlich besten Pfad (also die optimale Hypothese) zu verlieren. Auch dieses Verfahren kann jedoch keine bestimmte Verarbeitungsgeschwindigkeit garantieren.

3 Dynamische Suchraumeinschränkung

Die beschriebenen Verfahren lassen sich alle über jeweils einen Parameter (Schwellwert oder Länge des Zeitfenster) steuern. Durch Variierung der Parameter kann die Größe des Suchraums und damit die benötigte Rechenzeit verändert werden. Die Idee besteht nun darin, die Parameter und damit den Suchraum während des Erkennungsprozesses dynamisch anzupassen, um den Erkennungsprozess mit einer vorgegebenen Geschwindigkeit durchführen zu können und dabei die Erkennungsleistung (also die Wortakkuratheit) so hoch wie möglich zu halten.

3.1 Experimente

Zunächst wurden eine Reihe von Experimenten auf dem deutschen Verbmobil-Datensatz [5] durchgeführt, um den Einfluss der verschiedenen Parameter auf die Rechenzeit und die Wortakkuratheit zu bestimmen. Als Ausgangspunkt wurden dabei die Parameter gewählt, die standardmäßig für das Spracherkennungssystem verwendet werden. Diese Parameter (zur Festlegung der Schwellwerte für die Mischverteilungen und für das akustische Modell sowie die Länge des Zeitfenster für die Zusammenfassung benachbarter Suchbäume) wurden dann schrittweise erhöht bzw. erniedrigt. Das Ziel dabei ist es, einen Suchraum zu schaffen, in dem die optimale Parameterkombination für eine bestimmte Geschwindigkeitsanforderung zu finden ist.

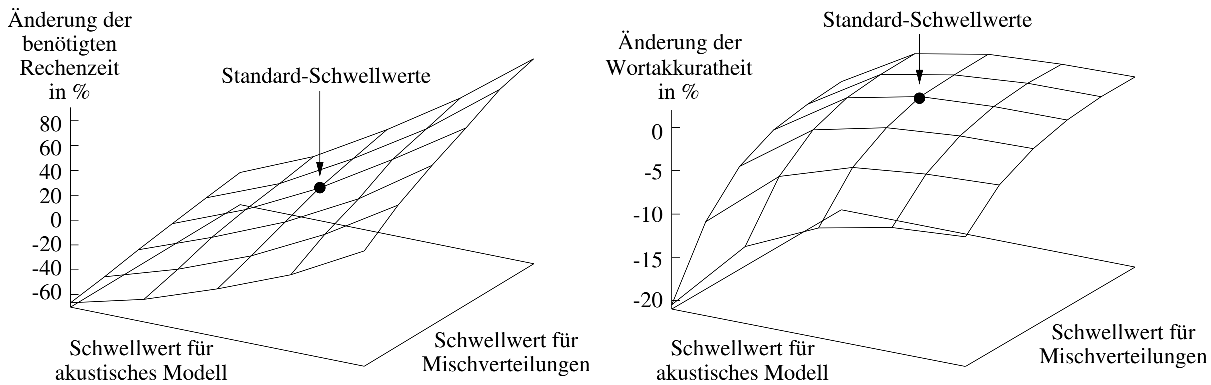

Abbildung 1 - Relative Änderung der Rechenzeit (links) und der Wortakkuratheit (rechts) in Abhängigkeit von verschiedenen Schwellwerten, ausgehend von den Standard-Schwellwerten.

In Abbildung 1 sind die Berechnungszeiten und Wortakkuratheiten (bzw. deren Änderungen gegenüber den Standard-Schwellwerten) bei verschiedenen Schwellwerten für die Mischverteilungen und für das akustische Modell dargestellt.[2]

Wie zu erwarten war, führen niedrigere Schwellwerte zu einer schnelleren Berechnung, allerdings nimmt auch die Wortakkuratheit ab. Auf der anderen Seite nimmt die Rechenzeit bei höheren Schwellwerten stark zu, wobei die Wortakkuratheit nicht mehr nennenswert erhöht wird (die Standard-Schwellwerte bilden also die näherungsweise optimalen Schwellwerte, wenn keine Zeitrestriktionen vorliegen).

[2] Alle Experimente wurden auf einer Compaq Alpha EV6 mit 500 MHz durchgeführt.

3.2 Adaption der Suchraum-Parameter

Basierend auf den beschriebenen Experimenten, die zur Schätzung der Rechenzeiten und Wortakkuratheiten bei bestimmten Parameterkombinationen dienen, wurde ein Mechanismus entwickelt, um die verschiedenen Parameter dynamisch zu adaptieren.

Zuerst wurde ein Gewichtungsparameter eingeführt, um die durch die Experimente erhaltenen Rechenzeiten und Wortakkuratheiten gewichtet zusammenfassen zu können. Jede Parameterkombination erhält auf der Basis des Gewichtungsparameters eine Bewertung, so dass die in Abbildung 1 dargestellten Suchräume zu einem Suchraum zusammengefasst werden, dessen Form vom Gewichtungsparameter abhängt.

Die Bewertung Q für die verschiedenen Parameterkombinationen wird mit Hilfe des Gewichtungsparameters α (Wert zwischen 0 und 1) wie folgt berechnet:

$$Q = \alpha \cdot \Delta T - (1 - \alpha) \cdot \Delta WA \cdot \beta \tag{1}$$

wobei ΔT die Änderung der Rechenzeit und ΔWA die Änderung der Wortakkuratheit gegenüber den Standard-Parametern ist. Der Faktor β (wird auf 20 festgesetzt) wird benötigt, um der Wortakkuratheit einen stärkeren Einfluss zu geben, da vor allem im Bereich um die Standard-Parameter die Änderung der Wortakkuratheit im Vergleich zur Änderung der Rechenleistung sehr viel geringer ausfällt (siehe Abbildung 1).

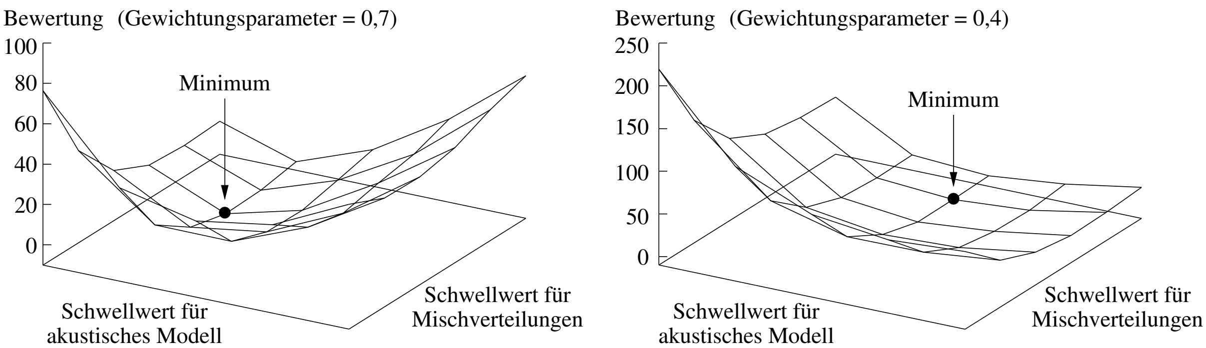

Abbildung 2 - Kombinierter Schwellwert-Suchraum für einen Gewichtungsparameter von 0,7 (links) bzw. 0,4 (rechts) mit dem jeweiligen Minimum.

Durch die Bewertungen kann nun ein kombinierter Suchraum erzeugt werden, dessen Form vom Gewichtungsparameter α abhängt (siehe Abbildung 2). Je größer α gewählt wird, desto stärker wird die Rechenzeit gewichtet und umgekehrt. Wird die Bewertung in Gleichung (1) minimiert, also das Minimum innerhalb des Suchraums ermittelt (siehe Abbildung 2), so lässt sich über den Gewichtungsparameter festlegen, ob Suchraum-Parameter mit eher geringer Rechenzeit oder eher hoher Wortakkuratheit verwendet werden. Durch dieses Vorgehen wird auch erreicht, dass im Fall von zwei Parameterkombinationen mit gleicher geschätzter Rechenzeit die Kombination gewählt wird, die die höhere geschätzte Wortakkuratheit erzielt, da diese Kombination eine niedrigere Bewertung erhält und somit eher das Minimum im Suchraum bildet.

Um nun die Parameter während des Erkennungsprozesses dynamisch anpassen zu können, wird laufend eine Zeitmessung vorgenommen, um festzustellen, ob die Zeitrestriktion eingehalten wird. Ergibt diese Zeitmessung, dass das Spracherkennungssystem zu langsam ist, wird der Gewichtungsparameter α erhöht. Dadurch wird der Suchraum für die möglichen Parameterkombinationen verändert (siehe Abbildung 2) und die Minimierung von Gleichung (1) liefert einen Parametersatz mit einer geringeren Rechenzeit, so dass der Erkennungsprozess beschleunigt wird. Ist umgekehrt das System schneller als gewünscht oder notwendig, wird der Gewichtungsparameter α gesenkt, um eine Parameterkombination mit einer höheren zu erwartenden Wortakkuratheit zu verwenden.

Um die dynamische Adaption des Gewichtungsparameters α zu erreichen, werden folgende Berechnungen vorgenommen:

- Zuerst wird die aktuelle Zeitdifferenz Δt zwischen der benötigten Rechenzeit t_r und der angestrebten Rechenzeit t_d berechnet:

$$\Delta t = t_r - t_d \tag{2}$$

- Diese Zeitdifferenz Δt bestimmt dann wie folgt die Änderung des Gewichtungsparameters α:

$$\Delta\alpha = \gamma \cdot (\Delta t - \Delta t_{old}) \tag{3}$$

  wobei γ eine Art Federkonstante und Δt_{old} die vorherige Zeitdifferenz ist, d.h. die Zeitdifferenz, die zum letzten Adaptionszeitpunkt berechnet wurde.

Die Gleichung (3) wird jedoch nur dann angewendet, wenn die aktuelle Zeitdifferenz

1. positiv und größer als die vorherige Zeitdifferenz oder
2. negativ und kleiner als die vorherige Zeitdifferenz ist.

Die Gründe für diese Bedingungen sind die folgenden:

a) Wenn die benötigte Rechenzeit größer (bzw. kleiner) als die angestrebte Rechenzeit, d.h. die aktuelle Zeitdifferenz positiv (bzw. negativ) ist, sollte, wie erläutert, der Gewichtungsparameter erhöht (bzw. gesenkt) werden.

b) Ist allerdings die aktuelle Zeitdifferenz zwar positiv, aber kleiner (bzw. negativ, aber größer) als die letzte Zeitdifferenz, ist das System bereits beschleunigt (bzw. verlangsamt) worden, und eine Adaption des Gewichtungsparameters α ist nicht notwendig.

Durch diesen Adaptionsmechanismus ist es möglich, den Gewichtungsparameter α in Abhängigkeit von der aktuellen Verarbeitungsgeschwindigkeit so anzupassen, dass der Erkennungsprozess, wenn notwendig, entsprechend beschleunigt oder verlangsamt wird. Durch die Konstante γ in Gleichung (3) kann außerdem festgelegt werden, wie schnell das System auf etwaige Zeitdifferenzen reagieren soll: Je kleiner die Konstante gewählt wird, desto geringer fällt die Änderung des Gewichtungsparameters in Gleichung (3) aus und desto langsamer reagiert das System damit auf Zeitdifferenzen.

Schließlich wurde das Spracherkennungssystem mit dem beschriebenen Verfahren der dynamischen Suchraumeinschränkung erweitert. Die Zeitrestriktion und damit die angestrebte Rechenzeit für Gleichung (2) wird durch einen Echtzeitfaktor (real time factor - RTF) angegeben, den das System einhalten soll:

$$\text{RTF} = \frac{\text{Rechenzeit}}{\text{Länge des verarbeiteten Sprachsignals}} \tag{4}$$

Zur Berechnung der Verarbeitungsgeschwindigkeit erfolgt eine Zeitmessung nach jedem zehnten Frame (die Framelänge beträgt 10 ms). Um die Parameterkombination für die Suchraumeinschränkung anzupassen, wird zuerst die aktuelle Zeitdifferenz Δt in Gleichungen (2) berechnet. Sind die Bedingungen für eine Änderung des Gewichtungsparameters α erfüllt, wird anschließend die Änderung nach Gleichung (3) berechnet und für den neuen Gewichtungsparameter die Bewertung in Gleichung (1) durch Gradientenabstieg, ausgehend von der aktuellen Parameterkombination, minimiert (d.h. es wird das Minimum innerhalb des neuen kombinierten Schwellwert-Suchraums berechnet). Die Parameterkombination mit der niedrigsten Bewertungszahl bildet dann die Suchraum-Schwellwerte für den weiteren Erkennungsprozess.

4 Evaluation

Das mit der dynamischen Suchraumeinschränkung erweiterte Spracherkennungssystem wurde auf zwei verschiedenen Datensätzen evaluiert.

Da die Änderungen von Rechenzeit und Wortakkuratheit auf dem deutschen *Verbmobil*-Datensatz gemessen wurden (siehe Abschnitt 3.1), erfolgte die erste Evaluation auch auf diesem Korpus. Der Datensatz besteht aus über 32 Stunden spontan gesprochenem Trainingsmaterial mit einem Wortschatz von 5336 Wortformen sowie etwa 35 Minuten Testmaterial [5].

Daneben erfolgte eine Evaluierung auf dem englischen *Wall-Street-Journal*-Datensatz (WSJ0). Hier stehen ca. 15 Stunden Trainings- und etwa 40 Minuten Testmaterial (vorgelesene Sätze aus der gleichnamigen Wirtschaftszeitung) zur Verfügung, der Wortschatz[3] für die Evaluierung umfasst 4986 Wortformen [6]. Das Verfahren der dynamischen Suchraumeinschränkung verwendet dabei die auf dem Verbmobil-Datensatz geschätzten Änderungen von Rechenzeit und Wortakkuratheit.

Für das Spracherkennungssystem werden 39 Merkmale verwendet, die sich aus der Gesamtenergie und den ersten 12 Cepstralkoeffizienten mit Vorwärtsmaskierung [7] sowie den ersten und zweiten Ableitungen dieser 13 Merkmale zusammensetzen. Die Klassifikation erfolgt auf der Basis von semikontinuierlichen HMMs mit linearer Struktur und variabler Zustandszahl. Als Wortuntereinheiten werden Triphone ohne Generalisierung, aber mit automatischem Zustandsclustering eingesetzt. Außerdem werden Bi-Gramm-Sprachmodelle (Perplexität: 64.4 auf Verbmobil, 109.6 auf WSJ0) verwendet. Eine detailliertere Beschreibung des Spracherkennungssystems ist in [8] zu finden.

Jede Evaluation wurde neunmal durchgeführt (jeweils dreimal auf drei verschiedenen Rechnern der gleichen Leistungsklasse, Compaq Alpha EV6 mit 500 MHz), um eine solide Abschätzung der Rechenzeiten (angegeben als Echtzeitfaktor RTF, siehe Gleichung (4)) und Wortakkuratheiten zu erhalten.

Bei den Evaluationen wurde neben dem Echtzeitfaktor für das gesamte Testmaterial (Gesamt-RTF) auch der individuelle Echtzeitfaktor für jede Äußerung berechnet, so dass die Varianz des Echtzeitfaktors (RTF-Varianz) innerhalb des bearbeiteten Testmaterials berechnet werden kann.

Verbmobil			
System mit	Gesamt-RTF	RTF-Varianz	Wortakkuratheit
Standard-Parametern	2,0	0,17	75,91 %
dynamischer Anpassung der Parameter	2,0	0,02	75,71 %
	1,5	0,02	74,12 %
	1,0	0,01	69,95 %

Tabelle 1 - Wortakkuratheiten des Systems mit dynamischer Suchraumeinschränkung unter verschiedenen Zeitrestriktionen (entsprechend dem Gesamt-RTF) auf Verbmobil. Die erste Zeile zeigt die Leistung des Systems mit Standard-Parametern.

In Tabelle 1 sind die Ergebnisse auf dem Verbmobil-Datensatz angegeben. Ohne dynamische Suchraumeinschränkung (also mit den Standard-Parametern) liegt der mittlere Echtzeitfaktor bei 2. Allerdings werden tatsächlich nur 13 % der Äußerungen mit einem Echtzeitfaktor von 2 und 39 % mit einem Echtzeitfaktor zwischen 1,9 und 2,1 verarbeitet, der Echtzeitfaktor für die verbleibenden 48 % weicht noch stärker vom durchschnittlichen Echtzeitfaktor ab. Das bedeutet, dass die Verarbeitungszeit bei Verwendung der Standard-Parameter sehr stark variiert.

[3] Die phonotypischen Transkriptionen des verwendeten Wortschatzes wurden auf der Basis des “Carnegie Mellon Pronouncing Dictionary” Version 0.6 erzeugt.

Wird dagegen die dynamische Suchraumeinschränkung verwendet und ein Echtzeitfaktor von 2 vorgegeben (d.h. die Verarbeitungsgeschwindigkeit darf der doppelten zeitlichen Länge der Äußerungen entsprechen), werden fast 63 % der Äußerungen mit einem Echtzeitfaktor von 2 und 90 % mit einem Echtzeitfaktor zwischen 1,9 und 2,1 verarbeitet. Die Schwankung in der Verarbeitungsgeschwindigkeit wird also erheblich reduziert.

Der geringe, jedoch nicht signifikante Verlust an Wortakkuratheit gegenüber dem System mit Standard-Parametern kommt dadurch zustande, dass der Suchraum während der Erkennungsprozesses manchmal stark eingeschränkt werden muss, um die vorgegebene Verarbeitungsgeschwindigkeit einhalten zu können. Der dabei auftretende Verlust an Wortakkuratheit (siehe auch Abbildung 1) kann jedoch nicht vollständig dadurch kompensiert werden, dass die dynamisch angepassten Parameter zwischenzeitlich einen größeren Suchraum zulassen als die Standard-Parameter, da ein größerer Suchraum als unter Standard-Parametern nur eine sehr geringe Verbesserung der Wortakkuratheit bewirkt (in Abschnitt 3.1 wurde bereits erläutert, dass die Standard-Parameter näherungsweise optimale Parameter darstellen).

Mit Hilfe der dynamische Suchraumeinschränkung ist das Spracherkennungssystem nun auch in der Lage, Eingaben in Echtzeit (RTF 1), also (annähernd) ohne zeitliche Verzögerung zu erkennen (siehe Tabelle 1, nur 6 % der Äußerungen wurden mit einem Echtzeitfaktor größer als 1,1 verarbeitet). Der relative Verlust an Wortakkuratheit gegenüber dem System mit Standard-Parametern, das die doppelte Verarbeitungszeit benötigt, beträgt dabei nur 7,9 %.

Um die auf dem Verbmobil-Datensatz erzielten Ergebnisse zu verifizieren, wurden zusätzlich Experimente auf dem Wall-Street-Journal-Datensatz durchgeführt. Dabei wurden die auf dem Verbmobil-Datensatz geschätzten Änderungen der Rechenzeit und Wortakkuratheit, wie sie in Abbildung 1 dargestellt sind, verwendet.

Wall-Street-Journal			
System mit	Gesamt-RTF	RTF-Varianz	Wortakkuratheit
Standard-Parametern	3.2	0.10	87.91 %
schnellere CPU	2.7	0.09	87.91 %
dynamischer Anpassung der Parameter	3.2	< 0.01	87.67 %
	2.0	< 0.01	85.31 %
	1.0	< 0.01	77.67 %
schnellere CPU	1.0	< 0.01	80.93 %

Tabelle 2 - Wortakkuratheiten des Systems mit dynamischer Suchraumeinschränkung unter verschiedenen Zeitrestriktionen (entsprechend dem Gesamt-RTF) auf Wall-Street-Journal. Die ersten beiden Zeilen zeigen die Leistung des Systems mit Standard-Parametern.

Die Ergebnisse in Tabelle 2 zeigen, dass das Spracherkennungssystem mit dynamischer Suchraumeinschränkung auch auf dem Wall-Street-Journal-Datensatz eingesetzt werden kann, ohne die zeitaufwendigen Experimente zur Schätzung der Änderung von Rechenzeit und Wortakkuratheit auf diesem Datensatz neu durchführen zu müssen. Da der Verlust an Wortakkuratheit bei einer Erkennung in Echtzeit (RTF 1) auf diesem Datensatz doch recht deutlich ausfällt (allerdings wird dabei das System immerhin um den Faktor 3,2 gegenüber den Standard-Parameter beschleunigt), wurde eine zusätzliche Evaluation auf einem leistungsstärkeren Rechner (schnellere CPU, Compaq Alpha EV6.7 mit 667 MHz) durchgeführt. Der dadurch verringerte Verlust an Wortakkuratheit unter Echtzeitbedingung (siehe letzten Zeile in Tabelle 2) zeigt, dass das Verfahren zur dynamischen Suchraumeinschränkung automatisch von der erhöhte Rechenleistung profitiert. Außerdem macht das Ergebnis deutlich, dass das Verfahren die vorgegebenen Zeitrestriktionen unabhängig von der zur Verfügung stehenden Rechenleistung einhalten kann.

5 Zusammenfassung

Mit Hilfe des vorgestellten Verfahrens der dynamischen Suchraumeinschränkung ist es möglich, Spracheingaben mit einer vorgegebenen Geschwindigkeit zur verarbeiten. Dieses Verfahren passt während des Erkennungsprozesses dynamisch die Parameter für verschiedene Verfahren zur Suchraumeinschränkung an, um einen vorgegebenen Echtzeitfaktor einzuhalten. Außerdem werden die Suchraum-Parameter im Hinblick auf die zu erwartende Wortakkuratheit optimiert. In diesem Beitrag wurden dabei nur die Parameter für einen Beamsearch auf der Ebene des akustischen Modells und der Mischverteilungen sowie zur Zusammenfassung von benachbarten Suchbäumen angepasst. Grundsätzlich kann dieses Verfahren bei der Optimierung jedoch alle Verfahren zur Suchraumeinschränkungen berücksichtigen, die sich durch veränderbare Parameter steuern lassen.

Die Evaluation des mit dem Verfahren erweiterten Spracherkennungssytems hat gezeigt, dass durch die dynamische Suchraumeinschränkung vorgegebene Zeitrestriktionen zuverlässig eingehalten werden können. Gegenüber dem System mit statischen Parametern wird die Varianz der Verarbeitungszeit deutlich reduziert, wobei nur ein minimaler, jedoch nicht signifikanter Verlust an Wortakkuratheit eintritt. Außerdem können Zeitrestriktionen unabhängig von der zur Verfügung stehenden Rechenleistung (zumindest in einem bestimmten Rahmen) eingehalten werden. Darüber hinaus kann das erweiterte System jederzeit so beschleunigt werden, dass eine Verarbeitung in Echtzeit, also ohne zeitlich Verzögerung, erfolgt.

Literatur

[1] B. T. Lowerre: *The HARPY Speech Recognition System*, Dissertation, Carnegie-Mellon University, Department of Computer Science, Pittsburg, 1976.

[2] H. Ney, S. Ortmanns: *Dynamic Programming Search for Continuous Speech Recognition*, *IEEE Signal Processing Magazine*, Bd. 16, Nr. 5, 1999, S. 64–83.

[3] V. Steinbiss, B.-H. Tran, H. Ney: *Improvements in Beam Search*, in *Proc. Int. Conf. on Spoken Language Processing*, Bd. 4, Yokohama, 1994, S. 2143–2146.

[4] G. A. Fink, C. Schillo, F. Kummert, G. Sagerer: *Incremental Speech Recognition for Multimodal Interfaces*, in *Proc. 24th Annual Conf. of the IEEE Industrial Electronics Society*, Aachen, 1998, S. 2012–2017.

[5] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, W. Thon: *Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0*, 11, Institut für Phonetik und digitale Sprachverarbeitung, Kiel, 1994.

[6] D. Paul, J. Baker: *The Design for the Wall Street Journal-based CSR Corpus*, in *Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

[7] S. Wendt, G. A. Fink, F. Kummert: *Forward Masking for Increased Robustness in Automatic Speech Recognition*, in *Proc. European Conf. on Speech Communication and Technology*, Bd. 1, Aalborg, 2001, S. 615–618.

[8] G. A. Fink: *Developing HMM-based Recognizers with ESMERALDA*, in V. Matoušek, P. Mautner, J. Ocelíková, P. Sojka (Hrsg.): *Lecture Notes in Artificial Intelligence*, Bd. 1692, Springer, Berlin Heidelberg, 1999, S. 229–234.