Dynamische Suchraumeinschränkung für automatische Spracherkennung unter Zeitrestriktionen

Sascha Wendt, Gernot A. Fink, Franz Kummert

Inhaltsverzeichnis

1 Einleitung **1**

2 Suchräume für die Hypothesengenerierung **1**

3 Allgemeine Verfahren zur Suchraumeinschränkung **2**

4 Dynamische Suchraumeinschränkung **3**

4.1 Experimente 4

4.2 Adaption der Suchraum-Parameter 5

5 Evaluation **8**

6 Zusammenfassung und Ausblick **11**

SFB 360 Report 02/02, Universitaet Bielefeld, 2002.

Zusammenfassung

Bei der automatischen Spracherkennung müssen während des Erkennungsprozesses komplexe Zustandsräume durchsucht werden. Durch Einschränkung dieser Suchräume kann die Bearbeitungszeit deutlich reduziert werden, wobei jedoch im allgemeinen auch die Wortakkuratheit (Erkennungsleistung) sinkt. Eine verstärkte Suchraumeinschränkung ist jedoch vor allem für zeitkritische Anwendungen, wie z.B. im Szenario des SFB 360 die Erkennung und Verarbeitung von Interventionen, notwendig. Deshalb wurde ein Verfahren zur dynamischen Suchraumeinschränkung entwickelt. Dieses Verfahren optimiert dynamisch verschiedene Suchraum-Parameter unter Berücksichtigung der Verarbeitungsgeschwindigkeit und der Wortakkuratheit. Mit Hilfe dieses Verfahrens ist es möglich, Sprachsignale mit einer vorgegebenen Geschwindigkeit zu verarbeiten. Im Gegensatz zu dem Spracherkennungssystem ohne dynamische Suchraumeinschränkung, bei dem die Verarbeitungsgeschwindigkeit zwischen verschiedenen Äußerungen stark variiert, zeigt das erweiterte System eine annähernd konstante Verarbeitungsgeschwindigkeit für alle Äußerungen, wobei die Wortakkuratheit bei gleicher verfügbarer Verarbeitungszeit nicht signifikant sinkt. Darüber hinaus kann das erweiterte System jederzeit so beschleunigt werden, dass eine automatische Spracherkennung ohne zeitliche Verzögerungen und damit eine reflexartige Reaktion der Robotik möglich ist.

1 Einleitung

Ein Ziel bei der Entwicklung einer Sprachperzeptionskomponente innerhalb des SFB 360 ist es, sprachliche Interventionen möglichst ohne zeitliche Verzögerungen erkennen zu können, um eine reflexartige Reaktion des Systems zu ermöglichen. Deshalb wurde ein Verfahren entwickelt, um zeitliche Restriktionen bei der automatischen Spracherkennung zu berücksichtigen. Dieses Verfahren passt die Größe des abzuarbeitenden Suchraums während des Erkennungsprozesses so an, dass dieser innerhalb eines vorgegebenen Zeitrahmens durchsucht werden kann. Dabei kombiniert das Verfahren dynamisch verschiedene Methoden zur Suchraumeinschränkung, um eine möglichst robuste Erkennung zu gewährleisten.

Im folgenden wird zunächst ein kurzer Überblick über die bei der automatischen Spracherkennung entstehenden Suchräume und verschiedene allgemeine Verfahren zur Suchraumeinschränkung gegeben. Dabei wird auch erläutert, warum keines dieser Verfahren für sich alleine wirklich geeignet ist, um zeitliche Restriktionen bei der automatischen Spracherkennung zu berücksichtigen. Anschließend wird das entwickelte Verfahren der dynamischen Suchraumeinschränkung vorgestellt und die mit diesem Verfahren erweiterte Sprachperzeptionskomponente evaluiert. Ein kurze Zusammenfassung sowie ein Ausblick beschließen diesen Report.

2 Suchräume für die Hypothesengenerierung

In automatischen Spracherkennungssystemen, die wie das zugrundeliegende System auf Hidden-Markov-Modellen (HMMs) basieren, müssen zur Generierung einer Äußerungshypothese komplexe Zustandsräume durchsucht werden (Dekodierung von Sprache, siehe z.B. [Sch95]). Um diese Zustandsräume möglichst gering zu halten und nur die vielversprechendsten Bereiche zu durchsuchen (eine vollständige Suche ist im allgemeinen aufgrund der Komplexität nicht möglich), gibt es verschiedene Verfahren. Mit diesen Verfahren kann die Erkennungsgeschwindigkeit deutlich erhöht werden, wobei im allgemeinen aber auch die Fehlerrate um so stärker steigt, je stärker der Suchraum eingeschränkt wird.

In Abbildung 1 ist ein solcher Zustandsraum schematisch dargestellt. Die linke Darstellung zeigt den Suchraum mit den bewerteten Zuständen und die entstehenden Suchpfade bei einer geringen Suchraumeinschränkung. Dabei werden die Pfade im Suchraum zeitlich von links nach rechts aufgebaut und verfolgt. Im links dargestellten Fall wird da-

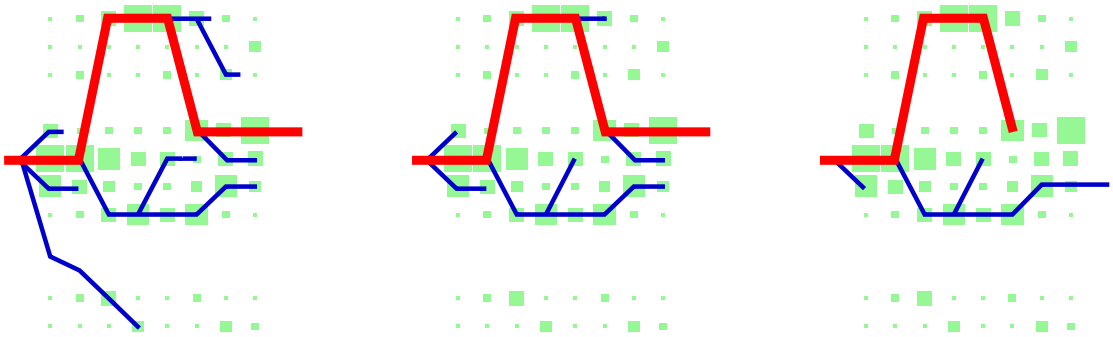

Abbildung 1: *Schematische Darstellung eines Suchraums, der während des Spracherkennungsprozesses zeitlich von links nach rechts durchsucht wird. Die drei verschiedenen Darstellungen zeigen die Auswirkung von zunehmenden Suchraumeinschränkungen.*

bei die optimale Hypothese (fett dargestellt) gefunden. Wird der Suchraum stärker eingeschränkt (in der Mitte dargestellt), müssen weniger Pfade verfolgt werden, wodurch die Erkennungsgeschwindigkeit steigt. Auch hier wird die optimale Hypothese noch gefunden. Wird der Suchraum jedoch zu stark eingeschränkt (rechte Darstellung), steigt die Geschwindigkeit zwar noch weiter, da noch weniger Pfade weiterverfolgt werden müssen. Allerdings wird in diesem Fall die optimale Hypothese aufgrund der Suchkriterien nicht mehr gefunden und eine falsche Hypothese generiert, d.h. die Wortakkuratheit des Systems nimmt ab.

Diese Darstellungen sollen verdeutlichen, dass ein Verfahren benötigt wird, um den Suchraum so einzuschränken, dass der Verlust an Wortakkuratheit möglichst gering gehalten wird.

3 Allgemeine Verfahren zur Suchraumeinschränkung

In der Literatur gibt es eine Vielzahl von Verfahren zur Suchraumeinschränkung, von denen einige im folgenden beschrieben werden, insbesondere diejenigen, die in der zugrundeliegenden Sprachperzeptionskomponente zum Einsatz kommen.

Ein sehr bekanntes und häufig eingesetztes Verfahren basiert auf dem sogenannte *Beamsearch* [Low76]: Innerhalb des Suchraums werden nur die Pfade weiter untersucht, deren Bewertung nur um einen bestimmten Wert (dem Schwellwert) von der Bewertung des aktuell besten Pfades abweichen. Dieser Schwellwert wird über einen einzelnen Parameter gesteuert, der jedoch heuristisch festgelegt werden muß und in der Regel während des gesamten Erkennungsprozesses konstant bleibt. Dadurch kann dieses Verfahren auch keine

bestimmte Verarbeitungsgeschwindigkeit garantieren, da im schlechtesten Fall alle Bewertungen sehr dicht beieinander liegen und somit keine Einschränkung des Suchraums erfolgt. Dieses Verfahren kann an verschiedenen Stellen innerhalb eines Spracherkennungssystem eingesetzt werden, z.B. auf der Ebene des akustischen Modells und des Sprachmodells (siehe dazu z.B. [NO99]) 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 [STN94]. 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 [FSKS98]): 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, wobei innerhalb dieser Gruppe nur der am besten bewertete Pfad weiter verfolgt wird. 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.

4 Dynamische Suchraumeinschränkung

Die beschriebenen Verfahren lassen sich alle über jeweils einen Parameter (Schwellwert oder Länge des Zeitfenster) steuern. Durch Variierung dieser Parameter kann der Suchraum 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 innerhalb eines vorgegebenen Zeitrahmens durchführen zu können und dabei die Erkennungsleistung (also die Wortakkuratheit) so hoch wie möglich zu halten.

4.1 Experimente

Zunächst wurden eine Reihe von Experimenten auf dem deutschen Verbmobil-Datensatz [KLP+94] 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 die Sprachperzeptionskomponente verwendet werden. Diese Parameter (zur Festlegung der Schwellwerte für Mischverteilungen, für das akustische Modell und das Sprachmodell sowie der Länge des Zeitfenster für die Zusammenfassung benachbarter Suchbäume) wurden dann schrittweise erhöht bzw. erniedrigt. Das Ziel dabei ist, einen Suchraum zu schaffen, in dem die optimale Parameterkombination für eine bestimmte Geschwindigkeitsanforderung zu finden ist.

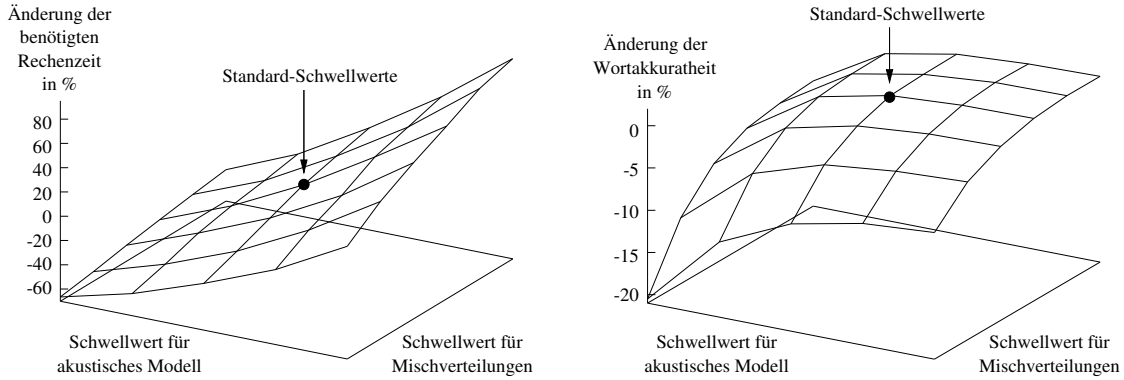

Abbildung 2: *Relative Änderung der Rechenzeit (linke Abbildung) und der Wortakkuratheit (rechte Abbildung) in Abhängigkeit von verschiedenen Schwellwerten, ausgehend von den Standard-Schwellwerten.*

In Abbildung 2 sind die Berechnungszeiten und Wortakkuratheiten (bzw. deren Änderungen gegenüber den Standardwerten) bei verschiedenen Schwellwerten für die Mischverteilungen und für das akustische Modell dargestellt (alle Experimente wurden auf einer Compaq Alpha EV6 mit 500 MHz durchgeführt).

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).

4.2 Adaption der Suchraum-Parameter

Basierend auf den beschriebenen Experimenten, die zur Schätzung der Berechnungszeiten und Wortakkuratheiten für bestimmte Parameter-Kombinationen dienen, wurde dann ein Mechanismus entwickelt, um die verschiedenen Parameter dynamisch zu adaptieren.

Zuerst wurde ein Gewichtungsparameter eingeführt, der dazu dient, die durch die Experimente erhaltenen Rechenzeiten und Wortakkuratheiten gewichtet zusammenzufassen. Dabei wird für jede Parameterkombination mit Hilfe des Gewichtungsparameters eine Bewertung berechnet. Auf diese Weise werden die in Abbildung 2 dargestellten Suchräume zu einem Suchraum zusammengefaßt, dessen Form vom Gewichtungsparameter abhängt.

Die Bewertung Q für die verschiedenen Parameterkombinationen wird mit Hilfe des Gewichtungsparameters α 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 2).

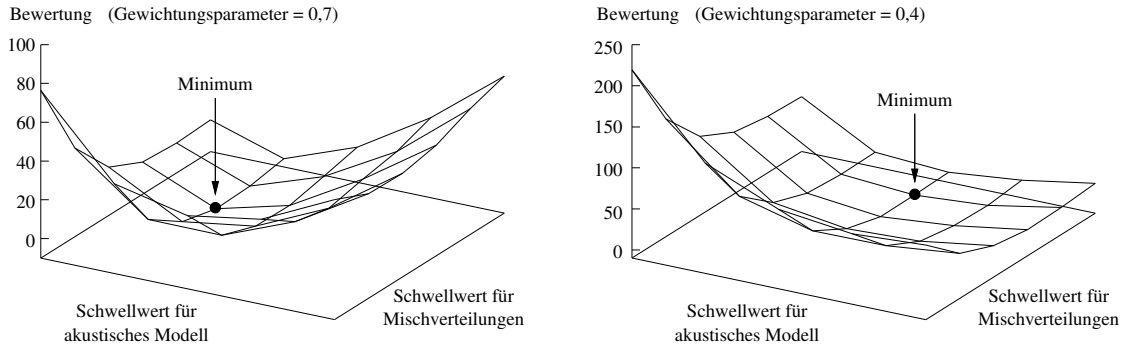

Abbildung 3: *Kombinierter Schwellwert-Suchraum für einen Gewichtungsparameter von 0,7 (links) bzw. 0,4 (rechts) mit dem jeweiligen Minimum.*

Mit Hilfe dieser Bewertungen kann nun ein kombinierter Suchraum erzeugt werden, dessen Form vom Gewichtungsparameter α (Wert zwischen 0 und 1) abhängt (siehe Abbildung 3). 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 3), so läßt sich über den Gewichtungsparameter bestimmen, ob Suchraum-Parameter mit eher geringer Rechenzeit oder eher hoher Wortakkuratheit verwendet werden sollen.

Durch dieses Vorgehen wird außerdem erreicht, dass im Fall von zwei Parameterkombinationen mit derselben (geschätzten) Rechenzeit diejenige Kombination gewählt wird, die die höhere (geschätzte) Wortakkuratheit erzielt, da diese Kombination eine niedrigere Bewertung erhält und damit 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 die Sprachperzeptionskomponente zu langsam ist (langsamer als gewünscht oder notwendig), wird der Gewichtungsparameter α erhöht. Dadurch wird der Suchraum für die möglichen Parameterkombinationen verändert (siehe Abbildung 3) und die Minimierung von Gleichung (1) liefert einen Parametersatz mit einer geringeren Rechenzeit, so dass der Erkennungsprozess beschleunigt wird. Ist umgekehrt die Sprachperzeptionskomponente schneller als gewünscht oder notwendig, wird der Gewichtungsparameter α gesenkt, um eine Parameterkombination mit einer höheren Rechenzeit, dafür aber auch mit einer höheren (geschätzten) Wortakkuratheit zu verwenden.

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

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

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

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

$$\Delta\alpha = \gamma \cdot (\Delta t_c - \Delta t_l) \tag{3}$$

  wobei γ eine Art Federkonstante und Δt_l 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 ist.
2. negativ und kleiner als die vorherige Zeitdifferenz ist.

Der 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, um eine gegebene Zeitrestriktion einzuhalten. Durch die Konstante γ in Gleichung (3) ist es außerdem möglich festzulegen, 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 die Sprachperzeptionskomponente mit dem beschriebenen Verfahren der dynamischen Suchraumeinschränkung erweitert. Um eine Zeitrestriktion und damit die angestrebte Rechenzeit für Gleichung (2)angeben zu können, wird der in der automatischen Spracherkennung übliche Echtzeitfaktor (real time factor - RTF) verwendet:

$$\text{RTF} = \frac{\text{benötigte Rechenzeit}}{\text{Länge der verarbeiteten Spracheingabe}} \tag{4}$$

Die Berechnung der notwendigen Zeitdifferenz erfolgt durch eine Zeitmessung nach jedem zehnten Frame (die Framelänge beträgt 10 ms). Um die Parameter-Kombination für die Suchraumeinschränkung anzupassen, wird zuerst die aktuelle Zeitdifferenz Δt_c 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 Parameter-Kombination mit der niedrigsten Bewertungszahl bildet dann die Suchraum-Schwellwerte für den weiteren Erkennungsprozess.

5 Evaluation

Die mit der dynamischen Suchraumeinschränkung erweiterte Sprachperzeptionskomponente wurde auf zwei verschiedenen Datensätzen evaluiert.

Da die Änderungen von Rechenzeit und Wortakkuratheit auf dem deutschen *Verbmobil*-Datensatz gemessen wurden (siehe Abschnitt 4.1), erfolgte die erste Evaluation konsequenter Weise auf diesem Corpus. Der Verbmobil-Datensatz besteht aus über 32 Stunden spontan gesprochenem Trainingsmaterial mit einem Wortschatz von 5336 Wortformen sowie etwa 35 Minuten Testmaterial [KLP+94].

Daneben erfolgte eine Evaluierung auf dem englischen *Wall-Street-Journal*-Datensatz (WSJ0) [KLP+94]. Hier stehen ca. 15 Stunden Trainings- und etwa 40 Minuten Testmaterial (vorgelesene Sätze aus der gleichnamigen Wirtschaftszeitung) zur Verfügung, der Wortschatz[1] für die Evaluierung umfaßt 4986 Wortformen [PB92]. 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 [WFK01] 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 [Fin99] zu finden.

Jede Evaluation wurde neun mal durchgeführt (jeweils dreimal auf drei verschiedenen

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

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	2,0	0,02	75,71 %
der Parameter	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 bedeuted, dass die Verarbeitungszeit sehr stark variiert, wenn die Standard-Parameter verwendet werden.

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 Abbildung 2) kann jedoch nicht vollständig dadurch kompen-

siert 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 4 wurde bereits erläutert, dass die Standard-Parameter näherungsweise optimale Parameter darstellen).

Mit Hilfe der dynamische Suchraumeinschränkung ist es der Sprachperzeptionskomponente nun auch möglich, 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 2 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	3.2	< 0.01	87.67 %
der Parameter	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 die Sprachperzeptionskomponente 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 (also mit einer schnelleren 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.

6 Zusammenfassung und Ausblick

Mit Hilfe des vorgestellten Verfahrens der dynamischen Suchraumeinschränkung ist es möglich, Spracheingaben innerhalb eines vorgegebenen Zeitrahmens zur verarbeiten. Gegenüber dem System mit statischen Parametern tritt bei gleicher Verarbeitungszeit nur ein minimaler, jedoch nicht signifikanter Verlust an Wortakkuratheit ein, dafür kann die Verarbeitungsgeschwindigkeit nahezu konstant gehalten werden. Darüber hinaus können Zeitrestriktionen unabhängig von der zur Verfügung stehenden Rechenleistung (zumindest in einem bestimmten Rahmen) eingehalten werden.

Mit diesem Verfahren können nun auch insbesondere Interventionen ohne zeitliche Verzögerungen, also in Echtzeit, erkannt werden, wodurch eine reflexartige Reaktion der Robotik ermöglicht wird. Der nächste Schritt wird nun darin bestehen, die erweiterte Sprachperzeptionskomponente in das Gesamtsystem des SFB 360 zu integrieren und situationsabhängige Zeitrestriktionen festzulegen.

Im ersten Ansatz soll zunächst nur der Zustand der Robotik ausgewertet werden: ist die Robotik gerade aktiv ist, wird eine Sprachverarbeitung in Echtzeit gefordert. Auf diese Weise wird es der Robotik ermöglicht, Interventionen reflexartig umzusetzen. Ist die Robotik dagegen nicht aktiv, kann die Zeitrestriktion gelockert werden.

Da jedoch eine Sprachverarbeitung in Echtzeit im allgemeinen zu einer Verschlechterung der Wortakkuratheit führt (zumindest bei der zur Zeit zur Verfügung stehenden Rechenleistung), stellt sich die Frage, ob der Zustand der Robotik nicht differenzierter betrachtet werden muss. Auch gilt es zu untersuchen, ob und welche Zeitrestriktionen neben dem Status der Robotik zu beachten sind.

Literatur

[Fin99] Gernot A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

[FSKS98] Gernot A. Fink, Christoph Schillo, Franz Kummert, and Gerhard Sagerer. Incremental speech recognition for multimodal interfaces. In *Proc. 24th Annual Conf. of the IEEE Industrial Electronics Society*, pages 2012–2017, Aachen, 1998.

[KLP+94] Klaus Kohler, Gloria Lex, Matthias Pätzold, Michael Scheffers, Adrian Simpson, and Werner Thon. Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0. Technical Report 11, Institut für Phonetik und digitale Sprachverarbeitung, Kiel, 1994.

[Low76] Bruce T. Lowerre. *The HARPY Speech Recognition System*. PhD thesis, Carnegie-Mellon University, Department of Computer Science, Pittsburg, 1976.

[NO99] Hermann Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999.

[PB92] D.B. Paul and J.M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language Workshop*. Morgan Kaufmann, 1992.

[Sch95] E. G. Schukat-Talamazzini. *Automatische Spracherkennung*. Vieweg, Wiesbaden, 1995.

[STN94] Volker Steinbiss, Bach-Hiep Tran, and Hermann Ney. Improvements in beam search. In *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 2143–2146, Yokohama, 1994.

[WFK01] Sascha Wendt, Gernot A. Fink, and Franz Kummert. Forward masking for increased robustness in automatic speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 615–618, Aalborg, 2001.