Ausnutzung zeitlicher Redundanzen der cepstralen Sprachmerkmale für die automatische Spracherkennung

Stefan Windmann

Institut für Elektrotechnik und Informationstechnik
Fachgebiet Nachrichtentechnik
Universität Paderborn

26. September 2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Sprache ist von herausragender Bedeutung für die zwischenmenschliche Kommunikation. Dennoch spielte sie in der Interaktion zwischen Mensch und Computer in der Vergangenheit eine eher untergeordnete Rolle. Dieser Umstand kann darauf zurückgeführt werden, dass das Erkennen gesprochener Sprache zu den schwierigsten Aufgaben der Signalverarbeitung gehört und bislang in den meisten Anwendungsfeldern noch nicht zuverlässig umgesetzt werden konnte. Die Aufgabe der Spracherkennung besteht darin, ein akustisches Sprachsignal in eine Textsequenz, typischerweise eine Wortsequenz, umzuwandeln [DO03]. Die Spracherkennung ist ein interdisziplinäres Arbeitsfeld, das sich über die Bereiche Computertechnik, Programmierung, Mustererkennung, künstliche Intelligenz, Phonetik und Linguistik erstreckt. Spracherkennungssysteme werden häufig in sprecherunabhängige und sprecherabhängige Systeme, in Systeme zur Erkennung kontinuierlich gesprochener Sprache und isolierter Spracheinheiten sowie bzgl. der Größe des verwendeten Wortschatzes unterteilt. Seit der Entwicklung der ersten Laborsysteme in den 50er Jahren, die für ein kleines Vokabular und isolierte Spracheinheiten ausgelegt und zumeist sprecherabhängige Systeme waren, haben sich bedeutende Veränderungen in der automatischen Spracherkennung ergeben. Neue Algorithmen, verbesserte Modellierungsarten und höhere Rechenleistung ermöglichen den zunehmenden Einsatz in kommerziellen Anwendungen, die teilweise sprecherunabhängig und für kontinuierliche Sprache sowie ein großes Vokabular ausgerichtet sind. Wichtige Anwendungsgebiete sind automatische Dialogsysteme, die beispielsweise für die Fahrplan- und Telefonauskunft eingesetzt werden und häufig mit Datenbankabfragen verknüpft sind. Weiterhin wird die Spracherkennung bereits für die sprachgesteuerte Erstellung von Dokumenten und Briefen, insbesondere in den Bereichen Medizin, Recht, Finanzen und Versicherungen, sowie für die Robotersteuerung in der Fertigung, in Freisprecheinrichtungen von Kraftfahrzeugen und in kleinen, portablen Geräten mit unhandlicher Benutzerschnittstelle eingesetzt. Ein weiteres mögliches Einsatzgebiet ist die Hausautomatisierung. Trotz der großen Fortschritte, die in den letzten Jahrzehnten auf dem Gebiet der automatischen Spracherkennung erzielt worden sind, bleiben die Fähigkeiten des menschlichen Gehirns in diesem Bereich noch weitgehend unerreicht. Ein Problem im Praxiseinsatz stellen insbesondere Szenarien dar, in denen deutliche Unterschiede zwischen den akustischen Trainings- und Testbedingungen auftreten.

In dieser Arbeit wird ein statistischer Spracherkennungsansatz aufgegriffen, der auf der Modellierung der Sprache mit Markovmodellen (HMMs) basiert und seit Mitte der

80er Jahre als Standard etabliert ist. Die Sprachdynamik wird in Markovmodellen mit versteckten Zuständen modelliert, denen Transitionswahrscheinlichkeiten zugeordnet sind. Zur robusten Repräsentation des Sprachsignals werden Merkmalsvektoren aus dem Sprachsignal extrahiert, die durch Emissionsverteilungen mit den HMM-Zuständen verknüpft sind.

Ein vielzitierter Schwachpunkt des HMM-basierten Modellierungsansatzes besteht darin, dass die Korrelationen zwischen zeitlich aufeinander folgenden Sprachmerkmalen nur indirekt über die versteckten Zustände der HMMs modelliert werden. An dieser Stelle setzt die vorliegende Arbeit an, in der die Integration von Inter-Frame Korrelationen in den skizzierten Ansatz untersucht wird.

Zunächst wird in Kapitel 2 der Stand der Forschung dargestellt. Nach einem kurzen Überblick über die geschichtliche Entwicklung der automatischen Spracherkennung wird der statistische Spracherkennungsansatz ausführlich beschrieben und anschließend eine Übersicht über bereits in der Literatur diskutierte Möglichkeiten zur Ausnutzung der Inter-Frame Korrelation, zur Schätzung des Umgebungsrauschens sowie zum Austausch von Informationen zwischen verschiedenen Stufen des Erkennungsprozesses gegeben.

Auf dieser Basis werden in Kapitel 3 die wissenschaftlichen Ziele, die der vorliegenden Arbeit zugrunde liegen, formuliert sowie die weitere Gliederung erläutert.

Diese Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter im Fachgebiet Nachrichtentechnik an der Universität Paderborn. Die Forschung wurde von der DFG Forschungsgruppe GK-693 des Paderborner Institutes für wissenschaftliches Rechnen (PaSCo) unterstützt. Besonders möchte ich mich bei meinem Betreuer, Herrn Professor Dr.-Ing. Reinhold Häb-Umbach, bedanken, dessen Unterstützung und Hilfsbereitschaft mir die Durchführung dieser Arbeit ermöglichten. Weiterhin möchte ich mich bei Herrn Professor Dr.-Ing. Tim Fingscheidt dafür bedanken, dass er die Zweitbetreuung meiner Arbeit übernommen hat. Mein Dank gilt auch den Mitarbeitern des Fachgebietes Nachrichtentechnik, deren Hilfsbereitschaft und Kollegialität einen guten Rahmen für das Zustandekommen dieser Arbeit darstellten. Schließlich möchte ich mich bei meiner Familie bedanken, die mich bei der Durchführung meiner Arbeit unterstützt hat.

Paderborn, im September 2008

Kapitel 2

Stand der Forschung

Die automatische Spracherkennung ist bereits seit vielen Jahren Gegenstand umfangreicher Untersuchungen und einer großen Anzahl an Publikationen. Daher können an dieser Stelle basierend auf [RJ93], [JH96] und [Mar96] nur die wichtigsten Entwicklungsrichtungen skizziert werden.

Die ersten Versuche zur Spracherkennung wurden bereits lange vor der Erfindung digitaler Computer von Alexander Graham Bell unternommen. Das erste rudimentäre Spracherkennungssystem wurde jedoch erst im Jahre 1942 als US-Patent zugelassen.

In den 50er Jahren entstanden einfache Systeme, die in der Regel sprecherabhängig waren und eine isolierte Ziffern- bzw. Phonemerkennung für ein kleines Vokabular durchführten [DBB52, OB56, FF59, Fry59]. Das vorherrschende Prinzip bestand in der Spektralanalyse von Sprachsignalausschnitten, die Vokalen zugeordnet werden konnten, und dem anschließenden Vergleich der Filterausgänge mit Vorlagen für mögliche Signalverläufe. Dabei wurde bereits die Verwendung eines rudimentären Sprachmodells [Fry59] und einer sprecherunabhängigen Erkennung untersucht [FF59].

Der Beginn der 60er Jahre wurde von zahlreichen Hardwarelösungen japanischer Laboratorien geprägt. Beispiele sind ein Spektralanalysator mit nachgeschalteter Logik und Mehrheitsvoting [SN61] oder ein Erkenner, der die Segmentierung der Sprache und eine anschließende Analyse der Nulldurchgänge durchführt [SD62]. Auch das Forschungsprogramm der NEC-Laboratorien auf dem Gebiet der Spracherkennung begann 1963 mit einem Hardwaresystem zur Zeichenerkennung [NKC63]. In den 60er Jahren sind vor allem drei wichtige Forschungsprojekte hervorzuheben. In den RCA-Laboratorien wurden Verfahren zur Anpassung der Zeitskalen von Sprachsignalen basierend auf der Erkennung von Wortanfängen und -enden in einem Strom kontinuierlicher Daten entwickelt [MNZ64]. Außerdem wurden bereits Algorithmen zur Zeitanpassung, die auf dynamischer Programmierung basieren, sowie rudimentäre Algorithmen zur Erkennung kontinuierlich gesprochener Sprache [Vin68] und Methoden zum dynamischen Tracking von Phonemen [Red66] entwickelt.

Die 70er Jahre brachten eine Reihe bedeutender technischer Entwicklungen mit sich. Insbesondere drei Neuerungen führten dazu, dass große Fortschritte in der Erkennung ganzer, isolierter Wörter verzeichnet wurden: die Übertragung von Methoden der Mustererkennung auf die Spracherkennung [VZ70], der Einsatz dynamischer Programmierung [SC78] sowie die Verwendung von LPC (Linear Predictive Coefficients)-Koeffizienten für die Spracherkennung [Ita75]. In dieses Jahrzehnt fallen auch der Be-

ginn der automatischen Spracherkennung mit großem Vokabular bei IBM [TDRC71], [JBM75], [Jel85], die Entwicklung sprecherunabhängiger Erkennungssysteme basierend auf Clusteralgorithmen [RLRW79] sowie erste Ansätze zur Erkennung kontinuierlich gesprochener Sprache ohne künstliche Pausen [Vin69].

Die Forschung der 80er Jahren wurde durch einen Wechsel von vorlagenbasierten Ansätzen zu statistischen Modellierungsmethoden, insbesondere dem Markovmodell, charakterisiert. Markovmodelle wurden bereits Ende der 60er Jahre und Anfang der 70er Jahre für die automatische Spracherkennung eingesetzt [Vin68], [VZ70], [SC71], [Bri73], [SC78]. Ihre große Verbreitung in nahezu allen Spracherkennungssystemen erfuhren sie aber erst in der Mitte der 80er Jahre. In den späten 80er Jahren konnten außerdem neuronale Netzwerke, die bereits in den 50er Jahren untersucht wurden, erfolgreich für die automatische Spracherkennung eingesetzt werden. Ein weiterer Forschungsschwerpunkt der 80er Jahre war die Erkennung kontinuierlich gesprochener Sprache. Hervorzuheben ist in diesem Zusammenhang das Forschungsprogramm der DARPA (Defense Advanced Research Projects Agency), das darauf abzielte, die Erkennungsrate für ein Vokabular mit 1000 Wörtern und kontinuierlich gesprochener Sprache zu maximieren.

Auch in den 90er Jahren und zu Beginn des 21. Jahrhunderts ist es nicht gelungen, Spracherkennungssysteme zu entwickeln, die unter allgemeinen Bedingungen die Erkennungsleistung des Menschen erreichen.

Die aktuelle Forschung beschäftigt sich im Rahmen der statistischen Spracherkennung auf der Grundlage der zur Zeit als Standard etablierten Markovmodelle u.a. mit der Verwendung geeigneter Sprachmerkmale, einer verbesserten Sprach- und Akustikmodellierung, der Erkennung mit großem Vokabular sowie der robusten Spracherkennung bei voneinander abweichenden Trainings- und Testbedingungen.

Im nächsten Abschnitt werden wichtige Aspekte der aktuellen Spracherkennungstechnologie dargestellt.

2.1 Statistische Spracherkennung

In der automatischen Spracherkennung ist die Verwendung statistischer Modelle gegenwärtig als Standard etabliert. Sie ermöglicht es, Entscheidungen trotz Unsicherheit und wenig Wissen über die gesprochene Wortsequenz zu treffen sowie ein automatisches Training auf großen Datenbanken durchzuführen. Den Ausgangspunkt für die statistische Spracherkennung stellen die akustischen Merkmalsvektoren $\mathbf{x}_t$ dar, die aus dem Sprachsignal extrahiert werden und dieses charakterisieren. Aus ihnen soll eine optimale Wortsequenz $\hat{w}_1^N = \hat{w}_1, \ldots, \hat{w}_N$ berechnet werden, die die a posteriori Wahrscheinlichkeit $P(w_1^N|\mathbf{x}_1^T)$ der Wortsequenz w_1^N bei einer gegebenen Merkmalsfolge $\mathbf{x}_1^T = \mathbf{x}_1 \ldots \mathbf{x}_T$ maximiert, wobei T die Länge der Merkmalsfolge bezeichnet:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}}\, P(w_1^N|\mathbf{x}_1^T) = \underset{w_1^N}{\operatorname{argmax}}\, \frac{P(w_1^N)p(\mathbf{x}_1^T|w_1^N)}{p(\mathbf{x}_1^T)} \tag{2.1}$$

Die Maximierung wird auch über die Länge N der Wortsequenz durchgeführt. Wie in der Literatur üblich, wird dies nicht explizit notiert. In der Praxis ist es erforderlich, das Sprachmodell gegenüber dem akustischen Modell mit einem Sprachmodell-Skalierungsfaktor S_β zu gewichten:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \frac{P(w_1^N)^{S_\beta} p(\mathbf{x}_1^T | w_1^N)}{p(\mathbf{x}_1^T)} \tag{2.2}$$

Das zugrunde gelegte Spracherkennungssystem ist in Abb. 2.1 dargestellt.

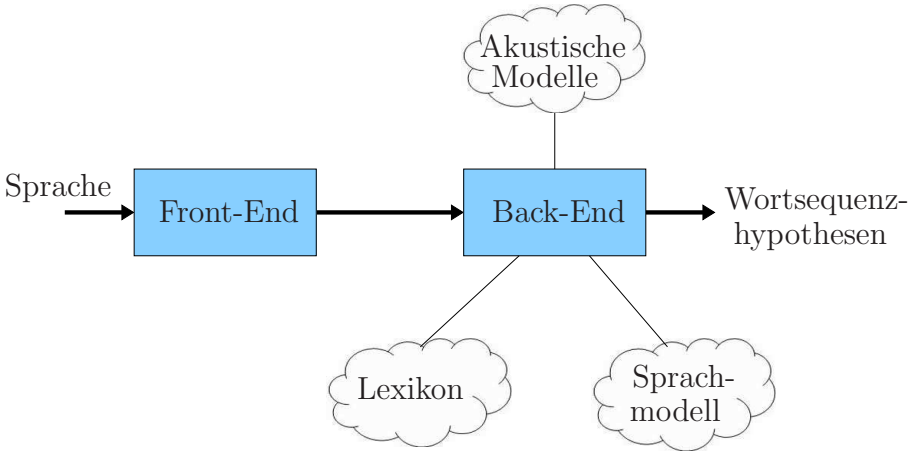

Abbildung 2.1: *Aufbau eines statistischen Spracherkennungssystems*

Um die optimale Wortfolge entsprechend der Entscheidungsregel in Gl. (2.1) zu bestimmen, wird nach der Merkmalsextraktion im Front-End eine Suche im Back-End des Spracherkenners durchgeführt. Dabei werden drei wesentliche Informationsquellen verwendet, nämlich ein Sprachmodell, aus dem sich die a priori Wahrscheinlichkeit $P(w_1^N)$ ergibt, ein akustisches Modell, das die Wahrscheinlichkeit $p(\mathbf{x}_t | w_1^N)$ bestimmt, sowie ein Lexikon, das den Wörtern des Spracherkennungssystems akustische Modelle zuordnet.

Das Lexikon legt die Zusammensetzung der Wörter aus kleineren akustischen Einheiten wie Phonemen fest. Im Folgenden wird es bei einem größeren Vokabular als Baumstruktur modelliert, in der alle Wörter den gleichen Ursprung besitzen und die sich jeweils an den Positionen verzweigt, an denen sich zwei Wörter in ihren akustischen Einheiten zu unterscheiden beginnen. Bei einem kleinen Vokabular, wie es beispielsweise in der AURORA2 Datenbank vorliegt, ist das Lexikon häufig eine 1:1-Abbildung zwischen dem Vokabular und den akustischen Modellen.

Das Sprachmodell ist ein statistisches Modell der Syntax, Semantik und Pragmatik der Sprache, das unabhängig von den akustischen Merkmalen $\mathbf{x}_t$ ist. Es enhält Informationen über zugelassene Wortsequenzen, indem es deren a priori Wahrscheinlichkeiten $P(w_1^N)$ modelliert. Durch die Einschränkung der möglichen Wortfolgen bei der Suche können eine drastische Reduktion des Rechenaufwandes sowie ein Rückgang der Fehlentscheidungen erreicht werden. Beispiele für Sprachmodelle sind Wortpaargrammatiken, in denen modelliert wird, welches Wort auf das aktuelle Wort folgt und

stochastische Sprachmodelle wie m-Gram Sprachmodelle, in denen möglichen Wortfolgen Wahrscheinlichkeiten zugeordnet werden. In m-Gram Sprachmodellen wird die Wahrscheinlichkeit

$$P(w_1^N) = P(w_1)P(w_2|w_1)\dots P(w_N|w_1^{N-1}) \tag{2.3}$$

durch den Markovprozess

$$P(w_1^N) \approx \prod_{n=1}^{N} P(w_n|w_{n-m+1}^{n-1}) \tag{2.4}$$

der Ordnung $m-1$ approximiert. Am weitesten verbreitet sind Trigram- ($m=3$) und Bigram- ($m=2$) Sprachmodelle. Das Sprachmodell wird in der Regel unabhängig von dem akustischen Modell auf großen Datenmengen geschriebenen Textes trainiert.

Die Merkmalsextraktion, die akustische Modellierung und die Suche werden ausführlich in den nächsten Abschnitten behandelt.

2.1.1 Merkmalsextraktion und -entrauschung

Die Aufgabe der Merkmalsextraktion besteht darin, eine charakteristische Repräsentation des Sprachsignals zu berechnen, die den Ausgangspunkt für die Suche der wahrscheinlichsten Wortfolge darstellt. Dabei sollen die extrahierten Merkmale robust gegenüber Rauschen, Verzerrungen und Variationen in der Artikulation sein, so dass auch bei voneinander abweichenden Trainings- und Testbedingungen eine zuverlässige Erkennung möglich ist und von Informationen über den Sprecher, den akustischen Kanal und die akustische Umgebung abstrahiert wird.

Eine Möglichkeit, dies zu erreichen, besteht in der Verwendung robuster Merkmale. Die am häufigsten eingesetzten Merkmale sind die PLP(Perceptual Linear Prediction)-Koeffizienten [Her90], bei deren Berechnung in der Regel eine Bandpaßfilterung im Spektrum durchgeführt wird (RASTA-PLP) [HM94], und die MFCC-Koeffizienten (Mel Frequency Cepstral Coefficients) [DM80]. In dieser Arbeit werden MFCC-Koeffizienten verwendet, die mit Hilfe des ETSI Standard Frontends (SFE) [ETS00] aus dem Sprachsignal extrahiert werden. Dabei wird das Energiedichtespektrum des Sprachsignals durch eine MEL-Filterung und Logarithmierung an das menschliche Gehör angepaßt und anschließend mit einer DCT-Matrix dekorreliert (siehe Anhang C.1).

Durch die zusätzliche Verwendung dynamischer Merkmale, die sich durch die Approximation der ersten und zweiten Ableitung der genannten statischen Merkmale ergeben, wird eine signifikante Erhöhung der Erkennungsrate erreicht. Die Untersuchungen in [YSFK07] zeigen, dass die alleinige Verwendung dynamischer Merkmale sogar zu einer höheren Erkennungsrate als die Verwendung statischer Merkmale führen kann. Um die Dimension des erweiterten Merkmalsvektors zu reduzieren, wird häufig eine lineare Diskriminantenanalyse (LDA) durchgeführt [HL89, HUN92, BWN95]. Eine weitere Möglichkeit, die Robustheit der Merkmale zu erhöhen, besteht in Normalisierungsmethoden wie der Mittelwertnormalisierung im Cepstrum (CMN) [Ata74],

der blinden Equalisierung (BEQ) [MMC+02] und der Histogrammnormalisierung bzw. -equalisierung [TSB+02], [MHN03]. Während durch die Anwendung der CMN und BEQ vor allem im Cepstrum additive Störungen entfernt werden, die durch eine Faltung im Zeitbereich (z.B. durch Kanalverzerrungen) hervorgerufen werden, kann u.a. mit den beiden zuletzt genannten Normalisierungsmethoden auch im Zeitbereich additives Rauschen kompensiert werden [JH96].

Der Unterschied zwischen Trainings- und Testbedingungen kann außerdem durch die Entrauschung der extrahierten Merkmale verringert werden, die prinzipiell auf jeder Stufe der Merkmalsextraktion möglich ist. Verbreitete Ansätze sind die spektrale Subtraktion [WAP74], [Bol79] und die Wiener-Filterung im Spektrum [Lim78], [Koo89], [BS91]. Im ETSI Advanced Front-End (AFE), das wegen seiner hohen Erkennungsleistung in vielen Arbeiten als Referenzverfahren angegeben wird, wird basierend auf dem SFE neben einer BEQ u.a. ein zweistufiges Wiener-Filter eingesetzt [ETS05] (vgl. Anhang C.4). In jüngerer Vergangenheit werden auch Methoden verwendet [HW04], die auf einer Singulär- oder Eigenwertzerlegung der Korrelationsmatrix des verrauschten Sprachsignals basieren [Bro92], [BWJ01]. Daneben existieren stereobasierte Ansätze wie RATZ [MRGS95] oder SPLICE [DAPH00], in denen Korrekturterme aus einer Trainingsdatenbank mit Stereodaten berechnet werden, die bei der Entrauschung von den cepstralen Merkmalen subtrahiert werden. Mit SPLICE wird, gemittelt über alle Testdaten der AURORA2 Datenbank, ungefähr die gleiche Erkennungsleistung wie mit dem AFE erreicht, wobei allerdings zu beachten ist, dass für diesen Ansatz eine Datenbank mit Stereopaaren und somit a priori Wissen über die Rauschbedingungen erforderlich ist.

In der vorliegenden Arbeit wird, wie in Kapitel 3 motiviert wird, ein Bayes'scher Ansatz für die Merkmalsentstörung herangezogen. Sprache und Rauschen werden als statistisch unabhängige Zufallsprozesse modelliert, für die ein additiver Zusammenhang im Zeitbereich angenommen wird. Das unverrauschte Sprachsignal ergibt sich in diesem Ansatz durch die Optimierung eines Fehlermaßes wie des mittleren quadratischen Fehlers (MMSE-Kriterium) oder der a posteriori Wahrscheinlichkeit (MAP-Kriterium) der geschätzten Sprachmerkmale. Dabei wird neben den verrauschten Sprachmerkmalen a priori Wissen über die Verteilung des Rauschens und der Sprache berücksichtigt. Ursprünglich mit dem Ziel der einkanaligen Sprachsignalverbesserung haben Ephraim und Malah einen MMSE-Schätzer im Kurzzeitspektrum unter der Annahme von Gaußverteilungen entwickelt [EM84]. Eine bessere Modellierung der Sprache wurde im Kurzzeitspektrum durch die Verwendung allgemeinerer Verteilungen wie Laplace- oder Gammaverteilungen erreicht [Mar02].

Im Log-Spektrum und Cepstrum ist das Beobachtungsmodell, das den Zusammenhang zwischen verrauschter und unverrauschter Sprache und Rauschen beschreibt, aufgrund der Logarithmierung bei der Merkmalsextraktion hochgradig nichtlinear, so dass in diesen Bereichen keine exakte MMSE- bzw. MAP-Schätzung möglich ist. Weiterhin beinhaltet das Beobachtungsmodell im Log-Spektrum und Cepstrum einen additiven Phasenterm, der in den meisten Literaturansätzen nicht exakt modelliert wird. Eine approximative MMSE-Schätzung im Log-Spektrum wurde erstmals von [VC89] und

[EW93a] durchgeführt und konnte in [EW93b] durch die Verwendung von Gaußmischungsverteilungen verbessert werden. Die modellbasierte Entrauschung der Sprachmerkmale im Cepstrum wurde zuerst in [Ace93] durchgeführt. Ein in vielen Veröffentlichungen erwähnter Ansatz ist das VTS-Verfahren von Moreno [Mor96]. Im VTS-Verfahren werden vor der MMSE-Schätzung der Sprachmerkmale im Log-Spektrum basierend auf einer Vektor-Taylorreihenentwicklung erster Ordnung des Beobachtungsmodells zunächst die Parameter des Messmodells und die Verteilung des Rauschens iterativ mit einem Expectation-Maximization(EM)-Algorithmus optimiert (vgl. Anhang C.3). Die Erweiterung der linearen Taylorreihenentwicklung auf Reihenglieder höherer Ordnung führte in [RGMS96] zu keiner signifikanten Verbesserung der Ergebnisse.

Die Übertragung des von Moreno eingeführten VTS-Verfahrens ins Cepstrum auf der Grundlage des in [Ace93] eingeführten Beobachtungsmodells wird in [KUK98] untersucht. Die Modellierung im Cepstrum ist, wie bereits in [Gal95] angeführt, vorteilhaft, da die Komponenten des cepstralen Merkmalsvektors weniger stark korreliert sind als die log-spektralen Komponenten. Außerdem stellen die cepstralen Merkmalsvektoren eine kompaktere Repräsentation mit weniger Parametern dar, wodurch die VTS-Näherung und das Training der Mischungsverteilungen robuster wird. Der Ansatz aus [Mor96] wird in [XRK06] um die zusätzliche Modellierung dynamischer Merkmale erweitert. In [STBP01] wird der VTS-Ansatz mit einer Bandpaßfilterung im Log-Spektrum kombiniert, wobei auf Kosten einer ungenauen Schätzung der Kovarianzmatrix des Rauschens auf den EM-Algorithmus verzichtet wird.

In [DDA03b] wird der blockweise EM-Algorithmus, mit dem die Parameter des Rauschmodells geschätzt werden, durch einen Bayes'schen Ansatz ersetzt, in dem für die einzelnen Sprachrahmen jeweils eine SNR-Variable iterativ optimiert wird, die den Ausgangspunkt für eine approximative MMSE-Schätzung der unverrauschten Sprachmerkmale darstellt (vgl. Anhang C.3). Eine exaktere statistische Methode, in der bei der MMSE-Schätzung im Gegensatz zu [DDA03b] auch Varianzen der a priori Verteilungen berücksichtigt werden, ist das ALGONQUIN-Verfahren [FDAK03]. Wie in Anhang E gezeigt wird, ist die Aktualisierung des Zustandsvektors aufgrund neuer Messwerte im ALGONQUIN-Verfahren äquivalent zu dem entsprechenden Schritt in einem iterativen erweiterten Kalman-Filter (IEKF). Die a priori Verteilungen von Sprache und Rauschen werden jedoch, anders als im IEKF, aus Trainingsdaten bzw. sprachfreien Signalabschnitten bestimmt. Auch in [SVhDW03] wird eine Linearisierung des Beobachtungsmodells und eine anschließende MMSE-Schätzung (LMMSE) angewendet, allerdings ohne die im ALGONQUIN-Verfahren durchgeführte Iteration, die auf der AURORA2 Datenbank zu einem signifikanten Anstieg der Erkennungsrate führt. In [Kim98c] wird eine statistische lineare Approximation durchgeführt, durch die sich ebenfalls das Beobachtungsmodell eines erweiterten Kalman-Filters (EKFs) ergibt.

Daneben sind Methoden untersucht worden, in denen die Taylorreihenentwicklung durch eine numerische Integration vermieden wird [MN03], [Afi05]. In diese Kategorie fallen auch die sequentiellen Monte-Carlo-Methoden, die in Abschnitt 2.3 beschrieben werden, und das Unscented Kalman-Filter (UKF), auf das in Abschnitt 2.2.2 einge-

gangen wird. In Arbeiten von Droppo und Deng [DAD02a, DDA02, DDA04a] wird die Berücksichtigung des weiter oben erwähnten, zusätzlichen Phasenterms untersucht. Dabei verwenden sie, um eine einfachere mathematische Behandlung zu erreichen, eine gegenüber dem SFE modifizierte Merkmalsextraktion, in der anstelle des Betragsspektrums das Energiespektrum des fouriertransformierten Signals berechnet wird. In [DAD02a] werden theoretische Vorteile nachgewiesen, die sich durch die Anwendung des um den Phasenterm erweiterten Beobachtungsmodells gegenüber der Durchführung einer spektralen Subtraktion ergeben. Ein MMSE-Schätzer auf der Basis einer Taylorreihenentwicklung zweiter Ordnung, in dem der Phasenterm berücksichtigt wird, führte in [DDA02] und [DDA04a] zu höheren Erkennungsraten als die spektrale Subtraktion. Gegenüber der MMSE-Schätzung mit dem ALGONQUIN-Verfahren konnte, wie in [DDA03b] angegeben, durch die zusätzliche Berücksichtigung des Phasenterms jedoch keine Erhöhung der Erkennungsrate erzielt werden. Auch in [SVhW05a] führte die Integration des Phasenterms zu keiner signifikanten Verbesserung. In [DDA04b] approximieren Deng und Droppo den Phasenterm daher durch eine gaußverteilte Zufallsvariable und leiten einen MMSE-Schätzer her, der dem ALGONQUIN-Algorithmus entspricht, den sie, wie in Abschnitt 2.2 beschrieben, jedoch um die Modellierung von Inter-Frame Korrelationen durch die Berücksichtigung der dynamischen Sprachmerkmale erweitern.

In [SVhDW03] wird das GMM, das zur Modellierung der Verteilung der Sprachmerkmale verwendet wird, basierend auf Arbeiten von [Gag93] und [SN94] durch ein HMM ersetzt, dessen Parameter mit einem Baum-Welch-Algorithmus trainiert werden. Bei der Filterung der Sprachmerkmale werden ein Vorwärts-Rückwärts-Algorithmus zur Berechnung der Zustandswahrscheinlichkeiten verwendet und die MMSE-Schätzwerte der einzelnen Zustände entsprechend dieser Wahrscheinlichkeiten kombiniert. Dadurch ergibt sich eine signifikante Verbesserung der Erkennungsleistung gegenüber dem GMM-Ansatz, die jedoch unter der Erkennungsleistung eines in [SVhW05b] eingeführten erweiterten Kalman-Filters liegt, in dem die statistischen Abhängigkeiten zwischen den Sprachmerkmalen direkt modelliert werden.

2.1.2 Akustische Modellierung

Das dynamische Verhalten der extrahierten Sprachmerkmale kann mit akustischen Modellen beschrieben werden. Gewöhnlich ist es wegen der begrenzten Trainingsdaten nicht möglich, akustische Modelle für komplette Wörter zu verwenden. Bei einem größeren Vokabular werden die Modelle daher in der Regel für kleinere Spracheinheiten wie Phoneme trainiert. In vielen Spracherkennungssystemen werden kontextabhängige Phonemmodelle eingesetzt, da die Dynamik der Merkmale, die einem Phonem zugeordnet werden können, häufig stark von den benachbarten Phonemen beeinflußt wird. In Triphonmodellen wird die Dynamik der Merkmale beispielsweise für jedes Phonem abhängig von dem vorangehenden und dem nachfolgenden Phonem im Sprachsignal modelliert. Bei kontextabhängigen Phonemmodellen kann unterschieden werden, ob die Abhängigkeit von den benachbarten Phonemen nur innerhalb eines Wortes oder

über die Wortgrenzen hinweg berücksichtigt wird. Die zweite Möglichkeit führt zwar in der Regel zu konsistent besseren Erkennungsergebnissen, ist aber mit einem deutlich höheren Rechenaufwand für die Suche verbunden. Zur Zeit ist das in Abb. 2.2 dargestellte Markovmodell der etablierte Standard für die akustische Modellierung. Es

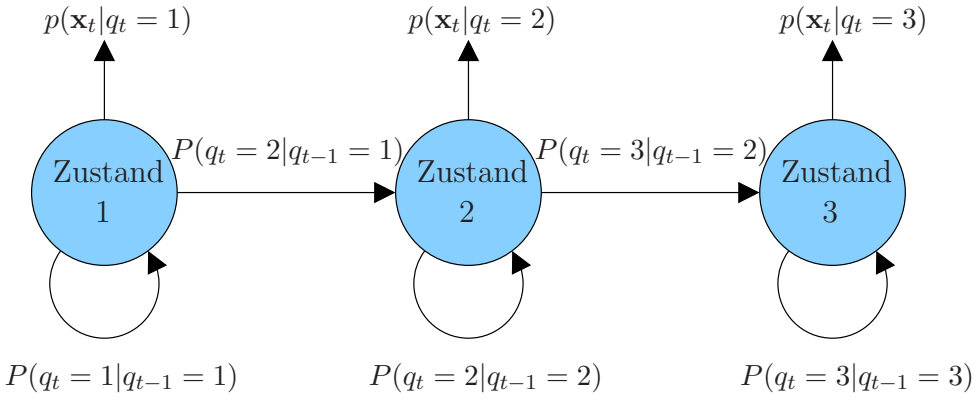

Abbildung 2.2: *Abbildung eines HMMs mit Links-Rechts-Topologie*

handelt sich dabei um einen stochastischen endlichen Automaten, dessen Zustände q_t stationäre Bereiche der Sprache charakterisieren [DO03]. Jedem Zustand ist eine akustische Emissionswahrscheinlichkeit $p(\mathbf{x}_t|q_t)$ für die Merkmale $\mathbf{x}_t$ zugeordnet. Dabei wird angenommen, dass eine statistische Abhängigkeit zwischen $\mathbf{x}_t$ und q_t besteht, während $\mathbf{x}_{t-1}$ und $\mathbf{x}_t$ nur indirekt über die HMM-Zustände voneinander abhängen. Zur Modellierung der Emissionswahrscheinlichkeiten sind vielfältige Möglichkeiten untersucht worden. Es können diskrete Wahrscheinlichkeitsverteilungen [Jel76], [Lip82], semi-kontinuierliche Wahrscheinlichkeitsverteilungen [Jel76], [Lip82] oder kontinuierliche Wahrscheinlichkeitsverteilungen [LRS83], [NN88] eingesetzt werden. Kontinuierliche Wahrscheinlichkeitsverteilungen können u.a. durch Kernel-Verteilungen, Neuronale Netze oder wie im Spracherkenner des Fachgebietes Nachrichtentechnik an der Universität Paderborn durch Mischungsverteilungen repräsentiert werden.

Zwischen den einzelnen Zuständen sind die Übergangswahrscheinlichkeiten $P(q_t|q_{t-1})$ definiert, für die in den folgenden Untersuchungen eine Links-Rechts-Topologie angenommen wird. Das bedeutet, dass abgesehen von Zustandsübergängen zwischen den Anfangs- und Endzuständen des HMMs nur Übergänge innerhalb eines HMM-Zustandes oder zwischen zwei unmittelbar aufeinander folgenden Zuständen erlaubt sind.

Die Integration des Markovmodells in das akustische Modell aus Gl. (2.1) läßt sich durch die Marginalisierung

$$p(\mathbf{x}_1^T|w_1^T) = \sum_{q_1^T} p(\mathbf{x}_1^T, q_1^T|w_1^T) \tag{2.5}$$

bzgl. der Zustandsfolge q_1^T des HMMs erreichen. Die Abhängigkeit von w_1^N wird dabei im Folgenden, wie in der Literatur üblich, weggelassen. Unter der Annahme eines

Markov-Prozesses erster Ordnung für die Zustandsvariable q_t und der statistischen Unabhängigkeit zwischen den Merkmalen $\mathbf{x}_1^T$ ergibt sich

$$\begin{aligned} p(\mathbf{x}_1^T|w_1^N) &= \sum_{q_1^T} p(\mathbf{x}_1^T, q_1^T) \\ &= \sum_{q_1^T} p(\mathbf{x}_1^T|q_1^T)P(q_1^T) \\ &= \sum_{q_1^T} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t)P(q_t|q_{t-1}). \end{aligned} \tag{2.6}$$

Gl. (2.6) kann mit Hilfe des Vorwärts-Rückwärts-Algorithmus ausgewertet werden [BP66]. Eine weniger aufwendige Lösung besteht darin, die sogenannte Viterbi-Approximation

$$\sum_{q_1^T} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t)P(q_t|q_{t-1}) \approx \max_{q_1^T} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t)P(q_t|q_{t-1}) \tag{2.7}$$

anzuwenden. Die Viterbi-Approximation ermöglicht es, Gl. (2.6) mit Methoden der dynamischen Programmierung auszuwerten [Bel57] [Vit67] (vgl. Abschnitt 2.1.3).

Der Standardalgorithmus für die Schätzung der Parameter der Markovmodelle ist der Baum-Welch-Algorithmus, in dem die gemeinsame Likelihood-Funktion von Trainingsdaten und Modellparametern mit Hilfe eines EM-Algorithmus maximiert wird [RJ93]. Bei einem großen Vokabular kann das Problem auftreten, dass nicht für alle HMM-Zustände eine ausreichende Anzahl an Trainingsdaten vorliegt. Für die im Rahmen dieser Arbeit durchgeführten Untersuchungen mit der AURORA4 Datenbank wird beispielsweise ein Vokabular mit 5000 Wörtern verwendet, für das sich mehrere Tausende verschiedener Triphonmodelle ergeben. Aus diesem Grund muß ein State-Tying durchgeführt werden, bei dem ähnliche Zustände verschiedener HMMs miteinander verknüpft werden.

Neben den in Abschnitt 2.1.1 beschriebenen Methoden, in denen die Sprachmerkmale an die Trainingsbedingungen angepaßt werden, besteht die Möglichkeit, den Unterschied zwischen Trainings- und Testbedingungen durch die Adaption der akustischen Modelle zu kompensieren. Beispiele sind die HMM-Dekomposition [VM90], die parallele Modellkombination (PMC) [Gal95], das MLLR (Maximum Likelihood Linear Regression)-Verfahren [LW95] und das JAC (Joint compensation of Additive and Convolutive distortions)-Verfahren [Gon03]. Die genannten Methoden ermöglichen zwar eine hohe Flexibilität bei der Adaption der Modelle an die Testbedingungen, weisen in der Regel aber eine wesentlich höhere Komplexität als die in Abschnitt 2.1.1 beschriebenen merkmalsbasierten Verfahren auf.

2.1.3 Suche

Die Aufgabe der Suche besteht darin, die Wortsequenz zu bestimmen, deren a posteriori Wahrscheinlichkeit für die gegebenen akustischen Merkmalsvektoren maximal ist.

Die Integration des HMMs aus Gl. (2.6) in die Bayes'sche Entscheidungsregel (Gl. (2.1)) führt auf das folgende Optimierungsproblem:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \left\{ \max_{q_1^T} \left\{ \prod_{t=1}^{T} p(\mathbf{x}_t|q_t, w_1^N) P(q_t|q_{t-1}, w_1^N) \right\} \prod_{n=1}^{N} P(w_n|w_{n-m+1}^{n-1})^{S_\beta} \right\} \tag{2.8}$$

Das Optimierungsproblem kann mit Hilfe dynamischer Programmierung effizient gelöst werden. Verbreitete Methoden sind die Viterbi-Suche und die A*-Suche. Das Prinzip der A*-Suche besteht darin, die noch nicht expandierten Teile der Zustandshypothesen optimistisch abzuschätzen und jeweils den Suchbaum der vielversprechendsten Hypothese weiter aufzuspannen. Bei der im Spracherkennungssystem an der Universität Paderborn eingesetzten Viterbi-Suche werden die Hypothesen dagegen zeitsynchron expandiert und ausgewertet. Die Suche wird an dieser Stelle für ein Bigram-Sprachmodell, das in den Untersuchungen mit der AURORA4 Datenbank verwendet wird, kurz beschrieben. Wie zu Beginn dieses Kapitels dargestellt, wird das Lexikon als Baum modelliert, der die akustischen Einheiten des Erkenners beinhaltet und dessen Blätter einzelnen Wörtern zugeordnet werden können. Bei der Suche werden Kopien des Lexikons für die möglichen Vorgänger v des aktuellen Wortes w angelegt. Die Verwendung von Baumkopien ermöglicht es, die Sprachmodellhistorie bei der Suche zu berücksichtigen. Den Zuständen q_t der Kopien v wird die Vorwärtsvariable

$$\alpha_t(v, q_t) = \max_{q_1^{t-1}} P(q_1^t, \mathbf{x}_1^t|v) \tag{2.9}$$

zugeordnet, die den Score des besten Pfades für die beobachtete Merkmalssequenz $\mathbf{x}_1^t$, der zum Zeitpunkt t im Zustand q_t der Baumkopie v endet, enthält. Daneben werden die Wortgrenzen $B_t(v, q_t)$ der entsprechenden Pfade festgehalten. Die Aktualisierung der akustischen Scores und Wortgrenzen ist in Abb. 2.3 schematisch für ein Vokabular mit den Wörtern A, B, C und Sil dargestellt.

In jedem Zeitschritt t werden innerhalb der Baumkopien v die Vorwärtsvariable $\alpha_{t-1}(v, q_{t-1})$ und die Wortgrenzen $B_{t-1}(v, q_{t-1})$ zunächst für die Startzustände $q_t = 0$ mit den Werten

$$\begin{aligned} \alpha_{t-1}(v, q_{t-1} = 0) &= H(v; t-1) \\ B_{t-1}(v, q_{t-1} = 0) &= t-1 \end{aligned} \tag{2.10}$$

initialisiert, wobei sich $H(v; t-1)$, wie weiter unten dargestellt, am Ende des Zeitschrittes $t-1$ ergibt. Anschließend werden $\alpha_t(v, q_t)$ und $B_t(v, q_t)$ für $q_t > 0$ mit der Vorwärtsiteration eines Viterbi-Algorithmus entsprechend der Transitions- und Emissionswahrscheinlichkeiten in den akustischen Modellen aktualisiert:

$$\begin{aligned} \alpha_t(v, q_t) &= \max_{q_{t-1}} \left\{ \alpha_{t-1}(v, q_{t-1}) P(q_t|q_{t-1}) \right\} p(\mathbf{x}_t|q_t) \\ B_t(v, q_t) &= B_{t-1}\left(v, \underset{q_{t-1}}{\operatorname{argmax}} \left\{ \alpha_{t-1}(v, q_{t-1}) P(q_t|q_{t-1}) \right\} \right) \end{aligned} \tag{2.11}$$

Zur Reduktion der Rechenzeit und des benötigten Speichers wird ein Pruning unwahrscheinlicher Hypothesen durchgeführt.

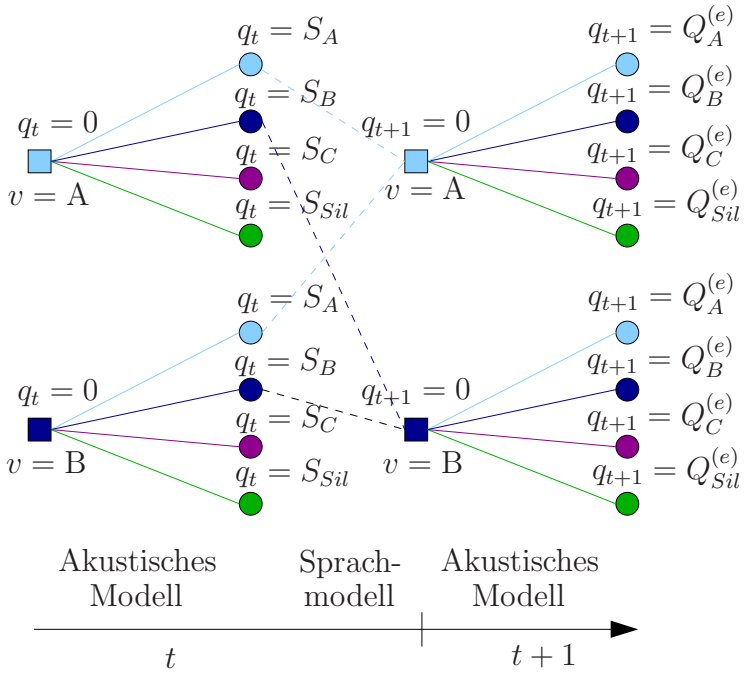

Abbildung 2.3: *Aktualisierung der Zustandswahrscheinlichkeiten (nach [ONA97])*

Nach der Aktualisierung der Baumkopien wird mit Hilfe des Sprachmodells für den Endzustand $Q_w^{(e)}$ jedes Wortes der beste Vorgänger

$$v_0(w;t) = \underset{v}{\operatorname{argmax}} \left\{ P(w|v)^{S_\beta} \alpha_t(v, q_t = Q_w^{(e)}) \right\} \tag{2.12}$$

in allen Baumkopien bestimmt und zusammen mit der Wortgrenze $B_t(v_0(w;t), q_t = Q_w^{(e)})$ gespeichert. Daneben werden die Vorwärtsvariablen

$$H(w;t) = \max_{v} \left\{ P(w|v)^{S_\beta} \alpha_t(v, q_t = Q_w^{(e)}) \right\} \tag{2.13}$$

berechnet, mit denen im nächsten Zeitschritt die Vorwärtsvariablen $\alpha_t(v, q_t = 0)$ initialisiert werden. Die Übergangswahrscheinlichkeiten $P(w|v)$ werden bei integrierter Sprachmodellvorschau bereits bei dem Eintritt in das letzte HMM eines Wortes w in dem akustischen Score $\alpha_t(v, q_t)$ berücksichtigt. Am Satzende wird mit Hilfe der gespeicherten Vorgänger $v_0(w;t)$ und Wortgrenzen $B_t\left(v_0(w;t), q_t = Q_w^{(e)}\right)$ die wahrscheinlichste Wortsequenz zurückverfolgt.

In einigen Anwendungen ist es von Bedeutung, nicht nur die wahrscheinlichste Wortsequenz zu ermitteln, sondern mehrere Wortsequenzen mit hohen Wahrscheinlichkeiten zu bestimmen. Ein Beispiel ist die Multiple-Pass-Suche, in der die Erkennung in mehreren Durchgängen abläuft. Im ersten Durchgang werden ein einfaches Akustik- und Sprachmodell verwendet, um die besten Wortsequenzen schnell ermitteln zu können. Diese Wortsequenzen können dann in einem oder mehreren weiteren Durchgängen mit rechenintensiveren Methoden verarbeitet werden. Zur Repräsentation der Worthypothesen existieren im Wesentlichen zwei Möglichkeiten, nämlich N-Best-Listen und Wortgraphen. In N-Best-Listen werden die N wahrscheinlichsten Wortsequenzen einzeln abgespeichert. In Wortgraphen werden gleiche Teile verschiedener Wortsequenzen

zusammengefaßt, so dass ein gerichteter, azyklischer Graph (DAG) entsteht, der eine kompaktere Darstellung der Wortsequenzen ist. Die Erzeugung von Wortgraphen wird in Kapitel 8 behandelt.

2.2 Ausnutzung von Inter-Frame Korrelationen

Ein vielzitierter Schwachpunkt des Markovmodells, das bei der akustischen Sprachmodellierung eingesetzt wird, besteht darin, dass die statistische Abhängigkeit zwischen den Sprachmerkmalen $\mathbf{x}_{t-1}$ und $\mathbf{x}_t$ aufeinander folgender Sprachrahmen nur indirekt über die HMM-Zustände q_{t-1} und q_t modelliert wird (Conditional-Independence-Annahme). Eine Möglichkeit, diese Abhängigkeit direkt zu beschreiben, stellen schaltende lineare Dynamikmodelle (SLDMs) dar. In diesem Modelltyp wird die Emissionswahrscheinlichkeit $p(\mathbf{x}_t|q_t)$ des HMMs durch die Übergangswahrscheinlichkeit $p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_t)$ ersetzt. Dabei ist zu beachten, dass ein konzeptioneller Unterschied zwischen den HMM-Zuständen und den SLDM-Zuständen besteht. Während die SLDM-Zustände Bereiche gleicher Dynamik innerhalb des Sprachsignals beschreiben, können die HMM-Zustände stationären Signalbereichen zugeordnet werden. Das segmentbasierte HMM ist ein anderer, in der Sprachverarbeitung bereits früher zur Modellierung von Inter-Frame Korrelationen verwendeter Ansatz. Es handelt sich dabei um eine Hybridform zwischen dem HMM und dem SLDM. In segmentbasierten HMMs wird die Dynamik der Sprachmerkmale für Segmente, in denen $q_{t-1} = q_t$ gilt, mit Trajektorienmodellen beschrieben, die es ermöglichen die Conditional-Independence-Annahme zu überwinden. An den Segmentgrenzen, d.h. für $q_{t-1} \neq q_t$, kann das dynamische Verhalten genauso wie im HMM mit den Übergangswahrscheinlichkeiten $P(q_t|q_{t-1})$ modelliert werden (State-Reset-Modelle). Daneben existieren segmentbasierte Ansätze, in denen die Trajektorienmodelle auch für $q_{t-1} \neq q_t$ definiert sind (State-Passed-Modelle), die Segmentgrenzen bei der Suche der optimalen Wortfolge jedoch in der Regel eine andere Behandlung als Bereiche innerhalb der Segmente erfahren.

Auch bei der Entstörung der Sprachmerkmale im Front-End des Erkenners können Inter-Frame Korrelationen berücksichtigt werden, wozu SLDMs einen erfolgversprechenden Ansatz darstellen (Abschnitt 2.2.2). Im Folgenden werden die versteckten Zustände unabhängig von dem zugrunde gelegten Dynamikmodell im Front-End mit s_t und im Back-End mit q_t bezeichnet.

In den folgenden beiden Unterabschnitten wird ein Überblick über Literaturansätze zur Ausnutzung von Inter-Frame Korrelationen bei der akustischen Sprachmodellierung im Back-End des Erkenners bzw. bei der Merkmalsentstörung im Front-End gegeben.

2.2.1 Ausnutzung von Inter-Frame Korrelationen bei der Suche der optimalen Wortsequenz

Segmentbasierte HMMs stellen eine Möglichkeit dar, die Conditional-Independence-Annahme zu überwinden, indem die Sprachdynamik innerhalb einzelner Sprachseg-

mente mit Trajektorienmodellen beschrieben wird. In [SG07] werden segmentbasierte HMMs in Hinblick auf die zugrunde gelegten Trajektorienmodelle in polynomielle Segmentmodelle, Buried Markovmodelle (BMMs), stochastische Segmentmodelle und Trajektorien-HMMs unterteilt. [SG07] klassifiziert auch SLDMs als segmentbasierte HMMs. Das BMM [Bil03] enthält zusätzlich zu den HMM-Zuständen eine verdeckte Variable, deren Zustand das Trajektorienmodell beeinflußt. In stochastischen Segmentmodellen [OR89] wird für die Beobachtungen $\mathbf{y}_{t_0}^{t_0+L-1}$ innerhalb eines Segmentes $p_{[t_0,t_0+L-1]}$ der Länge L mit Segmentbeginn t_0, das z.B. einem Phonem zugeordnet werden kann, die gemeinsame Emissionswahrscheinlichkeit $p(\mathbf{y}_{t_0}^{t_0+L-1}|p_{[t_0,t_0+L-1]})$ definiert. In diesem Ansatz wird die Merkmalsvektortrajektorie innerhalb des Segmentes in der Regel mit klassischen linearen Zustandsmodellen beschrieben. Beispiele für polynomielle Segmentmodelle sind [Dig92, MD04, FK07]. [Dig92] setzt phonemabhängige Zustandsmodelle der Form

$$\mathbf{x}_t = \mathbf{A}(p_t, I_t)\mathbf{x}_{t-1} + \mathbf{u}_t, \quad \mathbf{y}_t = \mathbf{H}(p_t, I_t)\mathbf{x}_t + \boldsymbol{\mu}(p_t, I_t) + \mathbf{w}_t, \tag{2.14}$$

ein, wobei für jedes Phonem p_t abhängig von der Verweildauer, die in Gl. (2.14) mit Hilfe der Indexfunktion I_t modelliert wird, eine festgelegte Sequenz von Modellen durchlaufen wird. In Gl. (2.14) wird, wie auch in den anderen in diesem Abschnitt beschriebenen Literaturansätzen, angenommen, dass die beobachteten Merkmale $\mathbf{y}_t$ verrauscht sind und über ein Messmodell mit den Zuständen $\mathbf{x}_t$ verknüpft sind. Die Modellparameter $\mathbf{A}(p_t, I_t)$, $\mathbf{H}(p_t, I_t)$ und $\boldsymbol{\mu}(p_t, I_t)$ sowie die Parameter des Zustandsrauschens $\mathbf{u}_t$ und des Messrauschens $\mathbf{w}_t$ werden mit Hilfe eines EM-Algorithmus aus Trainingsdaten bestimmt. Da anders als bei segmentabhängigen, schaltenden Modellen für jedes Segment nur eine mögliche Parametersequenz, also auch nur ein möglicher Wert der akustischen Likelihood, existiert, kann die Segmentierung und die Erkennung mit Hilfe dynamischer Programmierung durchgeführt werden. [MD04] verwendet für jedes Phonem p_t schaltende Modelle der Form

$$\mathbf{x}_t = \phi(p_t)\mathbf{x}_{t-1} + (1 - \phi(p_t))\mathbf{b}(p_t) + \mathbf{u}_t, \quad \mathbf{y}_t = \mathbf{H}(p_t)\mathbf{x}_t + \mathbf{w}_t, \tag{2.15}$$

die auch an den Segmentgrenzen definiert sind. Dabei bezeichnet $\mathbf{b}(p_t)$ einen phonemspezifischen Zielzustand, der für $t \to \infty$ erreicht wird, während $\phi(p_t) < 1$ die Konvergenzgeschwindigkeit gegen diesen Zustand beeinflußt. Bei der Decodierung der Sprache wird die Anzahl der möglichen Trajektorien durch ein Mixture-Path-Constraint, das für jedes Phonem nur eine einzelne Trajektorie zuläßt, begrenzt. Weiterhin wird ein exponentieller Anstieg der Trajektorienzahl durch mögliche Verzweigungen an den Phonemgrenzen dadurch verhindert, dass die Gaußverteilungen der M möglichen Trajektorien für jedes Phonem an den Übergängen jeweils zu einer einzelnen Gaußverteilung verschmolzen werden. [FK07] setzt phonemspezifische Zustandsmodelle der Form

$$\mathbf{x}_t = \mathbf{A}(p_t)\mathbf{x}_{t-1} + \mathbf{u}_t, \quad \mathbf{y}_t = \mathbf{H}(p_t)\mathbf{x}_t + \mathbf{w}_t \tag{2.16}$$

ein. Durch die Verwendung einer nichtquadratischen Beobachtungsmatrix $\mathbf{H}(p_t)$ des Meßmodells wird eine lineare Dimensionsreduktion des Zustandsraumes gegenüber dem

von den möglichen Beobachtungsvektoren aufgespannten Vektorraum durchgeführt. Bei der Erkennung mit dem beschriebenen Modell werden in [FK07] mit State-Passed-Modellen, in denen die Zustandsinformationen über die Phonem-Grenzen hinweg übergeben werden, bessere Ergebnisse erzielt als durch das Zurücksetzen der Zustandsinformationen an den Phonem-Grenzen. Der exponentielle Anstieg der Hypothesenzahl wird dabei mittels Pruning verhindert. Weitere Verbesserungen werden durch den Einsatz von Multi-Regime (MR)-Modellen erzielt, in denen die Zustandsmodelle sequentiell entsprechend der Verweildauer in dem Phonem ausgewählt werden.

In Trajektorien-HMMs [RH97, TZK03] wird die Trajektorie der Sprachmerkmale innerhalb des Segmentes $p_{[t_0,t_0+L-1]}$ explizit als Funktion der Verweildauer $t - t_0 + 1$ in dem Segment sowie der Beobachtungen $\mathbf{y}_{t_0}^{t_0+L-1}$ angegeben. [RH97] beschreibt die Dynamik der Sprachmerkmale z.B. mit einer linearen Trajektorie der Form

$$\mathbf{x}_t = \mathbf{x}_c(p_t) + \mathbf{m}(p_t)(t - t_0 - L/2), \tag{2.17}$$

wobei die Parameter $\mathbf{x}_c(p_t)$ und $\mathbf{m}(p_t)$ sowohl von allen beobachteten Merkmalen $\mathbf{y}_t$ in dem betrachteten Segment als auch von Parametern, die im Training bestimmt werden, abhängen. In [TZK03] wird ein Trajektorien-HMM eingesetzt, das sich durch die Überführung eines HMMs mit dynamischen Merkmalskomponenten in ein segmentbasiertes HMM, das nur statische Komponenten besitzt, ergibt.

[SZD03] verwendet ein HTHMM (Hidden Trajectory HMM) zur akustischen Modellierung, in dem den HMM-Zuständen neben den Mischungsverteilungen eine versteckte Trajektorie $\mathbf{G}_t$ zugeordnet ist. Die Likelihood der Beobachtungen wird in diesem Ansatz abhängig von der versteckten Trajektorie $\mathbf{G}_t$ und den Mischungsgewichten m_t des HTHMMs berechnet. Eine effiziente Suche mit Methoden der dynamischen Programmierung wird durch die Quantisierung von $\mathbf{G}_t$ erreicht, die das Training diskreter Übergangswahrscheinlichkeiten ermöglicht.

[Ros04] untersucht SLDMs und segmentbasierte Modelle, in denen das Verhalten innerhalb der einzelnen Segmente jeweils mit einem SLDM beschrieben wird. Ein Problem bei diesen Ansätzen besteht darin, dass das Maximum-Likelihood-Training der Parameter nicht effizient durchgeführt werden kann, da dazu die Berechnung von $p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_t)$ erforderlich ist, wobei alle in der Vergangenheit liegenden Pfade berücksichtigt werden müssen. In [Ros04] werden die Modellparameter durch ein Gibbs-Sampling approximiert. Mit dem in [Ros04] eingesetzten SLDM ergeben sich jedoch deutlich schlechtere Erkennungsraten als mit einem HMM mit vergleichbarer Parameterzahl.

Insgesamt haben sich schaltende lineare Dynamikmodelle und segmentielle HMMs bislang nicht als konkurrenzfähige Alternative zum HMM erwiesen, da sie in der Regel eine deutlich größere Parameteranzahl als ein HMM mit der gleichen Zustandszahl aufweisen, Approximationen bei der Decodierung der besten Zustandsfolge erfordern und in vielen Ansätzen kein zuverlässiges Maximum-Likelihood-Training der Modellparameter durchgeführt werden kann.

2.2.2 Ausnutzung von Inter-Frame Korrelationen bei der Merkmalsentstörung

Bei der einkanaligen Sprachsignalentstörung wurden Inter-Frame Korrelationen zunächst mit dem Ziel ausgenutzt, die perzeptuelle Sprachqualität zu verbessern. Eine ausführliche Darstellung dieser Ansätze wird z.B. in [BMC05] gegeben. Basierend auf dem AR-Modell, das zuerst von Lim und Oppenheimer [LO78] zur Sprachverbesserung eingesetzt wurde, verwendeten Paliwal und Basu [PB87] erstmals ein Kalman-Filter, um die Sprache zu entrauschen. Bei der Kalman-Filterung im Zeitbereich ist die exakte Bestimmung der AR-Parameter von großer Bedeutung. Diese kann entweder aus Trainingsdaten oder zur Laufzeit erfolgen. Ein bekanntes Beispiel für einen Online-Algorithmus ist ein in [GBW98] eingeführter EM-Algorithmus, in dem zwischen der Parameterschätzung und einer linearen Filterung iteriert wird. Durch die explizite Modellierung der AR-Parameter als Zufallsvariable ergeben sich nichtlineare Zustandsmodelle, die approximative Filtermethoden wie das erweiterte Kalman-Filter, das Unscented Kalman-Filter [GM03] oder Partikelfilteransätze [FG01, FGDW02, VADG02, WHU06b, WHU06a] erfordern. Neben den genannten Online-Methoden existieren auch Verfahren, in denen ein autoregressives Markovmodell (AR-HMM) für das unverrauschte Sprachsignal aus Trainingsdaten geschätzt und das Signal durch die Anwendung eines Bayes'schen Schätzers aus dem verrauschten Signal bestimmt wird. AR-HMMs wurden für die Entstörung von Sprachsignalen erstmals von Ephraim u.a. ([EMJ89], [Eph92]) eingesetzt. Die Verwendung interagierender Kalman-Filter auf der Grundlage von AR-HMMs wurde beispielsweise in [LS96] und [KLL00] untersucht. Neben den oben genannten Zeitbereichsverfahren existieren auch Methoden, in denen die Kalman-Filterung im Frequenzbereich durchgeführt wird [FA00, ZVY06]. Durch die Ausnutzung von Inter-Frame Korrelationen bei der Filterung des Sprachsignals konnte mit den genannten Ansätzen aus der Literatur eine hohe perzeptuelle Sprachqualität erzielt werden. Bei einer anschließenden Spracherkennung mit den auf diese Weise entstörten Sprachsignalen wurde die Erkennungsleistung anderer Verfahrensklassen wie beispielsweise der spektralen Wiener-Filterung bislang jedoch nicht erreicht.

Erfolgversprechende Ergebnisse wurden in den letzten Jahren hingegen trotz des hochgradig nichtlinearen Beobachtungsmodells durch die Ausnutzung von Inter-Frame Korrelationen bei der modellbasierten Entrauschung der Sprachmerkmale im Log-Spektrum bzw. Cepstrum erreicht. Die Inter-Frame Korrelationen werden in diesen Ansätzen entweder mit SLDMs ([DA04], [KLS05], [SVhW05b], [DBY06]) oder durch die Integration dynamischer Merkmale erster Ordnung in ein GMM [DDA03b] modelliert. In [DA04] wird ein SLDM zur Entrauschung der cepstralen Sprachmerkmale eingesetzt. Dabei wird ein GPB-Algorithmus angewendet, um die a posteriori Wahrscheinlichkeiten der Sprachmerkmale zu berechnen. In diesem Ansatz wird für jedes der linearen Dynamikmodelle ein Filter eingesetzt, in dem eine approximative MMSE-Schätzung der Sprachmerkmale berechnet wird. [DA04] verwendet zur MMSE-Schätzung in den einzelnen Filtern ein Verfahren aus [DDA03b], das auf der Iteration einer SNR-Variable basiert. Die Schätzwerte der einzelnen Filter werden anschließend entsprechend der Mo-

dellwahrscheinlichkeiten der zugehörigen Dynamikmodelle zu einer einzelnen Schätzung kombiniert. Die Parameter des Rauschens werden in [DA04] als konstant angenommen und aus sprachfreien Signalausschnitten trainiert. In [KLS05] wird der stationäre Anteil des Sprachsignals mit einer versteckten, nicht beobachtbaren Zustandsvariable im Log-Spektrum modelliert, deren Dynamik mit einem SLDM beschrieben wird. Da sowohl die versteckte Zustandsvariable als auch die Modellwahrscheinlichkeiten des SLDMs nicht beobachtbar sind, erfordert dieser Ansatz Heuristiken beim Training der Modellparameter. Die Rauschdynamik wird in [KLS05] mit einem Random-Walk-Zustandsmodell beschrieben. Um den Zusammenhang zwischen den unverrauschten Sprachmerkmalen, den verrauschten Merkmalen und dem Rauschen zu modellieren, werden stückweise lineare Beobachtungsmodelle herangezogen. Die a posteriori Verteilung des Rauschens wird mit einem IMM (Interacting Multiple Model)-Algorithmus gemeinsam mit der Verteilung der unverrauschten Sprachmerkmale berechnet. Verglichen mit [DA04] werden in [KLS05] bessere Ergebnisse bei instationärem Rauschen erzielt, während die Ergebnisse bei stationärem Rauschen schlechter sind. [DBY06] verwendet ebenfalls ein SLDM im Log-Spektrum, nimmt im Gegensatz zu [DA04] und [KLS05] jedoch an, dass die Modellwahrscheinlichkeiten des SLDMs in aufeinander folgenden Sprachrahmen statistisch abhängig sind. Das Zustandsmodell für die Dynamik des Rauschens wird aus Trainingsdaten bestimmt, die keine Sprache enthalten. Insgesamt führen diese Modifikationen auf einer Zifferndatenbank der Oregon Health & Science Universität zu besseren Ergebnissen als die Kombination des Zustandsmodells aus [DA04] mit dem Beobachtungsmodell aus [KLS05]. Eine weitere Verbesserung der Erkennungsergebnisse wird bei niedrigem Eingangs-SNR durch die Verwendung eines Markovmodells zweiter Ordnung für die Sprachdynamik erzielt. In [SVhW05b] werden parallele, nicht miteinander interagierende erweiterte Kalman-Filter im Cepstrum eingesetzt, deren Parameter durch eine Vektorquantisierung der Trainingsdaten ermittelt werden. Dabei wird die Rauschschätzung mit Hilfe einer Minimum-Statistik im Spektrum durchgeführt. Ein direkter Vergleich der Ergebnisse mit den zuvor beschriebenen Verfahren ist nicht möglich, da die in [SVhW05b] verwendete Baseline wesentlich höher als in den anderen Verfahren ist. Verglichen mit einem AR-Modell der Ordnung Null ergibt sich durch die Berücksichtigung von Inter-Frame Korrelationen in [SVhW05b] jedoch eine signifikante Erhöhung der Erkennungsrate. Die Erkennungsergebnisse können in [Sto06] durch die Verwendung von AR-Modellen höherer Ordnung weiter verbessert werden, wobei jedoch Schwierigkeiten hinsichtlich der Komplexität und Robustheit des Verfahrens auftreten. Durch den Einsatz von Unscented Kalman-Filtern ergibt sich in [Sto06] für ein AR-Modell der erster Ordnung keine signifikante Verbesserung gegenüber der Verwendung nicht-iterierter erweiterter Kalman-Filter gleicher Modellordnung. Bei einer höheren Modellordnung führt das UKF auf Kosten einer längeren Rechenzeit allerdings zu konsistent besseren Ergebnissen als das nicht-iterierte EKF.

In [DDA04b] wird das ALGONQUIN-Verfahren um die Modellierung von Inter-Frame Korrelationen erweitert, indem eine bedingte MMSE-Schätzung bei gegebenem Merkmalsvektor des letzten Sprachrahmens durchgeführt wird. Dabei wird neben der a priori Verteilung der statischen Sprachmerkmale, die als GMM modelliert wird, auch

ein GMM für die a priori Verteilung der dynamischen Merkmale trainiert, wobei beide Verteilungen als statistisch unabhängig angenommen werden. Im Gegensatz zum SLDM weist der beschriebene Ansatz zwei Schwachpunkte bzgl. der Modellierung der Sprachdynamik auf. Zum Einen wird das Dynamikmodell in Form der dynamischen Sprachmerkmale ohne die Berücksichtigung aktueller Messungen a priori aus Trainingsdaten festgelegt. Zum Anderen wird für den Merkmalsvektor des letzten Sprachrahmens eine Punktschätzung herangezogen, also keine Unsicherheit berücksichtigt. Trotz der genannten Nachteile führt die Modellierung von Inter-Frame Korrelationen in diesem Ansatz gegenüber der ausschließlichen Verwendung einer a priori Verteilung für die statischen Sprachmerkmale zu besseren Erkennungsergebnissen.

Ein weiterer Ansatz, der bereits erfolgreich bei der Kompensation von Paketverlusten in der verteilten Spracherkennung umgesetzt wurde [IHU08b], besteht in einer Quantisierung des Zustandsraumes, die es ermöglicht, diskrete Übergangswahrscheinlichkeiten zwischen den Quantisierungsklassen aufeinander folgender Sprachrahmen zu trainieren. Zur Merkmalsentstörung hat sich dieser Ansatz, der in [IHU08a] veröffentlicht wird, jedoch bislang als weniger robust als die Verwendung schaltender Modelle erwiesen, so dass in [IHU08a] nur Ergebnisse bei bekannter Verteilung des Rauschens angegeben werden können.

Die Erkennungsleistung des AFE konnte auf der AURORA2-Datenbank unter den in [ETS05] standardisierten Bedingungen durch die modellbasierte Entrauschung der mit dem SFE extrahierten Sprachmerkmale mit den oben genannten Verfahren bislang nicht erreicht werden. Allerdings ergibt sich ein konsistenter Rahmen, in dem sowohl im Front-End als auch im Back-End die Sprachmerkmale im cepstralen Bereich bzw. dem damit linear verknüpften log-spektralen Bereich statistisch modelliert werden.

In den nächsten beiden Abschnitten werden zwei weitere wichtige Aspekte für die im Rahmen dieser Arbeit durchgeführten Untersuchungen behandelt, nämlich die Bestimmung der Parameter des Rauschmodells sowie der Austausch von Informationen zwischen dem Front-End und dem Back-End.

2.3 Bestimmung der Parameter des Rauschmodells

Ein wichtiger Aspekt bei der Entrauschung der Sprachmerkmale ist die Bestimmung des Rauschens. In einigen Anwendungen ist es möglich, die Parameter des Rauschmodells aus sprachfreien Signalbereichen zu schätzen. Dabei kann das Rauschen als stationär angenommen werden oder wie in [STBP01] zwischen den Rauschparametern am Anfang und Ende des Satzes interpoliert werden. Die Verwendung einer Sprachaktivitätsdetektion (VAD), falls die sprachfreien Signalbereiche nicht a priori bekannt sind, hat sich insbesondere bei schlechtem Eingangs-SNR als unzuverlässig herausgestellt [MW97]. Bei instationärem Rauschen ist die Bestimmung des Rauschens aus sprachfreien Signalbereichen mit dem weiteren Nachteil verbunden, dass die Rauschschätzung nicht ständig aktualisiert wird. Eine mögliche Lösung für das zweite Problem stellt der Einsatz einer VAD mit weicher Entscheidung dar [SS98], [MCA99], [AH04],

deren Zuverlässigkeit allerdings ein kritischer Punkt in Hinblick auf die Qualität der Rauschschätzung bleibt.

[HE95] und [RD01] führen die Schätzung des Rauschens im Energiespektrum auf der Grundlage von Histogrammen durch. Die vorgeschlagenen Methoden erfordern jedoch einen großen Speicher- und Rechenaufwand und liefern bei einem niedrigen SNR ebenfalls unzuverlässige Schätzwerte.

Eine weitere Möglichkeit, die Parameter des Rauschens zu bestimmen, besteht in der Anwendung einer Minimum-Statistik [Mar94], [Mar01]. In diesem Ansatz werden die Minima des geglätteten Energiedichtespektrums innerhalb vorgegebener Fenster der einzelnen Frequenzbänder bestimmt. Dabei wird die Annahme zugrunde gelegt, dass das verrauschte Sprachsignal innerhalb der einzelnen Frequenzbänder auch während aktiver Sprachphasen auf den Rauschpegel zurückfällt. Die Minima werden nach einer Bias-Korrektur als Schätzwerte für den spektralen Rauschpegel verwendet. Eine verwandte Methode ist die in [CB01] eingeführte Minima-Controlled-Recursive-Averaging (MCRA)-Technik. In der MCRA-Technik wird die Rauschschätzung als gewichtete Summe des aktuellen Wertes des geglätteten Energiedichtespektrums und der Rauschschätzung für den letzten Sprachrahmen berechnet. Dabei wird aus den Minima des geglätteten Energiedichtespektrums der bei der Gewichtung eingesetzte Glättungsfaktor berechnet. Das IMCRA-Verfahren [CB02] stellt eine Erweiterung der MCRA-Technik dar, in der das Tracking der Minima und die Glättung des Energiedichtespektrums in zwei Iterationen durchgeführt werden. Ein schnelleres Tracking des Rauschpegels wurde in [RL06] durch eine andere Strategie bei der Berechnung der Sprachwahrscheinlichkeit und dem Tracking der Minima erreicht.

In [HJH07] wird zur Rauschschätzung eine Eigenwertzerlegung der Korrelationsmatrix zwischen den verrauschten DFT-Komponenten durchgeführt, wodurch eine schnelle Adaption der Schätzung erreicht wird. Dabei wird die spektrale Rauschvarianz als Mittelwert der kleinsten Eigenwerte der Korrelationsmatrix geschätzt, wobei angenommen wird, dass die entsprechenden Eigenwerte der Korrelationsmatrix des unverrauschten Sprachsignals den Wert Null haben. Die Anzahl der bei der Rauschschätzung berücksichtigten Eigenwerte ergibt sich in dieser Methode aus der Rauschschätzung des letzten Sprachrahmens.

Die Transformation der spektralen Rauschschätzungen ins Cepstrum wird in [SVhW06a] untersucht. Dabei ist eine Heuristik für die Varianz des Rauschens erforderlich, falls die spektralen Methoden nur Punktschätzungen des Rauschens liefern. In [SVhW06a] wird die Varianz der spektralen Rauschschätzung proportional zum Quadrat des Rauschens gewählt.

Neben den genannten Verfahren existiert eine Reihe von Ansätzen, um die Rauschschätzung direkt im Log-Spektrum oder im Cepstrum durchzuführen. Diese können in Methoden, in denen die Parameter des Rauschens mit einem EM-Algorithmus bestimmt werden, sowie Verfahren, in denen die Dynamik des Rauschens mit einem Zustandsmodell beschrieben wird, unterteilt werden. Ein Block-EM-Algorithmus, in dem gleichzeitig die Parameter des additiven und des Faltungsrauschens ermittelt werden, wird bereits in dem in [Mor96] eingeführten VTS-Verfahren verwendet. In [KKKK97] wird die

Rauschschätzung, die sich aus dem EM-Algorithmus ergibt, mit der a priori Verteilung des Rauschens aus repräsentativen Trainingsdaten kombiniert. Eine Modifikation des EM-Algorithmus wird in [FA04] untersucht. In diesem Ansatz werden für Sprache und Sprachpausen verschiedene GMMs trainiert. Das additive Rauschen wird in Sprachpausen aktualisiert, die sich aus einer harten VAD-Entscheidung auf der Grundlage der Modellwahrscheinlichkeiten der beiden GMMs ergeben, während die Parameter des Beobachtungsmodells, die das Faltungsrauschen modellieren, mit einem EM-Algorithmus geschätzt werden. Daneben sind auch sequentielle EM-Algorithmen entwickelt worden, die eine kausale Schätzung der Rauschparameter ermöglichen, indem die Parameterschätzung für jeden neuen Sprachrahmen aktualisiert wird [Kim98b, DDA03a, Afi05].

In den bereits in Abschnitt 2.2.2 angeführten Verfahren von Frey und Droppo [FDAK03, DA04] wird die a posteriori Verteilung des Rauschens für jeden Sprachrahmen einzeln innerhalb eines Bayes'schen Ansatzes optimiert.

Eine weitere Möglichkeit besteht in einer auf der Verwendung von Dynamikmodellen basierenden sequentiellen Schätzung des Rauschens. Dynamische Systeme wurden im Bereich der Spracherkennung erstmals in [VM90] zur Modellierung des Rauschens eingesetzt, wobei das Rauschen als Ausgangssignal eines HMMs modelliert wurde. In [Kim98a] wird das Rauschen mit einem Zustandsmodell der Form

$$\mathbf{n}_t = \mathbf{n}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{v}_t; \mathbf{0}, \mathbf{V}) \tag{2.18}$$

beschrieben, dessen Kovarianzmatrix $\mathbf{V}$ experimentell bestimmt wird. In diesem Ansatz wird ein IMM zur Zustandsschätzung eingesetzt, wobei die Sprache als Mischungsverteilung dargestellt wird, deren Komponenten jeweils ein stückweise lineares Beobachtungsmodell zugeordnet ist. Die Untersuchungen in [Kim02] zeigen jedoch, dass die Dynamik des Rauschens auf diese Weise nicht adäquat modelliert wird. Aus diesem Grund wird in [Kim02] eine theoretisch nicht motivierte Korrektur der Kalman-Verstärkung durch die Multiplikation mit einem konstanten Faktor durchgeführt. Daneben werden in [Kim02] die adaptive Schätzung von $\mathbf{V}$ sowie eine approximative Glättung der Zustandsschätzung innerhalb des IMM-Ansatzes untersucht. [SR03] verwendet ein AR-Modell erster Ordnung:

$$\mathbf{n}_t = \mathbf{D}\mathbf{n}_{t-1} + \boldsymbol{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{v}_t; \mathbf{0}, \mathbf{V}). \tag{2.19}$$

Dabei werden die Modellparameter $\mathbf{D}$ und $\mathbf{V}$ aus Trainingsdaten, die repräsentativ für die eingesetzten Rauschtypen sind, bestimmt. In diesem Ansatz, in dem die Verteilung der Sprache ebenfalls als GMM modelliert wird, wird die Linearisierung des Beobachtungsmodells durch die Verwendung eines SIR-Partikelfilters vermieden. Die Erweiterung auf ein AR-Modell höherer Ordnung wird in [RSS04] untersucht, wodurch jedoch keine signifikante Verbesserung erzielt wird.

In [YL04, SHU05, FN05, FN06] werden auf der Grundlage des in [SR03] und [RSS04] verwendeten Ansatzes verschiedene Partikelfiltertechniken untersucht, um die a posteriori Verteilung des Rauschens zu berechnen. Ein IMM-Algorithmus mit interagierenden Unscented Kalman-Filtern wird in [DDA05] angewendet.

Wie in Abschnitt 2.2.2 angeführt, wurden einige der genannten Methoden auch in Verbindung mit schaltenden Modellen für die Dynamik der Sprachmerkmale erprobt. Während [SVhW06a] die IMCRA-Methode zur spektralen Rauschschätzung verwendet, wird in [KLS05] das Rauschmodell aus Gl. (2.18) und in [DBY06] das Rauschmodell aus Gl. (2.19) eingesetzt.

2.4 Austausch von Informationen zwischen Front-End und Back-End

In den meisten Spracherkennungssystemen, die auf dem statistischen Ansatz basieren, der in Abschnitt 2.1 beschrieben wird, wird eine strikte Trennung zwischen dem Front-End, in dem die Sprachmerkmale extrahiert und entrauscht werden, und dem Back-End, in dem die Suche durchgeführt wird, vorgenommen. Dabei wird in vielen Ansätzen im Front-End eine Punktschätzung des unverrauschten Merkmalsvektors ermittelt, die im Back-End als Beobachtungsvektor verwendet wird. Wenn ein statistischer Ansatz im Front-End herangezogen wird, können jedoch bessere Erkennungsergebnisse erzielt werden, indem die gesamte a posteriori Verteilung der Sprachmerkmale, beispielsweise in Form einer Gaußverteilung oder Gaußmischungsverteilung, an das Back-End weitergegeben wird und somit die Unsicherheit der Schätzung berücksichtigt wird. Daneben sind in jüngerer Vergangenheit Ansätze veröffentlicht worden, in denen Informationen aus dem Back-End in das Front-End zurückgekoppelt werden, um so die Qualität der Sprachmerkmale zu verbessern.

Untersuchungen zum Uncertainty Decoding sind beispielsweise in [MBB01], [AC02], [KF02], [DAD02b], [MAC03], [SVhW04], [LG04], [BST+04], [DDA05], [LG06], [XRK06], [SVhW06b], [IHU08b] durchgeführt worden. [MBB01] verwendet anstelle einzelner Schätzwerte für die Merkmale $\mathbf{x}_t$ der Dimension N_c eine Evidenzfunktion $s'(\mathbf{x}_t)$ und berechnet im Erkenner statt der Likelihood $p(\mathbf{x}_t|q_t)$ des Schätzwertes $\mathbf{x}_t$ den Erwartungswert von $p(\mathbf{x}_t|q_t)$ bzgl. $s'(\mathbf{x}_t)$. Zu diesem Zweck wird das Integral

$$\int_{\mathbb{R}^{N_c}} p(\mathbf{x}_t|q_t)s'(\mathbf{x}_t)d\mathbf{x}_t \tag{2.20}$$

ausgewertet. Die Evidenzfunktion $s'(\mathbf{x}_t)$ wird in [MBB01] für jede Komponente $x_t^{(c)}$ des Merkmalsvektors $\mathbf{x}_t$ als Linearkombination zweier Verteilungen $f_1(x_t^{(c)})$ und $f_2(x_t^{(c)})$ modelliert, die entsprechend der Zuverlässigkeit von $\mathbf{x}_t$ gewichtet werden. In [AC02] wird die Likelihood $p(\mathbf{x}_t|q_t)$ mit der a posteriori Verteilung $p(\mathbf{x}_t|\mathbf{y}_t)$ anstelle der Evidenzfunktion $s'(\mathbf{x}_t)$ gewichtet. Dabei wird $p(\mathbf{x}_t|\mathbf{y}_t)$ als einzelne Gaußverteilung modelliert und im Front-End mittels einer aus Stereopaaren berechneten Abbildung aus den verrauschten Sprachmerkmalen bestimmt. [KF02] leitet aus der Formulierung des Optimierungsproblems für verrauschte Sprachdaten die Likelihood

$$p(\mathbf{y}_t|q_t) = \int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_t)p(\mathbf{x}_t|q_t)}{p(\mathbf{x}_t)} d\mathbf{x}_t \tag{2.21}$$

her, deren Parameter mit dem ALGONQUIN-Verfahren ermittelt werden. Die Berücksichtigung von Inter-Frame Korrelationen in der von [KF02] formulierten Decodierregel (Gln. 2.21) wird in [IHU08b] durch die Ersetzung von $p(\mathbf{x}_t|\mathbf{y}_t)$ durch $p(\mathbf{x}_t|\mathbf{y}_1^T)$ erreicht. In [DAD02b] wird die Likelihood $p(\mathbf{y}_t|q_t)$ berechnet, indem das zu Gl. 2.21 proportionale Integral

$$p(\mathbf{y}_t|q_t) = \int_{\mathbb{R}^{N_c}} p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|q_t)d\mathbf{x}_t \tag{2.22}$$

ausgewertet wird. Die Likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ wird in diesem Ansatz als Gaußmischungsverteilung modelliert, deren Parameter im Front-End mit dem SPLICE-Verfahren [DAPH00] bestimmt werden. [LG04] gibt neben der Ermittlung der Parameter der Verteilung $p(\mathbf{y}_t|\mathbf{x}_t)$ mit Hilfe des SPLICE-Verfahrens eine zweite Möglichkeit zur Parameterbestimmung an, das auf der Berechnung der gemeinsamen Verteilung von $\mathbf{x}_t$ und $\mathbf{y}_t$ basierende Joint-Schema. Wie [LG04] zeigt, können die Parameter des SPLICE- und des Joint-Schemas auch mit Front-End Techniken wie VTS, die nicht auf einem Training mit Stereodaten basieren, bestimmt werden. Die Anwendung des Joint-Schemas erfordert allerdings einen datenbasierten Ansatz bei der Berechnung der Kovarianzmatrix $\mathbf{\Sigma}_{xy}$ zwischen $\mathbf{x}_t$ und $\mathbf{y}_t$. Eine analytische Berechnungsmöglichkeit der Kovarianzmatrix $\mathbf{\Sigma}_{xy}$ wird in [XRK06] angegeben. [LG06] thematisiert die Instabilität des Verfahrens für die Singularität von $\mathbf{\Sigma}_{xy}$, die er durch eine Begrenzung der Parameter verhindert. In [SVhW04] werden zwei weitere Möglichkeiten des Uncertainty Decodings untersucht. Die eine Möglichkeit besteht darin, verschiedene Punktschätzungen an das Back-End zu übergeben und bei der Decodierung der Sprachmerkmale die Wahrscheinlichste auszuwählen. Daneben wird vorgeschlagen, die Varianz der a posteriori Wahrscheinlichkeit, die im Front-End berechnet wird, mit einem konstanten Faktor zu multiplizieren und zu den Varianzen der Mischungsverteilungen im Back-End zu addieren. In [SVhW06b] werden verschiedene Möglichkeiten untersucht, die Gewichte der in dem ersten Ansatz übergebenen Deltafunktionen mit Informationen aus dem Back-End anzupassen. Die Kombination eines Partikelfilteransatzes zur Verbesserung der Sprachmerkmale mit Uncertainty Decoding wird in [MAC03] untersucht. In [BST+04] wird das Konzept des Uncertainty Decodings für ein Wiener-Filter angewendet, wobei die Varianz der Sprachmerkmale durch eine Heuristik approximiert werden muß.

[FW06] und [YSW07] untersuchen in ihren Arbeiten die Rückkopplung von Erkennungsergebnissen in das Front-End. Zu diesem Zweck führen sie eine zweistufige Erkennung durch. In [FW06] wird ein Partikelfilter verwendet, um die a posteriori Verteilung der Sprachmerkmale zu berechnen. Die Dynamik des Rauschens wird in diesem Ansatz mit einem Zustandsmodell beschrieben. Die a priori Verteilung $p_1(\mathbf{x}_t) = p_{\text{train}}(\mathbf{x}_t)$ der Sprachmerkmale wird mit einem GMM modelliert, für dessen Parameter in der ersten Stufe der Erkennung vorberechnete Werte aus dem Training verwendet werden. In der zweiten Stufe werden die Erkennungsergebnisse zurückgekoppelt, um die a priori Verteilung $p_2(\mathbf{x}_t)$ abhängig von dem Phonem, das im Erkenner für den jeweiligen Sprachsignalabschnitt detektiert wurde, festzulegen. Da nur eine Hypothese über die wahrscheinlichste Phonemsequenz zurückgekoppelt wird, treten mit einer phonemspezifischen a priori Verteilung $p_2(\mathbf{x}_t) = p_{\text{phon}}(\mathbf{x}_t)$ starke Schwankungen der Parameter von

$p_2(\mathbf{x}_t)$ beim Schalten zwischen den Phonemen auf. Weiterhin wird die zurückgekoppelte Hypothese gegenüber anderen, möglichen Hypothesen zu stark gewichtet. Aus diesem Grund modelliert [FW06] die a priori Verteilung in der zweiten Erkennungsstufe als Mischungsverteilung

$$p_2(\mathbf{x}_t) = \alpha p_{\text{phon}}(\mathbf{x}_t) + (1-\alpha) p_{\text{train}}(\mathbf{x}_t) \tag{2.23}$$

mit festgelegtem Gewichtungsfaktor α. Mit dem Mischungsmodell ergeben sich auf einer Spezialdatenbank, die ungefähr 45 Minuten kontinuierlich gesprochener Sprache enthält, gegenüber dem a priori Modell etwas bessere Erkennungsergebnisse, wobei die Gewinne vor allem bei niedrigem SNR auftreten.

[YSW07] verwendet einen wortgraphbasierten Ansatz zur Entrauschung der Sprachmerkmale. In der ersten Stufe werden die Sprachmerkmale durch eine spektrale Subtraktion entrauscht [KSS06] und im Cepstrum mittels einer CMN normalisiert. Bei der anschließenden Decodierung wird anstelle einer einzelnen Hypothese über die beste Wortsequenz ein Wortgraph erzeugt. In der zweiten Stufe werden die Erwartungswerte der Sprachmerkmale für die einzelnen Hypothesen des Wortgraphens in den MEL-Frequenzbereich transformiert, wo sie neben einer Rauschschätzung zur Wiener-Filterung der verrauschten Sprachmerkmale eingesetzt werden. Die Decodierung erfolgt nach der Anwendung einer CMN in einem durch den Wortgraphen eingeschränkten Suchraum. Auf der AURORA2-Datenbank konnte mit diesem Verfahren im Vergleich zu der spektralen Subtraktion, die in der ersten Stufe durchgeführt wird, und einer GMM-basierten Merkmalsentrauschung eine deutlich höhere Erkennungsrate erzielt werden.

Kapitel 3

Wissenschaftliche Ziele

Der statistische Spracherkennungsansatz auf der Basis von HMMs, der in Kapitel 2 ausführlich beschrieben wird, wird zur Zeit in nahezu allen kommerziellen Anwendungen im Bereich der ASR eingesetzt. Wie in Abschnitt 2.2 ausgeführt wurde, besteht ein wesentlicher Schwachpunkt dieses Ansatzes in der fehlenden Ausnutzung statistischer Abhängigkeiten zwischen den Merkmalsvektoren aufeinander folgender Sprachrahmen. Das betrifft sowohl die akustische Modellierung mit HMMs im Back-End des Erkenners als auch eine modellbasierte Merkmalsentstörung auf der Basis von GMMs. An dieser Stelle setzt die vorliegende Arbeit an, deren Ziel die Berücksichtigung statistischer Abhängigkeiten zwischen den Sprachmerkmalen verschiedener Sprachrahmen innerhalb des dargestellten Spracherkennungssystems ist. Wie aus dem Literaturüberblick in Abschnitt 2.2.2 hervorgeht, erscheint die Modellierung der Sprachdynamik mit SLDMs als erfolgversprechender Ansatz, um Inter-Frame Korrelationen bei der Merkmalsentstörung im Front-End zu berücksichtigen. Die betrachteten Methoden führen auf der AURORA2 Datenbank zwar zu schlechteren Erkennungsraten als das AFE. Durch die statistische Modellierung der Sprachmerkmale im Cepstrum ergibt sich allerdings ein konsistenter Rahmen, der den Austausch von Informationen zwischen Front-End und Back-End des Erkenners erleichtert. Zunächst wird in Kapitel 4 die Entrauschung der Sprachmerkmale im Front-End unabhängig von der Erkennung im Back-End betrachtet. In Abschnitt 4.1 wird der statistische Modellierungsansatz eingeführt, der die Modellierung der Sprachdynamik mit schaltenden Modellen im Cepstrum, die Berücksichtigung der Rauschdynamik sowie das nichtlineare Beobachtungsmodell für die verrauschten Sprachmerkmale umfaßt. Auf dieser Grundlage soll in Abschnitt 4.2 die Berechnung der a posteriori Wahrscheinlichkeit der Sprachmerkmale untersucht werden. Dazu wird alternativ zu einem Filter, das in [DA04] entwickelt wurde (siehe Anhang C.3), die Filterung mit schaltenden EKFs betrachtet. Daneben soll das ALGONQUIN-Verfahren [FDAK03], das verglichen mit den in Abschnitt 2.1.1 betrachteten Methoden die besten Erkennungsergebnisse bei einer Sprachmodellierung mit GMMs liefert, in den SLDM-Ansatz integriert werden. Basierend auf der Vektortaylorreihenentwicklung eines in [Gal95] angegebenen Beobachtungsmodells soll dieser Ansatz in Abschnitt 4.3 um dynamische Merkmale erweitert werden, die eine kompakte Repräsentation statistischer Abhängigkeiten über mehr als einen Sprachrahmen darstellen. Um die Robustheit und Effizienz des Filters zu erhöhen, kann die Dimension des Subvektors, der die dynamischen Merkmale repräsentiert, mittels einer Principal

Component Analysis (PCA) reduziert werden. Die Glättung der Sprachmerkmale mit schaltenden Modellen wird in Abschnitt 4.4 betrachtet. Abschließend wird die modellbasierte Merkmalsentstörung in Abschnitt 4.5 experimentell untersucht und mit bekannten Ansätzen aus der Literatur verglichen.

Ein kritischer Punkt bei der Berechnung der a posteriori Wahrscheinlichkeit der Sprachmerkmale stellt die Bestimmung des Umgebungsrauschens dar. Dieses Problem wird in Kapitel 5 aufgegriffen. Basierend auf den in Abschnitt 2.3 dargestellten Ansätzen zur Rauschschätzung soll ein neues Rauschmodell eingeführt werden, das in den in Kapitel 4 untersuchten Ansatz integriert werden kann. Die Dynamik des Rauschens wird in diesem Modell mit einer versteckten, nicht-beobachtbaren Zufallsvariable mit kontinuierlichem Wertebereich modelliert. Zur Parameterschätzung sollen in Abschnitt 5.1 und Abschnitt 5.2 EM-Algorithmen hergeleitet werden. In Abschnitt 5.3 wird die Rauschschätzung experimentell untersucht.

In Kapitel 6 werden die Merkmalsentstörung und die anschließende Decodierung der Sprachmerkmale in einem gemeinsamen statistischen Rahmen betrachtet. Dazu soll das Optimierungsproblem, das in Abschnitt 2.1 für unverrauschte Sprachmerkmale formuliert wurde, auf der Grundlage eines gegenüber [KF02] und [IHU08b] modifizierten statistischen Modells, in dem zusätzlich Langzeitabhängigkeiten zwischen den Sprachmerkmalen modelliert werden, auf verrauschte Sprachmerkmale erweitert werden. Die Decodierung der Sprachmerkmale mit dem erweiterten Modell wird in Abschnitt 6.1 theoretisch betrachtet und in Abschnitt 6.3 experimentell untersucht. In Abschnitt 6.2 wird grundlegend für die Untersuchungen in Kapitel 7 und Kapitel 8 die Skalierung des Sprachmodells und des Akustikmodells, auf die bereits in Abschnitt 2.1 eingegangen wurde, ausführlicher behandelt.

Die Erweiterung des HMMs in Abschnitt 6.1 führt zu einer modifizierten Emissionsverteilung des HMMs, in der statistische Abhängigkeiten von den Sprachmerkmalen vorangehender Sprachrahmen berücksichtigt werden. Diese sollen in Kapitel 7 durch statistische Abhängigkeiten zwischen den Mischungskomponenten der Emissionsverteilung approximiert werden. Für das resultierende segmentielle HMM soll in Abschnitt 7.1 eine effiziente Suchstrategie entwickelt werden. Ein möglicher Ansatz besteht darin, die optimale Sequenz der HMM-Zustände unverändert mit einem Viterbi-Algorithmus zu berechnen und die Mischungsgewichte mit einem diskreten Filter im Vorwärtsschritt des Viterbi-Algorithmus zu aktualisieren. Der Speicheraufwand soll in Abschnitt 7.2 durch die Verwendung schaltender Dynamikmodelle bei der Decodierung reduziert werden. Die Ergebnisse der experimentellen Untersuchungen werden abschließend in Abschnitt 7.3 angegeben.

In Kapitel 8 sollen die Informationen aus den komplexen Sprach- und Akustikmodellen des Spracherkenners bei der Merkmalsentstörung im Front-End berücksichtigt werden. Zunächst werden in Abschnitt 8.1 Unterschiede zwischen dem SLDM und dem HMM herausgearbeitet, die eine Kombination der beiden Modelltypen erfolgversprechend erscheinen lassen. Anschließend sollen innerhalb eines statistischen Ansatzes (Abschnitt 8.2) Methoden hergeleitet werden, die es ermöglichen, Informationen in das Front-End zurückzukoppeln. Eine Möglichkeit, die in Abschnitt 8.3 untersucht wird, be-

steht darin, einen statistischen Zusammenhang zwischen den HMM-Zuständen und den schaltenden Dynamikmodellen im Front-End des Spracherkenners auszunutzen, um die a posteriori Verteilung der Sprachmerkmale genauer zu bestimmen. Daneben kann die Verteilung der Sprachmerkmale unmittelbar mit den Informationen aus dem Erkenner beeinflußt werden (Abschnitt 8.4). Neben diesen beiden Ansätzen ist es möglich, die Informationen zu einer verbesserten Rauschschätzung auszunutzen (Abschnitt 8.5). Bei der Rückkopplung von Informationen aus dem Back-End erscheint es als erfolgversprechend, anstelle der besten Zustandsfolge, die mit einem Viterbi-Algorithmus berechnet werden kann, die a posteriori Wahrscheinlichkeiten der HMM-Zustände zu ermitteln, um auf diese Weise die Unsicherheit der Schätzung zu berücksichtigen und die Verstärkung von Fehlentscheidungen zu vermeiden (Abschnitt 8.6). Die exakten a posteriori Wahrscheinlichkeiten der HMM-Zustände können mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene berechnet werden, der in Abschnitt 8.6.1 beschrieben wird. Außerdem soll in Abschnitt 8.6.2 untersucht werden, inwieweit die Berechnung der HMM-Zustände durch einen Vorwärts-Rückwärts-Algorithmus auf Wortebene beschleunigt werden kann. In Abschnitt 8.7 wird die Rückkopplung von Informationen in das Front-End des Spracherkenners experimentell untersucht.

Abschließend wird in Kapitel 9 eine Zusammenfassung der Forschungsergebnisse sowie ein Ausblick auf mögliche weiterführende Arbeiten gegeben.

Kapitel 4

Entrauschung der Sprachmerkmale mit schaltenden Modellen

In dem folgenden Kapitel wird die Entrauschung der Sprachmerkmale im Front-End des Spracherkenners betrachtet. Dabei soll, wie in Kapitel 3 motiviert wurde, ein statistischer, modellbasierter Ansatz zugrunde gelegt werden. Die Entrauschung soll jeweils für einen einzelnen Satz mit T Sprachrahmen durchgeführt werden. Das Ziel besteht in der Bestimmung der a posteriori Wahrscheinlichkeit $p(\mathbf{x}_t|\mathbf{y}_1^T)$ der unverrauschten, cepstralen Sprachmerkmale $\mathbf{x}_t$ bei gegebenen verrauschten Sprachmerkmalen $\mathbf{y}_1^T = \mathbf{y}_1 \dots \mathbf{y}_T$.

Zunächst wird unter vereinfachten Modellannahmen für ein unverrauschtes Sprachsignal untersucht, inwieweit es erfolgversprechend ist, Korrelationen zwischen aufeinander folgenden Sprachrahmen bei der statistischen Modellierung zu berücksichtigen.

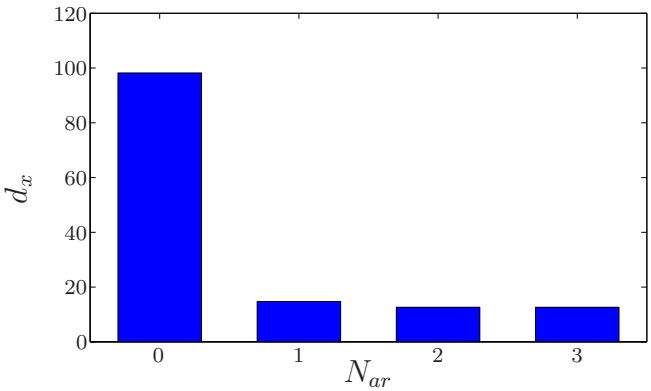

Abbildung 4.1: *Mittlerer quadratischer Prädiktionsfehler auf unverrauschten Sprachdaten der AURORA2 Datenbank für AR-Modelle der Ordnung N_{ar}*

Abb. 4.1 zeigt den mittleren quadratischen Prädiktionsfehler

$$d_x = \frac{1}{T_{set}} \sum_{t=1}^{T_{set}} \left(x_t^{(0)} - \sum_{l=1}^{N_{ar}} \mathbf{A}_l^{(0,0:N_c-1)} \mathbf{x}_{t-l} - \mathbf{b}^{(0)} \right)^2 \tag{4.1}$$

der Energiekomponente $x_t^{(0)}$ des N_c-dimensionalen Merkmalsvektors $\mathbf{x}_t$ für die unverrauschten Sprachdaten der T_{set} Sprachrahmen aus Sub-Set A der in Anhang A.1 beschriebenen AURORA2 Sprachdatenbank. Dabei wird die Sprachdynamik mit autoregressiven (AR-) Modellen der Ordnungen $N_{ar} = 0, \dots, 3$ mit den Parametern $\mathbf{A}_l^{(0,0:N_c-1)}$ und $\mathbf{b}^{(0)}$ beschrieben, die wie in Abschnitt 4.1.0.1 für $N_{ar} = 1$ dargestellt, aus den

rauschfreien Trainingsdaten der AURORA2 Datenbank ermittelt wurden. Aus der Abbildung wird ersichtlich, dass der mittlere quadratische Prädiktionsfehler unter den gegebenen Modellannahmen durch die Berücksichtigung von Inter-Frame Korrelationen deutlich reduziert wird.

Wie die experimentellen Untersuchungen in Abschnitt 4.5 zeigen, wird die Dynamik der Sprachmerkmale in der Praxis nur unzureichend mit einem einzelnen linearen Dynamikmodell beschrieben. Die Ausführungen in Abschnitt 2.2.2 lassen schaltende, lineare Dynamikmodelle (SLDMs) hingegen als erfolgversprechenden Ansatz zur Merkmalsentstörung erscheinen. In diesem Ansatz wird die Trajektorie der Sprachmerkmale durch einen Satz linearer Dynamikmodelle beschrieben, die zu jedem Zeitpunkt unter Berücksichtigung aktueller Beobachtungen des Merkmalsvektors gewichtet werden. Neben der Sprache wird auch das Rauschen modelliert, das mit den unverrauschten und verrauschten Sprachmerkmalen über ein hochgradig nichtlineares Beobachtungsmodell verknüpft ist (vgl. Abschnitt 4.1). Die a posteriori Verteilung des Zustandsvektors kann mit Filtern für die einzelnen linearen Dynamikmodelle aktualisiert werden, wobei die Verteilungen an den Filterausgängen zusammengefaßt werden müssen, um einen exponentiellen Anstieg des Rechenaufwandes mit der Anzahl der betrachteten Sprachrahmen zu verhindern (siehe Abschnitt 4.2). Zur Approximation der bedingten a posteriori Verteilungen der einzelnen Filter des SLDMs wird neben einer in [DDA03b] eingeführten Technik die Integration des ALGONQUIN-Algorithmus in den SLDM-Ansatz untersucht (Abschnitt 4.2.0.4). Die Grundlage dazu stellt ein ebenfalls in Abschnitt 4.2.0.4 beschriebenes erweitertes Kalman-Filter dar. In Abschnitt 4.3 wird ein verbessertes Zustands- und Beobachtungsmodell eingeführt, das auf der Ausnutzung von Informationen in den dynamischen Sprachmerkmalen erster und zweiter Ordnung basiert. Eine robuste und effiziente Integration der dynamischen Merkmale in das erweiterte Kalman-Filter wird durch die Anwendung einer Hauptachsentransformation (PCA) erreicht, mit der die dynamischen Merkmalsvektoren auf wenige zusätzliche Komponenten reduziert werden. Die Glättung der Sprachmerkmale wird in Abschnitt 4.4 beschrieben. In Abschnitt 4.5 werden experimentelle Untersuchungen durchgeführt. Diese umfassen qualitative Untersuchungen zu charakteristischen Systemeigenschaften, Auswirkungen der Parametrisierung, Systemvarianten sowie einordnende Vergleiche mit bekannten Verfahren aus der Literatur.

4.1 Statistische Modellierung

Das statistische Modell, auf dessen Grundlage die Entrauschung der Sprachmerkmale durchgeführt wird, ist in Abb. 4.2 graphisch dargestellt.

In schaltenden, linearen Dynamikmodellen wird das zeitliche Verhalten der Sprachmerkmale $\mathbf{x}_t$ mit Übergangswahrscheinlichkeiten $p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)$, die von einer diskreten Variable s_t abhängen und mit linearen Zustandsgleichungen beschrieben werden können, modelliert. Im Folgenden werden, genauso wie in [DA04], lineare Zustandsglei-

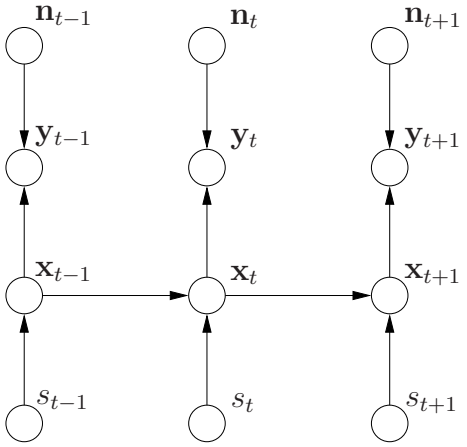

Abbildung 4.2: *Graphisches Modell des SLDMs*

chungen der Form

$$\mathbf{x}_t = \mathbf{A}(s_t)\mathbf{x}_{t-1} + \mathbf{b}(s_t) + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{u}_t; \mathbf{0}, \mathbf{C}(s_t)) \tag{4.2}$$

angenommen, wobei die Modellparameter $\mathbf{A}(s_t)$, $\mathbf{b}(s_t)$ und $\mathbf{C}(s_t)$ von der Zustandsvariable s_t abhängen. In Gl. (4.2) werden Modellierungsfehler durch das Zustandsrauschen $\mathbf{u}_t$ berücksichtigt. Zwischen den SLDM-Zuständen s_{t-1} und s_t wird keine direkte statistische Abhängigkeit angenommen. Der statistische Zusammenhang zwischen den unverrauschten Sprachmerkmalen $\mathbf{x}_t$, den verrauschten Sprachmerkmalen $\mathbf{y}_t$ und dem Rauschen $\mathbf{n}_t$ kann in Form eines Beobachtungsmodells angegeben werden.

In den folgenden Unterabschnitten werden das Training der Modellparameter, die Berücksichtigung der Dynamik des Rauschens in einem erweiterten Zustandsmodell, sowie das Beobachtungsmodell dargestellt.

4.1.0.1 Training der Modellparameter

Die Modellparameter können mittels eines EM-Algorithmus aus unverrauschten Trainingsdaten $\mathbf{x}_t$ geschätzt werden [DA04]. Dabei wird zwischen der Berechnung der Modellwahrscheinlichkeiten

$$\gamma_t^m = P(s_t = m|\mathbf{x}_1^T) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1}, m)P(m) = \mathcal{N}(\mathbf{x}_t; \mathbf{A}(m)\mathbf{x}_{t-1} + \mathbf{b}(m), \mathbf{C}(m))P(m) \tag{4.3}$$

für die Modelle $m = 1 \ldots M$ und der Schätzung der Modellparameter

$$\begin{aligned}
\mathbf{A}(m) &= (\langle \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\rangle_m - \langle \mathbf{x}_{t-1}\rangle_m \langle \mathbf{x}'_{t-1}\rangle_m)^{-1}(\langle \mathbf{x}_t\mathbf{x}'_{t-1}\rangle_m - \langle \mathbf{x}_t\rangle_m \langle \mathbf{x}'_{t-1}\rangle_m) \\
\mathbf{b}(m) &= \langle \mathbf{x}_t\rangle_m - \mathbf{A}(m)\langle \mathbf{x}_{t-1}\rangle_m \\
\mathbf{C}(m) &= \langle (\mathbf{x}_t - \mathbf{A}(m)\mathbf{x}_{t-1} - \mathbf{b}(m))(\mathbf{x}_t - \mathbf{A}(m)\mathbf{x}_{t-1} - \mathbf{b}(m))'\rangle_m \\
P(m) &= \frac{1}{T}\sum_{t=1}^{T} \gamma_t^m
\end{aligned} \tag{4.4}$$

iteriert. T bezeichnet in Gl. (4.4) die Anzahl der Sprachrahmen in allen Trainingssätzen. Wie in [DA04] wird die Kurznotation $< . >_m$ für den gewichteten Mittelwert über die

Trainingsdaten verwendet, z.B.

$$\langle \mathbf{x}_t \mathbf{x}'_{t-1} \rangle_m = \frac{\sum_{t=1}^{T} \gamma_t^m \mathbf{x}_t \mathbf{x}'_{t-1}}{\sum_{t=1}^{T} \gamma_t^m}. \tag{4.5}$$

Die Initialisierung der Modellparameter wird, wie in Anhang D.1 beschrieben, mit einem Splitting-Algorithmus durchgeführt.

4.1.0.2 Integration des Rauschmodells

Um die zeitvariante Charakteristik des Hintergrundrauschens berücksichtigen zu können, wird der Rauschprozess allgemein mit einem Zustandsmodell beschrieben:

$$\mathbf{n}_t = \mathbf{D}\mathbf{n}_{t-1} + \mathbf{e} + \mathbf{v}_t, \mathbf{v}_t \sim \mathcal{N}(\mathbf{v}_t; \mathbf{0}, \mathbf{V}). \tag{4.6}$$

In Gl. (4.6) bezeichnet $\mathbf{D}$ die Zustandsübergangsmatrix, $\mathbf{e}$ einen Bias und $\mathbf{v}_t$ das Zustandsrauschen mit der Kovarianzmatrix $\mathbf{V}$.

Für die Untersuchungen in diesem Kapitel werden die Parameter der a priori Verteilung $\mathcal{N}(\mathbf{n}_t; \boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$ aus sprachfreien Signalbereichen des aktuellen Satzes berechnet und die Parametrisierung

$$\mathbf{D} = \mathbf{0}, \quad \mathbf{e} = \boldsymbol{\mu_n}, \quad \mathbf{V} = \boldsymbol{\Sigma_n} \tag{4.7}$$

gewählt. Verschiedene Alternativen für dieses Rauschmodell werden in Kapitel 5 untersucht.

Um in den folgenden Abschnitten eine einfache mathematische Behandlung zu ermöglichen, wird das Rauschmodell in den Zustandsvektor

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{n}_t \end{bmatrix} \tag{4.8}$$

integriert, wodurch sich ein erweitertes Zustandsmodell der Form

$$\mathbf{z}_t = \mathbf{A}_z(s_t)\mathbf{z}_{t-1} + \mathbf{b}_z(s_t) + \begin{bmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{bmatrix} \tag{4.9}$$

mit

$$\begin{aligned} \mathbf{A}_z(s_t) &= \begin{bmatrix} \mathbf{A}(s_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}, \quad \mathbf{b}_z(s_t) = \begin{bmatrix} \mathbf{b}(s_t) \\ \mathbf{e} \end{bmatrix}, \\ \mathbf{C}_z(s_t) &= \begin{bmatrix} \mathbf{C}(s_t) & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{bmatrix}, \end{aligned} \tag{4.10}$$

ergibt, wobei $\mathbf{u}_t \sim \mathcal{N}(\mathbf{u}_t; \mathbf{0}, \mathbf{C}(s_t))$ und $\mathbf{v}_t \sim \mathcal{N}(\mathbf{v}_t; \mathbf{0}, \mathbf{V})$ das Zustandsrauschen der Dynamikmodelle für Sprache und Rauschen bezeichnen (vgl. Gl. (4.2) und Gl. (4.6)).

4.1.0.3 Beobachtungsmodell

Der Zusammenhang zwischen dem verrauschten Sprachsignal $y(k)$ und dem unverrauschten Sprachsignal $x(k)$ kann im Zeitbereich über die Beziehung

$$y(k) = x(k) * f(k) + n(k) \tag{4.11}$$

modelliert werden [Gal95]. Dabei werden das additive Rauschen $n(k)$ und das Faltungsrauschen $f(k)$ als Störquellen berücksichtigt. Aus Gl. (4.11) ergibt sich für die mit dem in Anhang C.1 beschriebenen ETSI Standard Front-End extrahierten cepstralen Sprachmerkmale der Zusammenhang [Ace93]

$$\mathbf{y}_t = \tilde{\mathbf{h}}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \mathbf{f}_t + \mathbf{M}_{DCT} \log\left(\mathbf{1} + e^{\mathbf{M}_{DCT}^{+}(\mathbf{n}_t - \mathbf{x}_t - \mathbf{f}_t)} + \boldsymbol{\varphi}_\alpha(\mathbf{x}_t, \mathbf{n}_t)\right) \tag{4.12}$$

mit

$$\boldsymbol{\varphi}_\alpha(\mathbf{x}_t, \mathbf{n}_t) = 2\alpha_t \mathbf{M}_{DCT} e^{\mathbf{M}_{DCT}^{+} \frac{\mathbf{n}_t - \mathbf{x}_t - \mathbf{f}_t}{2}} \tag{4.13}$$

und der gaußverteilten Zufallsvariable α_t. In Gl. (4.12) und Gl. (4.13) werden die Matrizenmultiplikationen auf den gesamten Vektor angewendet, während die anderen Operationen komponentenweise definiert sind. $\mathbf{M}_{DCT}$ bezeichnet die Matrix der diskreten Kosinustransformation (DCT), mit der die log-spektralen Sprachmerkmale ins Cepstrum transformiert werden und

$$\mathbf{M}_{DCT}^{+} = \begin{cases} (\mathbf{M}'_{DCT}\mathbf{M}_{DCT})^{-1}\mathbf{M}'_{DCT} & : \; N_c > N_l \\ \mathbf{M}_{DCT}^{-1} & : \; N_c = N_l \\ \mathbf{M}'_{DCT}(\mathbf{M}_{DCT}\mathbf{M}'_{DCT})^{-1} & : \; N_c < N_l \end{cases} \tag{4.14}$$

die (pseudo-)inverse Matrix der $N_c \times N_l$-Matrix $\mathbf{M}_{DCT}$.

In dem Zero-Variance-Model (ZVM) [DDA03b] wird der Phasenterm $\boldsymbol{\varphi}_\alpha(\mathbf{x}_t, \mathbf{n}_t)$, der nur für $\mathbf{x}_t \approx \mathbf{n}_t$ signifikanten Einfluß auf die Genauigkeit des Beobachtungsmodells hat, vernachlässigt. Der Parameter $\mathbf{f}_t$, der Faltungsrauschen im Zeitbereich modelliert, wird im Rahmen dieser Arbeit ebenfalls nicht berücksichtigt, da die experimentellen Untersuchungen ausschließlich mit additivem Rauschen durchgeführt werden. Das vereinfachte Beobachtungsmodell lautet somit:

$$\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, \mathbf{n}_t) = \mathbf{x}_t + \mathbf{M}_{DCT} \log\left(1 + e^{\mathbf{M}_{DCT}^{+}(\mathbf{n}_t - \mathbf{x}_t)}\right). \tag{4.15}$$

Der Zusammenhang $y_t^{(l)} = \log(e^{x_t^{(l)}} + e^{n_t^{(l)}})$, der sich im Log-Spektrum, das mit dem Cepstrum über eine lineare Transformation verknüpft ist, zwischen jeweils einer Komponente $x_t^{(l)}$, $n_t^{(l)}$ und $y_t^{(l)}$ der log-spektralen Merkmalsvektoren für Sprache, Rauschen und verrauschter Sprache ergibt, ist in Abb. 4.3 für $n_t^{(l)} = 0$ blau dargestellt. Daneben sind die Geraden, die sich jeweils durch eine Taylorreihenentwicklung um die Entwicklungspunkte $x_t^{(l)} = -0,5$ und $x_t^{(l)} + \delta x_t^{(l)} = -0,75$ ergeben, rot bzw. magenta dargestellt. Die Abbildung zeigt, dass das Beobachtungsmodell hochgradig nichtlinear ist und die Taylorreihe mit dem Entwicklungspunkt $x_t^{(l)}$ daher stark von dem tatsächlichen Kurvenverlauf $y_t^{(l)} = \log(e^{x_t^{(l)}} + e^{n_t^{(l)}})$ und der Taylorreihe um einen anderen Entwicklungspunkt $x_t^{(l)} + \delta x_t^{(l)}$ abweichen kann.

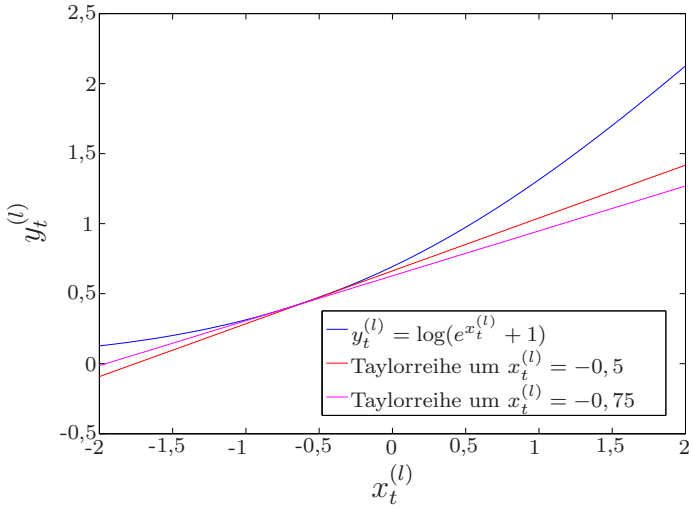

Abbildung 4.3: *Qualitativer Zusammenhang zwischen* $y_t^{(l)}$ *und* $x_t^{(l)}$ *für* $n_t^{(l)} = 0$

4.2 Berechnung der a posteriori Verteilung

Die a posteriori Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t)$ des Zustandsvektors $\mathbf{z}_t$ (und damit auch $p(\mathbf{x}_t|\mathbf{y}_1^t)$) kann dadurch ermittelt werden, dass für die Zustandsmodelle $s_t = 1 \ldots M$ bedingte a posteriori Wahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)$ berechnet werden und diese entsprechend der Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ zu der a posteriori Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t)$ zusammengefaßt werden [BSRLK01] (siehe Abb. 4.4).

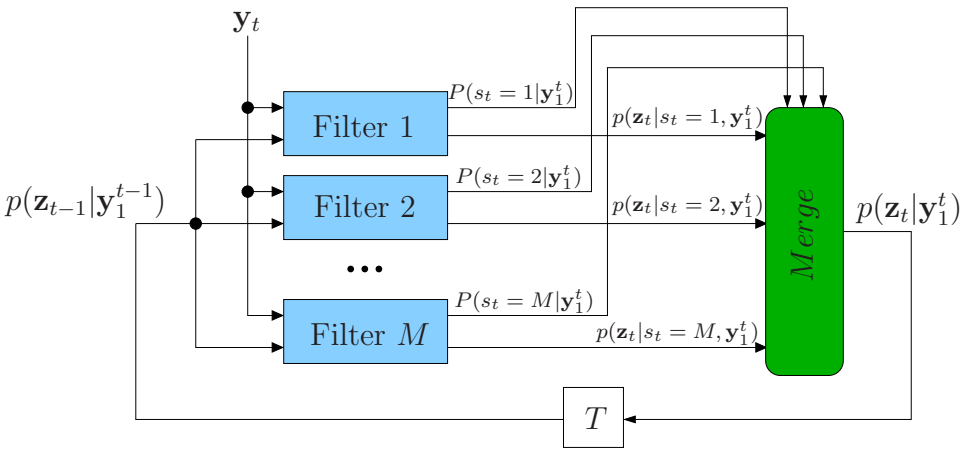

Abbildung 4.4: *Berechnung der a posteriori Wahrscheinlichkeit*

Die Berechnung der bedingten Wahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)$ wird in Abschnitt 4.2.0.4 untersucht. Dabei wird zur Vereinfachung angenommen, dass an den Ein- und Ausgängen der eingesetzten Filter jeweils Gaußverteilungen vorliegen. Das bedeutet, dass $p(\mathbf{z}_t|\mathbf{y}_1^t)$ eine Gaußmischungsverteilung ist, wobei die Zahl der überlagerten Normalverteilungen in jedem Zeitschritt um den Faktor M ansteigt. Mögliche Approximationen, in denen die Anzahl der Komponenten reduziert wird, sind der GPB-Algorithmus [AS72], der IMM-Algorithmus [BSRLK01] und das gauß'sche Summenfilter (GSF) [BSRLK01]. Während das GSF ein allgemeineres Konzept ist, in dem

die Anzahl der Moden nach verschiedenen Heuristiken reduziert wird, führt die Annahme der statistischen Unabhängigkeit $P(s_t|s_{t-1}) = P(s_t)$ für den GBP- und den IMM-Ansatz erster Ordnung zu einem Algorithmus, der in [DA04] für die Entrauschung von Sprachmerkmalen eingeführt wird (siehe Abb. 4.4). In diesem Ansatz wird $p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1})$ als unimodale Gaußverteilung modelliert, aus der M a posteriori Verteilungen $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)$ mit den Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ berechnet werden. Die resultierende Mischungsverteilung

$$p(\mathbf{z}_t|\mathbf{y}_1^t) = \sum_{s_t=1}^{M} p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)P(s_t|\mathbf{y}_1^t), \tag{4.16}$$

wird anschließend wieder zu einer unimodalen Gaußverteilung zusammengefaßt. Für den Erwartungswert $\mathbf{z}_{t|1:t}$ und die Kovarianzmatrix $\mathbf{P}_{t|1:t}$ dieser Verteilung erhält man die MMSE-Schätzwerte

$$\begin{aligned} \mathbf{z}_{t|1:t} &= \sum_{s_t} P(s_t|\mathbf{y}_1^t)\mathbf{z}_{t|1:t}(s_t) \\ \mathbf{P}_{t|1:t} &= \sum_{s_t} P(s_t|\mathbf{y}_1^t)(\mathbf{P}_{t|1:t}(s_t) + (\mathbf{z}_{t|1:t}(s_t) - \mathbf{z}_{t|1:t})(\mathbf{z}_{t|1:t}(s_t) - \mathbf{z}_{t|1:t})'), \end{aligned} \tag{4.17}$$

wobei $\mathbf{z}_{t|1:t}(s_t)$ und $P_{t|1:t}(s_t)$ die Momente der Einzelverteilungen $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)$ bezeichnen [BSRLK01].

Die Berechnung der bedingten a posteriori Wahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{y}_t, s_t)$ sowie der Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ ist Gegenstand der folgenden beiden Unterabschnitte.

4.2.0.4 Filterung

Die bedingten a posteriori Wahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{y}_t, s_t)$ für die schaltenden Modelle s_t mit den Zustandsübergangswahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{z}_{t-1}, s_t)$ können über die Rekursion

$$p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t) = \int_{\mathbb{R}^{2N_c}} p(\mathbf{z}_t|\mathbf{z}_{t-1}, s_t)p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}, s_t)d\mathbf{z}_{t-1} \tag{4.18}$$

$$p(\mathbf{z}_t|\mathbf{y}_1^t, s_t) \propto p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t)p(\mathbf{y}_t|\mathbf{z}_t) \tag{4.19}$$

berechnet werden [RSG04], wobei $p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}, s_t)$ für die einzelnen Filter mit

$$p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}, s_t) = p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}) \tag{4.20}$$

initialisiert wird [DA04]. Die praktische Durchführung der angegebenen Rekursion erfordert Approximationen, da das Beobachtungsmodell $p(\mathbf{y}_t|\mathbf{z}_t)$ in Gl. (4.19), das die unverrauschten Sprachmerkmale, das Rauschen und die verrauschten Sprachmerkmale miteinander verknüpft, hochgradig nichtlinear ist (vgl. Abschnitt 4.1.0.3). Im Zusammenhang mit schaltenden Modellen wurden in der Literatur ein auf der MMSE-Schätzung des VTS-Verfahrens [Mor96] beruhender Ansatz mit vorangehender Iteration einer SNR-Variable [DA04] sowie verschiedene Formen des erweiterten Kalman-Filters [KLS05, SVhW05b, DBY06] und des Uncented Kalman-Filters [SVhW05b]

untersucht. Umfangreichere Untersuchungen zur Berechnung der a posteriori Wahrscheinlichkeit wurden, wie in Abschnitt 2.1.1 ausgeführt, für die Sprachmodellierung mit GMMs durchgeführt, wobei das ALGONQUIN-Verfahren [FDAK03] zu hohen Erkennungsraten führte. Aus diesem Grund wird im Folgenden ein iteratives erweitertes Kalman-Filter (IEKF) eingeführt, das sich durch die Integration des ALGONQUIN-Verfahrens in den SLDM-Ansatz ergibt. Das IEKF basiert auf einem erweiterten Kalman-Filter, das sich ähnlich wie in [DBY06] aus der Kombination des Zustandsmodells in Gl. (4.2) und einer Approximation des Messmodells in Gl. (4.15), die in ähnlicher Form in [KLS05] verwendet wird, ergibt. Das in [DA04] eingeführte Verfahren, das den Ausgangspunkt für die in dieser Arbeit durchgeführten Untersuchungen darstellte, wird in Anhang C.3 beschrieben.

Der Einsatz eines erweiterten Kalman-Filters zur Berechnung der a posteriori Wahrscheinlichkeit erfordert die Linearisierung des Beobachtungsmodells in Gl. (4.15) um die Momente der a priori Verteilung des Zustandsvektors. Durch eine Vektortaylorreihenentwicklung um die Entwicklungspunkte $\mathbf{x}_t^{(0)}$ und $\mathbf{n}_t^{(0)}$ erhält man das linearisierte Beobachtungsmodell (vgl. [KLS05])

$$\mathbf{y}_t \approx \mathbf{h}(\mathbf{x}_t^{(0)}, \mathbf{n}_t^{(0)}) + \mathbf{H}_x(\mathbf{x}_t - \mathbf{x}_t^{(0)}) + \mathbf{H}_n(\mathbf{n}_t - \mathbf{n}_t^{(0)}) + \mathbf{w}_t. \tag{4.21}$$

mit der Einheitsmatrix $\mathbf{I}$ und den Jacobi-Matrizen

$$\mathbf{H}_x = \mathbf{M}_{DCT} \frac{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t^{(0)}}}{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t^{(0)}} + e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t^{(0)}}} \mathbf{M}_{DCT}^{+}, \quad \mathbf{H}_n = \mathbf{I} - \mathbf{H}_x. \tag{4.22}$$

Das Messrauschen wird in Gl. (4.21) wie in [KLS05] als Gaußverteilung

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{w}_t; \mathbf{0}, \mathbf{W}) \tag{4.23}$$

mit konstanter Kovarianzmatrix

$$\mathbf{W} = \mathbf{W}_c \tag{4.24}$$

modelliert. Im Folgenden werden, wie in der Literatur üblich, die Momente $\mathbf{x}_t^{(0)} = \mathbf{x}_{t|1:t-1}(s_t)$ und $\mathbf{n}_t^{(0)} = \mathbf{n}_{t|1:t-1}(s_t)$ der a priori Verteilungen von Sprache und Rauschen als Entwicklungspunkte verwendet, wodurch sich die modellabhängigen Jacobi-Matrizen $\mathbf{H}_x(s_t)$ und $\mathbf{H}_n(s_t)$ ergeben. Diese können zu der Matrix

$$\mathbf{H}_z(s_t) = [\mathbf{H}_x(s_t) \quad \mathbf{H}_n(s_t)] \tag{4.25}$$

zusammengefaßt werden.

Unter der Annahme gaußförmiger Verteilungen können die Momente der a posteriori Verteilung $p(\mathbf{x}_t|\mathbf{y}_t, s_t)$ elementar entsprechend der Zustandsgleichungen des EKFs aus der a posteriori Verteilung $p(\mathbf{x}_{t-1}|\mathbf{y}_{t-1})$ des letzten Sprachrahmens berechnet werden [BSRLK01]:

$$\begin{aligned}
\mathbf{z}_{t|1:t-1}(s_t) &= \mathbf{A}_z(s_t)\mathbf{z}_{t-1|1:t-1} + \mathbf{b}_z(s_t) \\
\mathbf{P}_{t|1:t-1}(s_t) &= \mathbf{A}_z(s_t)\mathbf{P}_{t-1|1:t-1}\mathbf{A}_z'(s_t) + \mathbf{C}_z(s_t) \\
\mathbf{K}_t(s_t) &= \mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z'(s_t)(\mathbf{H}_z(s_t)\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z'(s_t) + \mathbf{W})^{-1} \\
\mathbf{z}_{t|1:t}(s_t) &= \mathbf{z}_{t|1:t-1}(s_t) + \mathbf{K}_t(s_t)(\mathbf{y}_t - \mathbf{h}(\mathbf{z}_{t|1:t-1}(s_t))) \\
\mathbf{P}_{t|1:t}(s_t) &= (\mathbf{I} - \mathbf{K}_t(s_t)\mathbf{H}_z(s_t))\mathbf{P}_{t|1:t-1}(s_t).
\end{aligned} \tag{4.26}$$

In Gl. (4.26) bezeichnet $\mathbf{K}_t(s_t)$ die Kalman-Verstärkung. Die Variablen $\mathbf{z}_{t|1:t-1}(s_t)$, $\mathbf{P}_{t|1:t-1}(s_t)$, $\mathbf{z}_{t|1:t}(s_t)$ und $\mathbf{P}_{t|1:t}(s_t)$ bezeichnen die Momente der zustandsabhängigen Wahrscheinlichkeitsdichtefunktionen

$$\begin{aligned} p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t) &= \mathcal{N}(\mathbf{z}_t; \mathbf{z}_{t|1:t-1}(s_t), \mathbf{P}_{t|1:t-1}(s_t)) \\ p(\mathbf{z}_t|\mathbf{y}_1^{t}, s_t) &= \mathcal{N}(\mathbf{z}_t; \mathbf{z}_{t|1:t}(s_t), \mathbf{P}_{t|1:t}(s_t)) \end{aligned} \tag{4.27}$$

des Zustandsvektors $\mathbf{z}_t$ in Sprachrahmen t bei gegebenen Beobachtungen bis Sprachrahmen $t-1$ bzw. t.

Die Wahl des Entwicklungspunktes hat aufgrund der starken Nichtlinearität des Messmodells, wie in Abschnitt 4.1.0.3 ausgeführt, einen großen Einfluß auf die Genauigkeit der Taylorreihenentwicklung. Aus diesem Grund wird eine Relinearisierung des Messmodells mit einem iterativen erweiterten Kalman-Filter [BSRLK01] durchgeführt, die gleichzeitig die Neu-Schätzung der a priori Verteilungen von Sprache und Rauschen impliziert. Die Relinearisierung basiert auf der Maximierung der a posteriori Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t)$ mit dem Newton-Raphson-Verfahren (vgl. [BSRLK01]). Dabei ergeben sich die Iterationsschritte [BSRLK01]:

$$\begin{aligned} \mathbf{K}_t^{(i)}(s_t) &= \mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)'}(s_t)\left(\mathbf{H}_z^{(i)}(s_t)\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)'}(s_t) + \mathbf{W}\right)^{-1} \\ \mathbf{z}_{t|1:t}^{(i)}(s_t) &= \mathbf{z}_{t|1:t}^{(i-1)}(s_t) + \mathbf{K}_t^{(i)}(s_t)(\mathbf{y}_t - \mathbf{h}(\mathbf{z}_{t|1:t}^{(i-1)}(s_t)) \\ \mathbf{P}_{t|1:t}^{(i)}(s_t) &= \left(\mathbf{I} - \mathbf{K}_t^{(i)}(s_t)\mathbf{H}_z^{(i)}(s_t)\right)\mathbf{P}_{t|1:t-1}(s_t), \end{aligned} \tag{4.28}$$

wobei $i = 1 \ldots N_{it}$ die Iteration bezeichnet und der Startwert $\mathbf{z}_{t|1:t}^{(0)}(s_t) = \mathbf{z}_{t|1:t}(s_t)$ verwendet wird. Dabei wird die Jacobi-Matrix

$$\mathbf{H}_z^{(i)}(s_t) = [\mathbf{H}_x^{(i)}(s_t) \quad \mathbf{H}_n^{(i)}(s_t)] \tag{4.29}$$

vor jedem Iterationsschritt entsprechend Gl. (4.22) mit dem neuen Entwicklungspunkt

$$\mathbf{z}_{t|1:t}^{(i-1)}(s_t) = \begin{bmatrix} \mathbf{x}_{t|1:t}^{(i-1)}(s_t) \\ \mathbf{n}_{t|1:t}^{(i-1)}(s_t) \end{bmatrix} \tag{4.30}$$

der Taylorreihe aktualisiert.

Wie in Anhang E.1 gezeigt wird, entspricht die Iteration in Gl. (4.28) einem Schritt des ALGONQUIN-Verfahrens. Sie ist wegen der geringeren Anzahl an Matrixinversionen jedoch effizienter als die in [FDAK03] vorgeschlagene Rekursion.

4.2.0.5 Berechnung der Modellwahrscheinlichkeiten

Die Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ der Dynamikmodelle werden in dem in Abschnitt 4.2 dargestellten GPB-Algorithmus aus den Messungen $\mathbf{y}_1^t$ bestimmt. Entsprechend der Bayes'schen Formel gilt [DA04]

$$P(s_t|\mathbf{y}_1^t) \propto p(\mathbf{y}_t|\mathbf{y}_1^{t-1}, s_t)P(s_t), \tag{4.31}$$

wobei $P(s_t)$ beim Training der Modelle ermittelt wird (Gl. (4.4)) und $p(\mathbf{y}_t|\mathbf{y}_1^{t-1}, s_t)$ in den Filtern des SLDMs berechnet werden kann. Für das im letzten Abschnitt eingeführte erweiterte Kalman-Filter erhält man:

$$p(\mathbf{y}_t|\mathbf{y}_1^{t-1}, s_t) = \mathcal{N}(\mathbf{y}_t; \mathbf{h}(\mathbf{z}_{t|1:t-1}(s_t)), \mathbf{H}_z\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z' + \mathbf{W}). \tag{4.32}$$

4.3 Erweiterung des Beobachtungsmodells um dynamische Merkmale

In der HMM-basierten Spracherkennung werden in der Regel dynamische Merkmale in den Merkmalsvektor integriert, um die Modellierungsfehler, die durch die Annahme der statistischen Unabhängigkeit zwischen den Sprachmerkmalen hervorgerufen werden, zu kompensieren. Tatsächlich hat es sich herausgestellt, dass die Erkennungsleistung durch die Berücksichtigung der dynamischen Merkmale signifikant erhöht werden kann (vgl. Abschnitt 2.1.1). Da die dynamischen Sprachmerkmale aus mehreren Sprachrahmen berechnet werden, repräsentieren sie genauso wie AR-Modelle höherer Ordnung statistische Abhängigkeiten, die in einem AR-Modell erster Ordnung nicht berücksichtigt werden. In diesem Abschnitt wird eine Methode eingeführt, um die dynamischen Merkmale effizient in das erweiterte Kalman-Filter zu integrieren, wodurch sich ein robustes Modell der Sprachdynamik ergibt. Der Ansatz wurde erstmals in [WHU08a] veröffentlicht.

Das Zustandsmodell des EKFs wird um die in Anhang C.2 definierten dynamischen Sprachmerkmale erweitert, indem der Zustandsvektor $[\mathbf{x}_t' \; \mathbf{n}_t']'$ durch $[\mathbf{x}_t' \; \delta\mathbf{x}_t' \; \delta^2\mathbf{x}_t' \; \mathbf{n}_t' \; \delta\mathbf{n}_t' \; \delta^2\mathbf{n}_t']$ und der Messvektor $\mathbf{y}_t$ durch $[\mathbf{y}_t' \; \delta\mathbf{y}_t' \; \delta^2\mathbf{y}_t']'$ substituiert werden. Ein approximatives Beobachtungsmodell für die Komponenten $\delta x_t^{(l)}$, $\delta n_t^{(l)}$ und $\delta y_t^{(l)}$ bzw. $\delta^2 x_t^{(l)}$, $\delta^2 n_t^{(l)}$ und $\delta^2 y_t^{(l)}$ der log-spektralen δ- und δ^2-Merkmalsvektoren von Sprache $\mathbf{x}_t^{(l)}$, Rauschen $\mathbf{n}_t^{(l)}$ und verrauschter Sprache $\mathbf{y}_t^{(l)}$ wird in [Gal95] angegeben:

$$\begin{aligned}
\delta y_t^{(l)} &\approx \frac{e^{x_t^{(l)}}}{e^{x_t^{(l)}} + e^{n_t^{(l)}}}\delta x_t^{(l)} + \frac{e^{n_t^{(l)}}}{e^{x_t^{(l)}} + e^{n_t^{(l)}}}\delta n_t^{(l)} \\
\delta^2 y_t^{(l)} &\approx \frac{e^{x_t^{(l)}}}{e^{x_t^{(l)}} + e^{n_t^{(l)}}}\delta^2 x_t^{(l)} + \frac{e^{n_t^{(l)}}}{e^{x_t^{(l)}} + e^{n_t^{(l)}}}\delta^2 n_t^{(l)} \\
&\quad + \frac{e^{(x_t^{(l)}+n_t^{(l)})}}{(e^{x_t^{(l)}} + e^{n_t^{(l)}})^2}[(\delta x_t^{(l)})^2 + (\delta n_t^{(l)})^2 - 2\delta x_t^{(l)}\delta n_t^{(l)}].
\end{aligned} \tag{4.33}$$

Bei der Herleitung von (4.33) wurde ausgenutzt, dass die dynamischen Sprachmerkmale aus Gl. (C.2) die zeitlichen Ableitungen der statischen Sprachmerkmale approximieren, d.h. die Approximationen $\delta\mathbf{y}_t \approx \frac{\partial\mathbf{y}_t}{\partial t}$, $\delta^2\mathbf{y}_t \approx \frac{\partial^2\mathbf{y}_t}{\partial t^2}$, $\delta\mathbf{x}_t \approx \frac{\partial\mathbf{x}_t}{\partial t}$, $\delta^2\mathbf{x}_t \approx \frac{\partial^2\mathbf{x}_t}{\partial t^2}$ und $\delta\mathbf{n}_t \approx \frac{\partial\mathbf{n}_t}{\partial t}$, $\delta^2\mathbf{n}_t \approx \frac{\partial^2\mathbf{n}_t}{\partial t^2}$ getroffen werden können. Der letzte Term in Gl. (4.33) wirkt sich nur für $x_t^{(l)} \approx n_t^{(l)}$ signifikant auf die Beobachtung $\delta^2 y_t^{(l)}$ aus, da die Exponentialfunktionen

$e^{x_t^{(l)}-n_t^{(l)}}$ bzw. $e^{n_t^{(l)}-x_t^{(l)}}$ im Nenner von

$$\frac{e^{x_t^{(l)}+n_t^{(l)}}}{\left(e^{x_t^{(l)}}+e^{n_t^{(l)}}\right)^2}=\frac{1}{e^{x_t^{(l)}-n_t^{(l)}}+2+e^{n_t^{(l)}-x_t^{(l)}}} \tag{4.34}$$

diesen Term für $x_t^{(l)} >> n_t^{(l)}$ bzw. $x_t^{(l)} << n_t^{(l)}$ dominieren. Im Folgenden wird der letzte Term in Gl. (4.33) daher ähnlich wie der Phasenterm in Gl. (4.12) vernachlässigt, so dass sich ein lineares Beobachtungsmodell ergibt. Das linearisierte, log-spektrale Beobachtungsmodell kann mit der DCT-Matrix $\mathbf{M}_{DCT}$ und ihrer (Pseudo-)Inversen $\mathbf{M}_{DCT}^{+}$ unter Verwendung der Notation aus Gl. (4.12) näherungsweise als Funktion cepstraler Merkmalsvektoren geschrieben werden:

$$\begin{aligned}
\delta\mathbf{y}_t &\approx \mathbf{M}_{DCT}\frac{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}}{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}+e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}\mathbf{M}_{DCT}^{+}\delta\mathbf{x}_t+\mathbf{M}_{DCT}\frac{e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}+e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}\mathbf{M}_{DCT}^{+}\delta\mathbf{n}_t \\
\delta^2\mathbf{y}_t &\approx \mathbf{M}_{DCT}\frac{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}}{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}+e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}\mathbf{M}_{DCT}^{+}\delta^2\mathbf{x}_t+\mathbf{M}_{DCT}\frac{e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}{e^{\mathbf{M}_{DCT}^{+}\mathbf{x}_t}+e^{\mathbf{M}_{DCT}^{+}\mathbf{n}_t}}\mathbf{M}_{DCT}^{+}\delta^2\mathbf{n}_t.
\end{aligned} \tag{4.35}$$

In Matrizenschreibweise ergibt sich mit der Approximation aus Gl. (4.21) und dem durch eine Normalverteilung approximierten Fehlerterm $[\mathbf{w}_t' \ \delta\mathbf{w}_t' \ \delta^2\mathbf{w}_t']'$ somit das Beobachtungsmodell

$$\begin{bmatrix}\mathbf{y}_t\\ \delta\mathbf{y}_t\\ \delta^2\mathbf{y}_t\end{bmatrix}=\begin{bmatrix}\mathbf{h}(\mathbf{x}_t^{(0)},\mathbf{n}_t^{(0)})\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}+\mathbf{H}\begin{bmatrix}\mathbf{x}_t-\mathbf{x}_t^{(0)}\\ \delta\mathbf{x}_t\\ \delta^2\mathbf{x}_t\\ \mathbf{n}_t-\mathbf{n}_t^{(0)}\\ \delta\mathbf{n}_t\\ \delta^2\mathbf{n}_t\end{bmatrix}+\begin{bmatrix}\mathbf{w}_t\\ \delta\mathbf{w}_t\\ \delta^2\mathbf{w}_t\end{bmatrix} \tag{4.36}$$

mit der Jacobi-Matrix

$$\mathbf{H}=\begin{bmatrix}\mathbf{H}_x & \mathbf{0} & \mathbf{0} & \mathbf{H}_n & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}_x & \mathbf{0} & \mathbf{0} & \mathbf{H}_n & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{H}_x & \mathbf{0} & \mathbf{0} & \mathbf{H}_n\end{bmatrix}. \tag{4.37}$$

Die Kovarianzmatrizen $\delta\mathbf{W}$ und $\delta^2\mathbf{W}$ des Messrauschens $\delta\mathbf{w}_t \sim \mathcal{N}(\delta\mathbf{w}_t;\mathbf{0},\delta\mathbf{W})$ und $\delta^2\mathbf{w}_t \sim \mathcal{N}(\delta^2\mathbf{w}_t;\mathbf{0},\delta^2\mathbf{W})$ für die dynamischen Merkmale werden im Folgenden genauso wie $\mathbf{W}$ durch Diagonalmatrizen mit konstanten Werten approximiert.

Es ist zu beachten, dass die Zustandsvektoren $[\mathbf{x}_t' \quad \delta\mathbf{x}_t' \quad \delta^2\mathbf{x}_t']'$ und $[\mathbf{n}_t' \quad \delta\mathbf{n}_t' \quad \delta^2\mathbf{n}_t']'$ des erweiterten Modells jeweils 39 Komponenten enthalten. Die hohe Dimension des Zustandsvektors führt dazu, dass die Berechnungen mit dem erweiterten Kalman-Filter sehr zeitaufwendig sind. Außerdem besteht die Gefahr, dass das Training der Zustandsmatrizen anfällig gegenüber Irregularitäten in den Trainingsdaten ist. Aus diesem Grund wird der Subvektor $[\delta\mathbf{x}_t' \quad \delta^2\mathbf{x}_t']'$ mittels einer Hauptachsentransformation (PCA) in einen niedrig-dimensionalen Unterraum abgebildet:

$$\delta_{\mathbf{x}_t}=\mathbf{M}_{PCA}\begin{bmatrix}\delta\mathbf{x}_t\\ \delta^2\mathbf{x}_t\end{bmatrix}, \tag{4.38}$$

wobei für die Dimensionen der Vektorräume $\dim(\delta_{\mathbf{x}_t}) \leq \dim([\delta\mathbf{x}_t' \quad \delta^2\mathbf{x}_t']')$ gilt. Die Subvektoren $[\delta\mathbf{y}_t' \quad \delta^2\mathbf{y}_t']'$ und $[\delta\mathbf{n}_t' \quad \delta^2\mathbf{n}_t']'$ werden auf die gleiche Weise auf $\delta_{\mathbf{y}_t}$ und $\delta_{\mathbf{n}_t}$ reduziert. Dabei wird die PCA-Matrix M_{PCA} aus unverrauschten Trainingsdaten $[\delta\mathbf{x}_t' \quad \delta^2\mathbf{x}_t']'$ mit dem Mittelwert $[\boldsymbol{\mu}_{\delta\mathbf{x}}' \quad \boldsymbol{\mu}_{\delta^2\mathbf{x}}']'$ bestimmt, so dass sie die Eigenvektoren zu den $\dim(\delta_{\mathbf{x}_t})$ größten Eigenwerten der Streumatrix

$$\mathbf{S}_T = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \delta\mathbf{x}_t - \boldsymbol{\mu}_{\delta\mathbf{x}} \\ \delta^2\mathbf{x}_t - \boldsymbol{\mu}_{\delta^2\mathbf{x}} \end{bmatrix} \begin{bmatrix} \delta\mathbf{x}_t - \boldsymbol{\mu}_{\delta\mathbf{x}} \\ \delta^2\mathbf{x}_t - \boldsymbol{\mu}_{\delta^2\mathbf{x}} \end{bmatrix}' \tag{4.39}$$

enthält. Das Beobachtungsmodell für den modifizierten Zustands- und Beobachtungsvektor kann mit der (Pseudo-)Inversen $\mathbf{M}_{PCA}^{+}$ der PCA-Matrix als

$$\begin{bmatrix} \mathbf{y}_t \\ \delta_{\mathbf{y}_t} \end{bmatrix} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_t^{(0)}, \mathbf{n}_t^{(0)}) \\ \mathbf{0} \end{bmatrix} + \tilde{\mathbf{H}} \begin{bmatrix} \mathbf{x}_t - \mathbf{x}_t^{(0)} \\ \delta_{\mathbf{x}t} \\ \mathbf{n}_t - \mathbf{n}_t^{(0)} \\ \delta_{\mathbf{n}t} \end{bmatrix} + \begin{bmatrix} \mathbf{w}_t \\ \delta_{\mathbf{w}t} \end{bmatrix}, \tag{4.40}$$

mit $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_x \quad \tilde{\mathbf{H}}_n]$ und

$$\tilde{\mathbf{H}}_x = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA} \end{bmatrix} \begin{bmatrix} \mathbf{H}_x & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_x \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA}^{+} \end{bmatrix}, \tag{4.41}$$

$$\tilde{\mathbf{H}}_n = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA} \end{bmatrix} \begin{bmatrix} \mathbf{H}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_n \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{PCA}^{+} \end{bmatrix} \tag{4.42}$$

geschrieben werden. Dementsprechend wird das Zustandsübergangsmodell auf die Form

$$\begin{bmatrix} \mathbf{x}_t \\ \delta_{\mathbf{x}t} \end{bmatrix} = \tilde{\mathbf{A}} \begin{bmatrix} \mathbf{x}_{t-1} \\ \delta_{\mathbf{x}t-1} \end{bmatrix} + \tilde{\mathbf{b}} + \tilde{\mathbf{u}}_t, \quad \tilde{\mathbf{u}}_t \sim \mathcal{N}(\tilde{\mathbf{u}}_t; \mathbf{0}, \tilde{\mathbf{C}}(s_t)) \tag{4.43}$$

erweitert, wobei die Parameter $\tilde{\mathbf{A}}(s_t)$, $\tilde{\mathbf{b}}(s_t)$ und $\tilde{\mathbf{C}}(s_t)$, wie in Abschnitt 4.1.0.1 dargestellt, durch die Anwendung eines EM-Algorithmus aus Trainingsdaten bestimmt werden. Die Berechnung der a posteriori Wahrscheinlichkeiten mit dem erweiterten Zustandsmodell entspricht bei der Verwendung von IEKFs dem in Abschnitt 4.2 beschriebenen Vorgehen, wobei sich entsprechende Größen substituiert werden können.

4.4 Glättung

In den vorangehenden Abschnitten wurde die Filterung der Sprachmerkmale, d.h. die Berechnung der Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t)$ untersucht. Da die Wortfolge $w_1 \ldots w_N$ im Erkenner mit einem nicht-kausalen Algorithmus bestimmt wird, also zu Zeitpunkten T_{PTB}, $t \leq T_{PTB}$, aktualisiert wird, führt eine nicht-kausale Glättung, die die Eingaben $\mathbf{y}_1 \ldots \mathbf{y}_{T_{PTB}}$ berücksichtigt, bei verzögerten Eingaben neben der zusätzlichen Rechenzeit

für die Glättung zu keinen weiteren Verzögerungen in der Ausgabe. Im Folgenden wird angenommen, dass die Wortfolge im Back-End erst am Satzende bestimmt wird und alle Sprachrahmen bis $T_{PTB} = T$ bei der Glättung berücksichtigt werden können. Die exakte Berechnung der Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^T)$ ist wegen des exponentiellen Anstieges der möglichen Pfade s_1^T im SLDM nicht möglich. Aus diesem Grund wird in [DDA03b] eine Heuristik verwendet, in der $p(\mathbf{z}_t|\mathbf{y}_1^t) = \mathcal{N}(\mathbf{z}_t; \mathbf{z}_{t|1:t}, \mathbf{P}_{t|1:t})$ und $p(\mathbf{z}_t|\mathbf{y}_t^T) = \mathcal{N}(\mathbf{z}_t; \mathbf{z}_{t|t:T}, \mathbf{P}_{t|t:T})$ durch eine Filterung in Vorwärts- bzw. Rückwärtsrichtung bestimmt werden und die Momente von $p(\mathbf{z}_t|\mathbf{y}_1^T) = \mathcal{N}(\mathbf{z}_t; \mathbf{z}_{t|1:T}, \mathbf{P}_{t|1:T})$ durch

$$\begin{aligned} \mathbf{P}_{t|1:T}^{-1} &= \mathbf{P}_{t|1:t}^{-1} + \mathbf{P}_{t|t:T}^{-1} \\ \mathbf{z}_{t|1:T} &= \mathbf{P}_{t|1:T}(\mathbf{P}_{t|1:t}^{-1}\mathbf{z}_{t|1:t} + \mathbf{P}_{t|t:T}^{-1}\mathbf{z}_{t|t:T}) \end{aligned} \tag{4.44}$$

approximiert werden.

4.5 Experimentelle Untersuchungen

Die experimentellen Untersuchungen wurden auf Test-Set A und Test-Set B der AURORA2 Datenbank sowie verrauschten Daten des Wall Street Journal Task (AURORA4 Datenbank) durchgeführt. Eine ausführliche Beschreibung der Testdatenbanken und der jeweils verwendeten Konfigurationen des Spracherkenners ist in Anhang A zu finden. Die Rauschparameter wurden bei den modellbasierten Entstörungsansätzen, die in diesem Abschnitt untersucht werden, auf der AURORA2 Datenbank aus den ersten und letzten zehn Rahmen und auf der AURORA4 Datenbank aus den ersten und letzten 15 Rahmen des jeweiligen Satzes berechnet. Die Kovarianzmatrix $\mathbf{\Sigma}_n$ der a priori Verteilung des Rauschens wird, wie auch in den nachfolgenden Untersuchungen in dieser Arbeit, als Diagonalmatrix modelliert. Eine Übersicht über die Abkürzungen der untersuchten Entstörungsansätze ist in Anhang G zu finden. In Abb. 4.5 sind die Ergebnisse der Entrauschung qualitativ anhand eines Beispielsatzes der AURORA2 Datenbank dargestellt (Car-Rauschen, Signal-zu-Rausch-Verhältnis(SNR)=5dB).

In der oberen Hälfte von Abb. 4.5 sind die zeitlichen Verläufe der Energiekomponente $x^{(0)}$ des cepstralen Merkmalsvektors für das verrauschte Sprachsignal (blau), das entstörte Sprachsignal (rot) und das tatsächliche Sprachsignal (grün) abgebildet. Die Entrauschung wurde mit dem Baseline-SLDM aus [DA04] (vgl. Anhang C.3) und vier schaltenden Modellen durchgeführt (SLDM-M4). Man erkennt, dass der mittlere quadratische Fehler der geschätzten Werte der $x^{(0)}$-Komponente gegenüber den tatsächlichen Werten unter den beschriebenen Testbedingungen bereits für das Baseline-Verfahren deutlich geringer als der mittlere quadratische Fehler der verrauschten Merkmalskomponente ist.

Der zeitliche Verlauf der Zustandsvariable des SLDMs ist in der unteren Graphik dargestellt. Aus der Graphik wird ersichtlich, dass in näherungsweise statischen Signalausschnitten tendentiell Modelle mit einer kleinen Varianz $\mathbf{C}^{(0,0)}$ ausgewählt werden (gelb), während in Bereichen mit großen Änderungen der Sprachmerkmale Modelle mit großer Varianz (rot) wahrscheinlicher sind.

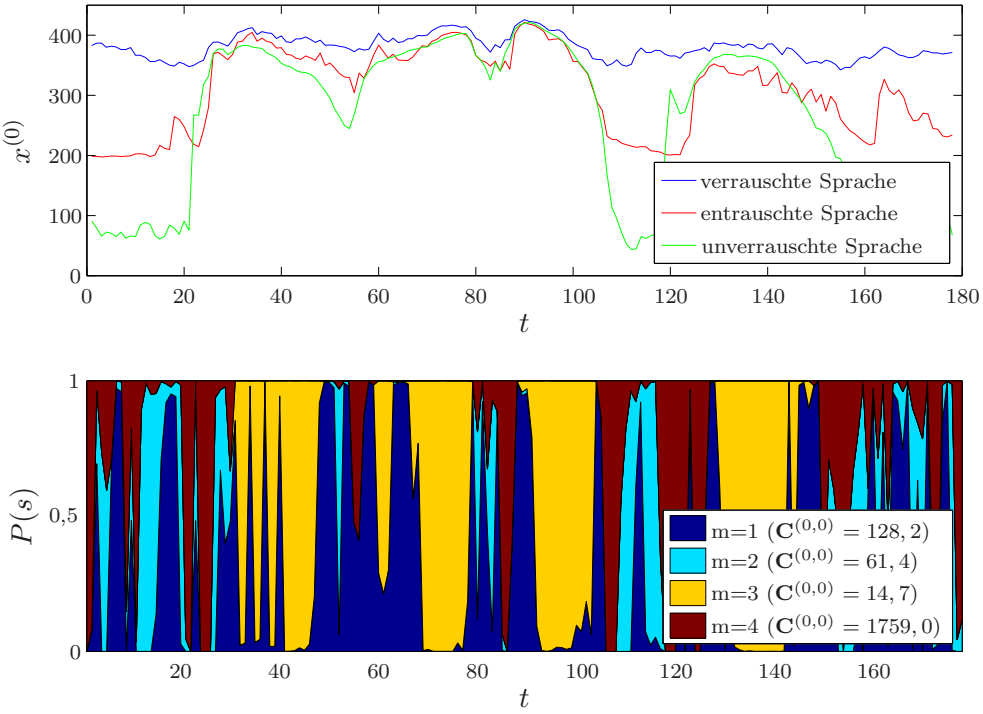

Abbildung 4.5: *Zeitliche Verläufe der $x^{(0)}$-Komponente und der Modellwahrscheinlichkeiten für einen Beispielsatz der AURORA2 Datenbank*

In Abb. 4.6 a) sind zum Vergleich die Kurvenverläufe für ein SLDM mit vier Modellen aufgetragen, in dem durch die zusätzliche Einschränkung $\mathbf{A}(s_t) = \mathbf{0}$ in Gl. (4.4), d.h. die Verwendung einer a priori Verteilung der Ordnung Null, keine statistischen Abhängigkeiten zwischen aufeinander folgenden Sprachrahmen bei der Entrauschung ausgenutzt werden (GMM-M4). Es ist zu erkennen, dass auch für das GMM-M4 der Verlauf der entrauschten $x^{(0)}$-Komponente tendentiell näher als der Verlauf der entsprechenden verrauschten Merkmalsvektorkomponente an dem tatsächlichen Verlauf dieser Komponente liegt. Im Gegensatz zum SLDM-M4 wirken sich Schwankungen des verrauschten Sprachsignals jedoch stark auf den entstörten Merkmalsvektor aus, was auf eine tendentiell größere Varianz der a priori Verteilung aufgrund der fehlenden Berücksichtigung des letzten Sprachrahmens zurückgeführt werden kann. Abb. 4.6 b) zeigt eine typische Trajektorie auf der AURORA4 Datenbank. Das betrachtete Signal schwankt bei der verwendeten Abtastrate von $8kHz$ wesentlich stärker und weist aufgrund des größeren Vokabulars der AURORA4 Datenbank eine facettenreichere Dynamik als ein typisches Signal der AURORA2 Datenbank auf. Starke Signaleinbrüche und -anstiege (in Abb. 4.6 b) z.B.: $t < 10$, $50 < t < 60$) können jedoch prinzipiell besser mit dem GMM als mit dem SLDM (hellblau) verfolgt werden, da im SLDM kleine Signaländerungen tendentiell wahrscheinlicher als große Änderungen sind, während im GMM die a priori Wahrscheinlichkeit ohne die Berücksichtigung des letzten Sprachrahmens ermittelt wird. Auf der anderen Seite weist die Zustandsschätzung

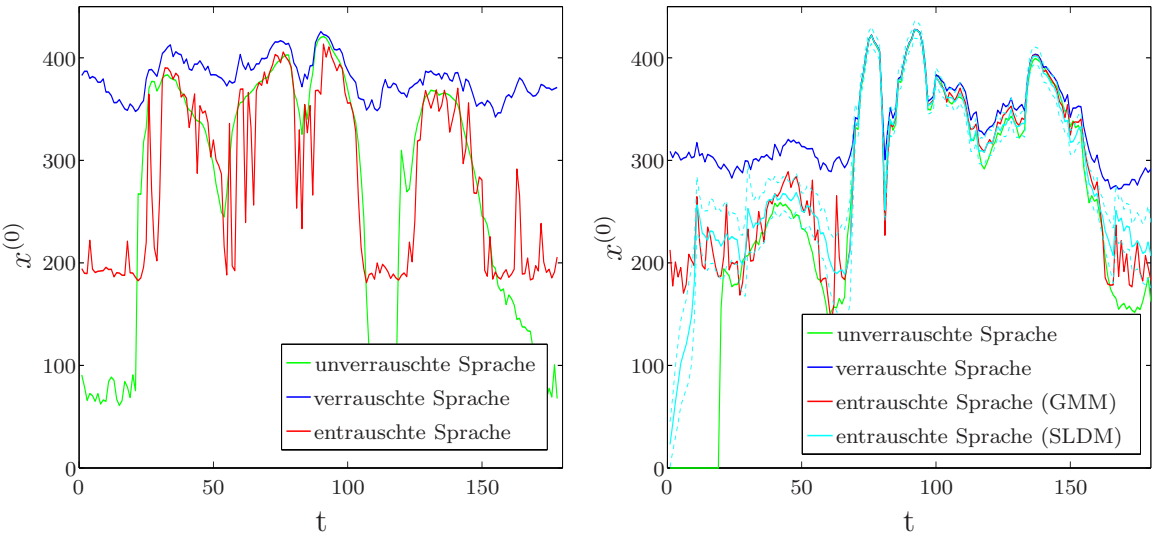

Abbildung 4.6: *Entstörung der $x^{(0)}$-Komponente ohne Ausnutzung von Inter-Frame Korrelationen: a) AURORA2 b) AURORA4*

$x^{(0,SLDM)}$ des SLDMs bei starken Signalschwankungen in der Regel eine größere Varianz $\mathrm{Var}(x^{(0,SLDM)})$ als in Bereichen mit kleinen Signaländerungen auf (gestrichelte Kurve: $x^{(0,SLDM)} \pm \sqrt{\mathrm{Var}(x^{(0,SLDM)})}$). Dies wird bei dem in Abschnitt 6.1 eingeführten Uncertainty Decoding ausgenutzt.

Die quantitativen Ergebnisse der Spracherkennung auf der AURORA4 Datenbank werden in Tab. 4.2 für einige Baseline-Verfahren angegeben. Für jedes Verfahren und die in Anhang A.2 spezifizierten Test-Sets sind jeweils der prozentuale Anteil der Ersetzungsfehler (S), der Einfügungen (I) sowie die Gesamtfehlerrate (E) aufgeführt. In Tab. 4.3 und Tab. 4.4 sind die Erkennungsergebnisse auf der AURORA2 Datenbank für das SFE, in dem keine Entstörung der Sprachmerkmale durchgeführt wird, sowie für die Entrauschung mit dem SLDM-M16 dargestellt.

Der Mittelwert Ø der Erkennungsrate wurde durch die Entrauschung der Merkmale mit dem SLDM-M16 auf Test-Set A der AURORA2 Datenbank um $19,50$ Prozentpunkte und auf Test-Set B um $22,29$ Prozentpunkte erhöht. Auf der AURORA4 Datenbank wurde der Mittelwert Ø der Fehlerrate um $10,8$ Prozentpunkte verringert. Tab. 4.1 und Tab. 4.2 zeigen, dass die Erkennungsrate sowohl auf Test-Set A der AURORA2 Datenbank als auch auf der AURORA4 Datenbank mit der Anzahl M der Modelle ansteigt. In informalen Experimenten wurde für $M < 16$ eine starke Abhängigkeit der Erkennungsrate von der Initialisierung des Zustandsvektors festgestellt.

Gegenüber dem GMM-M16, in dem, wie oben angegeben, eine a priori Verteilung der Ordnung Null verwendet wird, wurde die Erkennungsrate mit dem SLDM-M16 auf den beiden Test-Sets der AURORA2 Datenbank um $3,47$ Prozentpunkte und $2,22$ Prozentpunkte erhöht (siehe Tab. 4.5). Die Erkennungsrate des GMMs konnte durch die Verwendung einer größeren Anzahl Elementarverteilungen nicht weiter verbessert werden (GMM-M32: $76,38\%$ auf Test-Set A). Auf der AURORA4 Datenbank wurden mit dem GMM-M16 deutlich bessere Ergebnisse als mit dem SLDM-M16 erzielt, was

	Subway	Babble	Car	Exhib.	Ø
M=1	77,72	67,10	84,79	80,31	77,48
M=2	78,13	68,10	83,70	81,26	77,80
M=4	79,44	69,49	84,58	81,64	78,79
M=8	80,10	71,03	83,96	81,96	79,26
M=16	80,19	72,56	84,28	82,43	79,87

Tabelle 4.1: SLDM: Erkennungsrate auf Test-Set A der AURORA2 Datenbank abhängig von der Anzahl M der verwendeten Modelle

vermutlich auf die bei der Erläuterung von Abb. 4.6 b) angegebenen Gründe zurückzuführen ist. Durch eine Glättung (SLDM-S16) wurden sowohl auf der AURORA2 Datenbank als auch auf der AURORA4 Datenbank bessere Ergebnisse als mit dem SLDM-M16 erzielt (Tab. 4.6). Allerdings ist das Verfahren nicht kausal und erfordert ungefähr die doppelte Rechenzeit wie die Filterung mit dem SLDM-M16.

Das VTS-Verfahren [Mor96] ist ein bekanntes modellbasiertes Verfahren, das in Anhang C.3 kurz beschrieben wird. Die Ergebnisse wurden ohne die Anwendung eines EM-Algorithmus, um die Rauschparameter neu zu schätzen, also im Wesentlichen durch eine approximative MMSE-Schätzung erzielt. Es handelt sich somit um einen Spezialfall des GMM-M16, in dem die approximative MMSE-Schätzung in Gl. (C.14) ohne die iterative Verbesserung der SNR-Variable in Gl. (C.10) durchgeführt wird (vgl. Anhang C.3). Bei einer verbesserten Rauschschätzung ist daher eine Angleichung der Ergebnisse des GMM-M16 und des VTS-Verfahrens zu erwarten. Als a priori Verteilung wurde, wie in der Praxis üblich, eine Gaußmischungsverteilung mit 128 Komponenten verwendet. Auf der AURORA2 Datenbank ergaben sich mit dem SLDM-M16 deutlich bessere Ergebnisse als mit dem VTS-Verfahren (Tab. 4.7). Da das VTS-Verfahren genauso wie das GMM-M16 eine a priori Verteilung der Ordnung Null verwendet, wurden auf der AURORA4 Datenbank genauso wie mit dem GMM-M16 gegenüber den Verfahren, die auf schaltenden Modellen basieren, bessere Erkennungsergebnisse als auf der AURORA2 Datenbank erzielt. Im Vergleich zum SLDM-M16 waren die Ergebnisse mit dem VTS-Verfahren auf der AURORA4 Datenbank nur um $0,5$ Prozentpunkte schlechter.

Daneben sind die Ergebnisse des AFE angegeben, das auf einer zweistufigen Wiener-Filterung beruht und in Anhang C.4 kurz dargestellt wird. Die Erkennungsrate des AFE liegt sowohl auf der AURORA2 (Tab. 4.8) als auch auf der AURORA4 Datenbank deutlich über der Erkennungsrate der anderen Baseline-Verfahren, die auf einer modellbasierten Merkmalsentstörung basieren.

		Train	Airp.	Bab.	Car	Street	Rest.	Clean	Ø
SFE	S	37,2	37,0	39,7	26,4	36,2	37,1	9,1	28,8
	I	6,5	11,3	8,6	14,1	6,9	8,7	2,4	19,5
	E	63,0	60,7	60,2	44,2	58,6	59,8	12,7	51,3
SLDM-M1	S	33,2	38,1	35,3	14,8	31,7	38,9	9,0	28,7
	I	9,7	16,5	14,3	5,8	10,3	12,3	2,4	10,2
	E	48,1	59,9	54,3	22,9	47,1	57,3	12,6	43,2
SLDM-M2	S	32,4	38,3	35,4	14,4	31,5	38,5	9,1	28,5
	I	9,9	16,7	14,0	6,7	10,7	13,0	1,3	10,3
	E	47,6	60,1	54,0	23,6	47,6	57,3	12,7	43,3
SLDM-M4	S	31,4	38,0	34,1	13,4	30,9	37,9	9,2	27,8
	I	9,5	16,5	13,4	4,6	10,5	13,2	2,9	10,1
	E	45,5	59,1	51,2	20,1	46,0	56,8	13,4	41,7
SLDM-M8	S	31,2	38,8	33,2	13,4	29,7	39,3	9,3	27,8
	I	9,3	15,8	13,9	4,6	10,8	13,1	3,0	10,1
	E	44,8	59,0	50,8	20,2	45,2	57,8	13,5	41,6
SLDM-M16	S	29,9	37,9	32,2	12,7	29,9	37,2	9,5	27,0
	I	8,8	15,9	12,7	4,6	10,6	13,3	3,1	9,9
	E	42,6	58,2	48,8	19,5	45,0	56,0	13,7	40,5
SLDM-M32	S	29,7	39,5	32,3	13,0	29,2	38,1	9,2	26,7
	I	8,7	16,1	13,2	4,2	4,5	14,2	2,7	9,5
	E	42,3	59,7	49,3	19,2	44,2	57,4	13,1	40,7
GMM-M16	S	30,7	29,5	27,5	14,2	27,9	32,7	9,7	24,6
	I	6,8	11,7	10,1	4,8	8,1	10,2	3,2	7,8
	E	42,1	45,5	41,6	20,9	40,6	47,5	14,3	36,1
GMM-M32	S	31,3	29,1	27,5	13,0	28,1	32,3	10,5	24,5
	I	7,6	11,9	9,9	3,8	8,1	11,1	3,1	7,9
	E	43,0	45,7	40,9	18,9	40,9	47,7	15,0	36,0
SLDM-S16	S	28,1	32,9	28,6	12,3	28,6	34,3	8,8	33,4
	I	8,6	14,8	11,9	4,8	10,4	13,2	2,4	9,4
	E	41,8	52,7	44,5	19,2	44,3	53,2	12,6	38,3
VTS-M128	S	27,6	26,5	25,7	15,7	29,7	29,0	9,2	23,3
	I	1,9	3,0	2,1	2,4	2,7	3,1	2,4	2,9
	E	53,6	47,0	46,3	24,4	52,7	49,9	12,9	41,0
AFE	S	22,1	23,2	21,0	12,5	20,8	25,2	8,5	19,0
	I	3,1	7,2	4,3	2,8	4,0	6,2	2,5	4,3
	E	31,7	35,5	30,6	17,6	30,4	36,7	12,0	27,8

Tabelle 4.2: Fehlerraten auf der AURORA4 Datenbank für verschiedene Verfahren und Rauschbedingungen (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,72	99,61	99,61	99,75	99,67
20dB	98,40	89,84	98,39	98,52	96,29
15dB	94,35	72,13	90,34	94,48	87,83
10dB	78,72	44,20	65,20	79,88	67,00
5dB	47,50	16,84	31,73	49,68	36,44
0dB	21,34	0,67	14,20	21,07	14,32
-5dB	9,61	0,00	8,92	9,97	6,80
Ø	68,06	44,74	59,97	68,73	60,37

Set B	Res.	Str.	Air.	Tra.	Ø
clean	99,66	99,58	99,58	99,75	99,64
20dB	92,57	96,77	92,16	93,80	93,83
15dB	78,54	91,60	78,08	81,89	82,53
10dB	55,42	72,22	53,89	58,72	60,06
5dB	27,97	46,10	27,68	29,65	32,85
0dB	5,83	20,83	11,81	11,91	12,60
-5dB	1,69	10,58	7,49	8,08	6,96
Ø	52,07	65,50	52,72	55,19	56,37

Tabelle 4.3: SFE: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,42	99,49	99,55	99,60	99,52
20dB	98,19	96,28	98,96	98,61	98,01
15dB	95,64	91,84	98,27	96,64	95,60
10dB	90,30	81,74	94,51	90,56	89,28
5dB	73,90	62,12	82,49	76,46	73,74
0dB	42,92	30,80	47,18	49,89	42,70
-5dB	17,47	9,79	10,53	22,55	15,09
Ø	80,19	72,56	84,28	82,43	79,87

Set B	Res.	Str.	Air.	Tra.	Ø
clean	99,57	99,53	99,55	99,63	99,57
20dB	96,25	98,10	96,96	98,24	97,39
15dB	90,60	95,47	94,36	95,99	94,11
10dB	80,81	87,30	85,36	91,89	86,34
5dB	64,29	72,46	72,65	79,73	72,28
0dB	39,98	43,11	45,90	53,69	45,67
-5dB	15,78	13,97	20,01	16,60	16,59
Ø	74,39	79,29	79,05	83,91	79,16

Tabelle 4.4: SLDM-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
clean	99,54	99,33	99,46	99,44	99,44
20dB	97,02	98,25	98,78	97,99	98,01
15dB	94,26	95,16	97,76	94,94	95,53
10dB	84,92	84,31	91,32	85,68	86,56
5dB	68,50	58,77	72,41	63,90	65,90
0dB	39,79	28,17	38,14	37,86	35,99
-5dB	18,08	11,91	14,52	18,98	15,87
Ø	76,90	72,93	79,68	76,07	76,40

Set B	Res.	Str.	Air.	Tra.	Ø
clean	99,54	99,30	99,43	99,44	99,43
20dB	97,57	96,83	98,39	98,15	97,74
15dB	93,86	94,01	96,60	96,14	95,15
10dB	86,03	83,34	90,31	90,37	87,51
5dB	62,24	62,42	72,32	71,64	67,16
0dB	34,85	33,19	40,14	40,42	37,15
-5dB	13,48	13,33	17,00	16,04	14,96
Ø	74,91	73,96	79,55	79,34	76,94

Tabelle 4.5: GMM-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
clean	99,63	99,43	99,61	99,63	99,58
20dB	98,37	95,92	99,37	98,03	97,92
15dB	96,10	90,93	98,75	96,02	95,45
10dB	91,68	80,14	96,51	91,24	89,89
5dB	80,53	64,30	88,91	78,90	78,16
0dB	55,66	36,03	62,99	56,03	52,68
-5dB	23,95	9,46	16,19	25,33	18,73
Ø	84,47	73,46	89,31	84,04	82,82

Set B	Res.	Str.	Air.	Tra.	Ø
clean	99,63	99,43	99,61	99,63	99,58
20dB	95,30	98,34	96,96	97,69	97,07
15dB	89,50	96,34	94,53	95,93	94,08
10dB	80,69	90,33	86,04	93,03	87,52
5dB	65,43	77,75	74,56	82,91	75,16
0dB	40,37	49,24	50,79	59,46	49,97
-5dB	14,03	16,38	20,19	17,65	17,06
Ø	74,26	82,40	80,58	85,80	80,76

Tabelle 4.6: SLDM-S16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,60	99,46	99,55	99,60	99,55
20dB	97,67	97,79	97,64	97,50	97,65
15dB	94,01	94,38	94,15	92,44	93,75
10dB	82,35	83,68	80,85	78,49	81,34
5dB	55,97	58,04	48,82	49,83	53,17
0dB	26,53	25,97	15,45	18,85	21,70
-5dB	9,76	9,52	5,52	7,07	7,97
Ø	71,31	71,97	67,38	67,42	69,52

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,60	99,46	99,55	99,60	99,55
20dB	98,34	97,52	98,03	97,62	97,88
15dB	94,50	94,20	95,23	95,19	94,78
10dB	85,02	82,21	87,62	85,19	65,01
5dB	64,23	56,98	63,08	58,75	60,76
0dB	31,29	23,70	32,63	24,13	27,94
-5dB	12,31	9,22	11,21	7,87	10,15
Ø	74,68	54,92	75,32	72,18	69,27

Tabelle 4.7: VTS-M128: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,60	99,52	99,67	99,66	99,61
20dB	98,65	98,94	99,16	98,70	98,86
15dB	97,33	79,37	98,45	97,66	97,70
10dB	94,32	93,14	96,39	94,42	94,57
5dB	86,55	81,41	89,50	86,61	86,02
0dB	67,30	53,23	70,59	66,06	64,30
-5dB	37,06	18,41	35,07	36,32	31,72
Ø	88,83	84,82	90,82	88,69	88,29

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,60	99,52	99,67	99,66	99,61
20dB	98,37	98,61	98,99	98,83	98,70
15dB	96,19	97,61	97,97	97,38	97,29
10dB	91,28	93,59	94,39	95,06	93,58
5dB	78,39	85,28	85,27	86,27	83,80
0dB	52,32	63,75	62,60	66,00	61,17
-5dB	18,36	32,16	28,24	33,14	27,98
Ø	83,31	87,77	87,84	88,71	86,91

Tabelle 4.8: AFE: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Im Folgenden wird die Berechnung der a posteriori Verteilung der Sprachmerkmale mit den in diesem Kapitel behandelten erweiterten Kalman-Filtern untersucht. Auf Test-Set A und Test-Set B der AURORA2 Datenbank wurde mit einer Filterbank mit 16 EKFs (EKF-M16) insgesamt ungefähr die gleiche Erkennungsrate wie mit dem SLDM-M16 erzielt (vgl. Tab. 4.10). Auf der AURORA4 Datenbank ergaben sich, wie Tab. 4.9 zeigt, dagegen deutlich schlechtere Ergebnisse.

		Train	Airport	Babble	Car	Street	Rest.	Clean	Ø
	S	37,8	34,0	33,4	14,6	38,7	38,9	9,4	29,5
EKF	I	6,9	12,1	10,2	4,4	9,2	11,8	2,3	8,1
(M=16)	E	54,0	51,6	48,5	21,8	55,4	57,7	13,1	43,2
	S	31,5	32,5	28,7	13,2	31,6	35,7	9,3	26,1
IEKF	I	7,4	12,3	11,3	4,2	8,8	12,3	2,5	8,4
(M=16)	E	46,7	48,9	43,9	19,9	46,3	53,9	12,8	38,9
	S	29,2	31,9	28,3	12,5	27,8	33,7	9,1	24,6
EKF-d	I	8,2	14,6	12,1	4,4	9,1	14,1	2,5	9,3
(M=16)	E	41,5	50,8	44,9	18,9	41,6	52,3	12,9	37,6
	S	29,3	31,5	28,1	12,3	28,2	33,6	9,1	24,6
IEKF-d	I	7,4	13,4	11,2	4,2	8,4	12,0	2,7	8,5
(M=16)	E	41,0	49,4	42,7	18,6	41,2	50,4	13,1	36,6

Tabelle 4.9: Fehlerraten auf der AURORA4 Datenbank für verschiedene Verfahren und Rauschbedingungen (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

Eine mögliche Ursache liegt darin, dass im EKF die a priori Verteilungen von Sprache und Rauschen anders als im SLDM, wo die SNR-Variablen vor der MMSE-Schätzung mit Hilfe neuer Messungen aktualisiert werden, nicht neu geschätzt werden. Mit der in Gl. (4.28) angegebenen Iteration (IEKF-M16) wurden auf der AURORA4 Datenbank dagegen insgesamt bessere Ergebnisse als mit dem SLDM-M16 erzielt. Auf der AURORA2 Datenbank führte die Iteration insgesamt ebenfalls zu einem signifikanten Anstieg der Erkennungsrate (82, 35% auf Test-Set A und 78, 72% auf Test-Set B), wobei die Ergebnisse für zwei Rauschsorten aus Test-Set B durch die Iteration verschlechtert wurden (Tab. 4.11).

Die Verwendung dynamischer Merkmale wurde auf Test-Set A der AURORA2 Datenbank zunächst mit $M = 4$ Modellen und $\dim(\delta_{\mathbf{x}}) = 1$ (EKF-d-M4) getestet. Wie Tab. 4.12 zeigt, ergab sich bei einer deutlich reduzierten Rechenzeit bereits ein Anstieg der Erkennungsrate von 79, 87% auf 81, 71%. Bei der Verwendung eines höherdimensionalen Subvektors für die dynamischen Sprachmerkmale wurde, wie sich in informalen Experimenten herausgestellt hat, ein Teil der Gewinne, die sich bei der Verwendung einer einzelnen Komponente ergaben, wieder eingebüßt. Für $M = 16$ Modelle (EKF-d-M16) wurde auf Test-Set A (82, 45%) und Test-Set B (80, 98%) der AURORA2 Datenbank (Tab. 4.13) ungefähr die Erkennungsrate des IEKF-M16 erreicht, während diese auf der AURORA4 Datenbank bereits übertroffen wurde (Fehlerrate: 37, 6%). Mit dem

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,60	99,49	99,55	99,69	99,58
20dB	97,91	95,56	99,11	97,99	97,64
15dB	95,58	89,69	98,45	95,65	94,84
10dB	89,78	79,32	95,76	90,77	88,91
5dB	73,75	60,22	85,09	76,37	73,86
0dB	43,78	28,05	53,47	50,85	44,04
-5dB	18,58	8,16	11,66	23,63	15,51
Ø	80,16	70,57	86,38	82,33	79,86

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,48	99,46	99,46	99,66	99,52
20dB	97,51	97,88	97,79	97,93	97,78
15dB	93,80	95,83	96,33	95,99	95,49
10dB	86,64	88,60	89,80	92,69	89,43
5dB	69,76	67,47	76,32	78,19	72,94
0dB	35,16	29,84	43,87	42,05	37,73
-5dB	11,70	7,53	13,60	8,61	10,36
Ø	76,57	75,92	80,82	81,37	78,67

Tabelle 4.10: EKF-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,54	99,52	99,52	99,60	99,55
20dB	98,19	96,25	99,31	98,33	98,02
15dB	96,13	91,17	98,36	96,27	95,48
10dB	92,26	80,74	96,66	92,10	90,44
5dB	80,44	64,36	87,44	80,28	78,13
0dB	52,04	33,25	59,14	54,18	49,65
-5dB	22,87	10,55	14,32	24,90	18,16
Ø	83,81	73,15	88,18	84,23	82,35

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,57	99,55	99,43	99,54	99,52
20dB	96,84	98,13	97,49	98,18	97,66
15dB	92,63	95,53	95,65	96,14	94,99
10dB	84,1	89,21	89,47	93,21	89,00
5dB	66,75	66,90	74,56	80,41	72,16
0dB	35,89	31,59	44,63	47,12	39,81
-5dB	12,02	9,37	15,41	12,25	12,26
Ø	75,24	76,27	80,36	83,01	78,72

Tabelle 4.11: IEKF-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

	Sub.	Bab.	Car	Exh.	Ø
Clean	99,23	99,21	99,40	99,51	99,34
20dB	98,19	96,43	98,90	98,27	97,95
15dB	96,41	92,29	98,51	96,33	95,89
10dB	90,39	82,41	95,65	90,71	89,79
5dB	77,46	65,15	85,39	77,54	76,39
0dB	49,09	36,15	57,77	51,22	48,56
-5dB	22,20	12,88	19,18	24,47	19,68
Ø	82,31	74,49	87,24	82,81	81,71

Tabelle 4.12: EKF-d-M4: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,54	99,52	99,52	99,63	99,55
20dB	98,22	95,89	99,16	98,06	97,83
15dB	96,44	91,51	98,72	96,39	95,77
10dB	91,43	82,74	96,45	91,42	90,51
5dB	79,09	67,71	86,91	79,33	78,26
0dB	50,51	37,85	56,07	55,08	49,88
-5dB	19,80	11,28	10,47	22,74	16,07
Ø	83,14	75,14	87,46	84,06	82,45

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,54	99,52	99,52	99,63	99,55
20dB	96,01	98,58	97,20	97,69	97,37
15dB	90,60	96,37	95,05	95,59	94,40
10dB	83,91	90,87	87,92	92,87	88,89
5dB	86,50	76,18	75,96	82,38	75,76
0dB	44,15	42,96	52,49	54,24	48,46
-5dB	15,44	11,85	20,70	16,60	16,15
Ø	76,63	80,99	81,72	84,55	80,98

Tabelle 4.13: EKF-d-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,54	99,55	99,52	99,44	99,51
20dB	98,22	95,98	99,16	98,06	97,86
15dB	96,53	92,41	98,57	96,39	95,98
10dB	91,43	84,25	96,66	92,07	91,10
5dB	80,26	70,10	87,21	80,16	79,43
0dB	50,23	40,24	55,95	56,34	50,69
-5dB	19,01	11,12	9,78	23,02	15,73
Ø	83,33	76,60	87,51	84,60	83,01

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,54	99,56	99,52	99,57	99,55
20dB	96,13	98,61	97,44	97,84	97,51
15dB	91,56	96,46	95,47	96,02	94,88
10dB	85,78	91,54	89,08	93,34	89,94
5dB	71,14	77,21	77,18	82,66	77,05
0dB	45,38	42,47	53,83	53,90	48,90
-5dB	16,70	11,70	20,85	16,32	16,39
Ø	78,00	81,26	82,60	84,75	81,65

Tabelle 4.14: IEKF-d-M16: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

IEKF-d-M16 ergab sich sowohl auf der AURORA2 Datenbank (Tab. 4.14, Test-Set A: $83,01\%$, Test-Set B: $81,65\%$) als auch auf der AURORA4 Datenbank (Fehlerrate: $36,6\%$) eine weitere Verbesserung der Ergebnisse.

Tab. 4.15 zeigt für die wichtigsten Verfahren, die in diesem Kapitel untersucht wurden, die Komplexität der Merkmalsentstörung auf Test-Set A der AURORA2 Datenbank für ein SNR von $5dB$. Die angegebenen Laufzeiten sind auf die Laufzeit des SFEs normiert. Aus der Tabelle geht hervor, dass der modellbasierte Ansatz aus [DA04], der als Baseline für die Untersuchungen in dieser Arbeit verwendet wurde, für $M = 16$ Modelle eine wesentlich höhere Komplexität als das SFE und das AFE aufweist. Die Rechenzeit der modellbasierten Verfahren auf der Basis schaltender Modelle steigt bei konstanter Merkmalsvektordimension N_c ungefähr mit $O(M \times T)$ an, d.h. proportional zur Anzahl der Modelle M und der Sprachrahmen T. Die Laufzeit des GMMs unterscheidet sich nicht signifikant von der des SLDMs, da der Aufwand für die Prädiktion des Zustandsvektors gegenüber dem Aufwand des Beobachtungsschrittes vernachlässigt werden kann. Die Glättung mit dem SLDM-S16 führte nahezu zu einer Verdoppelung der Rechenzeit. Die Komplexität des EKF-M16 liegt unter der des Baseline-SLDMs,

Verfahren	Laufzeit
SFE	1
AFE	4
SLDM-M16	523
GMM-M16	523
SLDM-S16	1004
EKF-M16	461
IEKF-M16	692
IEKF-d-M16	816

Tabelle 4.15: Laufzeiten der untersuchten Verfahren auf der AURORA2 Datenbank normiert auf die Laufzeit des SFEs

während die Durchführung einer zweiten Iteration zu einer höheren Komplexität führte. Die Berücksichtigung dynamischer Sprachmerkmale, die mit einer Erhöhung der Merkmalsvektordimension N_c verbunden ist, führte zu einem weiteren Anstieg der Rechenzeit. Die relative Laufzeit für die Standarderkennung, die in Anhang A.1 spezifiziert ist, lag auf der AURORA2 Datenbank bei einem SNR von $5dB$ bei 34, d.h. deutlich unter dem Aufwand für die modellbasierte Merkmalsentstörung.

Zusammenfassung

In diesem Abschnitt wurde die Merkmalsentstörung mit SLDMs zunächst qualitativ und quantitativ anhand eines SLDMs aus [DA04] untersucht (SLDM-M4, SLDM-M16), wobei u.a. das Schaltverhalten und die Anzahl der benötigten Modelle betrachtet wurden. Im Vergleich zu einem GMM ergaben sich mit dem SLDM deutlich bessere Ergebnisse auf der AURORA2 Datenbank, während das betrachtete Baseline-Verfahren auf der AURORA4 Datenbank schlechtere Ergebnisse als das GMM erzielte. Durch die Verwendung der in Abschnitt 4.3 eingeführten dynamischen Sprachmerkmale (IEKF-d-M16) ergab sich auf beiden Datenbanken ein deutlicher Anstieg der Erkennungsrate gegenüber dem SLDM-M16 (AURORA2, Test-Set A: $83,01\%$, AURORA2, Test-Set B: $81,65\%$, Fehlerrate auf AURORA4: $36,6\%$). Die Erkennungsrate des AFE wurde jedoch auf beiden Testdatenbanken mit den betrachteten modellbasierten Ansätzen bei Weitem nicht erreicht. Der Rechenaufwand des modellbasierten Baseline-Verfahrens aus [DA04] liegt für $M = 16$ Modelle deutlich über der Komplexität des SFEs und des AFEs. Die Erweiterung dieses Ansatzes um dynamische Sprachmerkmale und die Verwendung iterativer, erweiterter Kalman-Filter führte zu einem Anstieg der Rechenzeit um den Faktor $1,6$.

Kapitel 5

Rauschschätzung

Die modellbasierte Merkmalsentstörung, die im vorangehenden Kapitel untersucht wurde, erfordert eine Schätzung der a priori Verteilung des Rauschens. Diese Schätzung kann einen großen Einfluß auf das Ergebnis der Entrauschung haben. Wenn neben den Parametern der a priori Verteilung auch das Faltungsrauschen und die Phase zwischen Rauschen und Sprache bekannt sind, ist es sogar möglich, die unverrauschten Sprachmerkmale mit Gl. (4.15) exakt aus den verrauschten Sprachmerkmalen zu bestimmen. Im vorangehenden Kapitel wurden konstante Parameter der a priori Veteilung angenommen, die aus sprachfreien Signalabschnitten am Anfang und Ende der einzelnen Sätze geschätzt wurden. Die experimentellen Untersuchungen in Abschnitt 5.3 zeigen jedoch, dass dieser Modellierungsansatz auch bei einer genaueren Parameterschätzung, in der die vorgegebenen, wahren Rauschwerte aller Sprachrahmen der jeweiligen Sätze berücksichtigt werden, zu erheblichen Erkennungsfehlern führen kann. In diesem Kapitel wird daher ein dynamischer Modellierungsansatz verfolgt, in dem die Dynamik des Rauschens mit einem Zustandsmodell beschrieben und auf diese Weise statistische Abhängigkeiten innerhalb des Rauschprozesses ausgenutzt werden. Wie in Abschnitt 2.3 ausgeführt wurde, werden in der Literatur zur modellbasierten Merkmalsentstörung in der Regel die Dynamikmodelle aus Gl. (2.18) und Gl. (2.19) verwendet. In diesen Ansätzen wird mit den Zustandsmodellen unmittelbar die Dynamik des in den Sprachpausen beobachteten Rauschprozesses $\tilde{\mathbf{n}}_t$ beschrieben. Es wird somit die Annahme getroffen, dass $\tilde{\mathbf{n}}_t$ der Zustandsvariable $\mathbf{n}_t$ entspricht.

In Abb. 5.1 ist der zeitliche Verlauf der Energiekomponente des cepstralen Rauschvektors für einen Beispielsatz bei Babble-Rauschen dargestellt. Wie die Abbildung zeigt, kann das beobachtete Rauschen (blaue Kurve) in der Praxis starken Schwankungen unterworfen sein, so dass es mit einem einfachen Zustandsmodell nicht exakt beschrieben werden kann. Unter der Annahme der Identität $\tilde{\mathbf{n}}_t = \mathbf{n}_t$ und eines Random-Walk-Modells

$$\mathbf{n}_t = \mathbf{n}_{t-1} + \mathbf{v}_t \tag{5.1}$$

mit dem Zustandsrauschen $\mathbf{v}_t$ ergibt sich für den Sprachrahmen $t = 30$ des dargestellten Satzes z.B. ungefähr der quadratische Prädiktionsfehler $\left(E[\mathbf{n}_t|\tilde{\mathbf{n}}_1^{t-1}] - \tilde{\mathbf{n}}_t\right)^2 \approx 900$. Eine große Varianz des Zustandsmodells kann dazu führen, dass die Rauschschätzung stark von der aktuellen Messung abhängt und dadurch unzuverlässig wird. Um eine zuverlässigere Rauschschätzung zu erreichen, wird das Rauschen im Folgenden mit einer verborgenen, unbeobachtbaren Zustandsvariable modelliert [WHU08a]. In Abb. 5.1

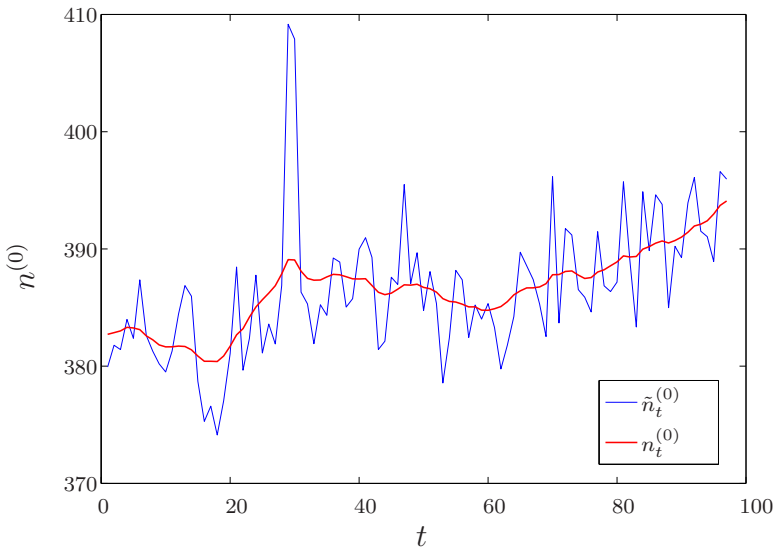

Abbildung 5.1: *Zeitlicher Verlauf der ersten Komponente des cepstralen Rauschvektors (Beispiel für Babble-Rauschen und SNR=-5dB)*

ist der Verlauf der Energiekomponente $n_t^{(0)}$ dieser Zustandsvariable rot dargestellt. Der dargestellte Verlauf wurde experimentell ermittelt, indem zehn Iterationen des in Abschnitt 5.1 eingeführten blockweisen EM-Algorithmus durchgeführt wurden. Danach ergab sich der Erwartungswert $\mathbf{n}_{t-1|1:T}$ des Zustandsvektors unter Berücksichtigung der Beobachtungen $\tilde{\mathbf{n}}_1^T$ als erster Subvektor des geglätteten Zustandsvektors $\boldsymbol{\eta}_{t-1|1:T}$ in Gl. (5.13). Die Abbildung zeigt, dass der Verlauf dieser Zustandsvariable wesentlich geringeren Schwankungen als der Verlauf der Beobachtungen unterworfen ist, so dass er mit dem in Gl. (5.1) angegebenen Random-Walk-Modell mit einer geringeren Varianz prädiziert werden kann. Das Zustandsrauschen $\mathbf{v}_t$ in Gl. (5.1) wird im Folgenden als Gaußverteilung $\mathbf{v}_t \sim \mathcal{N}(\mathbf{v}_t; \mathbf{0}, \mathbf{V})$ mit der Kovarianzmatrix $\mathbf{V}$ modelliert. Dies entspricht den Festlegungen

$$\mathbf{D} = \mathbf{I}, \quad \mathbf{e} = \mathbf{0} \tag{5.2}$$

in Gl. (4.6). Die verbleibende Fehlerwahrscheinlichkeit, die auf stark instationäre Rauschquellen sowie den Beobachtungsprozess (z.B. Varianz bei der Berechnung des Kurzzeitspektrums) zurückgeführt werden kann, wird mit einem Beobachtungsmodell beschrieben:

$$\tilde{\mathbf{n}}_t \;=\; \mathbf{n}_t + \mathbf{w}_t^{(n)} \tag{5.3}$$

mit $\mathbf{w}_t^{(n)} \sim \mathcal{N}(\mathbf{w}_t^{(n)}; \mathbf{0}, \mathbf{W}^{(n)})$, wobei $\mathbf{W}^{(n)}$ die Varianz des Beobachtungsrauschens bezeichnet. Es ist zu beachten, dass $\mathbf{n}_t$ im Beobachtungsmodell für die verrauschten Sprachmerkmale (Gl. (4.15)) durch das beobachtete Rauschen $\tilde{\mathbf{n}}_t$ substituiert werden muß. Man erhält somit das Beobachtungsmodell

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{M}_{DCT} \log\left(1 + e^{\mathbf{M}_{DCT}^{+}\left(\mathbf{n}_t - \mathbf{x}_t + \mathbf{w}_t^{(n)}\right)}\right) \tag{5.4}$$

für die verrauschten Sprachmerkmale, das gegenüber (4.15) zusätzlich die Zufallsvariable $\mathbf{w}_t^{(n)}$ enthält. Diese kann in dem linearisierten Beobachtungsmodell (4.21) berücksichtigt werden, indem die Taylorreihenentwicklung nicht nur bzgl. $\mathbf{x}_t$ und $\mathbf{n}_t$ sondern auch bzgl. $\mathbf{w}_t^{(n)}$ durchgeführt wird. Dadurch ergibt sich in Gl. (4.21) die gegenüber Gl. (4.24) erhöhte Varianz

$$\mathbf{W} = \mathbf{W}_c + \mathbf{H}_n \mathbf{W}^{(n)} \mathbf{H}_n'. \tag{5.5}$$

Der additive Term $\mathbf{H}_n \mathbf{W}^{(n)} \mathbf{H}_n'$ entspricht qualitativ dem in [Kim02] experimentell ermittelten Korrekturfaktor der Kalman-Verstärkung. Dieser führte bei der Festlegung der Varianz $\mathbf{V}$ als konstante Diagonalmatrix mit kleinen Werten zu einer deutlichen Verbesserung der Ergebnisse, was unter der Annahme eines beobachtbaren Rauschprozesses jedoch nicht theoretisch erklärt werden konnte.

In den nächsten beiden Abschnitten werden EM-Algorithmen für die Schätzung der Modellparameter $\mathbf{V}$ und $\mathbf{W}^{(n)}$ hergeleitet (vgl. [WHU08b]). Die EM-Algorithmen können auf sprachfreien Trainingsdaten (Abschnitt 5.1) oder zur Laufzeit mit verrauschten Sprachdaten (Abschnitt 5.2) durchgeführt werden.

5.1 Parameterschätzung aus Trainingsdaten

Im Folgenden wird die Schätzung der Parameter mit einem blockweisen EM-Algorithmus auf Trainingsdaten betrachtet. Dadurch können zusätzliche Unsicherheiten durch das Messmodel in Gl. (4.15), welches eine zuverlässige Schätzung der unverrauschten Sprachmerkmale erfordert, ausgeschlossen werden. Weiterhin ist es möglich, eine große Anzahl Iterationen des EM-Algorithmus sowie eine Glättung im E-Schritt ohne zusätzlichen Aufwand während der Laufzeit durchzuführen. Der Maximum-Likelihood(ML)-Schätzwert $\hat{\boldsymbol{\theta}}$ des Parametervektors $\boldsymbol{\theta} = [\mathbf{V} \quad \mathbf{W}^{(n)}]$ ergibt sich in Anlehnung an [GBW98], wo die AR-Parameter des Sprachsignals zum Zwecke der einkanaligen Sprachsignalverbesserung mit einem blockweisen EM-Algorithmus geschätzt werden, als Lösung des Optimierungsproblems

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \log p(\tilde{\mathbf{n}}_1^T; \boldsymbol{\theta}) \right\}. \tag{5.6}$$

Die unvollständigen, beobachtbaren Rauschwerte $\tilde{\mathbf{n}}_1^T$, aus denen $\boldsymbol{\theta}$ nicht eindeutig bestimmt werden kann, sind über das Beobachtungsmodell (5.3) mit den Zustandsvektoren $\mathbf{n}_0^T$ verknüpft, die zusammen mit $\tilde{\mathbf{n}}_1^T$ vollständige Daten für das Optimierungsproblem in Gl. (5.6) darstellen. Damit ergibt sich der EM-Algorithmus

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}) = E_{\hat{\boldsymbol{\theta}}^{(l)}}[\log p(\mathbf{n}_0^T, \tilde{\mathbf{n}}_1^T; \boldsymbol{\theta}) | \hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] \quad \text{(E-Schritt)} \tag{5.7}$$

$$\hat{\boldsymbol{\theta}}^{(l+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{ Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(l)}) \} \quad \text{(M-Schritt)}. \tag{5.8}$$

In dem Expectation(E)-Schritt wird in der Iteration $l+1$ die a posteriori Wahrscheinlichkeit $p(\mathbf{n}_0^T | \tilde{\mathbf{n}}_1^T; \boldsymbol{\theta})$ des beobachteten Rauschens mit einem RTS-Glätter, der die Parameter aus der Iteration l verwendet, ermittelt. Da der Maximization(M)-Schritt, wie

weiter unten gezeigt, die Berechnung von Kovarianzen zwischen $\mathbf{n}_{t-1}$ und $\mathbf{n}_t$ erfordert, wird für die Glättung der erweiterte Zustandsvektor $\boldsymbol{\eta}_t = [\mathbf{n}_t' \quad \mathbf{n}_{t-1}']'$ verwendet. Für $\boldsymbol{\eta}_t$ erhält man aus Gl. (5.1) das Zustandsmodell

$$\boldsymbol{\eta}_t = \mathbf{A}_\eta \boldsymbol{\eta}_{t-1} + \mathbf{v}_t^{(\eta)}, \quad \mathbf{v}_t^{(\eta)} \sim \mathcal{N}(\mathbf{v}_t^{(\eta)}; \mathbf{0}, \mathbf{C}_\eta) \tag{5.9}$$

mit

$$\mathbf{A}_\eta = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \text{und} \quad \mathbf{C}_\eta = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{5.10}$$

Aus Gl. (5.3) ergibt sich das Beobachtungsmodell

$$\tilde{\mathbf{n}}_t = \mathbf{H}_\eta \boldsymbol{\eta}_t + \mathbf{w}_t^{(n)}, \quad \mathbf{w}_t^{(n)} \sim \mathcal{N}(\mathbf{w}_t^{(n)}; \mathbf{0}, \mathbf{W}^{(n)}), \quad \mathbf{H}_\eta = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{5.11}$$

Die Wahrscheinlichkeitsdichte $p(\boldsymbol{\eta}_t | \tilde{\mathbf{n}}_1^T; \boldsymbol{\theta})$ erhält man durch Anwendung eines RTS-Glätters [BSRLK01]. Dazu wird zunächst eine Filterung in Vorwärtsrichtung durchgeführt, d.h. für die Zeitpunkte $t = 1, \ldots, T$ die Rekursion

$$\begin{aligned}
\boldsymbol{\eta}_{t|1:t-1} &= \mathbf{A}_\eta \boldsymbol{\eta}_{t-1|1:t-1} \\
\mathbf{P}_{t|1:t-1}^{(\eta)} &= \mathbf{A}_\eta \mathbf{P}_{t-1|1:t-1}^{(\eta)} \mathbf{A}_\eta' + \mathbf{C}_\eta \\
\mathbf{K}_t^{(\eta)} &= \mathbf{P}_{t|1:t-1}^{(\eta)} \mathbf{H}_\eta' (\mathbf{H}_\eta \mathbf{P}_{t|1:t-1}^{(\eta)} \mathbf{H}_\eta' + \mathbf{W}^{(n)})^{-1} \\
\boldsymbol{\eta}_{t|1:t} &= \boldsymbol{\eta}_{t|1:t-1} + \mathbf{K}_t^{(\eta)} (\tilde{\mathbf{n}}_t - \mathbf{H}_\eta \boldsymbol{\eta}_{t|1:t-1}) \\
\mathbf{P}_{t|1:t}^{(\eta)} &= (\mathbf{I} - \mathbf{K}_t^{(\eta)} \mathbf{H}_\eta) \mathbf{P}_{t|1:t-1}^{(\eta)}
\end{aligned} \tag{5.12}$$

angewendet. In Gl. (5.12) wird $\mathbf{V}$ in $\mathbf{C}_\eta$ durch $\hat{\mathbf{V}}^{(l)}$ und $\mathbf{W}^{(n)}$ durch $(\hat{\mathbf{W}}^{(n)})^{(l)}$ approximiert, die wie in Gl. (5.17) und Gl. (5.21) angegeben in der vorangehenden Iteration berechnet werden können. Anschließend erfolgt die Glättung in Rückwärtsrichtung. Für $t = T, \ldots, 1$ ergibt sich die Rekursion [BSRLK01]

$$\begin{aligned}
\boldsymbol{\eta}_{t-1|1:T} &= \boldsymbol{\eta}_{t-1|1:t-1} + \mathbf{S}_{t-1}^{(\eta)} (\boldsymbol{\eta}_{t|1:T} - \boldsymbol{\eta}_{t|1:t-1}) \\
\mathbf{P}_{t-1|1:T}^{(\eta)} &= \mathbf{P}_{t-1|1:t-1}^{(\eta)} + \mathbf{S}_{t-1}^{(\eta)} (\mathbf{P}_{t|1:T}^{(\eta)} - \mathbf{P}_{t|1:t-1}^{(\eta)}) \mathbf{S}_{t-1}^{(\eta)\,\prime}
\end{aligned} \tag{5.13}$$

mit der Hilfsvariable

$$\mathbf{S}_{t-1}^{(\eta)} = \mathbf{P}_{t-1|1:t-1}^{(\eta)} \mathbf{A}_\eta' (\mathbf{P}_{t|1:t-1}^{(\eta)})^{-1}. \tag{5.14}$$

Um den M-Schritt durchzuführen, wird der Erwartungswert

$$E[\log p(\mathbf{n}_0^T, \tilde{\mathbf{n}}_1^T; \boldsymbol{\theta}) | \hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] = E[\log p(\tilde{\mathbf{n}}_1^T | \mathbf{n}_0^T; \boldsymbol{\theta}) | \hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] + E[\log p(\mathbf{n}_0^T; \boldsymbol{\theta}) | \hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] \tag{5.15}$$

berechnet. Da $E[\log p(\tilde{\mathbf{n}}_1^T|\mathbf{n}_0^T;\boldsymbol{\theta})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T]$ unabhängig von $\mathbf{V}$ ist, erhält man $\hat{\mathbf{V}}^{(l+1)}$ durch die Maximierung von

$$\begin{aligned}
& E[\log p(\mathbf{n}_0^T;\boldsymbol{\theta})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
= & \sum_{t=1}^{T} E[\log p(\mathbf{n}_t|\mathbf{n}_1^{t-1};\boldsymbol{\theta})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \overset{(5.1)}{=} \sum_{t=1}^{T} E[\log p(\mathbf{n}_t|\mathbf{n}_{t-1};\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
= & \sum_{t=1}^{T} E\left[\log\left(\frac{1}{\sqrt{(2\pi)^{N_c}|\mathbf{V}|}} e^{-\frac{1}{2}(\mathbf{n}_t-\mathbf{n}_{t-1})'\mathbf{V}^{-1}(\mathbf{n}_t-\mathbf{n}_{t-1})}\right)\Big|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T\right] \\
= & -\frac{N_c T}{2}\log(2\pi) - \frac{T}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{t=1}^{T} E[(\mathbf{n}_t-\mathbf{n}_{t-1})'\mathbf{V}^{-1}(\mathbf{n}_t-\mathbf{n}_{t-1})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T].
\end{aligned} \tag{5.16}$$

Wie in Anhang E gezeigt wird, ergibt sich durch die Maximierung von Gl. (5.16) bzgl. $\mathbf{V}$ mit Lemma 3.2.3 aus [And84] der Schätzwert

$$\begin{aligned}
\hat{\mathbf{V}}^{(l+1)} &= \frac{1}{T}\sum_{t=1}^{T} E[(\mathbf{n}_t-\mathbf{n}_{t-1})(\mathbf{n}_t-\mathbf{n}_{t-1})'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
&= \frac{1}{T}\sum_{t=1}^{T}\Big\{E[\mathbf{n}_t\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] - E[\mathbf{n}_{t-1}\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
&\qquad - E[\mathbf{n}_t\mathbf{n}_{t-1}'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] + E[\mathbf{n}_{t-1}\mathbf{n}_{t-1}'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T]\Big\}.
\end{aligned} \tag{5.17}$$

Die Korrelationen in Gl. (5.17) ergeben sich aus den Momenten

$$\boldsymbol{\eta}_{t|1:T} = \begin{bmatrix}\mathbf{n}_{t|1:T}\\ \mathbf{n}_{t-1|1:T}\end{bmatrix}, \quad \mathbf{P}^{(\eta)}_{t|1:T} = \begin{bmatrix}\mathbf{P}^{(n)}_{t|1:T} & \mathbf{P}^{(n)}_{t,t-1|1:T}\\ \mathbf{P}^{(n)}_{t-1,t|1:T} & \mathbf{P}^{(n)}_{t-1|1:T}\end{bmatrix} \tag{5.18}$$

der a posteriori Wahrscheinlichkeit $p(\boldsymbol{\eta}_t|\tilde{\mathbf{n}}_1^T)$, die im E-Schritt berechnet werden:

$$\begin{aligned}
E[\mathbf{n}_t\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t|1:T}\mathbf{n}_{t|1:T}' + \mathbf{P}^{(n)}_{t|1:T} \\
E[\mathbf{n}_{t-1}\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t-1|1:T}\mathbf{n}_{t|1:T}' + \mathbf{P}^{(n)}_{t-1,t|1:T} \\
E[\mathbf{n}_t\mathbf{n}_{t-1}'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t|1:T}\mathbf{n}_{t-1|1:T}' + \mathbf{P}^{(n)}_{t,t-1|1:T} \\
E[\mathbf{n}_{t-1}\mathbf{n}_{t-1}'|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t-1|1:T}\mathbf{n}_{t-1|1:T}' + \mathbf{P}^{(n)}_{t-1|1:T} \quad .
\end{aligned} \tag{5.19}$$

Analog erhält man durch die Maximierung von

$$\begin{aligned}
& E[\log p(\tilde{\mathbf{n}}_1^T|\mathbf{n}_0^T;\boldsymbol{\theta})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
\overset{(5.3)}{=} & \sum_{t=1}^{T} E[\log p(\tilde{\mathbf{n}}_t|\mathbf{n}_t;\boldsymbol{\theta})|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T] \\
= & -\frac{N_c T}{2}\log(2\pi) - \frac{T}{2}\log|\mathbf{W}^{(n)}| - \frac{1}{2}\sum_{t=1}^{T} E[(\tilde{\mathbf{n}}_t-\mathbf{n}_t)'(\mathbf{W}^{(n)})^{-1}(\tilde{\mathbf{n}}_t-\mathbf{n}_t)|\hat{\boldsymbol{\theta}}^{(l)},\tilde{\mathbf{n}}_1^T]
\end{aligned} \tag{5.20}$$

den Schätzwert

$$\begin{aligned}(\hat{\mathbf{W}}^{(n)})^{(l+1)} &= \frac{1}{T}\sum_{t=1}^{T} E[(\tilde{\mathbf{n}}_t - \mathbf{n}_t)(\tilde{\mathbf{n}}_t - \mathbf{n}_t)'|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] \\ &= \frac{1}{T}\sum_{t=1}^{T} \tilde{\mathbf{n}}_t\tilde{\mathbf{n}}_t' - \tilde{\mathbf{n}}_t E[\mathbf{n}_t|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T]' - E[\mathbf{n}_t|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T]\tilde{\mathbf{n}}_t' + E[\mathbf{n}_t\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T]\end{aligned} \tag{5.21}$$

mit

$$\begin{aligned}E[\mathbf{n}_t|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t|1:T} \\ E[\mathbf{n}_t\mathbf{n}_t'|\hat{\boldsymbol{\theta}}^{(l)}, \tilde{\mathbf{n}}_1^T] &= \mathbf{n}_{t|1:T}\mathbf{n}_{t|1:T}' + \mathbf{P}_{t|1:T}^{(n)} \quad .\end{aligned} \tag{5.22}$$

In dem hergeleiteten EM-Algorithmus wird somit zwischen der Berechnung der Parameter entsprechend Gl. (5.17) und Gl. (5.21) und der RTS-Glättung mit den berechneten Parametern in Gl. (5.12) und Gl. (5.13) iteriert.

5.2 Adaption des Beobachtungsrauschens während der Laufzeit

In der Praxis kann es aufgrund wechselnder Rauschbedingungen notwendig sein, die Rauschparameter während der Filterung an die Umgebungsbedingungen anzupassen. Die wiederholte Filterung des Satzes mit schaltenden Modellen, die zur Durchführung eines blockweisen EM-Algorithmus erforderlich ist, ist sehr zeitaufwändig. Daher wird im Folgenden ein sequentieller EM-Algorithmus hergeleitet, der außerdem eine kausale Verarbeitung sowie die Anpassung der Parameter an Änderungen der Rauschbedingungen innerhalb eines einzelnen Satzes ermöglicht. Die experimentellen Untersuchungen in Abschnitt 5.3 zeigen, dass die Adaption des Zustandsrauschens mit einem blockweisen EM-Algorithmus auf Trainingsdaten die Durchführung einer großen Anzahl an Iterationen erfordert. Während der Laufzeit ist die Varianz des Beobachtungsmodells aufgrund der überlagerten Sprache in der Regel deutlich höher als mit den reinen Rauschdaten, die im Training verwendet werden, so dass eine schlechtere Adaption zu erwarten ist. Es erscheint daher unrealistisch, das Zustandsrauschens während der Laufzeit zu adaptieren, so dass im Folgenden nur das Beobachtungsrauschen $\mathbf{W}^{(n)}$ adaptiert wird. Das Zustandsrauschen wird, wie in Abschnitt 5.1 dargestellt, aus Trainingsdaten ermittelt. Um das Beobachtungsrauschen $\mathbf{W}^{(n)}$ zu schätzen, wird die kausale Zielfunktion

$$\begin{aligned}Q_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1}) = &E[\log p(\mathbf{y}_1^t|\mathbf{z}_0^t; \mathbf{W}^{(n)})|(\hat{\mathbf{W}}^{(n)})_1^{t-1}, \mathbf{y}_1^t] \\ &+ E[\log p(\mathbf{z}_0^t; \mathbf{W}^{(n)})|(\hat{\mathbf{W}}^{(n)})_1^{t-1}, \mathbf{y}_1^t]\end{aligned} \tag{5.23}$$

maximiert. Dabei werden die Schätzwerte $(\hat{\mathbf{W}}^{(n)})_1^{t-1} = \hat{\mathbf{W}}_1^{(n)}, \ldots, \hat{\mathbf{W}}_{t-1}^{(n)}$ anders als im blockweisen EM-Algorithmus aus den vorangehenden Sprachrahmen des jeweiligen Satzes berechnet. Da in Gl. (5.23) nur der Term $E[\log p(\mathbf{y}_1^t|\mathbf{z}_0^t; \mathbf{W}^{(n)})|(\hat{\mathbf{W}}^{(n)})_1^{t-1}, \mathbf{y}_1^t]$

von der Kovarianzmatrix $\mathbf{W}^{(n)}$ des Beobachtungsrauschens abhängt, hat der Term $E[\log p(\mathbf{z}_0^t; \mathbf{W}^{(n)})|(\hat{\mathbf{W}}^{(n)})_1^{t-1}, \mathbf{y}_1^t]$ keinen Einfluß auf das Ergebnis der Parameterschätzung und wird daher im Folgenden vernachlässigt. Damit ergibt sich zur Schätzung des Parameters $\mathbf{W}^{(n)}$ die zu optimierende Zielfunktion

$$\begin{aligned} Q_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1}) &= E[\log p(\mathbf{y}_1^t|\mathbf{z}_0^t; \mathbf{W}^{(n)})|(\hat{\mathbf{W}}^{(n)})_1^{t-1}, \mathbf{y}_1^t] \\ &\approx -\sum_{\tau=1}^{t} R_\tau(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{\tau-1}^{(n)}) \end{aligned}$$

mit

$$R_\tau(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{\tau-1}^{(n)}) = -E[\log p(\mathbf{y}_\tau|\mathbf{z}_\tau; \mathbf{W}^{(n)})|\hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau]. \tag{5.24}$$

In Gl. (5.24) werden nur Beobachtungen bis zum aktuellen Zeitpunkt τ berücksichtigt, um eine kausale Verarbeitung zu ermöglichen. Weiterhin wird die Likelihood $p(\mathbf{y}_\tau|\mathbf{z}_\tau; \mathbf{W}^{(n)})$ durch eine Gaußverteilung mit der Kovarianzmatrix $\mathbf{W}(\mathbf{W}^{(n)})$ aus Gl. (5.5) approximiert:

$$\begin{aligned} R_\tau(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{\tau-1}^{(n)}) \approx &\; E\left[\frac{1}{2}\log\left((2\pi)^{N_c}|\mathbf{W}(\mathbf{W}^{(n)})|\right) + \frac{1}{2}\Delta\mathbf{y}_\tau'\mathbf{W}(\mathbf{W}^{(n)})^{-1}\Delta\mathbf{y}_\tau \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] \\ = &\; \frac{1}{2}\log\left((2\pi)^{N_c}|\mathbf{W}(\mathbf{W}^{(n)})|\right) + \frac{1}{2}E\left[\Delta\mathbf{y}_\tau'\mathbf{W}(\mathbf{W}^{(n)})^{-1}\Delta\mathbf{y}_\tau \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right], \end{aligned} \tag{5.25}$$

mit

$$\Delta\mathbf{y}_\tau = \mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_\tau). \tag{5.26}$$

Der Erwartungswert in der letzten Zeile von Gl. (5.25) kann als

$$\begin{aligned} E\left[\Delta\mathbf{y}_\tau'\mathbf{W}(\mathbf{W}^{(n)})^{-1}\Delta\mathbf{y}_\tau \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] &= E\left[\mathrm{tr}\left(\mathbf{W}(\mathbf{W}^{(n)})^{-1}\Delta\mathbf{y}_\tau\Delta\mathbf{y}_\tau'\right) \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] \\ &= \mathrm{tr}\left(\mathbf{W}(\mathbf{W}^{(n)})^{-1}E\left[\Delta\mathbf{y}_\tau\Delta\mathbf{y}_\tau' \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right]\right) \end{aligned} \tag{5.27}$$

mit

$$\begin{aligned} E\left[\Delta\mathbf{y}_\tau\Delta\mathbf{y}_\tau' \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] &= E\left[(\mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_\tau))(\mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_\tau))' \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] \\ &= \mathbf{y}_\tau\mathbf{y}_\tau' - \mathbf{y}_\tau E\left[\mathbf{h}(\mathbf{z}_\tau) \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right]' \\ &\quad - E\left[\mathbf{h}(\mathbf{z}_\tau) \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right]\mathbf{y}_\tau' + E\left[\mathbf{h}(\mathbf{z}_\tau)\mathbf{h}(\mathbf{z}_\tau)' \,\middle|\, \hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau\right] \end{aligned} \tag{5.28}$$

geschrieben werden. Um die analytische Berechnung der Erwartungswerte in Gl. (5.28) zu ermöglichen, wird das Beobachtungsmodell durch die in Gl. (4.21) angegebene Vektortaylorreihenentwicklung

$$\mathbf{h}(\mathbf{z}_\tau) \approx \mathbf{h}(\mathbf{x}_\tau^{(0)}, \mathbf{n}_\tau^{(0)}) + \mathbf{H}_x(\mathbf{x}_\tau - \mathbf{x}_\tau^{(0)}) + \mathbf{H}_n(\mathbf{n}_\tau - \mathbf{n}_\tau^{(0)}) = \mathbf{h}(\mathbf{z}^{(0)}) + \mathbf{H}_z(\mathbf{z}_\tau - \mathbf{z}^{(0)}) \tag{5.29}$$

um den Entwicklungspunkt

$$\mathbf{z}^{(0)} = \mathbf{z}_{\tau|1:\tau} \tag{5.30}$$

linearisiert. Dadurch erhält man

$$\begin{aligned} E[\mathbf{h}(\mathbf{z}_\tau)|\hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau] &\approx E[\mathbf{h}(\mathbf{z}_{\tau|1:\tau}) + \mathbf{H}_z(\mathbf{z}_\tau - \mathbf{z}_{\tau|1:\tau})|\hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau] \\ &= \mathbf{h}(\mathbf{z}_{\tau|1:\tau}) \\ E[\mathbf{h}(\mathbf{z}_\tau)\mathbf{h}(\mathbf{z}_\tau)'|\hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau] &\approx E[(\mathbf{h}(\mathbf{z}_{\tau|1:\tau}) + \mathbf{H}_z(\mathbf{z}_\tau - \mathbf{z}_{\tau|1:\tau})) \\ &\quad \cdot (\mathbf{h}(\mathbf{z}_{\tau|1:\tau}) + \mathbf{H}_z(\mathbf{z}_\tau - \mathbf{z}_{\tau|1:\tau}))'|\hat{\mathbf{W}}_{\tau-1}^{(n)}, \mathbf{y}_1^\tau] \\ &= \mathbf{h}(\mathbf{z}_{\tau|1:\tau})\mathbf{h}(\mathbf{z}_{\tau|1:\tau})' + \mathbf{P}_{\tau|1:\tau}^{(y)} \end{aligned} \tag{5.31}$$

mit der Kovarianzmatrix

$$\mathbf{P}_{\tau|1:\tau}^{(y)} = \mathbf{H}_z \mathbf{P}_{\tau|1:\tau} \mathbf{H}_z'. \tag{5.32}$$

Der Erwartungswert $\mathbf{z}_{\tau|1:\tau}$ und die Kovarianzmatrix $\mathbf{P}_{\tau|1:\tau}$ ergeben sich aus der Zustandsschätzung des im letzten Kapitel beschriebenen SLDMs (Gl. (4.17)).

Einsetzten der Momente in Gl. (5.28) ergibt

$$E[\Delta\mathbf{y}_\tau \Delta\mathbf{y}_\tau'|\hat{\mathbf{W}}_{\tau-1}, \mathbf{y}_1^\tau] = \Delta\mathbf{y}_{\tau|1:\tau}\Delta\mathbf{y}_{\tau|1:\tau}' + \mathbf{P}_{\tau|1:\tau}^{(y)} \tag{5.33}$$

mit

$$\Delta\mathbf{y}_{\tau|1:\tau} = \mathbf{y}_\tau - \mathbf{h}(\mathbf{z}_{\tau|1:\tau}). \tag{5.34}$$

Insgesamt erhält man somit

$$\begin{aligned} R_\tau(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{\tau-1}^{(n)}) = &\frac{1}{2}\log((2\pi)^{N_c}) + \frac{1}{2}\log|\mathbf{W}(\mathbf{W}^{(n)})| \\ &+ \frac{1}{2}\mathrm{tr}\left(\mathbf{W}(\mathbf{W}^{(n)})^{-1}(\Delta\mathbf{y}_{\tau|1:\tau}\Delta\mathbf{y}_{\tau|1:\tau}' + \mathbf{P}_{\tau|1:\tau}^{(y)})\right). \end{aligned} \tag{5.35}$$

$Q_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1})$ kann rekursiv als

$$Q_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1}) = Q_{t-1}^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-2}) - R_t(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{t-1}^{(n)}) \tag{5.36}$$

geschrieben werden. Um den Einfluß fehlerhafter Messungen $\mathbf{y}_t$ auf die Zielfunktion $Q_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1})$ abzuschwächen, wird der neue Beitrag $R_t(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{t-1}^{(n)})$ zur Zielfunktion schwächer als $Q_{t-1}^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-2})$ gewichtet. Dadurch ergibt sich die modifizierte Zielfunktion $\tilde{Q}_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1})$, die über die Rekursion

$$\tilde{Q}_t^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1}) = \tilde{Q}_{t-1}^{(W)}(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-2}) - \gamma R_t(\mathbf{W}^{(n)}, (\hat{\mathbf{W}}^{(n)})_1^{t-1}) \tag{5.37}$$

definiert ist, wobei $\gamma < 1$ eine Konstante bezeichnet. Im Folgenden wird $\mathbf{W}^{(n)}$ als Diagonalmatrix $\mathrm{diag}(w_n^{(0)}, \ldots, w_n^{(i)}, \ldots, w_n^{(N_c-1)})$ mit den Komponenten $w_n^{(i)}$, $i = 0 \ldots N_c - 1$, modelliert. Der Diagonalvektor mit den Elementen $w_n^{(0)}, \ldots, w_n^{(i)}, \ldots, w_n^{(N_c-1)}$ wird im Folgenden mit $\mathrm{diag}(\mathbf{W}^{(n)})$ bezeichnet. Analog zu [DDA03a], wo die Zielfunktion ebenfalls in der in Gl. (5.37) angegebenen Form geschrieben werden kann, erhält man die

folgende Rekursion zur sequentiellen Aktualisierung der Rauschvarianz, die in [Tit84] aus dem Newton-Raphson-Verfahren abgeleitet wurde:

$$\mathrm{diag}(\hat{\mathbf{W}}_t^{(n)}) = \mathrm{diag}(\hat{\mathbf{W}}_{t-1}^{(n)}) + \gamma \mathbf{K}_t^{-1} \mathbf{s}_t \tag{5.38}$$

mit

$$\mathbf{s}_t = \left. \frac{\partial R_t}{\partial \mathrm{diag}(\mathbf{W}^{(n)})} \right|_{\mathbf{W}^{(n)}=\hat{\mathbf{W}}_{t-1}^{(n)}} \tag{5.39}$$

$$\mathbf{K}_t = - \left. \frac{\partial^2 \tilde{Q}_t^{(W)}}{\partial \mathrm{diag}(\mathbf{W}^{(n)}) \partial \mathrm{diag}(\mathbf{W}^{(n)})'} \right|_{\mathbf{W}^{(n)}=\hat{\mathbf{W}}_{t-1}^{(n)}} . \tag{5.40}$$

Für die Komponenten von $\mathbf{s}_t$ erhält man, wie in Anhang E.2.2 hergeleitet:

$$\begin{aligned} \mathbf{s}_t^{(i)} = \frac{\partial R_t(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{t-1}^{(n)})}{\partial w_n^{(i)}} = -\frac{1}{2} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} (2 - \delta_{kl}) \frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} \\ \cdot \mathbf{c}_k' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \left(\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)} - [\mathbf{W}(\mathbf{W}^{(n)})] \right) [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_l . \end{aligned} \tag{5.41}$$

In Gl. (5.41) bezeichnen $\mathbf{c}_l = [0 \ldots 0 \quad 1 \quad 0 \ldots 0]'$ einen Spaltenvektor, dessen l-te Komponente den Wert Eins besitzt, und

$$\delta_{kl} = \begin{cases} 1 & : \ k = l \\ 0 & : \ k \neq l \end{cases} \tag{5.42}$$

das Kronecker-Symbol. $\mathbf{W}(\mathbf{W}_{t-1}^{(n)})$ kann mit Gl. (5.5) berechnet werden, wobei sich $\mathbf{H}_n$ aus der Vektortaylorreihenentwicklung in Gl. (5.29) mit dem Entwicklungspunkt $\mathbf{n}_{t|t}$ ergibt. Für die partiellen Ableitungen der Komponenten $w^{(k,l)}$ der Matrix $\mathbf{W}(\mathbf{W}^{(n)})$ bzgl. $w_n^{(i)}$ erhält man aus der Beziehung

$$w^{(k,l)} = \mathbf{H}_n^{(k,1:N_c)} \mathbf{W}^{(n)} \mathbf{H}_n'^{\,(1:N_c,l)} = \sum_{i=1}^{N_c} h_n^{(k,i)} w_n^{(i)} h_n^{(l,i)} \tag{5.43}$$

mit dem k-ten Zeilenvektor $\mathbf{H}_n^{(k,1:N_c)}$ und l-ten Spaltenvektor $\mathbf{H}_n^{(1:N_c,l)}$ von $\mathbf{H}_n$:

$$\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} = h_n^{(k,i)} h_n^{(l,i)}, \tag{5.44}$$

wobei $h_n^{(k,i)}$ das Element von $\mathbf{H}_n$ in Zeile k und Spalte i bezeichnet. Aus der Ableitung von Gl. (5.37) ergibt sich für $\mathbf{K}_t$ die Rekursion

$$\mathbf{K}_t = \mathbf{K}_{t-1} - \gamma \mathbf{L}_t, \qquad \text{mit } \mathbf{L}_t = \left. \frac{\partial^2 R_t}{\partial \mathrm{diag}(\mathbf{W}^{(n)})^2} \right|_{\mathbf{W}^{(n)}=\mathbf{W}_{t-1}^{(n)}} . \tag{5.45}$$

Die Komponenten von $\mathbf{L}_t$ werden in Anhang E.2.2 hergeleitet:

$$\begin{aligned} L_t^{(i,j)} &= \frac{\partial^2 R_t(\mathbf{W}^{(n)}, \hat{\mathbf{W}}_{t-1}^{(n)})}{\partial w_n^{(i)} \partial w_n^{(j)}} \\ &= \frac{1}{2} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} \sum_{m=1}^{N_c} \sum_{n=1}^{N_c} (2-\delta_{kl})(2-\delta_{mn}) \frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} \frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \\ &\quad \cdot \Big[\mathbf{c}_k' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_m \mathbf{c}_n' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_l \\ &\quad + \mathbf{c}_k' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_m \mathbf{c}_n' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_l \\ &\quad - \mathbf{c}_k' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_m \mathbf{c}_n' [\mathbf{W}(\mathbf{W}^{(n)})]^{-1} \mathbf{c}_l \Big] . \end{aligned} \tag{5.46}$$

Eine Vereinfachung von Gl. (5.41) und Gl. (5.46) kann dadurch erreicht werden, dass $\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}'$ und die Kovarianzmatrizen $\mathbf{P}_{t|1:t}^{(y)}$ und $\mathbf{W}(\mathbf{W}^{(n)})$ als Diagonalmatrizen mit den Elementen $\Delta y_{t|1:t}^{(k)\,2}$, $(P_{t|1:t}^{(y)})^{(k,k)}$ und $w^{(k,k)}$ modelliert werden und außerdem die Terme $\frac{\partial w^{(k,k)}}{\partial w_n^{(i)}}$ für $i \neq k$ vernachlässigt werden. Dadurch werden die Summen in Gl. (5.41) und Gl. (5.46) jeweils auf einen einzelnen Term reduziert:

$$s_t^{(i)} = -\frac{1}{2} \frac{h_n^{(i,i)\,2}}{w^{(i,i)\,2}} \left(\Delta y_{t|1:t}^{(i)\,2} + (P_{t|1:t}^{(y)})^{(i,i)} - w^{(i,i)} \right) \tag{5.47}$$

$$L_t^{(i,i)} = \frac{h_n^{(i,i)\,4}}{w^{(i,i)\,3}} \left(\Delta y_{t|1:t}^{(i)\,2} + (P_{t|1:t}^{(y)})^{(i,i)} - \frac{1}{2} w^{(i,i)} \right) \tag{5.48}$$

$$L_t^{(i,j)} = 0, \quad \text{für } i \neq j. \tag{5.49}$$

Ein Schritt des sequentiellen EM-Algorithmus zum Zeitpunkt t besteht somit aus den folgenden Operationen:

Algorithmus 1 Schritt des sequentiellen EM-Algorithmus

1: **Für alle** s_t
2: Berechne $p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t; \hat{\mathbf{W}}_{t-1}^{(n)})$ mit Gl. (4.18).
3: Berechne $p(\mathbf{z}_t|\mathbf{y}_1^{t}, s_t; \hat{\mathbf{W}}_{t-1}^{(n)})$ mit Gl. (4.19).
4: Berechne $P(s_t|\mathbf{y}_1^{t}; \hat{\mathbf{W}}_{t-1}^{(n)})$ mit Gl. (4.31).
5: **Ende.**
6: Berechne $p(\mathbf{z}_t|\mathbf{y}_1^{t}; \hat{\mathbf{W}}_{t-1}^{(n)})$ mit Gl. (4.16).
7: Berechne $\mathbf{W}(\hat{\mathbf{W}}_{t-1}^{(n)})$ mit Gl. (5.5).
8: Berechne $\mathbf{P}_{t|1:t}^{(y)}$ mit Gl. (5.32).
9: Berechne $\Delta \mathbf{y}_{t|1:t}$ mit Gl. (5.34).
10: Berechne $\mathbf{s}_t$ mit Gl. (5.41) oder Gl. (5.47).
11: Berechne $\mathbf{L}_t$ mit Gl. (5.46) oder Gl. (5.48).
12: Aktualisiere $\mathbf{K}_t$ mit Gl. (5.45).
13: Aktualisiere $\hat{\mathbf{W}}_t^{(n)}$ mit Gl. (5.38).

$\hat{\mathbf{W}}_t^{(n)}$ kann für $t = 0$ aus sprachfreien Rahmen am Anfang des Satzes oder mit dem Schätzwert aus Gl. (5.21) initialisiert werden.

5.3 Experimentelle Untersuchungen

Die experimentellen Untersuchungen in diesem Kapitel können in drei Abschnitte unterteilt werden. Zunächst wird in Abschnitt 5.3.1 der Einfluß der Rauschschätzung auf die Erkennungsergebnisse untersucht und die Verwendung einer zeitvarianten Rauschschätzung motiviert. In Abschnitt 5.3.2 werden qualitative Untersuchungen des blockweisen und des sequentiellen EM-Algorithmus durchgeführt. Abschließend werden in Abschnitt 5.3.3 die Erkennungsergebnisse für das in diesem Kapitel eingeführte dynamische Rauschmodell und die entwickelten Methoden zur Parameterschätzung angegeben.

5.3.1 Einfluß der Rauschschätzung

Um die Bedeutung der Rauschschätzung bei der modellbasierten Merkmalsentstörung zu zeigen sowie obere und untere Schranken für die Erkennungsleistung zu erhalten, wurden experimentelle Untersuchungen auf der AURORA2 Datenbank und der AURORA4 Datenbank durchgeführt. In den durchgeführten Experimenten wurde die a priori Verteilung des Rauschens als Gaußverteilung $p(\mathbf{n}_t) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{n}_t}, \boldsymbol{\Sigma}_{\boldsymbol{n}})$ mit dem zeitvarianten Erwartungswert $\boldsymbol{\mu}_{\boldsymbol{n}_t}$ und der konstanten Kovarianzmatrix $\boldsymbol{\Sigma}_{\boldsymbol{n}}$ modelliert. Die Merkmalsentstörung wurde in den in diesem Abschnitt beschriebenen Experimenten mit dem SLDM-M1 aus [DA04] durchgeführt (vgl. Kapitel 4).

a)	Sub.	Bab.	Car	Exh.	Ø
20dB	96,38	93,92	98,99	97,99	96,82
15dB	93,71	86,70	97,41	95,03	93,21
10dB	86,40	73,82	92,54	88,80	85,39
5dB	69,11	54,90	81,06	73,40	69,62
0dB	42,98	26,15	53,95	46,34	42,36
-5dB	18,76	9,61	20,46	23,30	18,03
Ø	77,72	67,10	84,79	80,31	77,48

b)	Sub.	Bab.	Car	Exh.	Ø
20dB	97,76	95,16	99,25	98,49	97,67
15dB	96,22	90,02	98,63	96,82	95,42
10dB	91,10	80,47	95,65	92,07	89,82
5dB	79,15	65,02	87,18	81,21	78,14
0dB	56,89	37,58	65,20	58,35	54,51
-5dB	29,20	15,90	28,72	33,48	26,83
Ø	84,22	73,65	89,18	85,39	83,11

Tabelle 5.1: SLDM-M1: Erkennungsraten auf der AURORA2 Datenbank bei einer Rauschschätzung aus a) den Rauschwerten der ersten und letzten zehn Rahmen eines Satzes b) den wahren Rauschwerten des gesamten Satzes

Tab. 5.1 zeigt die Erkennungsergebnisse auf Test-Set A der AURORA2 Datenbank unter der Annahme konstanter Parameter der a priori Verteilung $p(\mathbf{n}_t)$ innerhalb eines Satzes. In Tab. 5.1-a) wurden die Rauschparameter wie bei den Untersuchungen in Abschnitt 4.5 aus den ersten und letzten zehn Rahmen des jeweiligen Satzes berechnet, während in Tab. 5.1-b) die als bekannt vorausgesetzten, wahren Rauschwerte des gesamten Satzes zur Berechnung der Parameter verwendet wurden. Die Erkennungsrate

in Tab. 5.1 a) liegt deutlich unter der Erkennungsrate in Tab. 5.1 b), die eine Abschätzung für das Potential einer konstanten Rauschschätzung darstellt. In [STBP01] wird angegeben, dass der Erwartungswert der Rauschverteilung bereits aus zehn Rahmen hinreichend genau geschätzt werden kann, d.h. die Abweichung der Schätzung von dem tatsächlichen Erwartungswert sich unter der Annahme konstanter Rauschparameter und unter ähnlichen Testbedingungen wie bei den Untersuchungen in diesem Abschnitt nicht signifikant auf die Erkennungsrate auf der AURORA2 Datenbank auswirkt, die Varianz des Rauschens aus diesen Daten jedoch nicht zuverlässig ermittelt werden kann.

Die Ergebnisse auf der AURORA4 Datenbank sind in Tab. 5.2 angegeben. Für

		Train	Airport	Babble	Car	Street	Rest.	Clean	Ø
	S	37,8	34,0	33,4	14,6	38,7	38,9	9,4	29,5
SLDM-M1	I	6,9	12,1	10,2	4,4	9,2	11,8	2,3	8,1
	E	54,0	51,6	48,5	21,8	55,4	57,7	13,1	43,2
	S	29,5	34,7	32,3	12,7	29,4	35,1	9,3	26,1
SLDM-M1-α	I	8,7	15,3	12,6	4,8	9,6	11,8	2,5	10,8
$\alpha = 0$	E	43,4	54,6	49,1	19,8	43,8	52,2	12,8	39,4
	S	27,3	31,1	28,7	12,3	27,7	33,0	9,3	24,2
SLDM-M1-α	I	5,4	10,4	8,8	3,7	9,3	10,1	2,5	7,2
$\alpha = 0,2$	E	40,5	45,4	41,5	18,3	41,8	47,7	12,8	35,4
	S	25,1	25,6	23,2	12,3	37,0	28,7	9,3	23,0
SLDM-M1-α	I	6,5	7,0	5,8	2,7	6,9	7,5	2,5	5,6
$\alpha = 0,4$	E	35,7	36,5	32,7	17,4	37,0	40,5	12,8	30,4
	S	17,4	14,8	16,7	12,4	17,8	16,7	9,3	15,0
SLDM-M1-α	I	4,1	3,0	4,3	2,9	3,9	3,9	2,5	3,5
$\alpha = 1$	E	24,5	20,0	24,0	17,4	25,0	23,2	12,8	21,0

Tabelle 5.2: Fehlerraten auf der AURORA4 Datenbank abhängig von der Rauschschätzung (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

die Ergebnisse in der ersten Zeile von Tab. 5.2 (SLDM-M1) wurden die Parameter der a priori Verteilung ähnlich wie in Tab. 5.1-a) innerhalb eines Satzes als konstant angenommen und aus den ersten und letzten 15 Rahmen des jeweiligen Satzes ermittelt. Die Erkennungsergebnisse in den übrigen Zeilen (SLDM-M1-α) ergaben sich dadurch, dass für jeden Sprachrahmen der als bekannt vorausgesetzte, wahre Rauschwert $\mathbf{n}_t^{(true)}$ und der aus den wahren Rauschwerten des jeweiligen Satzes geschätzte Erwartungswert

$$\boldsymbol{\mu}_{\boldsymbol{n}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{n}_t^{(true)} \tag{5.50}$$

mit dem Faktor α, $0 \leq \alpha \leq 1$, gewichtet wurden:

$$\boldsymbol{\mu}_{\boldsymbol{n}_t} = \alpha \mathbf{n}_t^{(true)} + (1 - \alpha) \boldsymbol{\mu}_{\boldsymbol{n}}. \tag{5.51}$$

Die Kovarianzmatrix $\boldsymbol{\Sigma_n}$ der a priori Verteilung wurde als MMSE-Schätzwert aus den resultierenden Rauschwerten $\boldsymbol{\mu}_{\boldsymbol{n}_t}$, $t = 1 \ldots T$, berechnet und als konstant für den jeweiligen Satz angenommen. Die Komponenten von $\boldsymbol{\Sigma_n}$ wurden zur Berücksichtigung von Modellierungsfehlern mit den Schwellwerten 0.5 für die Diagonalkomponenten $\boldsymbol{\Sigma_n}^{(1,1)}, \ldots \boldsymbol{\Sigma_n}^{(12,12)}$ und mit dem Schwellwert 5 für die Energiekomponente $\boldsymbol{\Sigma_n}^{(0,0)}$ nach unten begrenzt. Wie aus Tab. 5.2 ersichtlich ist, ergaben sich durch die Verwendung der wahren Rauschwerte zur Schätzung des konstanten Rauschvektors $\boldsymbol{\mu_n}$ (SLDM-M1-α, $\alpha = 0$) auf der AURORA4 Datenbank genauso wie auf der AURORA2 Datenbank bessere Ergebnisse als mit der Rauschschätzung aus Sprachrahmen am Anfang und Ende des Satzes (SLDM-M1).

Die Ergebnisse in Tab. 5.3 und in Tab. 5.2 zeigen das Potential einer zeitvarianten Rauschschätzung auf der AURORA2 Datenbank und der AURORA4 Datenbank. Die

	$\alpha = 0$	$\alpha = 0,2$	$\alpha = 0,4$	$\alpha = 1$
20dB	97,67	98,34	98,83	99,22
15dB	95,42	97,02	97,70	98,86
10dB	89,82	92,87	94,29	97,87
5dB	78,14	83,16	86,09	94,15
0dB	54,51	62,13	67,44	82,96
-5dB	26,83	32,03	36,97	59,65
Ø	83,11	86,70	88,87	94,61

Tabelle 5.3: SLDM-M1: Auswirkung der Stationaritätsannahme auf der AURORA2 Datenbank

letzte Spalte in Tab. 5.3 und die letzte Zeile in Tab. 5.2 ($\alpha = 1$) entsprechen dem (hypothetischen) Fall, dass der instantane, cepstrale Rauschvektor zu jedem Zeitpunkt bekannt ist und dienen daher als obere Schranke für das Potential der Merkmalsentstörung mit dem gewälten Ansatz. Die verbleibenden Fehler können auf vereinfachte Modellannahmen bei der Merkmalsentstörung wie die Vektortaylorreihenentwicklung in Gl. (4.21) sowie Vereinfachungen des Beobachtungsmodells, z.B. die Vernachlässigung des Phasenterms zwischen unverrauschter Sprache und Rauschen, zurückgeführt werden. Weiterhin ist zu beachten, dass die Fehlerrate auf der AURORA4 Datenbank selbst für unverrauschte Sprachmerkmale aufgrund einer zu ungenauen statistischen Modellierung im Back-End des Spracherkenners ungefähr bei 13% liegt.

Die großen Unterschiede zwischen den Erkennungsergebnissen in der ersten und letzten Zeile von Tab. 5.3 bzw. in der zweiten und letzten Spalte von Tab. 5.2 zeigen, dass sich ein hohes Potential durch die Verwendung einer guten instantanen Rauschschätzung ergibt. Somit erscheint der Einsatz eines dynamischen Rauschmodells, der in diesem Kapitel untersucht wird, als erfolgversprechender Ansatz gegenüber der Verwendung konstanter Rauschparameter wie in den experimentellen Untersuchungen des vorangehenden Kapitels.

5.3.2 Qualitative Untersuchung der Parameterschätzung

Im folgenden Abschnitt werden die entwickelten Methoden zur Parameterschätzung qualitativ untersucht. In Abb. 5.2 ist die Schätzung des Rauschens mit dem blockweisen EM-Algorithmus, der in Abschnitt 5.1 beschrieben wird, für die Energiekomponente einer künstlichen Rauschtrajektorie dargestellt. Die versteckte Trajektorie des Rauschens (grüne, durchgezogene Linie) wurde zufällig entsprechend dem Zustandsmodell in Gl. (5.1) mit der Varianz $\mathbf{V}^{(0,0)} = 5$ der cepstralen Energiekomponente erzeugt. Anschließend wurde Beobachtungsrauschen entsprechend dem Beobachtungsmodell in Gl. (5.3) mit der Varianz $(\mathbf{W}^{(n)})^{(0,0)} = 200$ hinzugefügt, so dass sich die Beobachtungen $\tilde{n}_t^{(0)}$ des Rauschprozesses (blaue, strichpunktierte Linie) ergaben. Weiterhin ist die geschätzte Trajektorie $\hat{n}_t^{(0)}$ der versteckten Zustandsvariable $n_t^{(0)}$ nach 20 Iterationen des blockweisen EM-Algorithmus in Abb. 5.2 dargestellt (magenta, gestrichelte Kurve). Die Abbildung zeigt, dass die geschätzte Zustandstrajektorie dem tatsächlichen Verlauf

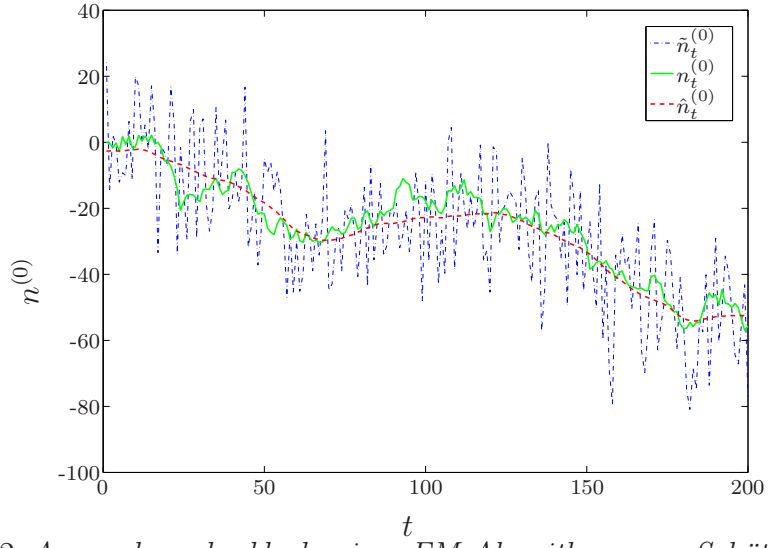

Abbildung 5.2: *Anwendung des blockweisen EM-Algorithmus zur Schätzung einer künstlichen Rauschtrajektorie*

folgt, wobei die geschätzte Trajektorie glatter als die tatsächliche Trajektorie der versteckten Zustandsvariable ist. Ein ähnlicher Verlauf der versteckten Zustandsvariable ergibt sich, wie Abb. 5.1 zeigt, für reale Rauschdaten (Babble Rauschen, SNR$=-5dB$).

In Abb. 5.3.2 ist das Konvergenzverhalten des blockweisen EM-Algorithmus für die Energiekomponente $(\hat{\mathbf{V}}^{(l)})^{(0,0)}$ des Zustandsrauschen aus Gl. (5.17) dargestellt. Die Abbildung zeigt $(\hat{\mathbf{V}}^{(l)})^{(0,0)}$ als Funktion der Iteration l des blockweisen EM-Algorithmus. Da keine verrauschten Trainingsdaten vorlagen, wurden für jede Rauschsorte jeweils die gesamten zur Verfügung stehenden Testdaten des entsprechenden Test-Sets zur Parameterschätzung verwendet.
Man erkennt, dass die Parameterschätzung für alle Rauschsorten gegen einen Grenzwert konvergiert, der abhängig von der Rauschsorte jedoch teilweise erst für $l \approx 50$ näherungsweise erreicht wird.

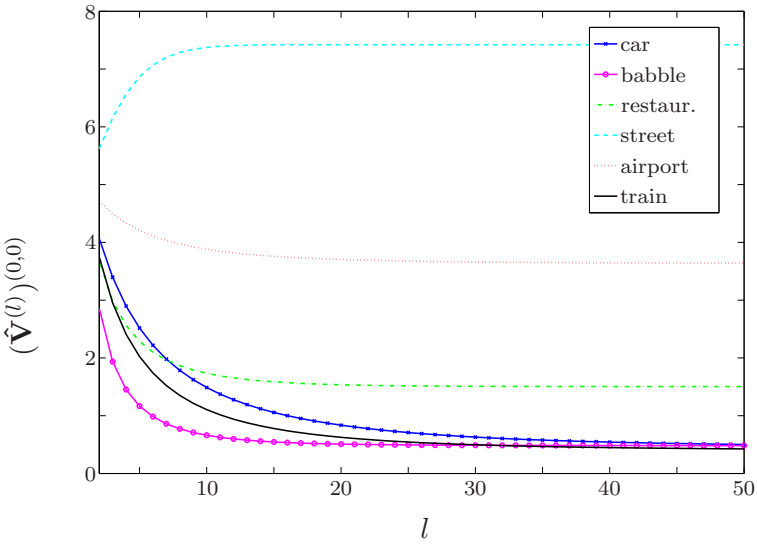

Abbildung 5.3: *Konvergenz des blockweisen EM-Algorithmus*

Anschließend wurde die Merkmalsentstörung mit dem sequentiellen EM-Algorithmus untersucht. Die Cepstra des Rauschens wurden zufällig entsprechend Gl. (5.1) und Gl. (5.3) erzeugt. Um eine von der Schätzung des Zustandsrauschens unabhängige Untersuchung der Parameterschätzung für das Beobachtungsrauschen zu ermöglichen, wurde die Varianz des Zustandsrauschens bei der Generierung des Rauschens und der Merkmalsentstörung auf den Wert $\mathbf{V} = \mathbf{0}$ gesetzt, wodurch sich eine gaußverteilte a priori Verteilung $p(\mathbf{n}_t)$ des Rauschens mit dem konstantem Erwartungswert $\boldsymbol{\mu_n}$ und der Kovarianzmatrix $\boldsymbol{\Sigma_n} = \mathbf{W}^{(n)}$ ergab. Bei der Erzeugung des Rauschens wurde die Energiekomponente des Vektors $\boldsymbol{\mu_n}$ auf den Wert $\mu_n^{(0)} = E[n_t^{(0)}] = 300$ gesetzt. Die Beobachtungsvarianz wurde in den Experimenten für jede Merkmalsvektorkomponente als quadratisch mit der Zeit abfallende Funktion $\mathbf{W}^{(n)} = \mathbf{W}_t^{(n)}$ vorgegeben. Das cepstrale Rauschen wurde entsprechend dem Beobachtungsmodell in Gl. (4.15) zu der Trajektorie der unverrauschten Sprachmerkmale (Abb. 5.4, grüne durchgezogene Linie) hinzugefügt, wodurch sich die blaue, gestrichelte Trajektorie der verrauschten Sprachmerkmale ergab. Die Merkmalsentstörung wurde, wie in Kapitel 4 beschrieben, mit einem SLDM mit $M = 16$ dynamischen Modellen und erweiterten Kalman-Filtern mit dynamischen Sprachmerkmalen (EKF-M16-d) durchgeführt. Die Ergebnisse der Merkmalsentstörung für das EKF-d mit einer sequentiellen Adaption der Beobachtungsvarianz sind in Abb. 5.4 als rote, strichpunktierte Linie dargestellt. In diesem Experiment wurde die Konstante γ auf den Wert $1,0$ festgelegt. Weiterhin wurde eine einfache Detektion der Sprachaktivität (VAD) durchgeführt, indem die Beobachtungsvarianz des Rauschmodells nur für Sprachrahmen mit $|\mathbf{H}_n| > 0,1$ aktualiert wurde.

Abb. 5.5 zeigt die Adaption der Beobachtungsvarianz. Der tatsächliche Verlauf der Energiekomponente der Beobachtungsvarianz, die von 1600 auf 16 reduziert wurde, ist in Abb. 5.5 als durchgezogene rote Linie dargestellt. Die Beobachtungsvarianz wird in Signalabschnitten mit wenig Sprache näherungsweise richtig adaptiert, während die Adaption in Bereichen, in denen die Sprache überwiegt, nur langsam oder überhaupt

nicht erfolgt. Dies ist auf die harte VAD, die sich aus der Vorbedingung $|\mathbf{H}_n| > 0,1$ für die Adaption ergibt, sowie den multiplikativen Term $\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} = h_n^{(k,i)} h_n^{(l,i)}$ in Gl. (5.41) und Gl. (5.46) zurückzuführen, der eine implizite Soft-VAD darstellt.

5.3.3 Erkennungsergebnisse

Das in diesem Kapitel eingeführte Rauschmodell wurde auf der AURORA2 Datenbank und der AURORA4 Datenbank unter den in Anhang A.2 spezifizierten Bedingungen getestet. Auf der AURORA2 Datenbank konnten mit dem neuen Rauschmodell keine signifikanten Verbesserungen erzielt werden. Dies ist wahrscheinlich darauf zurückzuführen, dass die Sätze zu kurz für eine zuverlässige Adaption der Rauschschätzung sind. Die Ergebnisse auf der AURORA4 Datenbank sind in Tabelle 5.4 angegeben. Die Fehlerrate des Baseline-Verfahrens (EKF-d-M16, vgl. Kapitel 4) unter der Annahme stationären Rauschens, d.h. für das Zustandsrauschen $\mathbf{V} = \mathbf{0}$ und konstantes Beobachtungsrauschen $\mathbf{W}^{(n)}$, und einer Parameterschätzung aus den ersten und letzten 15 Sprachrahmen des jeweiligen Satzes beträgt $37,6\%$. Die Rauschschätzung wurde zunächst mit den Approximationen in Gl. (5.47) und Gl. (5.48) durchgeführt. Die Gewichtungskonstante γ im sequentiellen EM-Algorithmus wurde in informalen Experimenten auf den Wert $\gamma = 0,01$ festgelegt. Für $\mathbf{V} = \mathbf{0}$, wodurch der Erwartungswert $\boldsymbol{\mu_n}$ der a priori Rauschverteilung wie im EKF-d-M16 konstant ist, ergab sich durch die Adaption der Rauschvarianz $\mathbf{W}^{(n)}$ mit dem sequentiellen EM-Algorithmus (EKF-d-M16-dyn) eine Verbesserung der Fehlerrate um $1,4$ Prozentpunkte. Durch die Schätzung von $\mathbf{V}$ mit 50 Iteration des blockweisen EM-Algorithmus, wobei die a priori Verteilung des Rauschens für den ersten Sprachrahmen und das Beobachtungsrauschen $\mathbf{W}^{(n)}$ weiterhin aus den ersten und letzten 15 Sprachrahmen des jeweiligen Satzes berechnet wurden,

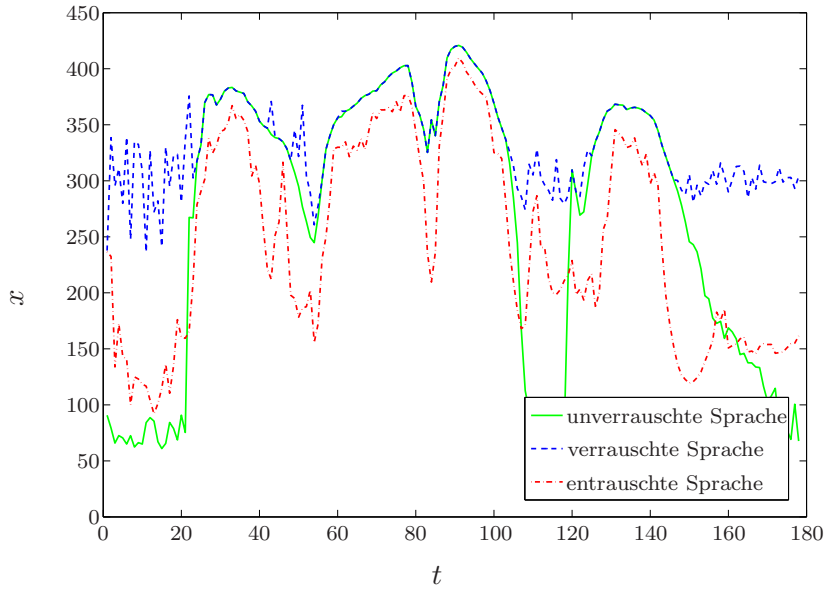

Abbildung 5.4: *Zeitlicher Verlauf der ersten Komponente des cepstralen Merkmalsvektors für künstlich erzeugtes Rauschen (abnehmende Rauschvarianz)*

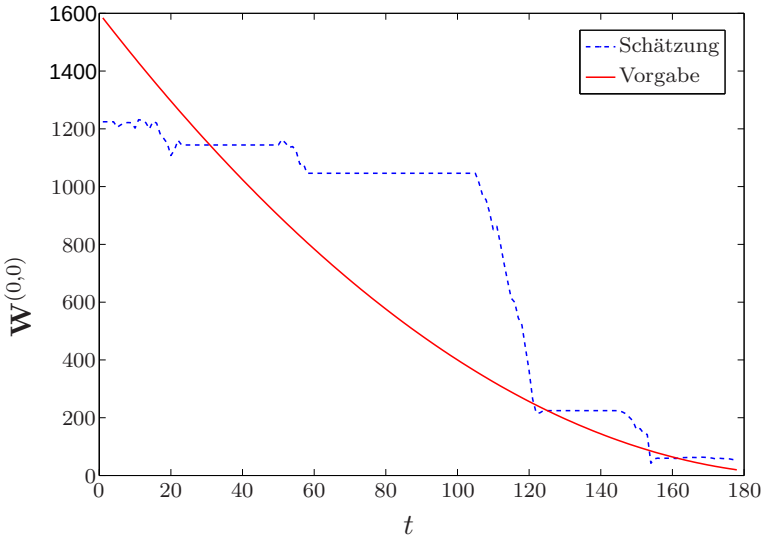

Abbildung 5.5: *Sequentielle Adaption der Rauschvarianz*

wurde die Fehlerrate auf $35,5\%$ reduziert, während sich bei 10 und 25 Iterationen des blockweisen EM-Algorithmus geringere Verbesserungen ergaben.

Die Ergebnisse, die sich für 50 Iterationen des blockweisen EM-Algorithmus bei der Auswertung der vollständigen Summen in Gl. (5.41) und Gl. (5.46) (EKF-d-M16-dyn-f) ergeben, sind in der letzten Zeile der Tabelle angegeben. Die Implementierungsunterschiede gegenüber dem sequentiellen EM-Algorithmus, die sich durch die Vereinfachungen in Gl. (5.41) und Gl. (5.46) ergeben, führen somit zu einem Unterschied in der Erkennungsrate von $0,1$ Prozentpunkten, der statistisch nicht signifikant ist. Insgesamt wurde die Erkennungsrate auf der AURORA4 Datenbank durch die dynamische Rauschschätzung um $2,1$ Prozentpunkte verbessert.

Zum Vergleich sind in Tab. 5.4 auch die Ergebnisse für zwei Dynamikmodelle aus der Literatur angegeben (Comp1, Comp2). In Comp1 wurde die Beobachtungsvarianz $\mathbf{W}^{(n)}$ in Gl. (5.3) ähnlich wie in [SR03] auf den Wert $\mathbf{0}$ gesetzt, während $\mathbf{V}$ aus jeweils 15 Sprachrahmen am Anfang und Ende jedes Satzes geschätzt wurde. In Comp2 wurde $\mathbf{W}^{(n)}$ ebenfalls auf den Wert $\mathbf{0}$ gesetzt, während $\mathbf{V}$ wie in [Kim98a] als kleine Konstante gewählt wurde. Wie in Tab. 5.4 dargestellt, ergab sich mit Comp1 insgesamt eine Fehlerrate von $36,3\%$, während Comp2 zu einer Fehlerrate von $36,6\%$ führte.

Der Rechenaufwand wird durch den sequentiellen EM-Algorithmus gegenüber der modellbasierten Merkmalsentstörung mit dem EKF-d-M16 auch bei der vollständigen Implementierung der Summen in Gl. (5.41) und Gl. (5.46) nicht signifikant erhöht.

		Train	Airp.	Babb.	Car	Street	Rest.	Clean	Ø
	S	29,2	31,9	28,3	12,5	27,8	33,7	9,1	24,6
EKF-d-M16	I	8,2	14,6	12,1	4,4	9,1	14,1	2,5	9,3
	E	41,5	50,8	44,9	18,9	41,6	52,3	12,9	37,6
	S	28,0	31,7	26,7	12,2	27,2	32,9	9,1	24,0
EKF-d-M16-dyn	I	7,6	14,1	11,7	4,3	9,0	12,6	2,5	9,5
$\mathbf{V} = \mathbf{0}$	E	39,7	50,1	41,8	18,4	40,6	49,7	12,9	36,2
	S	31,7	29,5	27,6	12,8	31,3	33,8	9,1	25,1
EKF-d-M16-dyn	I	7,0	9,0	6,8	3,7	6,5	10,4	2,5	9,5
10 Iterationen	E	45,0	44,8	38,3	18,7	43,9	50,1	12,9	36,1
	S	29,9	29,9	27,6	12,7	29,8	34,3	9,1	24,8
EKF-d-M16-dyn	I	7,3	9,2	6,6	3,5	6,8	10,2	2,5	6,6
25 Iterationen	E	43,0	44,2	37,9	18,4	42,7	49,9	12,9	35,6
	S	29,9	29,9	27,6	12,7	29,8	34,3	9,1	24,8
EKF-d-M16-dyn	I	7,3	9,2	6,6	3,5	6,8	10,2	2,5	6,6
50 Iterationen	E	42,4	44,2	37,9	18,4	42,7	49,9	12,9	35,5
	S	29,9	29,7	27,9	12,5	30,1	34,1	9,1	24,8
EKF-d-M16-dyn-f	I	7,3	9,1	6,8	3,3	7,3	9,8	2,5	6,6
50 Iterationen	E	42,7	43,9	38,5	17,8	43,5	49,7	12,9	35,6
	S	31,0	30,1	27,6	12,5	30,7	35,2	9,1	25,2
Comp1	I	7,5	9,2	8,1	2,1	6,2	10,0	2,5	6,5
	E	43,3	44,5	39,7	18,6	44,7	51,0	12,9	36,3
	S	33,3	29,4	27,9	13,4	31,2	33,3	9,1	25,4
Comp2	I	6,8	9,5	7,4	3,7	6,7	10,3	2,5	6,7
	E	47,1	44,4	39,4	19,2	44,3	49,2	12,9	36,6

Tabelle 5.4: Fehlerraten auf der AURORA4 Datenbank für verschiedene Verfahren und Rauschbedingungen (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

Kapitel 6

Optimierungsproblem für verrauschte Sprachmerkmale

In dem Spracherkennungsansatz, der in Abschnitt 2.1 eingeführt wurde, wird die Decodierung im Back-End auf ein Optimierungsproblem für die unverrauschten Sprachmerkmale $\mathbf{x}_1^T$ zurückgeführt. Bei verrauschten Eingangsdaten wurden die Sprachmerkmale, die im Front-End entrauscht werden, bislang durch Punktschätzungen approximiert, d.h. bei einer modellbasierten Merkmalsentstörung durch die Erwartungswerte der Verteilungen $p(\mathbf{x}_t|\mathbf{y}_1^T)$. Den weiteren Ausführungen soll ein strikterer Bayes'scher Ansatz zugrunde gelegt werden, in dem das Optimierungsproblem entsprechend [KF02] für die verrauschten Eingangsdaten $\mathbf{y}_1^T = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ formuliert und somit die Unterteilung zwischen Front-End und Back-End aufgehoben wird:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}}\, p(\mathbf{y}_1^T|w_1^N)P(w_1^N). \tag{6.1}$$

Dabei wird das statistische Modell in Abb. 6.1 zugrunde gelegt, das gegenüber dem in Abschnitt 2.1 beschriebenen HMM um statistische Abhängigkeiten zwischen den unverrauschten Sprachmerkmalen $\mathbf{x}_1 \ldots \mathbf{x}_T$ erweitert wurde. Anders als in [IHU08b]

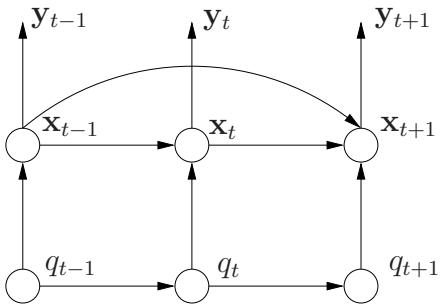

Abbildung 6.1: *Statistisches Modell für die Sprachdynamik*

wird angenommen, dass $\mathbf{x}_t$ von den Sprachmerkmalen aller vorangehenden Sprachrahmen abhängt. Diese Annahme ist notwendig, um die Integration des segmentiellen HMMs, das in Kapitel 7 eingeführt wird, in diesen Ansatz zu ermöglichen. Im Rahmen der Sprachdecodierung im Back-End bezeichnen $\mathbf{x}_t$ und $\mathbf{y}_t$ Merkmalsvektoren der Dimension $N_c = 39$, die neben den statischen Merkmalsvektorkomponenten dynamische Merkmale erster und zweiter Ordnung enthalten und aus den 13-dimensionalen

Merkmalsvektoren im Front-End, wie in Anhang C.2 angegeben, berechnet werden können. Die Lösung des Optimierungsproblems durch die Umformung von $p(\mathbf{y}_1^T|w_1^N)$ wird in Abschnitt 6.1 untersucht. In der Praxis ist es genauso wie für unverrauschte Eingangsdaten erforderlich, das Sprachmodell $P(w_1^N)$ und das Akustikmodell $p(\mathbf{y}_1^T|w_1^N)$ gegeneinander zu gewichten. Auf die Gewichtung der Modelle, die grundlegend für die Algorithmen ist, die in Kapitel 7 und Kapitel 8 entwickelt werden, wird ausführlicher in Abschnitt 6.2 eingegangen.

6.1 Uncertainty Decoding

Basierend auf den Arbeiten von [KF02] und [IHU08b] wird die Lösung des in Gl. (6.1) formulierten Optimierungsproblems für das in Abb. 6.1 dargestellte Modell untersucht. Wie in [KF02] angegeben, kann $p(\mathbf{y}_1^T|w_1^N)$ als

$$p(\mathbf{y}_1^T|w_1^N) = \sum_{q_1^T} p(\mathbf{y}_1^T|q_1^T) \prod_{t=1}^{T} P(q_t|q_{t-1}) \approx \max_{q_1^T} p(\mathbf{y}_1^T|q_1^T) \prod_{t=1}^{T} P(q_t|q_{t-1}) \tag{6.2}$$

geschrieben werden. Unter der Modellannahme in Abb. 6.1 erhält man:

$$\begin{aligned}
p(\mathbf{y}_1^T|q_1^T) &= \int_{\mathbb{R}^{N_c T}} p(\mathbf{y}_1^T|\mathbf{x}_1^T) p(\mathbf{x}_1^T|q_1^T) d\mathbf{x}_1^T && (6.3)\\
&= \int_{\mathbb{R}^{N_c T}} p(\mathbf{y}_1^T|\mathbf{x}_1^T) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t) d\mathbf{x}_1^T && (6.4)\\
&\propto \int_{\mathbb{R}^{N_c T}} \frac{p(\mathbf{x}_1^T|\mathbf{y}_1^T)}{p(\mathbf{x}_1^T)} \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t) d\mathbf{x}_1^T && (6.5)\\
&= \int_{\mathbb{R}^{N_c T}} \prod_{t=1}^{T} \frac{p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, \mathbf{y}_1^T)}{p(\mathbf{x}_t|\mathbf{x}_1^{t-1})} p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t) d\mathbf{x}_1^T. && (6.6)
\end{aligned}$$

Gl. (6.6) kann durch die Vertauschung der Integration und des Produktes vereinfacht werden, wenn die Abhängigkeit von $\mathbf{x}_1^{t-1}$ vernachlässigt wird:

$$p(\mathbf{y}_1^T|q_1^T) \approx \int_{\mathbb{R}^{N_c T}} \prod_{t=1}^{T} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t) d\mathbf{x}_1^T = \prod_{t=1}^{T} \int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t) d\mathbf{x}_t. \tag{6.7}$$

Dies entspricht einer Approximation der Likelihood in Gl. (6.6), die in [IHU08b] für die Kompensation von Paketverlusten im Rahmen einer verteilten Spracherkennung hergeleitet wurde. In der Decodierregel von [KF02], die in Abschnitt 2.4 angegeben wird, wird anstelle der Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^T)$ die Verteilung $p(\mathbf{x}_t|\mathbf{y}_t)$ berechnet. Bei der Kompensation von Paketverlusten werden durch diese weitere Vereinfachung jedoch schlechtere Ergebnisse erzielt [IHU08b]. Eine Verallgemeinerung von Gl. (6.7), in der die zusätzliche Abhängigkeit von $\mathbf{x}_1^{t-1}$ berücksichtigt wird, wird in Kapitel 7 untersucht.

Wie in [IHU08b] gezeigt wird, unterscheidet sich das Optimierungsproblem, das man durch das Einsetzen von Gl. (6.7) in Gl. (6.2) und die anschließende Substitution von

Gl. (6.2) in Gl. (6.1) erhält, von dem Optimierungsproblem für unverrauschte Sprachmerkmale, das sich durch das Einsetzen von Gl. (2.7) in Gl. (2.1) ergibt, ausschließlich bzgl. der Likelihood

$$p_{LH}(\mathbf{y}_1^T|q_t) = \int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t) d\mathbf{x}_t, \tag{6.8}$$

die ein Integral über die Likelihood $p(\mathbf{x}_t|q_t)$ des HMMs aus Abschnitt 2.1 darstellt. Die Viterbi-Approximation kann mit der modifizierten Likelihood somit auch für das verallgemeinerte Optimierungsproblem angewendet werden. Durch die Verwendung der modifizierten Likelihood wird ein sogenanntes Uncertainty Decoding (vgl. Abschnitt 2.4) durchgeführt, da neben dem Erwartungswert der Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^T)$ auch deren Varianz berücksichtigt wird, die ein Maß für die Unsicherheit der Punktschätzung $E[\mathbf{x}_t|\mathbf{y}_1^T]$ darstellt. Die Momente der Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^T)$ können, wie in Kapitel 4 beschrieben, im Front-End des Spracherkenners berechnet werden, während

$$p(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \tag{6.9}$$

die a priori Verteilung der Sprache ist, deren Momente $\boldsymbol{\mu}_{\boldsymbol{x}}$ und $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ aus Trainingsdaten ermittelt werden können. In dem folgenden Unterabschnitt wird auf die Auswertung des Integrals in Gl. (6.8) eingegangen.

6.1.1 Auswertung der Decodierregel unter der Annahme gaußförmiger Verteilungen

In Gl. (6.8) wird die Emissionsverteilung $p(\mathbf{x}_t|q_t)$ des HMMs als Gaußmischungsverteilung

$$p(\mathbf{x}_t|q_t) = \sum_m p(\mathbf{x}_t|m_t = m) P(m_t = m|q_t) \tag{6.10}$$

mit den Einzelverteilungen $p(\mathbf{x}_t|m_t = m) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$ mit Erwartungswerten $\boldsymbol{\mu}^{(m)}$ und diagonalen Kovarianzmatrizen $\boldsymbol{\Sigma}^{(m)}$ sowie den Mischungsgewichten $P(m_t = m|q_t)$ modelliert. Die Verteilung $p(\mathbf{x}_t)$ und die a posteriori Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^T)$ des SLDMs werden, wie in der Literatur üblich, durch Gaußverteilungen $p(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ und $p(\mathbf{x}_t|\mathbf{y}_1^T) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t|1:T}, \mathbf{P}_{t|1:T}^{(x)})$ mit diagonalen Kovarianzmatrizen repräsentiert. Für die Berechnung des Integrals in Gl. (6.8) ergibt sich damit der Zusammenhang [DAD02b, IHU08b]:

$$\int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|q_t) d\mathbf{x}_t = \sum_m A_t P(m_t = m|q_t) \mathcal{N}(\boldsymbol{\mu}_{e_t}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)} + \boldsymbol{\Sigma}_{e_t}) \tag{6.11}$$

mit

$$\begin{aligned} \boldsymbol{\Sigma}_{e_t}^{-1} &= (\boldsymbol{\Sigma}_{t|1:T}^{(x)})^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{x}}{}^{-1}, \\ \boldsymbol{\Sigma}_{e_t}^{-1} \boldsymbol{\mu}_{e_t} &= (\boldsymbol{\Sigma}_{t|1:T}^{(x)})^{-1} \mathbf{x}_{t|1:T} - \boldsymbol{\Sigma}_{\boldsymbol{x}}{}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}, \\ A_t &= \frac{\mathcal{N}(\mathbf{0}; \mathbf{x}_{t|1:T}, \boldsymbol{\Sigma}_{t|1:T}^{(x)})}{\mathcal{N}(\mathbf{0}; \boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \mathcal{N}(\mathbf{0}; \boldsymbol{\mu}_{e_t}, \boldsymbol{\Sigma}_{e_t})}. \end{aligned} \tag{6.12}$$

In [DAD02b] und [LG08] wird angegeben, dass eine Begrenzung der Varianz $\mathbf{\Sigma}_{\boldsymbol{x}}$ bzw. $\mathbf{\Sigma}_{e_t}$ bei der Anwendung von Uncertainty Decoding zu einer Erhöhung der Erkennungsrate führen kann. Dies wird in [LG08] damit begründet, dass die Likelihoods $p(\mathbf{x}_t|q_t)$ verschiedener HMM-Zustände für einen großen Wert von $\mathbf{\Sigma}_{e_t}$, d.h. für $\mathbf{\Sigma}^{(x)}_{t|1:T} \approx \mathbf{\Sigma}_{\boldsymbol{x}}$, ähnliche Werte aufweisen und es daher zu einer großen Anzahl an Einfügungsfehlern kommen kann. In den experimentellen Untersuchungen in dieser Arbeit wird aus diesem Grund die Schwellwertbegrenzung

$$\mathbf{\Sigma}^{(x)}_{t|1:T} = (1 - T_\alpha)\mathbf{\Sigma}_{\boldsymbol{x}}, \quad \text{für} \quad \mathbf{\Sigma}^{(x)}_{t|1:T} > (1 - T_\alpha)\mathbf{\Sigma}_x, \tag{6.13}$$

mit der Konstanten T_α durchgeführt.

6.2 Akustischer Skalierungsfaktor

Wie eingangs erwähnt, können die Ergebnisse der Spracherkennung durch die richtige Gewichtung von Sprachmodell und Akustikmodell verbessert werden. Eine allgemeinere Formulierung des Optimierungsproblems in Gl. (6.1), die verschiedene Skalierungsfaktoren berücksichtigt, wird in [DVCW02] angegeben:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}}\, P(w_1^N)^{S_\alpha S_\beta} p(\mathbf{y}_1^T|w_1^N)^{S_\alpha} S_\gamma^N. \tag{6.14}$$

Die Notwendigkeit der Skalierung mit S_α und S_β kann auf verschiedene Ursachen zurückgeführt werden. Brown [Bro87] führt an, dass das akustische Modell und das Sprachmodell Approximationen der tatsächlichen Modelle sind, deren Parameter auf unterschiedlichen Trainingsdaten bestimmt werden. Rubio et al. [RDVGS97] geben die fehlende Abstimmung im Erkennungsprozess als Ursache an, die dadurch zustande kommt, dass die Sprachmodellwahrscheinlichkeiten nur an den Wortgrenzen integriert werden, während die akustischen Wahrscheinlichkeiten auf Zustandsebene berechnet werden. In [EW00] werden die Skalierungsfaktoren auf die fehlende Modellierung der Inter-Frame Korrelationen zurückgeführt. Bei der Viterbi-Decodierung kann der akustische Skalierungsfaktor S_α vernachlässigt werden, da die Maximierung invariant gegenüber einer konstanten Potenz ist. Die Vernachlässigung von S_α bei der Berechnung von a posteriori Wahrscheinlichkeiten für die HMM-Zustände führt jedoch dazu, dass die wahrscheinlichste Zustandshypothese zu stark gewichtet wird. In [Wes02] wurde experimentell nachgewiesen, dass S_α nahe bei dem inversen Sprachmodellskalierungsfaktor liegt. Der Gewichtsfaktor S_γ bewirkt, dass lange Wörter stärker als kurze Wörter gewichtet werden. In den folgenden Untersuchungen werden die vereinfachenden Annahmen $S_\gamma \approx 1$ und $S_\alpha \approx 1/S_\beta$ getroffen, d.h. die Wahrscheinlichkeitsdichte $p(\mathbf{y}_1^T|w_1^N)$ wird durch eine skalierte Wahrscheinlichkeitsdichte

$$p_\alpha(\mathbf{y}_1^T|w_1^N) \propto p(\mathbf{y}_1^T|w_1^N)^{S_\alpha} \tag{6.15}$$

ersetzt. Der optimale akustische Skalierungsfaktor auf der AURORA2 Datenbank wird in Abschnitt 8.7 experimentell bestimmt.

6.3 Experimentelle Ergebnisse

Abb. 6.2 zeigt exemplarisch den zeitlichen Verlauf der Energiekomponente $x_t^{(0)}$ des Merkmalsvektors für einen Beispielsatz der AURORA2 Datenbank und das in Anhang A.1 beschriebenen Setup für ein SNR von $5dB$. Neben dem Verlauf der verrauschten (blau) und der unverrauschten (grün) Merkmalsvektorkomponente sind die MMSE-Schätzwerte der Verteilungen $p(x_t^{(0)}|\mathbf{y}_1^t)$ (rot) sowie deren Standardabweichungen (senkrechte Linien), die sich bei der Entrauschung der Sprachmerkmale mit dem SLDM-M16 ergeben, als Funktion der Zeit t dargestellt.

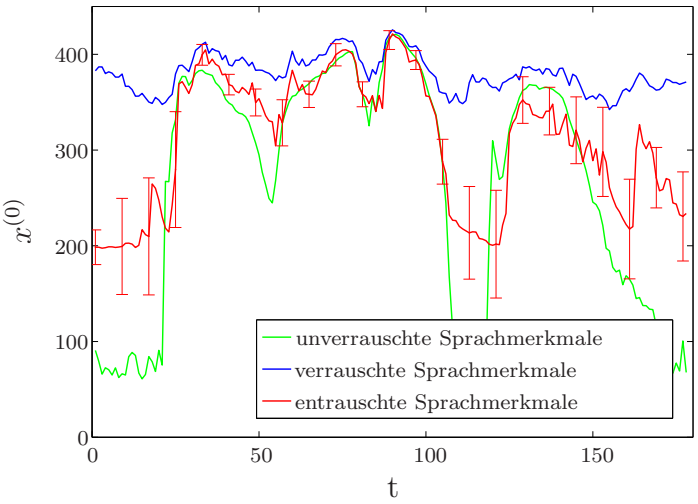

Abbildung 6.2: *Zeitlicher Verlauf der Merkmalsvektorkomponente $x_t^{(0)}$ für einen Beispielsatz der AURORA2 Datenbank*

Aus der Abbildung wird ersichtlich, dass die Zuverlässigkeit der MMSE-Schätzung in verschiedenen Signalbereichen unterschiedlich groß ist, was sich in der Standardabweichung der MMSE-Schätzung ausdrückt. Das Ziel des Uncertainty Decodings besteht darin, die Unsicherheit der Schätzung zu berücksichtigen, indem über $p(\mathbf{x}_t|\mathbf{y}_1^T)$ integriert wird statt Punktschätzungen zu verwenden. Im Folgenden werden verschiedene Uncertainty-Decodierregeln untersucht, die sich dadurch ergeben, dass die Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^T)$ in Gl. (6.8) durch $p(\mathbf{x}_t|\mathbf{y}_t)$ und $p(\mathbf{x}_t|\mathbf{y}_1^t)$ approximiert wird. Bei der Auswertung der Decodierregeln wird in Gl. (6.13) jeweils der Schwellwert $T_\alpha = 0,95$ verwendet.

Die Decodierregel aus [KF02] ergibt sich für den Fall, dass keine Inter-Frame Korrelationen bei der Entrauschung der Sprachmerkmale ausgenutzt werden, d.h. die Verteilung $p(\mathbf{x}_t|\mathbf{y}_t)$ berechnet wird. Wie in Tab. 6.1 a) dargestellt, konnte die Erkennungsrate mit dem GMM-M16 (vgl. Tab. 4.5) durch die Anwendung der Uncertainty-Decodierregel aus [KF02] auf Test-Set A der AURORA2 Datenbank von $76,40\%$ auf $78,29\%$ erhöht werden.

In Tab. 6.1 b) sind die Ergebnisse für das SLDM-M16, das sich wie in Abschnitt 4.5 beschrieben, durch die Integration von Inter-Frame Korrelationen in das GMM-M16 ergibt, dargestellt. Mit der Decodierregel, die in [IHU08b] für die Kompensation von

a)	Sub.	Bab.	Car	Exh.	Ø
clean	99,51	99,27	99,46	99,44	99,42
20dB	97,45	98,31	99,02	98,55	98,33
15dB	94,87	95,74	98,15	95,62	96,10
10dB	86,80	86,70	93,20	85,87	88,14
5dB	71,42	64,90	72,62	65,13	68,52
0dB	43,14	34,16	45,54	38,57	40,35
-5dB	20,08	12,73	16,94	18,36	17,03
Ø	78,74	75,96	81,71	76,75	78,29

b)	Sub.	Bab.	Car	Exh.	Ø
clean	99,63	99,55	99,58	99,66	99,61
20dB	97,61	95,56	99,22	97,59	97,50
15dB	95,92	90,33	98,45	95,59	95,07
10dB	90,36	78,02	95,38	90,00	88,44
5dB	77,89	60,37	87,35	77,91	75,88
0dB	51,95	31,20	61,41	54,37	49,73
-5dB	22,87	9,64	16,34	25,83	18,67
Ø	82,75	71,10	88,36	83,09	81,32

Tabelle 6.1: a) GMM-M16 b) SLDM-M16 - Ergebnisse auf Test-Set A der AURORA2 Datenbank mit der Uncertainty-Decodierregel aus [KF02] bzw. [IHU08b]

Set A	Sub.	Bab.	Car	Exh.	Ø
clean	99,63	99,54	99,55	99,65	99,59
20dB	98,10	95,65	99,43	98,15	97,83
15dB	96,22	90,69	98,84	96,17	95,48
10dB	92,17	79,53	96,69	92,19	90,15
5dB	82,47	64,36	89,89	80,35	79,27
0dB	57,81	35,25	64,27	58,41	53,94
-5dB	24,01	9,49	15,21	26,20	18,73
Ø	85,35	67,79	80,55	78,73	83,33

Tabelle 6.2: SLDM-S16 - Ergebnisse auf Test-Set A der AURORA2 Datenbank mit der Uncertainty-Decodierregel aus [IHU08b]

Paketverlusten eingeführt wurde, wurde die Erkennungsrate auf Test-Set A von $79,87\%$ auf $81,32\%$ erhöht (vgl. Tab. 4.4).

Tab. 6.2 zeigt die Erkennungsergebnisse für die Glättung der Sprachmerkmale (SLDM-S16) und die anschließende Anwendung der Decodierregel aus [IHU08b]. Auf Test-Set A wurde durch die Anwendung von Uncertainty Decoding für das SLDM-S16 eine leichte Verbesserung der Erkennungsrate von $82,82\%$ auf $83,33\%$ erreicht (vgl. Tab. 4.6). Diese liegt bei $N = 65766$ Wörtern des betrachteten Test-Sets mit dem in Anhang B.1 angegebenen Konfidenzmaß an der Grenze des statistisch signifikanten Bereiches ($WER_2 + 1,96\sigma_{WER_2} = 83,10\% > 83,07\% = WER_1 - 1,96\sigma_{WER_1}$).

In Tab. 6.3 sind die Ergebnisse mit Uncertainty Decoding auf der AURORA4 Datenbank angegeben. Auf der AURORA4 Datenbank ergaben sich für das SLDM-M16, das GMM-M16 und das SLDM-S16 Fehlerraten zwischen $33,0\%$ und $33,4\%$, d.h. keine statistisch signifikanten Unterschiede zwischen den einzelnen Verfahren. Die großen Gewinne beim SLDM-M16 durch Uncertainty Decoding können, wie in Abschnitt 4.5 ausgeführt wird, dadurch erklärt werden, dass das SLDM tendentiell in Signalbereichen, in denen die Zustandsschätzung eine große Varianz aufweist, deren Einfluß auf

den akustischen Score durch das Uncertainty Decoding somit reduziert wird, schlechter als das GMM zur Beschreibung der Sprachdynamik geeignet ist, während es in Bereichen, in denen eine zuverlässige Zustandsschätzung durchgeführt werden kann, tendentiell eine genauere Prädiktion des Signalverlaufs ermöglicht. Neben den genannten Verfahren wurden auch das in Kapitel 4 eingeführte IEKF-d-M16 und die in Kapitel 5 eingeführte adaptive Rauschschätzung mit Uncertainty Decoding getestet. Mit diesen Verfahren wurden Fehlerraten von $32,5\%$ und $31,8\%$ erreicht.

		Train	Airp.	Babble	Car	Street	Rest.	Clean	Ø
	S	30,7	29,5	27,5	14,2	27,9	32,7	9,7	24,6
GMM-M16	I	6,8	11,7	10,1	4,8	8,1	10,2	3,2	7,8
	E	42,1	45,5	41,6	20,9	40,6	47,5	14,3	36,1
	S	28,8	27,2	25,2	12,4	26,6	29,5	9,1	22,7
GMM-M16	I	6,2	11,5	9,0	3,3	6,7	10,3	2,3	7,0
+ UD	E	39,1	43,0	37,8	17,8	37,7	44,2	12,9	33,2
	S	29,9	37,9	32,2	12,7	29,9	37,2	9,5	27,0
SLDM-M16	I	8,8	15,9	12,7	4,6	10,6	13,3	3,1	9,9
	E	42,6	58,2	48,8	19,5	45,0	56,0	13,7	40,5
	S	25,9	25,5	25,6	12,1	26,2	29,7	10,0	22,1
SLDM-M16	I	3,9	8,2	6,8	3,8	6,7	8,7	2,1	5,7
+ UD	E	37,8	39,3	37,7	17,3	38,9	45,6	14,5	33,0
	S	28,1	32,9	28,6	12,3	28,6	34,3	8,8	24,8
SLDM-S16	I	8,6	14,8	11,9	4,8	10,4	13,2	2,4	9,4
	E	41,8	52,7	44,5	19,2	44,3	53,2	12,6	38,3
	S	25,8	27,3	24,2	11,8	26,6	29,5	9,2	22,1
SLDM-S16	I	5,4	13,0	9,4	3,6	6,7	11,5	2,0	7,4
+ UD	E	36,6	45,1	37,9	17,4	37,7	46,2	12,9	33,4
	S	29,3	31,5	28,1	12,3	28,2	33,6	9,1	24,6
IEKF-d-M16	I	7,4	13,4	11,2	4,2	8,4	12,0	2,7	8,5
	E	41,0	49,4	42,7	18,6	41,2	50,4	13,1	36,6
	S	25,9	27,2	24,3	11,9	25,4	29,8	8,8	21,9
IEKF-d-M16	I	5,3	10,1	8,3	3,6	6,5	10,6	2,3	6,7
+ UD	E	36,5	42,0	36,5	18,0	37,1	45,3	12,3	32,5
	S	26,9	25,2	24,9	11,6	26,2	31,3	8,8	22,1
EKF-d-M16-dyn	I	3,6	6,6	3,8	2,9	4,1	6,7	2,3	4,3
+ UD	E	37,2	38,8	33,9	17,0	37,7	45,6	12,3	31,8

Tabelle 6.3: Fehlerraten auf der AURORA4 Datenbank (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

Kapitel 7

Modellierung statistischer Abhängigkeiten im Back-End des Spracherkenners

Bislang wurde ein Markovmodell zur akustischen Modellierung im Back-End des Spracherkenners verwendet. Einen wesentlichen Schwachpunkt dieses Ansatzes stellt, wie bereits in Abschnitt 2.2 ausgeführt wurde, die Annahme der bedingten statistischen Unabhängigkeit zwischen aufeinander folgenden Sprachrahmen dar. Auf der anderen Seite sind segmentielle HMMs, die die Abschwächung dieser Annahme ermöglichen, in der Regel mit einer großen Anzahl an Parametern, einem unzuverlässigen Maximum-Likelihood(ML)-Training der Parameter sowie einer aufwendigen Suche verbunden (vgl. Abschnitt 2.2.1). Ein möglicher Ansatz für effiziente Algorithmen ergibt sich jedoch durch die Quantisierung des Merkmalsraumes [SZD03]. In [SZD03] wird, wie in Abschnitt 2.2.1 dargestellt, ein Hidden Trajectory HMM (HTHMM) für die akustische Modellierung verwendet, in dem die Sprachmerkmale neben der Emissionsverteilung des HMMs mit einer versteckten, quantisierten Trajektorie beschrieben werden. Im Folgenden soll die Quantisierung des Merkmalsraumes im Standard-Training des HMMs ausgenutzt werden, um eine genauere Näherung der Verteilung $p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t)$ in Gl. (6.6) zu ermöglichen, die im HMM durch die Emissionsverteilung $p(\mathbf{x}_t|q_t)$ approximiert wird. Den Ausgangspunkt dazu stellt die verbreitete Modellierung der Emissionsverteilung als Gaußmischungsverteilung

$$p(\mathbf{x}_t|q_t) = \sum_{m_t} p(\mathbf{x}_t|m_t)P(m_t|q_t), \tag{7.1}$$

mit den Mischungsgewichten $P(m_t|q_t)$ und den Einzelverteilungen $p(\mathbf{x}_t|m_t)$ dar (Abb. 7.1 a)). Es wird angenommen, dass die Mischungskomponenten m_t in disjunkte Mengen $\mathcal{M}(q_t)$, $q_t = 1 \ldots N_q$, unterteilt werden können, wobei N_q die Anzahl der HMM-Zustände bezeichnet:

$$\mathcal{M}(q_t) = \{m_t|P(m_t|q_t) \neq 0\}. \tag{7.2}$$

Das bedeutet, dass der Wert des Mischungsindexes m_t eindeutig den Wert der verborgenen Zustandsvariable q_t spezifiziert:

$$P(q_t|m_t) = \delta(q_t - \hat{q}_t(m_t)). \tag{7.3}$$

In Gl. (7.3) bezeichnet $\hat{q}_t(m_t)$ den HMM-Zustand, der m_t zugeordnet ist.

Das statistische Modell, das im Folgenden eingeführt und als CMHMM (Continuous Mixture HMM) bezeichnet wird [WHUL08], ist in Abb. 7.1 b) dargestellt. Im CMHMM

wird zusätzlich zu der statistischen Abhängigkeit zwischen den HMM-Zuständen q_{t-1} und q_t in aufeinander folgenden Sprachrahmen die Abhängigkeit $P(m_t|m_{t-1})$ zwischen den Mischungskomponenten m_{t-1} und m_t modelliert und auf diese Weise die statistische Abhängigkeit zwischen $\mathbf{x}_{t-1}$ und $\mathbf{x}_t$ approximiert (Abb. 7.1-c).

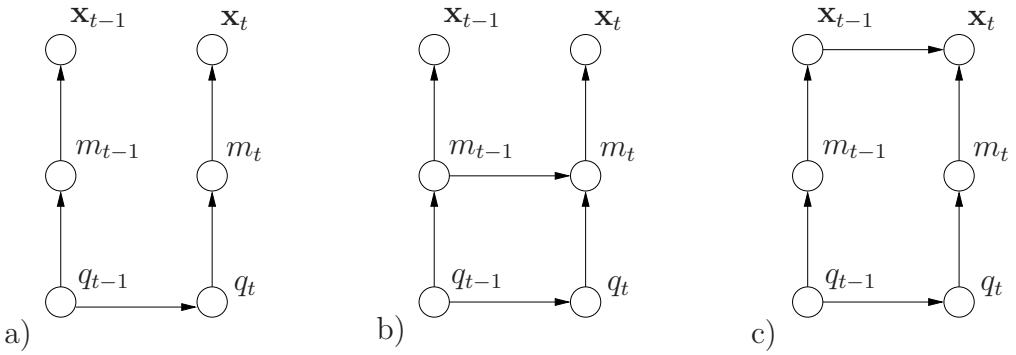

Abbildung 7.1: *Statistisches Modell: a) HMM b) CMHMM c) Direkte Modellierung der Inter-Frame Korrelationen*

Daneben werden Einschränkungen für die statistische Abhängigkeit zwischen m_t und q_t formuliert. Es wird angenommen, dass das Tupel (q_t, m_t) bei gegebenen weiteren Zufallsvariablen $\mathbf{X}$ genauso wahrscheinlich wie m_t ist, sobald die Bedingung in (7.3) erfüllt ist. Dies wird über die Bedingung

$$P(q_t, m_t|\mathbf{X}) = \begin{cases} P(m_t|\mathbf{X}) & : \; q_t = \hat{q}_t(m_t) \\ 0 & : \; \text{sonst} \end{cases} \tag{7.4}$$

ausgedrückt. Zusätzlich wird die Annahme getroffen, dass der Einfluß von q_t in bedingten Wahrscheinlichkeiten der Form $P(m_t|\mathbf{X}, q_t)$ gegenüber weiteren Zufallsvariablen $\mathbf{X}$ vernachlässigt werden kann, sobald m_t von $\mathbf{X}$ abhängt:

$$P(m_t|\mathbf{X}, q_t) = \begin{cases} P(m_t|\mathbf{X}) & : \; q_t = \hat{q}_t(m_t) \\ 0 & : \; \text{sonst.} \end{cases} \tag{7.5}$$

Die a priori Wahrscheinlichkeit $P(m_t|q_t)$ ergibt sich aus dem HMM-Training.

Das statistische Modell in Abb. 6.1 wurde durch die Berücksichtigung des Zusammenhanges zwischen $\mathbf{x}_t$ und den Sprachmerkmalen der vorangehenden Sprachrahmen $\mathbf{x}_1^{t-1}$ so allgemein gewählt, dass die in Gl. (6.6) hergeleitete Beziehung für $p(\mathbf{y}_1^T|q_1^T)$ auch für das CMHMM gültig ist. Um die Vertauschung der Integration und der Summation in Gl. (6.6) zu ermöglichen, wird die Approximation

$$p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t) \approx p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t) \tag{7.6}$$

mit $\hat{\mathbf{x}}_{t-1} = E[\mathbf{x}_{t-1}|\mathbf{y}_1^{t-1}]$ vorgenommen. Durch Einsetzen von Gl. (7.6) in Gl. (6.6) erhält man:

$$p(\mathbf{y}_1^T|q_1^T) \approx \int_{\mathbb{R}^{N_c T}} \prod_{t=1}^{T} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t) d\mathbf{x}_1^T = \prod_{t=1}^{T} \int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t) d\mathbf{x}_t. \tag{7.7}$$

Bei der Viterbi-Decodierung, die Abschnitt 2.1.3 beschrieben wird, ergibt sich somit die modifizierte Likelihood

$$p_{LH}(\mathbf{y}_1^T|q_t) \approx \int_{\mathbb{R}^{N_c}} \frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)} p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t) d\mathbf{x}_t, \tag{7.8}$$

die formal der Likelihood in Gl. (6.8) entspricht, so dass die Berechnung des Integrals wie in Abschnitt 6.1.1 beschrieben durchgeführt werden kann.

7.1 Suche im CMHMM

Das CMHMM erfordert eine geeignete Suchstrategie, da die Anzahl der Mischungskomponenten in den akustischen Modellen in der Regel um ein Vielfaches größer als die Anzahl der HMM-Zustände ist, so dass die Anwendung des Viterbi-Algorithmus auf der Ebene der Mischungsgewichte nicht effizient ist. Ein signifikanter Anstieg der Rechenzeit kann dadurch vermieden werden, dass die optimale HMM-Zustandssequenz mit einem Viterbi-Algorithmus auf der Ebene der HMM-Zustände berechnet wird, während die Mischungsgewichte mit einem kausalen Filter im Vorwärtsschritt des Viterbi-Algorithmus unter Berücksichtigung der beobachteten Sprachmerkmale sowie der Übergangswahrscheinlichkeiten zwischen den Mischungsgewichten aktualisiert werden. Vor der Darstellung des erweiterten Viterbi-Algorithmus werden zunächst die benötigten Beziehungen hergeleitet.

Die Mischungsgewichte m_t können in Gl. (7.8) unter der Modellannahme in Abb. 7.1-b durch die Marginalisierung

$$p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t) = \sum_{m_t} p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1}, q_t, m_t) P(m_t|\mathbf{x}_1^{t-1}, q_t) \overset{(7.5)}{=} \sum_{m_t \in \mathcal{M}(q_t)} p(\mathbf{x}_t|m_t) P(m_t|\hat{\mathbf{x}}_1^{t-1}) \tag{7.9}$$

berücksichtigt werden. Die a posteriori Wahrscheinlichkeit $P(m_t|\hat{\mathbf{x}}_1^{t-1})$ der Mischungsgewichte kann als

$$P(m_t|\hat{\mathbf{x}}_1^{t-1}) = \sum_{q_{t-1}} P(q_{t-1}, m_t|\hat{\mathbf{x}}_1^{t-1}) \tag{7.10}$$

mit der Zustandsübergangsgleichung

$$\begin{aligned} P(q_{t-1}, m_t|\hat{\mathbf{x}}_1^{t-1}) &= \sum_{m_{t-1}} P(m_t|q_{t-1}, m_{t-1}, \hat{\mathbf{x}}_1^{t-1}) P(q_{t-1}, m_{t-1}|\hat{\mathbf{x}}_1^{t-1}) \\ &\overset{(7.4)}{=} \sum_{m_{t-1} \in \mathcal{M}(q_{t-1})} P(m_t|m_{t-1}) P(m_{t-1}|\hat{\mathbf{x}}_1^{t-1}) \end{aligned} \tag{7.11}$$

berechnet werden. In Gl. (7.11) wird ausgenutzt, dass m_t bei gegebenem m_{t-1} unabhängig von $\hat{\mathbf{x}}_1^{t-1}$ ist. Einsetzen von (7.9) und (7.10) in die Viterbi-Approximation (2.11) mit der modifizierten Likelihood (7.8) führt auf

$$B_t(v, q_t) = B_{t-1}(v, \hat{q}_{t-1}(q_t)) \tag{7.12}$$

mit

$$\begin{aligned}
&\hat{q}_{t-1}(q_t) \\
&\overset{(7.8)}{=} \underset{q_{t-1}}{\operatorname{argmax}}\{\alpha_{t-1}(v,q_{t-1})P(q_t|q_{t-1})\int_{\mathbb{R}^{N_c}}\frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)}p(\mathbf{x}_t|\hat{\mathbf{x}}_1^{t-1},q_t)d\mathbf{x}_t\} \\
&\overset{(7.9)}{=} \underset{q_{t-1}}{\operatorname{argmax}}\{\alpha_{t-1}(v,q_{t-1})P(q_t|q_{t-1})\int_{\mathbb{R}^{N_c}}\frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)}\sum_{m_t\in\mathcal{M}(q_t)}p(\mathbf{x}_t|m_t)P(m_t|\hat{\mathbf{x}}_1^{t-1})d\mathbf{x}_t\} \\
&\overset{(7.10)}{=} \underset{q_{t-1}}{\operatorname{argmax}}\{\alpha_{t-1}(v,q_{t-1})P(q_t|q_{t-1})\sum_{m_t\in\mathcal{M}(q_t)}p(\mathbf{y}_1^T|m_t)\sum_{\tilde{q}_{t-1}}P(\tilde{q}_{t-1},m_t|\hat{\mathbf{x}}_1^{t-1})\}
\end{aligned} \tag{7.13}$$

und

$$p(\mathbf{y}_1^T|m_t) = \int_{\mathbb{R}^{N_c}}\frac{p(\mathbf{x}_t|\mathbf{y}_1^T)}{p(\mathbf{x}_t)}p(\mathbf{x}_t|m_t)d\mathbf{x}_t. \tag{7.14}$$

Mit Hinblick auf die Viterbi-Approximation (2.11) ist es konsequent, die Summation über die Vorgängerzustände $\tilde{q}_{t-1}$ in Gl. (7.13) durch den Beitrag des Summanden für den "besten" Vorgängerzustand zu ersetzen:

$$\hat{q}_{t-1}(q_t) = \underset{q_{t-1}}{\operatorname{argmax}}\{\alpha_{t-1}(v,q_{t-1})P(q_t|q_{t-1})\sum_{m_t\in\mathcal{M}(q_t)}p(\mathbf{y}_1^T|m_t)P(q_{t-1},m_t|\hat{\mathbf{x}}_1^{t-1})\}. \tag{7.15}$$

Um $\alpha_t(v,q_t)$ zu erhalten, muß die argmax()-Operation in Gl. (7.13) und Gl. (7.15) durch eine Maximierung ersetzt werden. Mit dem bekannten Argument $\hat{q}_{t-1}(q_t)$, das Gl. (7.15) maximiert, ergibt sich

$$\alpha_t(v,q_t) \overset{(7.13)}{=} \alpha_{t-1}(v,\hat{q}_{t-1}(q_t))P(q_t|\hat{q}_{t-1}(q_t))\sum_{m_t\in\mathcal{M}(q_t)}p(\mathbf{y}_1^T|m_t)P(m_t|\hat{\mathbf{x}}_1^{t-1}) \tag{7.16}$$

$$\overset{(7.15)}{\approx} \alpha_{t-1}(v,\hat{q}_{t-1}(q_t))P(q_t|\hat{q}_{t-1}(q_t))\sum_{m_t\in\mathcal{M}(q_t)}p(\mathbf{y}_1^T|m_t)P(\hat{q}_{t-1}(q_t),m_t|\hat{\mathbf{x}}_1^{t-1}). \tag{7.17}$$

Der Vergleich von Gl. (7.16) und Gl. (7.17) zeigt, dass die Approximation in Gl. (7.17) die Approximation

$$P(m_t|\hat{\mathbf{x}}_1^{t-1}) \approx P(\hat{q}_{t-1}(q_t),m_t|\hat{\mathbf{x}}_1^{t-1}). \tag{7.18}$$

impliziert. Der zweite und letzte Schritt bei der Aktualisierung der Mischungsgewichte ist die Beobachtungsgleichung

$$P(m_t|\hat{\mathbf{x}}_1^t) \propto P(m_t|\hat{\mathbf{x}}_1^{t-1})p(\hat{\mathbf{x}}_t|m_t). \tag{7.19}$$

Da die modifizierte Suchstrategie auf der Ebene der Mischungsgewichte die Berechnung von Summen erfordert, ist es anders als bei der Viterbi-Suche (vgl. Abschnitt 2.1.3), wo eine Maximum-Approximation durchgeführt wird, erforderlich, die skalierten Wahrscheinlichkeiten entsprechend Gl. (6.15) zu verwenden:

$$P(m_t|\hat{\mathbf{x}}_1^t) \propto P(m_t|\hat{\mathbf{x}}_1^{t-1})p(\hat{\mathbf{x}}_t|m_t)^{S_\alpha}. \tag{7.20}$$

Die Berücksichtigung der Zustandsgleichung (7.11) und der Messgleichung (7.20) im Viterbi-Algorithmus, der in Abschnitt 2.1.3 beschrieben wird, führt zu der folgenden modifizierten Viterbi-Suche. Zum Zeitpunkt $t = 0$ werden die Mischungsgewichte $m_0 \in \mathcal{M}(q_0)$ für alle HMM-Zustände q_0 initialisiert:

$$P(m_0|\hat{\mathbf{x}}_1^0) = P(m_0|\hat{q}_0(m_0)). \tag{7.21}$$

Die Vorwärtsiteration innerhalb von Wörtern in der Standard-Viterbi-Suche (2.11) wird durch die folgenden Schritte ersetzt:

Algorithmus 2 Vorwärtsiteration im CMHMM

Für alle q_t
  Für alle q_{t-1}
    Für alle $m_t \in \mathcal{M}(q_t)$
      Berechne $P(q_{t-1}, m_t|\hat{\mathbf{x}}_1^{t-1})$ mit Gl. (7.11).
    Ende.
  Ende.
  Berechne $\hat{q}_{t-1}(q_t)$ mit Gl. (7.15).
  Berechne $\alpha_t(v, q_t)$ mit Gl. (7.17).
  Für alle $m_t \in \mathcal{M}(q_t)$
    Aktualisiere $P(m_t|\hat{\mathbf{x}}_1^{t-1})$ mit Gl. (7.18).
    Aktualisiere die Mischungsgewichte mit Gl. (7.20).
  Ende.
Ende.

7.2 Speichereffiziente Durchführung der Zustandsübergänge

Das CMHMM erfordert die Berechnung der Übergangswahrscheinlichkeiten $P(m_t|m_{t-1})$ anstelle der Mischungsgewichte $P(m_t|q_t)$. Dabei ist eine Speicherreduktion möglich, indem $P(m_t|m_{t-1})$ nur für $\hat{q}_t(m_t) = \hat{q}_{t-1}(m_{t-1})$ bestimmt wird, während die Gewichte an den Grenzen der HMM-Zustände zurückgesetzt werden, d.h. für $\hat{q}_{t-1}(m_{t-1}) \neq \hat{q}_t(m_t)$ die Mischungsgewichte $P(m_t|q_t)$ verwendet werden. Auch mit diesem Ansatz liegt der Speicheraufwand für die Übergangswahrscheinlichkeiten eines HMM-Zustandes mit N_m Mischungskomponenten mit $O(N_m \times N_m)$ in der Größenordnung des Speichers $O((2N_c + 1) \times N_m)$ für die übrigen Modellparameter. Eine Möglichkeit, den Speicheraufwand zu reduzieren ohne die Mischungsgewichte an den HMM-Grenzen zurückzusetzen, besteht in der Verwendung der in Kapitel 4 für die Merkmalsentstörung eingeführten linearen Zustandsmodelle, die z.B. in stochastischen Segmentmodellen (z.B. [Dig92, MD04, FK07]) und SLDMs [RH97] auch für die Sprachdecodierung eingesetzt werden. Dazu werden die Mischungsgewichte in den Merkmalsraum transformiert und der resultierende Merkmalsvektor zunächst mit einem Zustandsmodell prädiziert und anschließend auf die Mischungsgewichte des nächsten Sprachrahmens projiziert.

Es ergeben sich anstelle der Zustandsgleichung (7.11) somit die folgenden Schritte, um die Mischungsgewichte m_t zu aktualisieren:

$$E[\mathbf{x}_{t-1}, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}] = \sum_{m_{t-1}\in\mathcal{M}(q_{t-1})} P(m_{t-1}|\hat{\mathbf{x}}_1^{t-1}) E[\mathbf{x}_{t-1}|m_{t-1}] \tag{7.22}$$

$$E[\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}] = \sum_{s_t} P(s_t|q_t)(\mathbf{A}(s_t)E[\mathbf{x}_{t-1}, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}] + \mathbf{b}(s_t)) \tag{7.23}$$

$$\begin{aligned} P(q_{t-1}, m_t|\hat{\mathbf{x}}_1^{t-1}) &= \int_{\mathbb{R}^{N_c}} p(\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}) P(m_t|\mathbf{x}_t, \hat{\mathbf{x}}_1^{t-1}, q_{t-1}) d\mathbf{x}_t \\ &= P(m_t|E[\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}]) \\ &\propto P(m_t) p(E[\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}]|m_t), \end{aligned} \tag{7.24}$$

wobei $p(E[\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1}]|m_t)$ in Gl. (7.24) genauso wie die Likelihood in Gl. (7.20) mit dem akustischen Skalierungsfaktor S_α potenziert wird. In Gl. (7.22)-(7.24) wurden die Verteilungen $p(\mathbf{x}_{t-1}, q_{t-1}|\hat{\mathbf{x}}_1^{t-1})$ und $p(\mathbf{x}_t, q_{t-1}|\hat{\mathbf{x}}_1^{t-1})$ des Merkmalsvektors jeweils durch einen Dirac-Stoß approximiert. Für die Prädiktion des Merkmalsvektors in Gl. (7.23) wurde ein lineares Zustandsmodell angenommen. Dieses kann über die in Abschnitt 8.3 eingeführte statistische Abhängigkeit $P(s_t|q_t)$ für jeden HMM-Zustand aus den Dynamikmodellen $(\mathbf{A}(s_t), \mathbf{b}(s_t), \mathbf{C}(s_t))$, die in Abschnitt 4.1 eingeführt wurden, bestimmt werden. Die Projektion des Merkmalsvektors auf die Gewichte in Gl. (7.24) ergibt sich aus der Bayes'schen Formel.

7.3 Experimentelle Ergebnisse

Das CMHMM wurde auf der AURORA2 Datenbank mit dem in Anhang A.1 spezifizierten Setup getestet. Zur Aktualisierung der Mischungsgewichte wurden $M = 16$ Zustandsmodelle mit dynamischen Merkmalen erster und zweiter Ordnung, d.h. insgesamt 39 Komponenten, trainiert und den HMM-Zuständen mit der in Kapitel 8 eingeführten Zustandstabelle zugeordnet. Die Extraktion der Sprachmerkmale im Front-End wurde mit dem in Kapitel 4 eingeführten SLDM durchgeführt.

Zunächst wurde das CMHMM ohne Uncertainty Decoding, d.h. unter der Annahme $p(\mathbf{x}_t|\mathbf{y}_1^T) = \delta(\mathbf{x}_t - E[\mathbf{x}_t|\mathbf{y}_1^T])$ getestet. Die Erkennungsergebnisse sind in Tabelle 7.1 dargestellt. Gegenüber dem HMM-Decoder mit festen Mischungsgewichten ergab sich auf Test-Set A der AURORA2 Datenbank eine Verbesserung der Erkennungsrate von $79,87\%$ auf $81,20\%$ und auf Test-Set B eine Verbesserung von $79,16\%$ auf $79,67\%$. Die Anwendung von Uncertainty Decoding durch die Modellierung von $p(\mathbf{x}_t|\mathbf{y}_1^T)$ als Gaußverteilung (Tab. 7.2) führte auf der AURORA2 Datenbank zu einem leichten Anstieg der Erkennungsrate (Test-Set A: $81,50\%$, Test-Set B: $79,77\%$), der nicht statistisch signifikant ist.

Durch die Verwendung des CMHMMs wurde der Speicheraufwand gegenüber dem Standard-HMM um den Aufwand für die Zustandstabelle erhöht, der bei N_q HMM-Zuständen in der Größenordnung $O(N_q \times M)$ liegt. Die Rechenzeit wurde mit der vorgeschlagenen Suchstrategie nicht signifikant erhöht.

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,57	99,52	99,60	99,60	99,57
20dB	97,73	96,95	99,02	97,69	97,85
15dB	96,32	92,14	98,15	96,02	95,66
10dB	89,99	82,32	95,02	90,00	89,33
5dB	77,56	65,21	84,73	76,06	75,89
0dB	49,95	34,43	54,58	51,00	47,49
-5dB	20,20	12,27	13,96	24,22	17,66
Ø	82,31	74,21	86,30	82,15	81,24

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,54	99,52	99,52	99,57	99,54
20dB	96,50	98,16	97,02	98,34	97,51
15dB	90,88	95,68	94,48	96,08	94,28
10dB	81,46	88,06	85,86	92,22	86,90
5dB	65,00	73,55	73,25	80,22	73,01
0dB	40,90	44,47	46,79	54,49	46,66
-5dB	15,97	14,27	18,64	16,78	16,42
Ø	74,95	79,98	79,48	84,27	79,67

Tabelle 7.1: SLDM-M16 mit Gewichtsupdate im Erkenner: Testergebnisse auf Test-Set A und B der AURORA2 Datenbank

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,60	99,58	99,58	99,66	99,61
20dB	97,73	95,86	99,22	97,69	97,63
15dB	95,95	90,72	98,14	95,65	95,12
10dB	90,33	78,54	95,44	90,25	88,64
5dB	78,45	61,61	87,59	77,97	76,41
0dB	51,73	31,59	60,93	54,61	49,72
-5dB	23,12	10,46	16,02	27,15	19,19
Ø	82,84	71,66	88,26	83,23	81,50

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,60	99,61	99,52	99,66	99,60
20dB	95,79	98,00	96,93	98,09	97,20
15dB	89,75	95,65	94,24	96,39	94,01
10dB	80,07	87,94	85,03	92,66	86,43
5dB	64,69	74,61	73,16	81,67	73,53
0dB	40,41	45,89	48,11	56,34	47,69
-5dB	15,32	14,87	18,52	18,45	16,79
Ø	74,14	80,42	79,49	85,03	79,77

Tabelle 7.2: SLDM-M16 mit UD + Gewichtsupdate im Erkenner: Testergebnisse auf Test-Set A und B der AURORA2 Datenbank

Kapitel 8

Rückkopplung der Erkennungsergebnisse in das Front-End

In den vorangehenden Kapiteln wurde die Spracherkennung auf der Grundlage eines gegenüber dem HMM verallgemeinerten statistischen Modells, das in Abb. 6.1 dargestellt ist, untersucht. Unter der allgemeineren Modellannahme ergab sich eine Decodierregel für die verrauschten Sprachmerkmale, die u.a. die a posteriori Verteilung der Sprachmerkmale, deren Berechnung in Kapitel 4 beschrieben wird, berücksichtigt. Bei der Berechnung dieser Verteilung wurden in Kapitel 4 wiederum vereinfachte Modellannahmen getroffen. Im SLDM werden zwar wie in dem statistischen Modell, das in Abb. 6.1 dargestellt ist, statistische Abhängigkeiten zwischen den Sprachmerkmalen verschiedener Sprachrahmen ausgenutzt, nicht jedoch die Abhängigkeiten zwischen den HMM-Zuständen und den Sprachmerkmalen. Die Informationen in den komplexen akustischen Modellen und den Sprachmodellen des Erkenners, der bei großem Vokabular einige tausend HMM-Zustände enthalten kann, werden somit erst bei der Erkennung im Back-End berücksichtigt, während die a posteriori Wahrscheinlichkeit der Sprachmerkmale auf der Grundlage statistischer Modelle mit vergleichsweise wenigen Parametern berechnet wird.

Ausgehend von dem statistischen Modell, das in Abb. 6.1 dargestellt ist, sollen in diesem Kapitel daher Ansätze hergeleitet werden, in denen die a posteriori Wahrscheinlichkeiten der HMM-Zustände bereits bei der Merkmalsentstörung berücksichtigt werden. Dazu wird eine zweistufige Erkennung durchgeführt, in der Informationen, die in der ersten Erkennungsstufe im Back-End berechnet werden, in der zweiten Stufe in das Front-End zurückgekoppelt werden. In dem folgenden Abschnitt wird die Kombination des SLDMs mit dem HMM zunächst anhand eines qualitativen Vergleiches der beiden Modellierungsansätze motiviert. Anschließend werden auf der Grundlage eines statistischen Ansatzes Methoden zur Rückkopplung der Informationen hergeleitet.

8.1 Vergleich der Modelle im Front-End und Back-End des Erkenners

Abb. 8.1 zeigt jeweils für ein SLDM ($x_t^{(0,SLDM)}$) und ein HMM ($x_t^{(0,HMM)}$) mögliche Trajektorien einer geschätzten Merkmalsvektorkomponente. Dabei wird zur Vereinfa-

chung angenommen, dass keine Beobachtungen $\mathbf{y}_t$ vorliegen, mit denen der Verlauf der Trajektorien korrigiert wird.

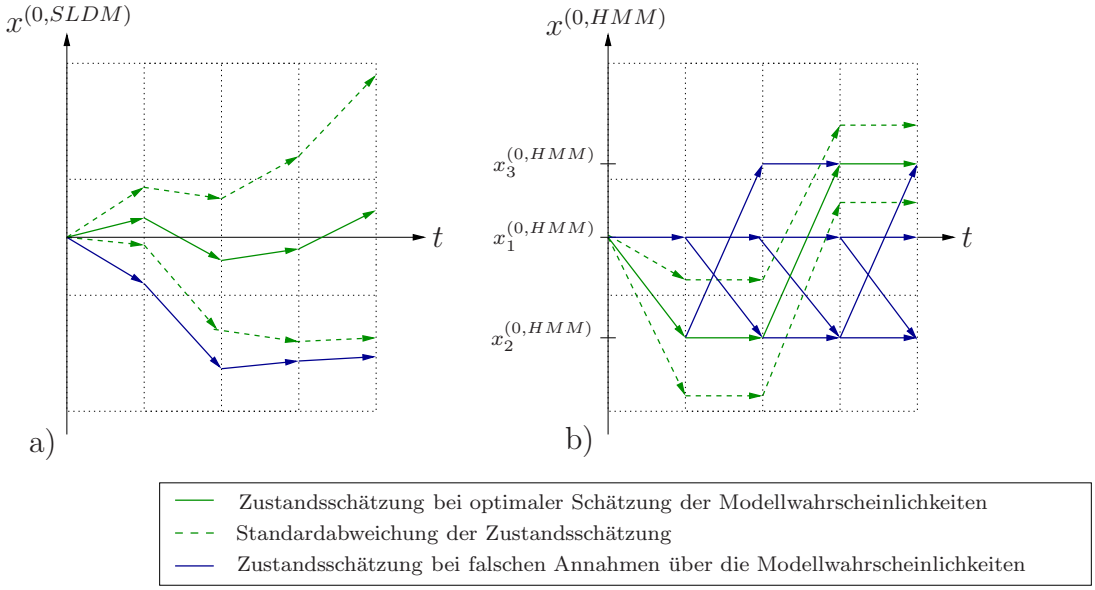

Abbildung 8.1: *Mögliche Trajektorien einer geschätzten Merkmalsvektorkomponente: a) SLDM b) HMM*

Die grün dargestellte, durchgezogene Linie in Abb. 8.1 a) soll eine typische Trajektorie der Zustandsschätzung zeigen, für die die Modelle zu jedem Zeitpunkt so gewichtet werden, dass die Sprachdynamik am besten beschrieben wird. Ohne eine Korrektur der Zustandsschätzung aufgrund neuer Beobachtungen $\mathbf{y}_t$ nimmt die Varianz der Zustandsschätzung mit der Zeit zu, da die Varianz der modellabhängigen Zustandsschätzungen für alle Zustandsmodelle s_t anwächst, so dass auch die Varianz der gewichteten Summe entsprechend Gl. (4.17) größer wird. Die Trajektorien, die sich durch die Addition bzw. Subtraktion der Standardabweichung der Zustandsschätzung zu dieser Schätzung ergeben, sind in Abb. 8.1 als gestrichelte Linien dargestellt. Eine weitere Unsicherheit ergibt sich aufgrund der Gewichtung der schaltenden Modelle. Die blaue Kurve ist ein Beispiel für eine Kurve bei falschen Annahmen über die Modellwahrscheinlichkeiten.

Abb. 8.1 b) zeigt schematisch alle möglichen Trajektorien des Zustandsvektors

$$x_t^{(0,HMM)} = x_{q_t}^{(0,HMM)} = E[x_t^{(0)}|q_t], \quad q_t = 1, \ldots, 3, \tag{8.1}$$

für ein HMM mit drei Zuständen und einer Links-Rechts-Topologie (blaue Kurven), die unter der Annahme fehlender Beobachtungen die Schätzung der Merkmalsvektorkomponente darstellen. Im Gegensatz zum SLDM nimmt die Varianz $\mathrm{Var}(x_t^{(0)}|q_t)$ der Schätzung einer gegebenen Zustandsfolge nicht mit der Zeit t zu und hängt nur vom aktuellen Zustand q_t ab. In Abb. 8.1 b) wird dies durch die durchgezogenen und gestrichelten grünen Linien dargestellt, die $x_{q_t}^{(0,HMM)}$ und $x_{q_t}^{(0,HMM)} \pm \sqrt{\mathrm{Var}(x_t^{(0)}|q_t)}$ darstellen sollen.

Weiterhin erfordern beide Modelltypen einen unterschiedlichen Ansatz zur Schätzung der Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ bzw. der Zustandswahrscheinlichkeiten

$P(q_t|\mathbf{y}_1^T)$. Wie in Abschnitt 4.2.0.5 herausgestellt, werden die Modellwahrscheinlichkeiten im SLDM proportional zu $p(\mathbf{y}_t|\mathbf{y}_1^{t-1}, s_t)$ gewählt, wobei zukünftige Beobachtungen zur Vermeidung von Verzögerungen und zusätzlichem Rechenaufwand vernachlässigt werden, was zu einer starken Abhängigkeit von der aktuellen Beobachtung $\mathbf{y}_t$ führt. Im Gegensatz dazu können die beste Zustandsfolge $\hat{q}_1^T$ oder die a posteriori Wahrscheinlichkeiten $P(q_t|\mathbf{y}_1^T)$ der HMM-Zustände effizient mit einem Viterbi-Algorithmus bzw. einem Vorwärts-Rückwärts-Algorithmus berechnet werden (siehe Abschnitt 8.6). Dabei werden Langzeitabhängigkeiten im Sprachsignal ausgenutzt, die z.B. durch die Einschränkung der erlaubten Zustandssequenzen aufgrund der Akustik- und Sprachmodelle berücksichtigt werden.

In dem folgenden Experiment wird der Einfluß des Sprach- und Akustikmodells auf die statistische Abhängigkeit zwischen zwei Zuständen q_0 und q_τ für verschiedene zeitliche Abstände τ zwischen den beiden Zuständen untersucht.

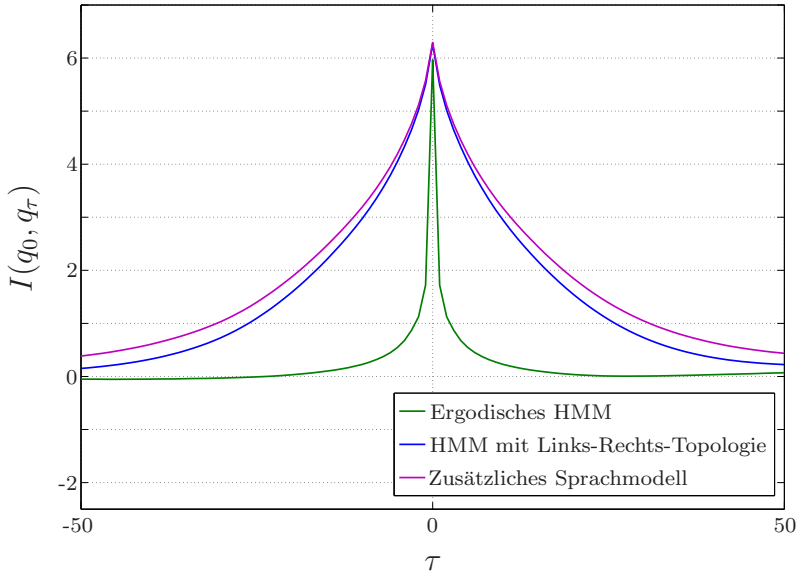

Abbildung 8.2: *Mittlere wechselseitige Information zwischen zwei HMM-Zuständen q_0 und q_τ abhängig vom zeitlichen Abstand τ zwischen den Zuständen*

Als Maß für die Abhängigkeit zwischen den HMM-Zuständen wird die mittlere wechselseitige Information

$$I(q_0, q_\tau) = -\sum_{q_0=1}^{N_q} P(q_0)\log(P(q_0)) + \sum_{q_\tau=1}^{N_q}\sum_{q_0=1}^{N_q} P(q_0, q_\tau)\log(P(q_\tau|q_0)) \tag{8.2}$$

herangezogen. Die HMM-Zustände wurden aus verrauschten Testdaten der AURORA2 Datenbank (SNR $= 5dB$) ermittelt, wozu die Merkmale mit dem SFE extrahiert wurden und anschließend eine Erkennung, wie in Abschnitt 2.1 beschrieben, durchgeführt wurde.

Die Ergebnisse für ein ergodisches HMM ohne Sprachmodell sind als grüne Kurve dargestellt. Es ist zu erkennen, dass die mittlere wechselseitige Information zwischen

den einzelnen Zuständen sehr schnell mit $|\tau|$ abfällt, also kaum Langzeitinformationen modelliert werden.

Durch die Verwendung einer Links-Rechts-Topologie, wodurch die Anzahl der möglichen Zustandsübergänge im HMM stark eingeschränkt wird, ergibt sich ein langsamerer Abfall von $I(q_0, q_\tau)$ (blaue Kurve).

Um den Einfluß des Sprachmodells qualitativ nachzuweisen, wurde unter Verwendung der HTK-Tools [YEH+02] jeweils auf unverrauschten Test-Sets mit 20 Sätzen der AURORA2 Testdatenbank ein Bigram-Sprachmodell trainiert. Die durchschnittliche Perplexität, gemittelt über alle N_{TS} Test-Sets TS, betrug dabei

$$\overline{PP} = \frac{1}{N_{TS}} \sum_{TS=1}^{N_{TS}} [\prod_{n=1}^{N} P(w_{TS,n}|w_{TS,n-1})]^{-\frac{1}{N}} \approx 4,3, \tag{8.3}$$

wobei $P(w_{TS,n}|w_{TS,n-1})$ die Übergangswahrscheinlichkeiten zwischen den Wörtern $n = 1, \ldots, N$ der Test-Sets TS bezeichnen. Das Sprachmodell wurde bei der Erkennung der gleichen 20 Sätze, jedoch bei einem SNR von $5dB$ im Rahmen der in Abschnitt 2.1.3 beschriebenen Viterbi-Decodierung zur Bestimmung der Übergangswahrscheinlichkeiten zwischen zwei Baumkopien eingesetzt. Aus der Abbildung wird ersichtlich (rote Kurve), dass die mittlere wechselseitige Information durch die Berücksichtigung des Sprachmodells für kleine $|\tau|$ nur geringfügig erhöht wird. Für einen zunehmenden Betrag von τ wird der Einfluß des Sprachmodells auf die mittlere wechselseitige Information größer, wobei jedoch gleichzeitig die mittlere wechselseitige Information stark abfällt.

Wie die vorangehende Analyse zeigt, wächst die Varianz der prädizierten Merkmalsvektorfolge bei gegebenen Modellwahrscheinlichkeiten im SLDM anders als im HMM monoton mit der Zeit an. Weiterhin wurde plausibel gemacht, dass die Modellwahrscheinlichkeit s_t, falls sie wie in Abschnitt 4.2.0.5 geschätzt wird, stark von der aktuellen Messung $\mathbf{y}_t$ abhängt, während die Zustandsfolge q_1^T über den gesamten Satz optimiert wird. Auf der anderen Seite ermöglicht das SLDM die direkte Modellierung von Inter-Frame Korrelationen. Das SLDM ist somit tendentiell besser zur Modellierung linearer statistischer Abhängigkeiten zwischen unmittelbar aufeinander folgenden Sprachrahmen geeignet, während das HMM sich eher zur Beschreibung nichtlinearer Langzeitabhängigkeiten eignet. Die Kombination der beiden Modellierungsansätze wird in den folgenden Abschnitten betrachtet.

8.2 Merkmalsentstörung unter Berücksichtigung der HMM-Zustände

Um die HMM-Zustände bei der Merkmalsentstörung im Front-End zu berücksichtigen, wird das statistische Modell in Abb. 8.3 angenommen. Das Modell stellt eine Erweiterung des statistischen Modells in Abb. 6.1 um die SLDM-Zustände s_t, die bei der Merkmalsentstörung im Front-End berücksichtigt werden, bzw. des SLDMs in Abb. 4.2 um

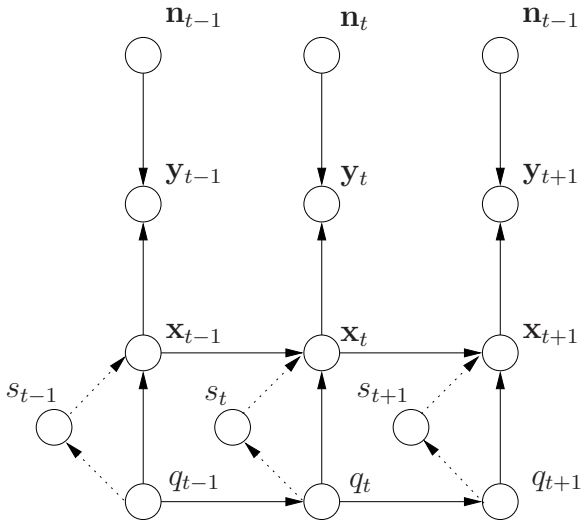

Abbildung 8.3: *Statistisches Modell der Sprachdynamik für die Merkmalsentstörung unter Berücksichtigung der HMM-Zustände*

die HMM-Zustände q_t dar. Die Abhängigkeit zwischen q_t und $\mathbf{x}_t$ kann entweder direkt oder über den gestrichelten Pfad $q_t \leftrightarrow s_t \leftrightarrow \mathbf{x}_t$ modelliert werden. Zunächst werden die SLDM-Zustände s_t nicht explizit berücksichtigt, d.h. der gestrichelte Pfad $q_t \leftrightarrow s_t \leftrightarrow \mathbf{x}_t$ wird vernachlässigt. Die HMM-Zustände können durch eine Marginalisierung bei der Berechnung der a posteriori Wahrscheinlichkeit der Zustandsvariable

$$\mathbf{z}_t = (\mathbf{x}_t, \mathbf{n}_t) \tag{8.4}$$

berücksichtigt werden:

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) = \sum_{q_t} p(\mathbf{z}_t, q_t|\mathbf{y}_1^t; \lambda_{q_1^t}), \tag{8.5}$$

wobei die Summation über alle HMM-Zustände des statistischen Modells in Abb. 8.3 durchgeführt wird. Der Parameter $\lambda_{q_1^t}$ bezeichnet das zugrunde liegende statistische Modell. Während im SLDM, das in Abb. 4.2 dargestellt ist, die Abhängigkeit zwischen q_1^T und $\mathbf{x}_1^T$ vernachlässigt wird, bedeutet die Angabe des Parameters

$$\lambda_{q_1^t} = \lambda_{q_1} \ldots \lambda_{q_t}, \tag{8.6}$$

dass die Abhängigkeiten $q_1 \leftrightarrow \mathbf{x}_1, \ldots, q_t \leftrightarrow \mathbf{x}_t$ (bzw. die Abhängigkeiten $q_1 \leftrightarrow s_1, \ldots$, $q_t \leftrightarrow s_t$ bei Vernachlässigung der direkten Abhängigkeiten $q_1 \leftrightarrow \mathbf{x}_1, \ldots, q_t \leftrightarrow \mathbf{x}_t$) in Abb. 8.3 zusätzlich bei der Merkmalsentstörung berücksichtigt werden. Die Umformung von Gl. (8.5) als

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) = \sum_{q_t} p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t}) P(q_t|\mathbf{y}_1^t; \lambda_{q_1^t}) \tag{8.7}$$

zeigt, dass $p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ eine Funktion der Wahrscheinlichkeit $P(q_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ ist, die durch die Lösung des Optimierungsproblems in Gl. (6.1) approximiert werden kann.

In Abb. 8.4 ist der Faktorgraph für einen Ausschnitt des statistischen Modells in Abb. 8.3 mit den Zufallsvariablen q_t, q_{t+1}, $\mathbf{x}_t$ und $\mathbf{x}_{t+1}$ dargestellt, der sich aus der Faktorisierung

$$\begin{aligned} p(\mathbf{x}_{t+1}, \mathbf{x}_t, q_{t+1}, q_t) &= p(\mathbf{x}_{t+1}|\mathbf{x}_t, q_{t+1}, q_t)p(\mathbf{x_t}|q_{t+1}, q_t)P(q_{t+1}|q_t)P(q_t) \\ &= p(\mathbf{x}_{t+1}|\mathbf{x}_t, q_{t+1})p(\mathbf{x}_t|q_t)P(q_{t+1}|q_t)P(q_t) \end{aligned} \tag{8.8}$$

ergibt. In dem Faktorgraphen werden die Zufallsvariablen durch Variablenknoten (◯) und die Funktionen dieser Zufallsvariablen durch Funktionsknoten (•) repräsentiert [SP04]. Da der Faktorgraph des Modellausschnittes eine zyklische Abhängigkeiten zwi-

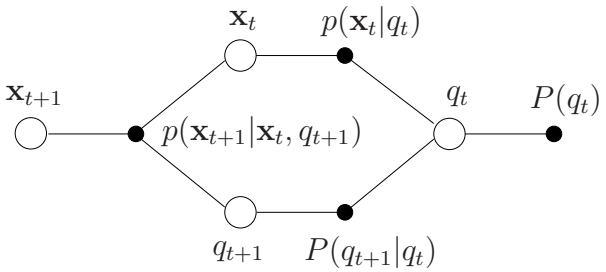

Abbildung 8.4: *Faktorgraph für einen Ausschnitt des statistischen Modells in Abb. 8.3*

schen q_t, q_{t+1} und $\mathbf{x}_t$ enthält, kann die a posteriori Wahrscheinlichkeit der beteiligten Zufallsvariablen prinzipiell nur iterativ berechnet werden (vgl. [SP04]). Es werden daher Ansätze hergeleitet, in denen $P(q_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ in einer vorangehenden Erkennungsstufe berechnet wird. Im Folgenden wird dies durch die Notation $P^{(-)}(q_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ ausgedrückt. Im Falle einer iterativen Merkmalsentstörung und Spracherkennung ist auch eine nichtkausale Verarbeitung akzeptabel. Daher wird $P^{(-)}(q_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ durch $P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ ersetzt:

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) = \sum_{q_t} p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t}) P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T}). \tag{8.9}$$

Auf ähnliche Weise kann der Term $p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t})$, der im SLDM berechnet wird, durch die geglättete Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^T, q_t; \lambda_{q_1^T})$ substituiert werden, wodurch sich allerdings die Komplexität der Merkmalsentstörung deutlich erhöht.

In den folgenden beiden Abschnitten werden zwei Approximationen für Gl. (8.9) untersucht. Daneben ist es möglich, die Informationen aus dem Back-End bei der Rauschschätzung auszunutzen (Abschnitt 8.5). Die Berechnung der a posteriori Wahrscheinlichkeiten der HMM-Zustände wird in Abschnitt 8.6 untersucht.

8.3 Rückkopplung von Informationen über das Zustandsmodell

Die erste Methode, Informationen aus dem Back-End zurückzukoppeln, die erstmals in [WHU07] veröffentlicht wurde, setzt bei der Berechnung der Modellwahrscheinlichkeiten im Front-End an. Wie bereits angeführt, besteht ein Nachteil des SLDMs darin, dass

die in Abschnitt 4.2.0.5 eingeführten Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ in der Regel stark von der Messung $\mathbf{y}_t$ abhängen. Daher sollen stattdessen die Modellwahrscheinlichkeiten $P^{(-)}(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ aus den a posteriori Wahrscheinlichkeiten $P^{(-)}(q_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ der HMM-Zustände ermittelt werden, die im Back-End unter Berücksichtigung des Akustik- und Sprachmodells berechnet werden. Es ist zu beachten, dass keine direkte Äquivalenz zwischen den HMM- und SLDM-Zuständen besteht, da die HMM-Zustände stationären Bereichen des Sprachsignals zugeordnet werden können, während die SLDM-Zustände Bereichen gleicher Dynamik entsprechen. Aufgrund der Links-Rechts-Topologie des HMM-Modells wird mit den HMM-Zuständen jedoch zugleich die Dauer des Wortes modelliert. Abb. 4.5 läßt eine Abhängigkeit zwischen der Position im Wort und den Zustandsmodellen plausibel erscheinen. Am Beginn und Ende einer Ziffer bzw. eines Wortes scheinen z.B. Modelle mit großer Varianz wahrscheinlicher als Modelle mit kleiner Varianz zu sein. Es erscheint daher naheliegend, eine statistische Abhängigkeit $P(s_t|q_t;\lambda_{q_t})$ zwischen s_t und q_t anzunehmen. Um die Existenz einer solchen Abhängigkeit nachzuweisen, wurde das folgende Experiment durchgeführt.

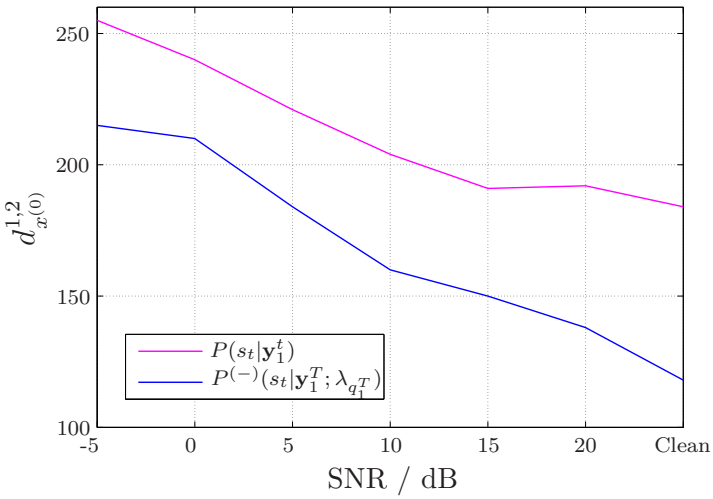

Abbildung 8.5: *Statistische Abhängigkeit zwischen HMM- und SLDM-Zuständen*

Auf den Testdaten aus Test-Set A der AURORA2 Datenbank wurden bei SNR-Pegeln zwischen $-5dB$ und $20dB$ die Wahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ und $P^{(-)}(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ der SLDM-Zustände berechnet. $P(s_t|\mathbf{y}_1^t)$ ergibt sich aus Gl. (4.31), während $P^{(-)}(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$, wie im nächsten Abschnitt beschrieben, über die statistische Abhängigkeit $P(s_t|q_t;\lambda_{q_t})$ aus den a posteriori Wahrscheinlichkeiten $P^{(-)}(q_t|\mathbf{y}_1^T;\lambda_{q_1^T}) = \delta(q_t - \hat{q}_t)$ der HMM-Zustände, die in einer vorangehenden Erkennungsstufe durch eine Viterbi-Erkennung mit verrauschten Sprachdaten des jeweiligen SNR-Pegels ermittelt werden, berechnet wird (Gl. (8.18)). Sowohl für die mit dem SLDM ermittelten Wahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t)$ als auch für die aus dem HMM zurückgekoppelten Wahrscheinlichkeiten $P^{(-)}(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ wurde auf den unverrauschten Testdaten $\mathbf{x}_1^T$ aus Test-

Set A der AURORA2 Datenbank der Prädiktionsfehler

$$d^{1,2}_{x^{(0)}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_t^{(0)} - \sum_{s_t} P_{1,2}(s_t)(\mathbf{A}^{(0,0:N_c-1)}(s_t)\mathbf{x}_{t-1}^{(0)} + b^{(0)}(s_t))\right)^2} \tag{8.10}$$

eines SLDMs mit den Modellparametern $\mathbf{A}(s_t)$ und $\mathbf{b}(s_t)$ sowie den Modellwahrscheinlichkeiten $P_1(s_t) = P(s_t|\mathbf{y}_1^t)$ und $P_2(s_t) = P(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ für die Energiekomponente $x_t^{(0)}$ berechnet. Die resultierenden Kurven für $P(s_t|\mathbf{y}_1^t)$ (magenta) und $P^{(-)}(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ (blau) sind in Abb. 8.5 dargestellt. Aus der Abbildung ist ersichtlich, dass bei allen Rauschpegeln durch die Rückkopplung der HMM-Zustände ein niedrigerer Prädiktionsfehler als mit der in Abschnitt 4.2.0.5 beschriebenen Methode erzielt wird.

8.3.1 Einbettung der Rückkopplungsmethode in den statistischen Ansatz

Wie aus den vorangehenden Ausführungen hervorgeht, soll die Abhängigkeit zwischen q_t und $\mathbf{x}_t$ in diesem Ansatz indirekt über den in Abb. 8.3 gestrichelt dargestellten Pfad $q_t \leftrightarrow s_t \leftrightarrow \mathbf{x}_t$ berücksichtigt werden. Die direkte Abhängigkeit $q_t \leftrightarrow \mathbf{x}_t$ wird stattdessen vernachlässigt. Die SLDM-Zustände können in Gl. (8.5) durch eine Marginalisierung bzgl. s_t berücksichtigt werden:

$$\begin{aligned} p(\mathbf{z}_t|\mathbf{y}_1^t;\lambda_{q_1^t}) &= \sum_{s_t=1}^{M} p(\mathbf{z}_t, s_t|\mathbf{y}_1^t;\lambda_{q_1^t}) = \sum_{s_t=1}^{M}\sum_{q_t} p(\mathbf{z}_t, s_t, q_t|\mathbf{y}_1^t;\lambda_{q_1^t}) \\ &= \sum_{s_t=1}^{M}\sum_{q_t} p(\mathbf{z}_t|\mathbf{y}_1^t, s_t, q_t;\lambda_{q_1^t}) P(s_t|\mathbf{y}_1^t, q_t;\lambda_{q_1^t}) P(q_t|\mathbf{y}_1^t;\lambda_{q_1^t}). \end{aligned} \tag{8.11}$$

Man beachte, dass Gl. (8.11) der Marginalisierung (4.16) im SLDM entspricht. Da die unmittelbare Abhängigkeit zwischen q_t und $\mathbf{x}_t$ und damit auch zwischen q_t und $\mathbf{z}_t$ in dem zugrunde liegenden statistischen Modell vernachlässigt wird, kann $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t, q_t;\lambda_{q_1^t})$ als

$$p(\mathbf{z}_t|\mathbf{y}_1^t, s_t, q_t;\lambda_{q_1^t}) = p(\mathbf{z}_t|\mathbf{y}_1^t, s_t;\lambda_{q_1^{t-1}}) \tag{8.12}$$

geschrieben werden, d.h. bei gegebenen s_t liefert q_t und damit auch die Modellerweiterung λ_{q_t} um die statistische Abhängigkeit $q_t \leftrightarrow s_t$ keine zusätzlichen Informationen über $\mathbf{x}_t$. Die bedingten a posteriori Wahrscheinlichkeiten $p(\mathbf{z}_t|\mathbf{y}_1^t, s_t;\lambda_{q_1^{t-1}})$ der einzelnen Filter können wie im SLDM, das in Abb. 4.2 dargestellt ist, aus der a posteriori Wahrscheinlichkeit $p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1};\lambda_{q_1^{t-1}})$ des vorangehenden Sprachrahmens berechnet werden. Dazu wird analog zu Gl. (4.20), Gl. (4.18) und Gl. (4.19) die Rekursion

$$p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}, s_t;\lambda_{q_1^{t-1}}) = p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1};\lambda_{q_1^{t-1}}) \tag{8.13}$$

$$p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t;\lambda_{q_1^{t-1}}) = \int_{\mathbb{R}^{2N_c}} p(\mathbf{z}_t|\mathbf{z}_{t-1}, s_t) p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}, s_t;\lambda_{q_1^{t-1}}) d\mathbf{z}_{t-1} \tag{8.14}$$

$$p(\mathbf{z}_t|\mathbf{y}_1^t, s_t;\lambda_{q_1^{t-1}}) \propto p(\mathbf{z}_t|\mathbf{y}_1^{t-1}, s_t;\lambda_{q_1^{t-1}}) p(\mathbf{y}_t|\mathbf{z}_t) \tag{8.15}$$

durchgeführt, bei der im Unterschied zu den entsprechenden Gleichungen für das SLDM aus Abb. 4.2 die statistischen Abhängigkeiten $\lambda_{q_1^{t-1}}$ berücksichtigt werden.

Die Informationen über die HMM-Zustände, die im Back-End ermittelt werden, werden dazu verwendet, die Wahrscheinlichkeiten $P(s_t|\mathbf{y}_1^T)$ durch die entsprechenden Terme in Gl. (8.11) zu ersetzen. Unter der Annahme, dass ein stärkerer statistischer Zusammenhang zwischen q_t und s_t als zwischen q_t und $\mathbf{y}_1^t$ besteht, kann die Approximation

$$P(s_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t}) \approx P(s_t|q_t; \lambda_{q_1^t}) = P(s_t|q_t; \lambda_{q_t}) \tag{8.16}$$

vorgenommen werden. $P(s_t|q_t; \lambda_{q_t})$ wird, wie in Abschnitt 8.3.2 beschrieben, aus unverrauschten Trainingsdaten gelernt und in einer Zustandstabelle gespeichert. Weiterhin wird $P(q_t|\mathbf{y}_1^t; \lambda_{q_1^t})$, wie eingangs erläutert, durch $P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ ersetzt.

Damit ergibt sich die a posteriori Wahrscheinlichkeit

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) = \sum_{s_t=1}^{M} \sum_{q_t} p(\mathbf{z}_t|\mathbf{y}_1^t, s_t; \lambda_{q_1^{t-1}}) P(s_t|q_t; \lambda_{q_t}) P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T}). \tag{8.17}$$

Ein Vergleich von (8.17) und (4.16) zeigt, dass anstelle der Wahrscheinlichkeit $P(s_t|\mathbf{y}_1^t)$, die im Standard-SLDM zur Gewichtung der Modelle verwendet wird, die Wahrscheinlichkeit

$$P^{(-)}(s_t|\mathbf{y}_1^T; \lambda_{q_1^T}) = \sum_{q_t} P(s_t|q_t; \lambda_{q_t}) P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T}) \tag{8.18}$$

berechnet wird. Die Schätzung von $P^{(-)}(s_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ als Funktion der HMM-Wahrscheinlichkeiten $P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$, die in einer vorangehenden Erkennungsstufe berechnet werden, führt auf die in Abb. 8.6 dargestellte Iteration.

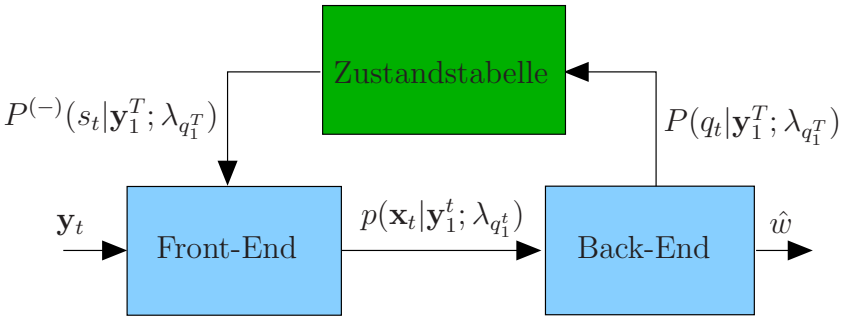

Abbildung 8.6: *Berechnung der SLDM-Zustände im HMM*

Die verrauschten Sprachmerkmale werden im Front-End mit einem SLDM entstört, das die a posteriori Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ des Zustandsvektors $\mathbf{z}_t$ und damit die a posteriori Wahrscheinlichkeit $p(\mathbf{x}_t|\mathbf{y}_1^t; \lambda_{q_1^t})$ der unverrauschten Sprachmerkmale $\mathbf{x}_t$ liefert. In der ersten Iteration existiert kein Ausgangssignal des HMMs, so dass die Abhängigkeit $P(s_t|q_t; \lambda_{q_t})$ vernachlässigt und die Entrauschung der Sprachmerkmale wie in Kapitel 4 beschrieben durchgeführt wird, d.h. die Modelle des SLDMs, wie in Abschnitt 4.2.0.5 beschrieben, mit $P(s_t|\mathbf{y}_1^t)$ gewichtet werden. Der HMM-Decoder liefert die a posteriori Wahrscheinlichkeiten $P(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$, $t = 1, \ldots, T$. Anschließend wird

mit Hilfe der Zustandstabelle die Wahrscheinlichkeit $P(s_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ entsprechend Gl. (8.18) berechnet, mit der die Modelle bei der Entrauschung der Sprachmerkmale in der nächsten Iteration gewichtet werden.

8.3.2 Training der Zustandstabelle

Die bedingte Wahrscheinlichkeit $P(s|q)$ wird aus unverrauschten Trainingsdaten gelernt. Dazu wird der Schätzwert $\hat{P}(s|q)$ der Abbildung $P(s|q)$ als

$$\hat{P}(s|q) = \frac{\hat{P}(s,q)}{\hat{P}(q)} \tag{8.19}$$

geschrieben. Der Zähler kann aus der a posteriori Wahrscheinlichkeit der Zustände bestimmt werden:

$$\hat{P}(s,q) = \sum_{t=1}^{T} P(s_t = s|\mathbf{x}_1^T, q_t = q)P(q_t = q|\mathbf{x}_1^T), \tag{8.20}$$

wobei die Summe über alle Sprachrahmen $t = 1 \ldots T$ der Datenbank gebildet wird und $\mathbf{x}_1^T$ alle Merkmalsvektoren bezeichnet. Da ein statistischer Zusammenhang zwischen s_t und q_t a priori nicht bekannt ist, sind vereinfachende Annahmen in Gl. (8.20) erforderlich. Es wird angenommen, dass $P(s_t|\mathbf{x}_1^T, q_t = q)$ bei gegebenen unverrauschten Sprachmerkmalen $\mathbf{x}_1^T$ hinreichend genau durch

$$P(s_t|q_t, \mathbf{x}_1^T) \approx P(s_t|\mathbf{x}_1^T) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)P(s_t) \tag{8.21}$$

approximiert werden kann. Der Term $P(q_t = q|\mathbf{x}_1^T)$ in Gl. (8.20) kann durch eine harte Entscheidung approximiert werden:

$$P(q_t = q|\mathbf{x}_1^T) \approx \delta(q_t - q) = \begin{cases} 1, \text{ falls Zustand } q_t \text{ in Sprachrahmen } t \text{ aktiv ist} \\ 0, \text{ sonst.} \end{cases} \tag{8.22}$$

Auf ähnliche Weise ergibt sich der Nenner in Gl. (8.19):

$$\hat{P}(q) = \sum_{t=1}^{T} P(q_t = q|\mathbf{x}_1^T). \tag{8.23}$$

8.4 Rückkopplung von Informationen über die Verteilung der Sprachmerkmale

In diesem Ansatz werden die Informationen aus dem Back-End zu einer genaueren Approximation der a posteriori Verteilung $p(\mathbf{z}_t|\mathbf{y}_1^t;\lambda_{q_1^t})$ nach dem Aktualisierungsschritt (4.16) im SLDM verwendet. Dabei wird in dem statistischen Modell aus Abb. 8.3 anstelle der statistischen Abhängigkeit zwischen q_t und s_t, die in Abschnitt 8.3 ausgenutzt wurde, die Abhängigkeit zwischen q_t und $\mathbf{x}_t$ berücksichtigt. Es wird angenommen, dass

q_t und die statistische Abhängigkeit λ_{q_t} gegenüber $\mathbf{y}_1^t$ und $\lambda_{q_1^{t-1}}$ vollständig redundanzfreie Informationen über $\mathbf{x}_t$ (und damit über $\mathbf{z}_t$) liefern. Diese Annahme erscheint besser als die alternative Approximation $p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t}) \approx p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})$, die sich aus der Vernachlässigung der statistischen Abhängigkeit von q_t ergibt. Die a posteriori Wahrscheinlichkeit des Zustandes q_t wird nämlich über den ganzen Satz optimiert und enthält somit nichtlineare Langzeitabhängigkeiten (vgl. Abschnitt 8.1), die ansonsten unberücksichtigt blieben. Unter der getroffenen Annahme kann $p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t})$ in Gl. (8.9) faktorisiert werden:

$$p(\mathbf{z}_t|\mathbf{y}_1^t, q_t; \lambda_{q_1^t}) \approx \frac{p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})p(\mathbf{z}_t|q_t; \lambda_{q_t})}{p(\mathbf{z}_t)} = \frac{p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})p(\mathbf{x}_t|q_t; \lambda_{q_t})}{p(\mathbf{x}_t)}. \tag{8.24}$$

In Gl. (8.24) wurden weiterhin die Beziehungen

$$p(\mathbf{z}_t|q_t; \lambda_{q_t}) = p(\mathbf{n}_t|q_t; \lambda_{q_t})p(\mathbf{x}_t|q_t; \lambda_{q_t}) = p(\mathbf{n}_t)p(\mathbf{x}_t|q_t; \lambda_{q_t}) \tag{8.25}$$

$$p(\mathbf{z}_t) = p(\mathbf{n}_t)p(\mathbf{x}_t) \tag{8.26}$$

ausgenutzt, die aus dem statistischen Modell in Abb. 8.3 folgen. Die Verteilung

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}}) = \sum_{s_t=1}^{M} p(\mathbf{z}_t|\mathbf{y}_1^t, s_t; \lambda_{q_1^{t-1}})P(s_t|\mathbf{y}_1^t) \tag{8.27}$$

kann mit Gl. (4.31) und Gln. (8.13)-(8.15) aus der a posteriori Wahrscheinlichkeit $p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}; \lambda_{q_1^{t-1}})$ des vorangehenden Sprachrahmens berechnet werden. Die Verteilung $p(\mathbf{x}_t|q_t; \lambda_{q_t})$ ist die Emissionsverteilung des Akustikmodells, die in der Regel als Gaußmischungsverteilung modelliert wird (vgl. Abschnitt 6.1.1). Verglichen mit Gl. (4.20) führt das Einsetzen von Gl. (8.24) in Gl. (8.9) zu der modifizierten a posteriori Verteilung

$$p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) = \frac{p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T; \lambda_{q_1^T})}{p(\mathbf{x}_t)} \tag{8.28}$$

mit

$$p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T; \lambda_{q_1^T}) = \sum_{q_t} p(\mathbf{x}_t|q_t; \lambda_{q_t})P^{(-)}(q_t|\mathbf{y}_1^T; \lambda_{q_1^T}), \tag{8.29}$$

wobei die Wahrscheinlichkeitsdichte $p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ zur Komplexitätsreduktion entsprechend Gl. (4.17) zu einer unimodalen Gaußverteilung zusammengefaßt wird.

In der resultierenden Rückkopplungsschleife wird $p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T; \lambda_{q_1^T})$, wie in Abb. 8.7 dargestellt, in das Front-End zurückgekoppelt und, wie in Gl. (8.28) angegeben, mit der a posteriori Wahrscheinlichkeit $p(\mathbf{z}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})$ des SLDMs kombiniert.

Gl. (8.28) kann unter der Annahme der statistischen Unabhängigkeit zwischen $\mathbf{x}_t$ und $\mathbf{n}_t$ in eine Gleichung für die Sprachmerkmale $\mathbf{x}_t$ und eine zweite Gleichung für das Rauschen $\mathbf{n}_t$ aufgespalten werden:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) &= \frac{p(\mathbf{x}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}})p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T; \lambda_{q_1^T})}{p(\mathbf{x}_t)} \\ p(\mathbf{n}_t|\mathbf{y}_1^t; \lambda_{q_1^t}) &= p(\mathbf{n}_t|\mathbf{y}_1^t; \lambda_{q_1^{t-1}}). \end{aligned} \tag{8.30}$$

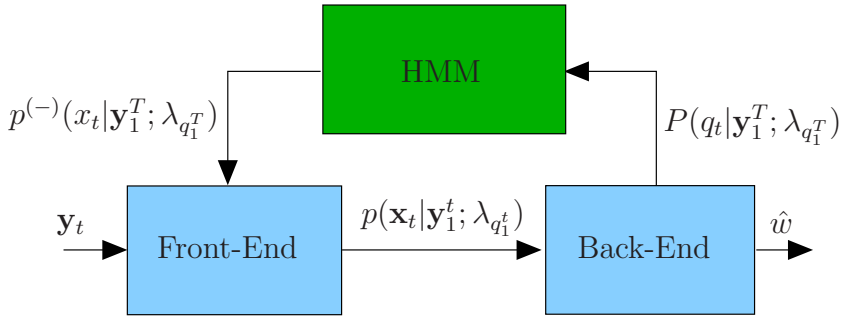

Abbildung 8.7: *Rückkopplung von Informationen über die Verteilung der Sprachmerkmale*

In Anhang E.3 wird gezeigt, dass $p(\mathbf{x}_t|\mathbf{y}_1^t;\lambda_{q_1^t})$ unter der Annahme gaußförmiger Ausgangsverteilungen

$$p(\mathbf{x}_t|\mathbf{y}_1^t;\lambda_{q_1^{t-1}}) = \mathcal{N}(\mathbf{x}_t;\mathbf{x}_{t|1:t;\lambda_1^{t-1}},\mathbf{P}^{(x)}_{t|1:t;\lambda_1^{t-1}}) \tag{8.31}$$

$$p^{(-)}(\mathbf{x}_t|\mathbf{y}_1^T;\lambda_{q_1^T}) = \mathcal{N}(\mathbf{x}_t;\mathbf{x}^{(-)}_{t|1:T},\mathbf{P}^{(x-)}_{t|1:T}) \tag{8.32}$$

$$p(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_{\boldsymbol{x}},\boldsymbol{\Sigma}_{\boldsymbol{x}}) \tag{8.33}$$

mit den Momenten $\mathbf{x}_{t|1:t;\lambda_1^{t-1}}$, $\mathbf{P}^{(x)}_{t|1:t;\lambda_1^{t-1}}$, $\mathbf{x}^{(-)}_{t|1:T}$, $\mathbf{P}^{(x-)}_{t|1:T}$, $\boldsymbol{\mu}_{\boldsymbol{x}}$ und $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ proportional zu einer Gaußverteilung

$$p(\mathbf{x}_t|\mathbf{y}_1^t;\lambda_{q_1^t}) = \mathcal{N}(\mathbf{x}_t;\mathbf{x}_{t|1:t;\lambda_1^t},\mathbf{P}^{(x)}_{t|1:t;\lambda_1^t}) \tag{8.34}$$

mit den Momenten

$$\mathbf{x}_{t|1:t;\lambda_1^t} = \mathbf{P}^{(x)}_{t|1:t;\lambda_1^t}\left((\mathbf{P}^{(x)}_{t|1:t;\lambda_1^{t-1}})^{-1}\mathbf{x}_{t|1:t;\lambda_1^{t-1}} + (\mathbf{P}^{(x-)}_{t|1:T})^{-1}\mathbf{x}^{(-)}_{t|1:T} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}}\right) \tag{8.35}$$

$$(\mathbf{P}^{(x)}_{t|1:t;\lambda_1^t})^{-1} = (\mathbf{P}^{(x)}_{t|1:t;\lambda_1^{t-1}})^{-1} + (\mathbf{P}^{(x-)}_{t|1:T})^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{8.36}$$

ist. Im Gegensatz zu der Rückkopplung von Informationen über das Zustandsübergangsmodell ist die Rückkopplung von Informationen über die Verteilung der Sprachmerkmale nicht auf schaltende Modelle beschränkt, da bei der Faktorisierung in Gl. (8.24) keine Annahmen über die Modellwahrscheinlichkeiten des SLDMs getroffen werden.

8.5 Verwendung der Zustandswahrscheinlichkeiten bei der Rauschschätzung

Eine weitere Möglichkeit, die a posteriori Wahrscheinlichkeiten $P^{(-)}(q_t|\mathbf{y}_1^T;\lambda_{q_1^T})$ der HMM-Zustände im Front-End des Spracherkenners auszunutzen, besteht in einer verbesserten Rauschschätzung. Dazu ist es möglich, eine Soft-VAD-Variable

$$P_{sil} = \sum_{q_t \epsilon Q_{sil}} P^{(-)}(q_t|\mathbf{y}_1^T;\lambda_{q_1^T}) \tag{8.37}$$

zu berechnen, wobei Q_{sil} die Menge der HMM-Zustände, die Sprachpausen zugeordnet sind, bezeichnet. Im Folgenden wird in Gl. (4.6) ein Rauschmodell der Form

$$\mathbf{D} = \mathbf{0}, \quad \mathbf{e} = \boldsymbol{\mu}_{\mathbf{n}_t}, \quad \mathbf{F} = \boldsymbol{\Sigma}_{\boldsymbol{n}} \tag{8.38}$$

angenommen, d.h.

$$\mathbf{n}_t = \mathcal{N}(\mathbf{n}_t; \boldsymbol{\mu}_{\mathbf{n}_t}, \boldsymbol{\Sigma}_{\boldsymbol{n}}) \tag{8.39}$$

wird als a priori Verteilung der Ordnung Null mit zeitvariantem Erwartungswert $\boldsymbol{\mu}_{\mathbf{n}_t}$ und konstanter Varianz $\boldsymbol{\Sigma}_n$ modelliert. Die Rauschschätzung kann mit den Merkmalsvektoren aus Sprachpausen am Anfang und Ende des Satzes initialisiert werden. Anschließend wird der Erwartungswert $\boldsymbol{\mu}_{\mathbf{n}_t}$ über die Beziehung

$$\boldsymbol{\mu}_{\mathbf{n}_t} = (1 - \alpha P_{sil})\boldsymbol{\mu}_{\mathbf{n}_{t-1}} + \alpha P_{sil}\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t} \tag{8.40}$$

aktualisiert, wobei α einen Gewichtsfaktor bezeichnet, der den Einfluß der aktuellen Rauschschätzung $\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}$ beeinflußt. In informalen Experimenten wurde für α der Wert $0,05$ festgelegt. Falls die Wahrscheinlichkeit $P_{sil} = 0$ ist, wird keine Aktualisierung der Zustandsschätzung durchgeführt. Die instantane Rauschschätzung $\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}$ wird in Sprachpausen als $\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}^{(sil)} = \mathbf{y}_t$ und in Phasen mit Sprachaktivität als

$$\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}^{(speech)} = \mathbf{n}_{t|1:t-1} \approx \mathbf{n}_{t-1|1:t-1} \tag{8.41}$$

berechnet, wobei $\mathbf{n}_{t-1|1:t-1}$ bei der Filterung in Gl. (4.17) berechnet wird. Damit ergibt sich für die instantane Rauschschätzung in Gl. (8.40)

$$\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t} = P_{sil}\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}^{(sil)} + (1 - P_{sil})\tilde{\boldsymbol{\mu}}_{\mathbf{n}_t}^{(speech)}. \tag{8.42}$$

8.6 Berechnung der Zustandswahrscheinlichkeiten

In dem HMM-Erkenner, der in Abschnitt 2.1 beschrieben wird, wird eine harte Entscheidung bzgl. der besten Zustandsfolge $\hat{q}_1^T$ getroffen. Es ist möglich, diese Zustandsfolge als Ausgangspunkt für die Rückkopplung zu verwenden und $P(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ in Gl. (8.18) und Gl. (8.29) somit durch den Dirac-Stoß

$$P(q_t|\mathbf{y}_1^T; \lambda_{q_1^T}) \approx \delta(q_t - \hat{q}_t) \tag{8.43}$$

zu approximieren, wobei $\hat{q}_t$ den Zustand der Zustandsfolge $\hat{q}_1^T$ zum Zeitpunkt t bezeichnet. Bei einem solchen Vorgehen ist es allerdings möglich, dass alternative Zustandsschätzungen nicht hinreichend berücksichtigt und die Auswirkungen von Fehlentscheidungen verstärkt werden. Auf die Problematik der zu starken Gewichtung einer Hypothese bei der Rückkopplung von Informationen in das Front-End wird bereits in [FW06] hingewiesen (vgl. Abschnitt 2.4). Um die exakten Zustandswahrscheinlichkeiten $P(q_t|\mathbf{y}_1^T; \lambda_{q_1^T})$ zu berechnen, kann ein Vorwärts-Rückwärts-Algorithmus auf Zustandsebene angewendet werden (Abschnitt 8.6.1). Daneben werden in Abschnitt 8.6.2 zwei Möglichkeiten aufgezeigt, die Zustandswahrscheinlichkeiten mit einem Vorwärts-Rückwärts-Algorithmus auf Wortebene zu approximieren, um eine Reduktion des Speicher- und Rechenaufwandes zu erreichen. Auf Implementierungsdetails wird in Anhang D eingegangen. Der Modellparameter $\lambda_{q_1^T}$ wird im Folgenden aus Gründen der Übersichtlichkeit vernachlässigt.

8.6.1 Vorwärts-Rückwärts-Algorithmus auf Zustandsebene

Die exakten a posteriori Wahrscheinlichkeiten $\gamma_t(i) = P(q_t = i|\mathbf{y}_1^T)$ der HMM-Zustände können mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene ermittelt werden, der in vielen Spracherkennern in ähnlicher Form beim Training der Modellparameter als Bestandteil des Baum-Welch-Algorithmus eingesetzt wird [RJ93]. Anders als bei der Suche der optimalen Zustandsfolge $\hat{q}_1^T$, die in Abschnitt 2.1.3 beschrieben wird, wird in diesem Abschnitt ein lineares Lexikon angenommen, in dem die akustischen Einheiten jedem Wort einzeln zugeordnet werden. Der Ansatz basiert auf der Berechnung der Hilfsvariablen

$$\alpha_t(i) = p(\mathbf{y}_1^t, q_t = i), \quad \beta_t(i) = p(\mathbf{y}_{t+1}^T|q_t = i), \tag{8.44}$$

deren Produkt proportional zu $\gamma_t(i)$ ist:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N_q} \alpha_t(j)\beta_t(j)}, \tag{8.45}$$

mit der Anzahl N_q der HMM-Zustände. Die Werte der Variablen α_t und β_t können rekursiv in Vorwärts- bzw. Rückwärtsrichtung bestimmt werden:

$$\begin{aligned} \alpha_t(j) &= \left(\sum_{i=1}^{N_q} \alpha_{t-1}(i)\alpha_{ij} \right) p_{LH}(\mathbf{y}_1^T|q_t = j)^{S_\alpha} \\ \beta_t(i) &= \left(\sum_{j=1}^{N_q} \beta_{t+1}(j)\alpha_{ij} p_{LH}(\mathbf{y}_1^T|q_{t+1} = j)^{S_\alpha} \right) \end{aligned} \tag{8.46}$$

mit der Likelihood $p_{LH}(\mathbf{y}_1^T|q_t)$ des HMM-Zustandes q_t entsprechend Gl. (6.8). Der akustische Skalierungsfaktor $S_\alpha < 1$ verhindert, dass die beste Hypothese im Vergleich zu den anderen Hypothesen zu stark gewichtet wird (vgl. Abschnitt 6.2). Die Vorwärtsvariable α_t wird, wie in [RJ93] angegeben, mit

$$\alpha_1(i) = \pi_i p_{LH}(\mathbf{y}_1^T|q_1)^{S_\alpha} \tag{8.47}$$

initialisiert, wobei π_i die a priori Wahrscheinlichkeit des Zustandes $q_t = i$ bezeichnet, die Rückwärtsvariable $\beta_t(i)$ mit dem Wert $\beta_T(i) = 1$. Die Berücksichtigung eines Bigram-Sprachmodells (Gl. (2.4)) führt auf die gegenüber [RJ93] modifizierten Übergangswahrscheinlichkeiten

$$\alpha_{ij} = \begin{cases} P(w_t(q_t)|w_{t-1}(q_t - 1)) & : \quad q_t = Q_{w_t}^{(s)}, \quad q_{t-1} = Q_{w_{t-1}}^{(e)} \\ P(q_t|q_{t-1}) & : \quad \text{sonst}, \end{cases} \tag{8.48}$$

wobei w_{t-1} und w_t die Wörter der Zustände $q_{t-1} = i$ und $q_t = j$, $Q_{w_t}^{(s)}$ den Startzustand des Wortes w_t und $Q_{w_{t-1}}^{(e)}$ den Endzustand des Wortes w_{t-1} bezeichnen.

8.6.2 Verwendung eines Wortgraphen bei der Berechnung der Zustandswahrscheinlichkeiten

In diesem Abschnitt werden im Vergleich zu dem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene effizientere Möglichkeiten untersucht, um die a posteriori Wahrscheinlichkeiten der HMM-Zustände zu berechnen, wozu die wahrscheinlichsten Wortsequenzen betrachtet werden. Wie in Abschnitt 8.6.2.2 herausgestellt wird, stellen Wortgraphen eine geeignete Repräsentation der Wortsequenzen dar. Ein Wortgraph ist ein gerichteter, azyklischer Graph (DAG), dessen Knoten $\tau_i \in \{1, \ldots, T\}$, $i = 1, \ldots, N_\tau$, diskrete Zeitpunkte bezeichnen und dessen Kanten $e = [w; \tau_s, \tau_e]$ Worthypothesen entsprechen. Dabei bezeichnen w das Wort, τ_s den Startknoten und τ_e den Endknoten der Hypothese. Den Kanten sind die akustischen Wahrscheinlichkeiten $p(\mathbf{y}_{\tau_s}^{\tau_e}|[w; \tau_s, \tau_e])$ als Gewichte zugeordnet. In den folgenden Unterabschnitten werden die Konstruktion des Wortgraphen, die Berechnung der a posteriori Wahrscheinlichkeiten der Wörter sowie die Verwendung des Wortgraphen zur effizienten Berechnung der a posteriori Wahrscheinlichkeiten der HMM-Zustände behandelt. Zu den beiden ersten Punkten existieren verschiedene Literaturansätze, auf die in Abschnitt 2.1.3 kurz eingegangen wurde, deren ausführliche Untersuchung jedoch nicht Gegenstand dieser Arbeit ist. Im Folgenden wird die Konstruktion des Wortgraphen basierend auf den Ausführungen in [ONA97] durchgeführt, während für die Berechnung der a posteriori Wahrscheinlichkeiten für die Worthypothesen der Ansatz in [Wes02] zugrunde gelegt wird.

8.6.2.1 Konstruktion des Wortgraphen

Die Grundlage für die Konstruktion des Wortgraphen stellt der in Abschnitt 2.1.3 beschriebene Viterbi-Algorithmus dar, der durch die Substitution der Likelihood $p(\mathbf{x}_t|q_t)$, wie in Abschnitt 6.1 beschrieben, auf verrauschte Eingangsdaten $\mathbf{y}_1^T$ erweitert wurde. Dieser wird unter der Annahme eines Bigram-Sprachmodells entsprechend den Ausführungen in [ONA97] und [Wes02] modifiziert. Dazu wird Gl. (2.12) im Vorwärtsschritt des Viterbi-Algorithmus so abgeändert, dass für ein Wort w zum Zeitpunkt t anstelle eines einzelnen Vorgängers

$$v_0(w; t) = \underset{v}{\operatorname{argmax}}\{P(w|v)^{S_\beta}\alpha_t(v, q_t)\} \tag{8.49}$$

mit der Wortgrenze $B_t(v_0(w; t), q_t = Q_w^{(e)})$ die Wortgrenzen $\tau(t; v, w) = B_t(v, q_t = Q_w^{(e)})$ der besten Vorgänger $v \in \mathcal{V}(w; t)$ abgespeichert werden. Um die Elemente von $\mathcal{V}(w; t)$ zu bestimmen, wird ein Schwellwert-Pruning durchgeführt, d.h.

$$\mathcal{V}(w; t) = \left\{v|P(w|v)^{S_\beta}\alpha_t(v, q_t) > \max_{\tilde{v}}\{P(w|\tilde{v})^{S_\beta}\alpha_t(\tilde{v}, q_t)\} - T_{wg}\right\} \tag{8.50}$$

mit dem Schwellwert $T_{wg} > 0$. Daneben wird den Wortpaaren (v, w) zum Zeitpunkt t die Likelihood

$$A(t; v, w) = \alpha_t(v, q_t = Q_w^{(e)})/H(v; \tau) \tag{8.51}$$

mit der Vorwärtsvariable $H(v; \tau)$ aus Gl. (2.13) zugeordnet.

Der Wortgraph wird am Satzende erzeugt, indem ausgehend von dem Endknoten $t_{N_\tau} = T$ des Wortgraphen unter Verwendung der Wortgrenzen $\tau(t; v, w)$ rekursiv alle Vorgänger v des wahrscheinlichsten Endwortes

$$\hat{w}_T = \underset{w}{\operatorname{argmax}}\{H(w; T)\} \tag{8.52}$$

zurückverfolgt werden:

Algorithmus 3 Erzeugung des Wortgraphen

1: Gegeben: Wort w, Wortgrenze τ_e.
2: Erzeuge einen Knoten τ_e.
3: **Für alle** $v \in \mathcal{V}(w; \tau_e)$
4: $\tau_s = \tau(\tau_e; v, w)$.
5: Erzeuge einen Knoten τ_s.
6: Erzeuge eine Kante $e = [w; \tau_s, \tau_e]$ mit der Likelihood $p(\mathbf{y}_{\tau_s}^{\tau_e}|e) = A(\tau_e; v, w)$.
7: Wiederhole die Rekursion für $w := v$ und $\tau_e := \tau_s - 1$.
8: **Ende.**

Entsprechend [Wes02] wird eine Wortgraph-Optimierung durchgeführt. Dazu werden während der rekursiven Zurückverfolgung der Wortgrenzen Knoten, die identischen Zeitpunkten zugeordnet werden können, die also in Schritt 5 von Algorithmus 3 mehrfach vorhanden sind, zu einem einzelnen Knoten zusammengefaßt. Nach der Konstruktion des Wortgraphen werden die Knoten und Kanten entsprechend der Startzeitpunkte sortiert und jeweils zwei Kanten $[w^{(i)}; \tau_s^{(i)}, \tau_e^{(i)}]$ und $[w^{(j)}; \tau_s^{(j)}, \tau_e^{(j)}]$, für die

$$w^{(i)} = w^{(j)}, \quad \tau_s^{(i)} = \tau_s^{(j)} \quad \text{und} \quad \tau_e^{(i)} = \tau_e^{(j)} \tag{8.53}$$

gilt, auf eine einzelne Kante mit der akustischen Wahrscheinlichkeit

$$p(\mathbf{y}_{\tau_s}^{\tau_e}|[w; \tau_s, \tau_e]) = \max\left\{p(\mathbf{y}_{\tau_s^{(i)}}^{\tau_e^{(i)}}|[w; \tau_s, \tau_e]^{(i)}), p(\mathbf{y}_{\tau_s^{(j)}}^{\tau_e^{(j)}}|[w; \tau_s, \tau_e]^{(j)})\right\} \tag{8.54}$$

reduziert.

Abschließend soll auf eine Problematik hingewiesen werden, die dadurch entsteht, dass in der Vorwärtsiteration unverändert die Viterbi-Approximation (2.11) angewendet wird, um einen exponentiellen Anstieg der Hypothesen bei der Suche zu verhindern. Die Anwendung der Viterbi-Approximation führt dazu, dass die Wortgrenze $B_t(v, q_t = Q_w^{(e)})$ zwischen den Wörtern v und w nur für den wahrscheinlichsten Vorgänger des Wortes v, der im Folgenden als a bezeichnet wird, richtig bestimmt wird (siehe Abb. 8.8). Aufgrund der Viterbi-Approximation wird in Gl. (2.11) nämlich nur die Endhypothese von Baumkopie a in den Startzustand von Baumkopie v propagiert. Das bedeutet, dass der Zeitpunkt der Wortgrenze zwischen Wort v und Wort w aufgrund der Historie $a-v$ berechnet wird, obwohl sich aufgrund der Hypothese $b-v$ oder $c-v$ unter Umständen ein anderer Startzeitpunkt ergäbe, da in Gl. (2.10) ein anderer Wert der Vorwärtsvariable in den Startzustand propagiert würde. In dem klassischen Viterbi-Algorithmus, wo nur die Hypothese $a - v - w$ zurückverfolgt wird, hat dies

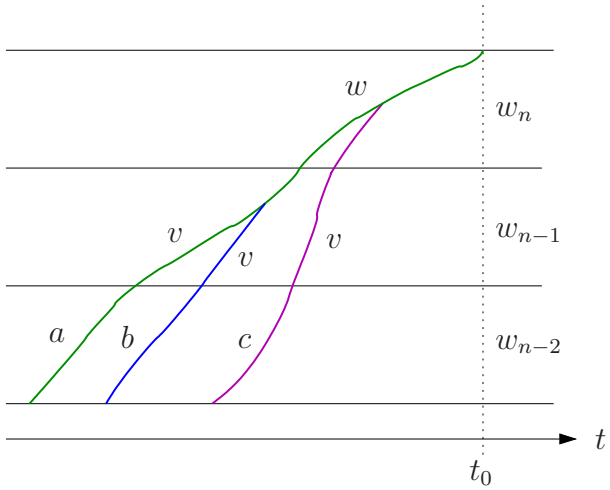

Abbildung 8.8: *Wortpaarapproximation*

keine Auswirkungen, während sich im Wortgraph-Algorithmus unter der Berücksichtigung anderer Historien eine falsche Wortgrenze ergibt. An dieser Stelle greift, wie in [ONA97] erläutert, die in Abb. 8.8 dargestellte Wortpaarapproximation [SA91], die besagt, dass die Wortgrenze zwischen zwei Wörtern unabhängig von den vorhergehenden Wörtern ist. Unter der Voraussetzung eines Bigram-Sprachmodells und hinreichend langer Wörter wird angenommen, dass unabhängig von der Historie irgendwann der gleiche Zustand im akustischen Modell des Wortes $w_{n-1} = v$ für alle hypothetischen Vorgängerwörter am wahrscheinlichsten ist und der Übergang zu Wort $w_n = w$ somit unabhängig von Wort w_{n-2} ist. In Abb. 8.8 ist die Wortpaarapproximation für die Historie $b - v$ erfüllt, während sie für die Historie $c - v$ verletzt wird.

8.6.2.2 Berechnung der a posteriori Wahrscheinlichkeiten für Wörter

Der Wortgraph, dessen Konstruktion im letzten Abschnitt beschrieben wurde, enthält die akustischen Likelihoods $p(\mathbf{y}_{\tau_s}^{\tau_e}|[w;\tau_s,\tau_e])$ der Wortgraphkanten $[w;\tau_s,\tau_e]$ (vgl. Algorithmus 3). Um die Wortalternativen bewerten zu können, werden jedoch die a posteriori Wahrscheinlichkeiten $P([w;\tau_s,\tau_e]|\mathbf{y}_1^T)$ der Wortgraphkanten benötigt, aus denen, wie weiter unten angegeben, die a posteriori Wahrscheinlichkeiten der Wörter w_t berechnet werden können. Es ist möglich, eine N-Best-Liste mit den Likelihoods $p(\mathbf{y}_1^T|([w;\tau_s,\tau_e]_1^{N^{(j)}})^{(j)})$ für alle im Wortgraphen enthaltenen Wortfolgen

$$([w;\tau_s,\tau_e]_1^{N^{(j)}})^{(j)} = [w;\tau_s,\tau_e]_1^{(j)} \ldots [w;\tau_s,\tau_e]_n^{(j)} \ldots [w;\tau_s,\tau_e]_{N^{(j)}}^{(j)}, \quad j = 1 \ldots N_{hyp}, \tag{8.55}$$

der Länge $N^{(j)}$ aufzustellen. Anschließend kann die a posteriori Wahrscheinlichkeit der Wortgraphkante $e = [w;\tau_s,\tau_e]$ als Summe

$$P(e|\mathbf{y}_1^T) = \sum_{j=1}^{N_{hyp}} \sum_{n=1}^{N^{(j)}} p(\mathbf{y}_1^T|(w_1^{N^{(j)}})^{(j)})^{S_\alpha} P((w_1^{N^{(j)}})^{(j)}) \delta(e - e_n^{(j)}) \tag{8.56}$$

der a posteriori Wahrscheinlichkeiten der Wortfolgen, die die Wortgraphkante e enthalten, berechnet werden. Wie in Abschnitt 6.2 erläutert, muß die akustische Likelihood

$p(\mathbf{y}_1^T|(w_1^N)^{(j)})$ in Gl. (8.56) mit dem akustischen Skalierungsfaktor S_α gegenüber der a priori Wahrscheinlichkeit $P((w_1^{N^{(j)}})^{(j)})$ der Wortfolge $(w_1^{N^{(j)}})^{(j)}$ gewichtet werden. Bei einem Wortgraphen mit $N = 10$ aufeinander folgenden Wörtern und drei Möglichkeiten an jeder Stelle, der also $3N = 30$ Kanten besitzt, ergeben sich mit dem beschriebenen Ansatz allerdings bereits $N_{hyp} = 3^{10}$ Wortfolgen. Daher ist es effizienter, die a posteriori Wahrscheinlichkeiten der Wortgraphkanten direkt auf dem Wortgraphen zu berechnen, wozu ein in [Wes02] beschriebener Wortgraph-Algorithmus herangezogen wurde.

Die folgende Darstellung orientiert sich an den Ausführungen in [Wes02] für ein allgemeines m-Gram Sprachmodell. In diesem Zusammenhang bezeichnet $h_1^{m-1}(e) = h_1(e), \ldots, h_{m-1}(e)$ die Historie der Wortgraphkante $e = [w; \tau_s, \tau_e]$, wobei $h_{m-1}(e)$ der unmittelbare Vorgänger von e ist. Im Folgenden wird die Abhängigkeit der Historie von e wie auch in [Wes02] der Einfachheit halber vernachlässigt, insofern sie aus dem Kontext ersichtlich ist. Weiterhin wird angenommen, dass für alle Indizes i, j gilt: $P(w|h_i^j) = P(w|h_1^j)$, falls $i < 0$, und $P(w|h_i^j) = P(w)$, falls $j < i$. Analog dazu werden die $m-1$ Nachfolger f_1^{m-1} der Wortgraphkante e definiert.

Die a posteriori Wahrscheinlichkeiten der Wörter können ähnlich wie die a posteriori Wahrscheinlichkeiten der HMM-Zustände in Abschnitt 8.6.1 aus einer Vorwärts- und einer Rückwärtsvariable berechnet werden. Die Vorwärtsvariable $\phi(h_2^{m-1}; [w; \tau_s, \tau_e])$, die eine Funktion der Sprachmodellhistorie h_2^{m-1} und der Wortgraphkante $[w; \tau_s, \tau_e]$ ist, ergibt sich aus der Iteration [Wes02]

$$\phi(h_2^{m-1}; [w; \tau_s, \tau_e]) = p(\mathbf{y}_{\tau_s}^{\tau_e}|[w; \tau_s, \tau_e])^{S_\alpha} \sum_{h_1} \sum_{\tau_s'} \phi(h_1^{m-2}; [h_{m-1}; \tau_s', \tau_s - 1]) P(w|h_1^{m-1}). \tag{8.57}$$

In Gl. (8.57) bezeichnet τ_s den Startpunkt des aktuellen Wortes, während $\tau_s - 1$ der Endpunkt des letzten Wortes ist. Die Definition der Rückwärtsvariable wurde gegenüber [Wes02] leicht verändert, indem $\hat{\psi}([w; \tau_s, \tau_e]; f_1^{m-2})$ aus [Wes02] durch

$$\psi([w; \tau_s, \tau_e]; f_1^{m-2}) = \hat{\psi}([w; \tau_s, \tau_e]; f_1^{m-2}) / p(\mathbf{y}_{\tau_s}^{\tau_e}|[w; \tau_s, \tau_e])^{S_\alpha} \tag{8.58}$$

ersetzt wurde, wodurch in Gl. (8.60) die Division durch $p(\mathbf{y}_{\tau_s}^{\tau_e}|[w; \tau_s, \tau_e])^{S_\alpha}$ entfällt. Die Rückwärtsvariable $\psi([w; \tau_s, \tau_e]; f_1^{m-2})$, die eine Funktion der Wortgraphkante e und der Nachfolger f_1^{m-2} des Wortes w ist, kann nach der Substitution in Gl. (8.58) über die Iteration

$$\psi(e; f_1^{m-2}) = \sum_{f_{m-1}} \sum_{\tau_e'} \psi([f_1; \tau_e + 1, \tau_e']); f_2^{m-1}) p(\mathbf{y}_{\tau_e+1}^{\tau_e'}|[f_1; \tau_e + 1, \tau_e'])^{S_\alpha} P(f_{m-1}|w, f_1^{m-2}) \tag{8.59}$$

berechnet werden. Insgesamt ergibt sich entsprechend [Wes02] mit der Modifikation in Gl. (8.58) die a posteriori Wahrscheinlichkeit

$$\begin{aligned} P(e|\mathbf{y}_1^T) &= \sum_{h_2^{m-1}} \sum_{f_1^{m-2}} \left\{ \frac{\phi(h_2^{m-1}; [w; \tau_s, \tau_e]) \psi([w; \tau_s, \tau_e]; f_1^{m-2})}{p(\mathbf{y}_1^T)} \prod_{n=1}^{m-2} P(f_n|h_{n+1}^{m-1}, w, f_1^{n-1}) \right\} \\ &\propto \sum_{h_2^{m-1}} \sum_{f_1^{m-2}} \left\{ \phi(h_2^{m-1}; [w; \tau_s, \tau_e]) \psi([w; \tau_s, \tau_e]; f_1^{m-2}) \prod_{n=1}^{m-2} P(f_n|h_{n+1}^{m-1}, w, f_1^{n-1}) \right\} \end{aligned} \tag{8.60}$$

der Kante $e = [w; \tau_s, \tau_e]$, wobei das Produkt über die Wahrscheinlichkeiten $P(f_n|h_{n+1}^{m-1}, w, f_1^{n-1})$ in Gl. (8.60), wie in [Wes02] angegeben, unter der Annahme eines Unigram- oder Bigram-Sprachmodells vernachlässigt werden kann.

Die a posteriori Wahrscheinlichkeiten $P(e|\mathbf{y}_1^T)$ der Wortgraphkanten $e = [w; \tau_s, \tau_e]$, für die $\tau_s \leq t \leq \tau_e$ gilt, können zum Zeitpunkt t elementar durch eine Normierung von Gl. (8.60) auf

$$\sum_{[w;\tau_s,\tau_e]|\tau_s \leq t \leq \tau_e} P(e|\mathbf{y}_1^T) = 1, \tag{8.61}$$

bestimmt werden, wodurch die explizite Berechnung der Verteilung $p(\mathbf{y}_1^T)$, für die in [Wes02] ein Zusammenhang angegeben wird, nicht erforderlich ist.

Wie in [Wes02] angegeben, kann die a posteriori Wahrscheinlichkeit $P(w_t|\mathbf{y}_1^T)$ des Wortes w_t zum Zeitpunkt t durch die Aufsummierung der a posteriori Wahrscheinlichkeiten $P([w; \tau_s, \tau_e]|\mathbf{y}_1^T)$, die zum Zeitpunkt t verschieden von Null sind, bestimmt werden:

$$P(w_t|\mathbf{y}_1^T) = \sum_{[w;\tau_s,\tau_e]|\tau_s \leq t \leq \tau_e} \delta(w_t - w) P([w; \tau_s, \tau_e]|\mathbf{y}_1^T). \tag{8.62}$$

In der vorliegenden Arbeit wurde abweichend von [Wes02] eine Transducer-basierte Umsetzung der Vorwärts- und Rückwärtsiteration in Gl. (8.57) und Gl. (8.59) gewählt, die im Folgenden kurz erläutert wird. In dem beschriebenen Ansatz wird die Konsistenz mit dem Sprachmodell mit einer Zustandsmaschine (FSM) sichergestellt, deren Zustände den möglichen Sprachmodellhistorien $q^{(FSM)} = h_1^{m-1}$ entsprechen und deren Übergangswahrscheinlichkeiten $P(w|h_1^{m-1})$ sich aus dem zugrunde gelegten m-Gram Sprachmodell ergeben. Wie in Abb. 8.9 dargestellt, werden einer Kante des

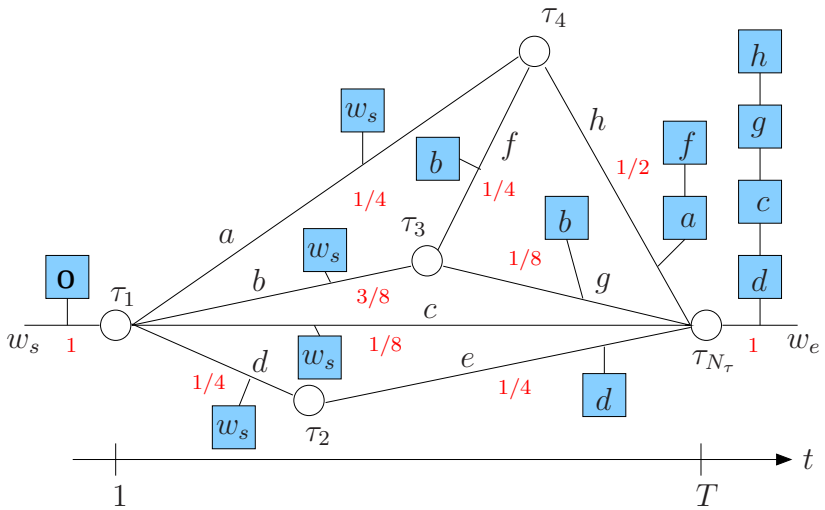

Abbildung 8.9: *Berechnung der a posteriori Wahrscheinlichkeiten (rot) für einen Wortgraphen mit den Wörtern a ... h*

Wortgraphen neben dem Wort w alle mit der FSM erreichbaren Sprachmodellhistorien zugeordnet (blaue Rechtecke). Dadurch ist es möglich, bei der Vorwärts- und Rückwärtsiteration nur Historien zu berücksichtigen, die mit dem Sprachmodell vereinbar sind, so dass beispielsweise nicht über Übergänge zwischen SIL-Kanten summiert wird.

Vor der Durchführung der Vorwärts- und Rückwärtsiteration wird der Wortgraph um eine Kante $[w_s;\, -\infty,\, 0]$, die in den Startknoten $\tau_1 = 1$ mündet, und eine zweite Kante $[w_e,\, T+1,\, \infty]$, die aus dem Endknoten herausführt, erweitert.

In der Vorwärtsiteration werden der Kante $[w_s;\, -\infty,\, 0]$ der Startzustand der FSM $q_0^{(FSM)}$ und der Wert der Vorwärtsvariable

$$\phi(q_0^{(FSM)}; [w_s; -\infty, 0]) = 1 \tag{8.63}$$

zugeordnet. Anschließend werden die Knoten rekursiv beginnend mit dem Startknoten $\tau_s = \tau_1 = 1$ abgearbeitet. Dabei werden für jeden Knoten τ_s des Wortgraphen die Hypothesen der wegführenden Kanten $[w;\, \tau_s,\, \tau_e]$ mit Wort w und Endknoten τ_e aktualisiert, sobald die Aktualisierung aller hinführenden Kanten $[h_{m-1};\, \tau_s',\, \tau_s - 1]$ mit den Startknoten τ_s' und den Wörtern h_{m-1} abgeschlossen ist:

Algorithmus 4 Vorwärtsiteration im Wortgraphen

1: Gegeben: Knoten τ_s.
2: **Für alle** wegführenden Kanten $[w; \tau_s, \tau_e]$
3: **Für alle** hinführenden Kanten $[h_{m-1}; \tau_s', \tau_s - 1]$
4: **Für alle** FSM-Zustände h_1^{m-1} der hinführenden Kante
5: Ermittle $P(w|h_1^{m-1})$ durch einen FSM-Übergang.
6: **Falls** $P(w|h_1^{m-1}) > 0$ **dann**
7: Erhöhe $\phi(h_2^{m-1}; [w; \tau_s, \tau_e])$ um den Summanden (h_1, τ_s') (Gl. (8.57)).
8: Erweitere $[w; \tau_s, \tau_e]$ um den FSM-Zustand $h_1^{m-1} := (h_2^{m-1}, w)$.
9: **Ende.**
10: **Ende.**
11: **Ende.**
12: **Falls** alle Vorgängerkanten von $\tau_e + 1$ aktualisiert sind **dann**
13: Wiederhole die Rekursion für $\tau_s := \tau_e + 1$.
14: **Ende.**
15: **Ende.**

In der Rückwärtsiteration werden die Knoten in umgekehrter Reihenfolge abgearbeitet, wobei der Rückwärts-Variable für die Kante $[w_e, T+1, \infty]$ der Wert

$$\psi([w_e; T+1, \infty]; f_1^{m-2}) = 1 \tag{8.64}$$

zugewiesen wird, die Iteration mit dem Knoten $\tau_e = \tau_{N_\tau} = T$ begonnen wird und die Rückwärtsvariable entsprechend Gl. (8.59) aktualisiert wird (Algorithmus 5).
Der prinzipielle Unterschied zu der Vorwärtsiteration besteht in der Aktualisierung der Historien h_1^{m-1} bzw. f_1^{m-1}, da die Übergänge der FSM in Rückwärtsrichtung nicht eindeutig sind. Die Übergangswahrscheinlichkeiten $P(f_1|(h_2^{m-1}, w))$ können effizient berechnet werden, indem in Schritt 5 der Rückwärtsiteration die Historien (h_2^{m-1}, w) verwendet werden, die in Schritt 8 der Vorwärtsiteration im Wortgraphen gespeichert werden.

Algorithmus 5 Rückwärtsiteration im Wortgraphen

Gegeben: Knoten τ_e.
Für alle hinführenden Kanten $[w; \tau_s, \tau_e]$
  Für alle wegführenden Kanten $[f_1; \tau_e + 1, \tau'_e]$ mit Wort f_1 und Endknoten τ'_e
    Für alle FSM-Zustände (h_2^{m-1}, w) der hinführenden Kante
      Ermittle $P(f_1|(h_2^{m-1}, w))$ durch einen FSM-Übergang.
      Falls $P(f_1|(h_2^{m-1}, w)) > 0$ **dann**
        Erhöhe $\psi([w; \tau_s, \tau_e]; f_1^{m-2})$ um den Summanden (f_{m-1}, τ'_e) (Gl. (8.59)).
      Ende.
    Ende.
  Ende.
  Falls alle Nachfolgerkanten von $\tau_e + 1$ aktualisiert sind **dann**
    Wiederhole die Rekursion für $\tau_e := \tau_s - 1$.
  Ende.
Ende.

8.6.2.3 Berechnung der Zustandswahrscheinlichkeiten unter Berücksichtigung des Wortgraphen

Nachfolgend werden zwei Ansätze eingeführt, in denen der Wortgraph dazu verwendet wird, die a posteriori Wahrscheinlichkeiten der HMM-Zustände effizient zu berechnen.

Ansatz 1: Berechnung der optimalen HMM-Zustände für die Wortgraphkanten

Eine Möglichkeit, die a posteriori Wahrscheinlichkeiten $P(e|\mathbf{y}_1^T)$ der Wortgraphkanten bei der Berechnung der Zustandswahrscheinlichkeiten $P(q_t|\mathbf{y}_1^T)$ zu berücksichtigen, besteht in der Marginalisierung

$$P(q_t|\mathbf{y}_1^T) = \sum_e P(q_t|e, \mathbf{y}_1^T) P(e|\mathbf{y}_1^T) \tag{8.65}$$

bzgl. der Wortgraphkanten e. Die Wahrscheinlichkeiten $P(q_t|e, \mathbf{y}_1^T)$ können während der Konstruktion des Wortgraphen ermittelt werden. Dazu wird $P(q_t|e, \mathbf{y}_1^T)$ durch einen diskreten Dirac-Stoß

$$P(q_t|e, \mathbf{y}_1^T) = \delta(q_t - \hat{q}_t(e)) \tag{8.66}$$

approximiert, wobei $\hat{q}_t(e)$ den wahrscheinlichsten Zustand der Wortgraphkante e bezeichnet. Um $\hat{q}_t(e)$ zu berechnen, wird zunächst im Vorwärtsschritt des Wortgraph-Algorithmus die beste Zustandsfolge $\hat{Q}_t(v, q_t) = \hat{q}_1, \ldots, \hat{q}_{t-1}, q_t$ zusammen mit der Vorwärtsvariable $\alpha_t(v, q_t)$ aktualisiert. Dazu wird die Initialisierung der Baumkopien in Gl. (2.10) um die Beziehung

$$\hat{Q}_{t-1}(v, q_{t-1} = 0) = \emptyset \tag{8.67}$$

erweitert. $\hat{Q}_t(v, q_t)$ wird während der Vorwärtsiteration in Gl. (2.11) um den besten Vorgängerzustand erweitert:

$$\hat{q}_{t-1}(q_t) = \underset{q_{t-1}}{\text{argmax}}\{\alpha_{t-1}(v, q_{t-1})P(q_t|q_{t-1})\} \tag{8.68}$$

$$\hat{Q}_t(v, q_t) = \hat{Q}_{t-1}(v, \hat{q}_{t-1}(q_t)) \circ q_t. \tag{8.69}$$

Bei der Erzeugung des Wortgraphen mit Algorithmus 3 wird zu jeder Wortgraphkante $e = [w; \tau_s, \tau_e]$ mit dem Wort w die Zustandsfolge $\hat{Q}_t(v_0(w), q_t = Q_e^{(w)})$ mit den Zuständen $\hat{q}_t(e) \in \hat{Q}_t(v_0(w), q_t = Q_e^{(w)})$ gespeichert.

Ansatz 2: Einschränkung der möglichen Zustände im Vorwärts-Rückwärts-Algorithmus auf Zustandsebene

Ein zweiter Ansatz besteht darin, die a posteriori Wahrscheinlichkeiten $P(w_t|\mathbf{y}_1^T)$ der Wörter w_t zur Beschränkung der möglichen Zustände im Vorwärts-Rückwärts-Algorithmus auf Zustandsebene (siehe Abschnitt 8.6.1) zu verwenden. Um den Suchraum einzuschränken, wird zunächst ein Schwellwert-Pruning der Wahrscheinlichkeiten $P(w_t|\mathbf{y}_1^T)$ durchgeführt:

$$P(w_t|\mathbf{y}_1^T) = 0, \quad \text{für} \quad P(w_t|\mathbf{y}_1^T) < T_{wp}, \tag{8.70}$$

mit konstantem Schwellwert T_{wp}. Daneben wird ein Histogramm-Pruning

$$P(w_t|\mathbf{y}_1^T) = 0, \quad \text{für} \quad n_w(w_t) > N_{wp}, \tag{8.71}$$

mit dem Schwellwert N_{wp} durchgeführt, wobei die Indizes $n_w(w_t) = 1 \ldots N_w$, mit der Anzahl N_w der Wörter, sich aus der absteigenden Sortierung der Wörter w_t nach der Wahrscheinlichkeit $P(w_t|\mathbf{y}_1^T)$ ergeben.

Anschließend werden die möglichen Zustände bei der Berechnung der Vorwärts- und Rückwärts-Variablen auf Wörter mit positiver a posteriori Wahrscheinlichkeit begrenzt:

$$\alpha_t(q_t) = 0, \quad \beta_t(q_t) = 0, \quad \text{für} \quad P(w_t(q_t)|\mathbf{y}_1^T) = 0, \tag{8.72}$$

so dass $\alpha_t(q_t)$ und $\beta_t(q_t)$ nur für die HMM-Zustände mit $P(w_t(q_t)|\mathbf{y}_1^T) > 0$ berücksichtigt werden müssen. Die Menge dieser Zustände wird im Folgenden als

$$Q_t^{(wp)} = \{q_t | P(w_t(q_t)|\mathbf{y}_1^T) > 0)\} \tag{8.73}$$

bezeichnet. In Gl. (8.46) ergibt sich somit die Vereinfachung

$$\begin{aligned} \forall j \in Q_t^{(wp)}: \quad \alpha_t(j) &= \left(\sum_{i=1, i \in Q_{t-1}^{(wp)}}^{N_q} \alpha_{t-1}(i)\alpha_{ij} \right) p_{LH}(\mathbf{y}_1^T|q_t = j)^{S_\alpha} \\ \forall i \in Q_t^{(wp)}: \quad \beta_t(i) &= \left(\sum_{j=1, j \in Q_{t+1}^{(wp)}}^{N_q} \beta_{t+1}(j)\alpha_{ij} p_{LH}(\mathbf{y}_1^T|q_{t+1} = j)^{S_\alpha} \right). \end{aligned} \tag{8.74}$$

Daneben werden durch die Einschränkung in Gl. (8.72) auch bei der Initialisierung der Vorwärts- und Rückwärtsvariablen und der Berechnung der a posteriori Wahrscheinlichkeit in Gl. (8.45) nur die HMM-Zustände aus $Q_t^{(wp)}$ berücksichtigt.

8.7 Experimentelle Ergebnisse

Die Rückkopplung der Erkennungsergebnisse in das Front-End wurde mit verrauschten Sprachdaten der TI-Digits Datenbank (AURORA2) und des Wallstreet Journal Task (AURORA4) getestet (siehe Anhang A). Zur Entstörung der mit dem SFE extrahierten Sprachmerkmale wurde das SLDM-M16 verwendet (vgl. Abschnitt 4.5). Dieses wurde entsprechend der Ausführungen in Abschnitt 8.3 und Abschnitt 8.4 modifiziert, um die Rückkopplung von Informationen in das Front-End zu ermöglichen. Im Folgenden wird die Rückkopplungsmethode aus Abschnitt 8.3 mit FB1 und die Rückkopplungsmethode aus Abschnitt 8.4 mit FB2 bezeichnet.

Zunächst wurde die Auswirkung der Rückkopplung qualitativ untersucht. Abb. 8.10 zeigt den Verlauf der $x^{(0)}$-Komponente des Merkmalsvektors für einen Beispielsatz der AURORA2 Datenbank. Der Verlauf der verrauschten und unverrauschten Merkmals-

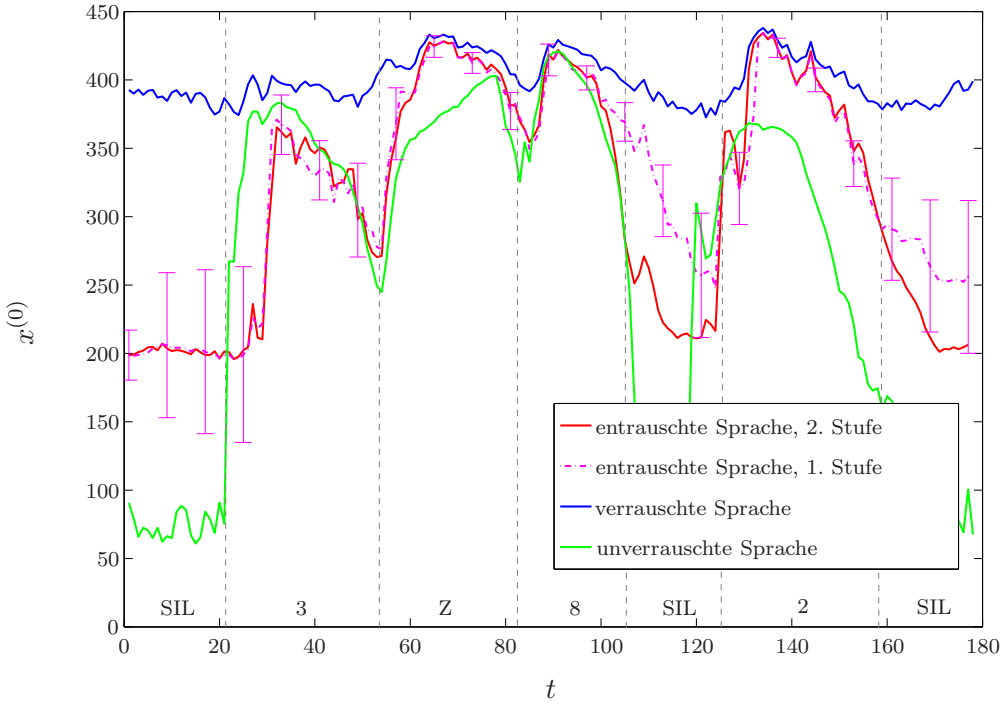

Abbildung 8.10: *Zeitliche Verläufe der $x^{(0)}$-Komponente für einen Beispielsatz der AURORA2 Datenbank*

vektorkomponente sind in Abb. 8.10 als blaue bzw. grüne Linie eingezeichnet. Daneben ist der Verlauf der entstörten Merkmalsvektorkomponente nach der Entstörung in der ersten (gestrichelte, magenta Linie) und zweiten (durchgezogene, rote Linie) Erkennungsstufe dargestellt. Man erkennt, dass das Rauschen in der zweiten Stufe in Sprachpausen tendentiell stärker unterdrückt wird. Außerdem entsprechen die Verläufe an den Wortgrenzen tendentiell besser den wahren Signalverläufen. In dem Beispielsatz sind diese Effekte insbesondere in der Sprachpause zwischen den Rahmen $t = 105$ und

$t = 125$ zu beobachten, in der der Verlauf der Merkmalsvektorkomponente in der zweiten Stufe dem tatsächlichen Kurvenverlauf deutlich besser entspricht als der Verlauf in der ersten Stufe.

Tab. 8.1 zeigt das Potential des Rückkopplungsansatzes für das SLDM-M16. Um eine obere Grenze für die erreichbare Erkennungsrate zu erhalten, wurde die optimale Zustandsfolge des HMMs auf unverrauschten Sprachdaten berechnet und anschließend in das Front-End zurückgekoppelt. Tab. 8.1 zeigt, dass auf Test-Set A der AURORA2

FB1	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,49	99,66	99,61
20dB	98,68	97,76	99,16	98,92	98,63
15dB	97,21	94,95	98,72	97,07	96,99
10dB	93,46	87,45	96,66	92,19	92,44
5dB	83,60	74,49	89,44	83,40	82,73
0dB	61,25	45,41	68,18	62,85	59,42
-5dB	32,91	19,59	29,41	36,10	29,50
Ø	86,84	80,01	90,43	86,89	86,04

FB2	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,58	99,49	99,57	99,58
20dB	99,17	98,37	99,31	98,83	98,92
15dB	98,83	96,98	99,31	98,27	98,35
10dB	97,73	94,23	98,90	97,96	97,21
5dB	96,38	92,41	98,21	95,50	95,63
0dB	95,98	89,87	97,35	94,72	94,48
-5dB	95,89	90,05	97,46	95,46	94,72
Ø	97,62	94,37	98,62	97,06	96,92

Tabelle 8.1: Potential der Rückkopplung auf Test-Set A der AURORA2 Datenbank a) SLDM-FB1opt b) SLDM-FB2opt

Datenbank mit verrauschten Eingangsdaten, aber durch die Rückkopplung der optimalen Zustandsfolge, die aus den unverrauschten Sprachdaten berechnet wurde, Erkennungsraten von $86,04\%$ (SLDM-FB1opt) bzw. $96,92\%$ (SLDM-FB2opt) für die beiden Rückkopplungsarten erreicht werden können. Bei bekannten HMM-Zuständen ist es somit besser, unmittelbar die a posteriori Verteilung $p(\mathbf{x}_t|\mathbf{y}_1^t, \lambda_{q_1^t})$ des Merkmalsvektors $\mathbf{x}_t$ anstelle der Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_1^t, \lambda_{q_1^t})$ durch die Rückkopplung zu beeinflussen. Das Ergebnis bestätigt allerdings, dass ein statistischer Zusammenhang zwischen s_t und q_t besteht und durch die Rückkopplung der Modellwahrscheinlichkeiten bei hinreichend genauer Schätzung der HMM-Zustände bessere Ergebnisse als mit den Modellwahrscheinlichkeiten $P(s_t|\mathbf{y}_t)$, die im SLDM berechnet werden, erzielt werden können.

Im Folgenden wird die Rückkopplung einer einzelnen, mit dem Viterbi-Algorithmus aus verrauschten Eingangsdaten ermittelten Zustandsfolge untersucht. Wie Tab. 8.2 zu entnehmen ist, ergab sich durch die Festlegung der Modellwahrscheinlichkeiten des SLDMs aufgrund der detektierten Zustandsfolge des HMMs (SLDM-M16-FB1) auf der AURORA2 Datenbank ein Anstieg der Erkennungsrate von $79,87\%$ auf $80,77\%$ auf Test-Set A und ein Anstieg von $79,16\%$ auf $80,18\%$ auf Test-Set B. Der Anstieg der Erkennungsrate war für niedrige SNR-Pegel größer als für hohe SNR-Pegel. Dieses Ergebnis ist vermutlich auf den zunehmenden Einfluß des Meßmodells aufgrund der abnehmenden Beobachtungsvarianz bei ansteigenden SNR-Pegel zurückzuführen. Wie in Abb. 8.5 dargestellt, ist die Abnahme des Prädiktionsfehlers auf den unverrauschten Testdaten aufgrund der Rückkopplung für alle SNR-Pegel ungefähr konstant. Mit der

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,49	99,60	99,59
20dB	97,94	96,13	99,08	99,00	98,04
15dB	95,79	90,57	98,36	96,15	95,22
10dB	89,68	79,96	94,84	90,16	88,66
5dB	75,38	60,91	84,70	77,38	74,59
0dB	48,63	30,99	57,05	52,79	47,37
-5dB	21,46	12,15	19,13	24,41	19,29
Ø	81,48	71,71	86,81	83,10	80,77

Set B	Res.	Str.	Air.	Tra.	Ø
Clean	99,53	99,54	99,42	99,57	99,52
20dB	96,21	97,97	97,12	98,27	97,39
15dB	90,67	95,62	94,30	96,76	94,34
10dB	81,30	88,30	86,04	92,87	87,13
5dB	66,10	74,61	72,62	81,95	73,82
0dB	41,20	46,55	48,88	56,18	48,20
-5dB	32,91	19,59	29,41	36,10	29,50
Ø	75,10	80,61	79,79	85,21	80,18

Tabelle 8.2: SLDM-M16-FB1: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

zweiten Rückkopplungsmethode wurde durch die Rückkopplung einer einzelnen Zustandsfolge auf der AURORA2 Datenbank keine signifikante Verbesserung gegenüber dem SLDM-M16 erzielt.

In Tab. 8.3 wurde untersucht, inwieweit die Erkennungsergebnisse durch die Verwendung eines Sprachmodells verbessert werden können. Das Bigram-Sprachmodell wurde, wie in Abschnitt 8.1 beschrieben, jeweils auf 20 Sätzen trainiert und bei der Berechnung der Zustandswahrscheinlichkeiten in der ersten Erkennungsstufe eingesetzt. In

	Sub.	Bab.	Car	Exh.	Ø
Clean	99,54	99,55	99,46	99,51	99,52
20dB	97,73	95,92	99,14	98,83	97,91
15dB	95,89	90,96	98,63	96,17	95,41
10dB	89,96	80,11	95,17	90,43	88,92
5dB	76,92	62,48	86,25	78,28	75,98
0dB	49,12	31,74	60,42	54,35	48,91
-5dB	21,22	12,21	19,32	27,65	20,10
Ø	81,92	72,24	87,92	83,61	81,43

Tabelle 8.3: SLDM-M16-FB1 mit Bigram-Sprachmodell in der ersten Erkennungsstufe: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

der zweiten Erkennungsstufe wurde das SLDM-M16-FB1 zur Merkmalsentstörung verwendet und anschließend eine Erkennung ohne Sprachmodell durchgeführt. Durch die Verwendung des Sprachmodells in der ersten Erkennungsstufe ergab sich ein statistisch signifikanter Anstieg der Erkennungsrate von $80,77\%$ auf $81,43\%$.

Auf der AURORA4 Datenbank wurde die Fehlerrate durch die Rückkopplung der besten Zustandsfolge von $40,5\%$ (SLDM-M16, vgl. Tab. 4.2) auf $36,6\%$ (SLDM-M32-FB1) bzw. $36,9\%$ (SLDM-M16-FB2) reduziert (siehe Tab. 8.4). Für die Rückkopplungsmethode FB1 wurden $M = 32$ Modelle verwendet, da $M = 16$ nur optimal ist,

		Train	Airp.	Babble	Car	Street	Rest.	Clean	Ø
SLDM-M16	S	29,9	37,9	32,2	12,7	29,9	37,2	9,5	27,0
	I	8,8	15,9	12,7	4,6	10,6	13,3	3,1	9,9
	E	42,6	58,2	48,8	19,5	45,0	56,0	13,7	40,5
SLDM-M32-FB1	S	29,6	32,0	27,8	12,5	27,6	33,0	9,3	24,5
	I	6,5	12,9	9,4	4,4	8,0	10,9	3,1	7,9
	E	41,4	50,2	41,4	18,9	41,0	49,8	13,7	36,6
SLDM-M16-FB2	S	28,9	31,5	27,8	12,3	28,1	32,8	9,1	24,4
	I	7,8	13,4	10,4	5,2	9,7	11,6	2,4	8,6
	E	41,6	49,8	42,3	19,4	42,6	50,2	12,7	36,9
SLDM-M16+UD	S	25,9	25,5	25,6	12,1	26,2	29,7	10,0	22,1
	I	3,9	8,2	6,8	3,8	6,7	8,7	2,1	5,7
	E	37,8	39,3	37,7	17,3	38,9	45,6	14,5	33,0
SLDM-M32-FB1 +UD	S	25,9	25,5	24,6	12,3	25,0	28,4	11,5	22,0
	I	3,9	8,2	6,3	3,0	6,5	6,6	3,0	5,4
	E	37,8	39,3	37,1	17,6	36,5	42,1	13,3	32,0
SLDM-M16-FB2 +UD	S	26,5	24,9	24,2	11,9	25,5	28,9	9,3	24,5
	I	4,4	8,6	6,7	3,5	5,3	8,0	3,1	7,9
	E	37,8	39,4	36,3	17,6	36,9	44,4	14,4	32,4

Tabelle 8.4: Rückkopplung der besten Zustandsfolge: Fehlerraten auf der AURORA4 Datenbank (S: Ersetzungsfehler, I: Einfügungen, E: Gesamtfehlerrate)

wenn die Modellwahrscheinlichkeiten im SLDM berechnet werden, während die Genauigkeit der Prädiktion bei der Berechnung der Modellwahrscheinlichkeiten aus den a posteriori Wahrscheinlichkeiten der HMM-Zustände prinzipiell mit der Anzahl der schaltenden Modelle zunimmt. Daneben wurde die Anwendung von Uncertainty Decoding untersucht, da sich dadurch eine deutlich höhere Baseline ergab (vgl. Tab. 6.3). Das Uncertainty Decoding wird in den folgenden Untersuchungen durchgeführt, indem anstelle der Emissionsverteilung des HMMs die Likelihood in Gl. (6.8) bei der Decodierung der Sprachmerkmale verwendet wird. Die Fehlerrate mit dem SLDM-M16+UD beträgt ohne die Rückkopplung $33,0\%$. Mit den Rückkopplungsmethoden, die in diesem Kapitel eingeführt wurden, wurden unter Anwendung der Uncertainty-Decodierregel Fehlerraten von $32,0\%$ (FB1) und $32,4\%$ (FB2) erzielt (vgl. Tab. 8.4).

Weiterhin wurde die Berücksichtigung der Unsicherheit bei der Rückkopplung von Informationen in das Front-End auf der AURORA2 Datenbank untersucht. Wie in Abschnitt 8.6.1 ausgeführt wurde, können die exakten a posteriori Wahrscheinlichkeiten der HMM-Zustände mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene berechnet werden. Dieser Ansatz wird im Folgenden durch die Endung “state” bei der Angabe des Verfahrens gekennzeichnet. Der akustische Skalierungsfaktor, der zur Gewichtung des Akustikmodells gegenüber dem Sprachmodell verwendet wird (siehe Abschnitt 6.2), wurde experimentell aus den Sprachdaten aus Test-Set A ermittelt,

während Test-Set B ausschließlich zur Evaluation verwendet wird. Um den optimalen Skalierungsfaktor experimentell zu bestimmen, wurden die a posteriori Wahrscheinlichkeiten der HMM-Zustände für verschiedene Skalierungsfaktoren S_α mit dem ersten Ansatz aus Abschnitt 8.6.2.3 berechnet, in dem für jede Wortgraphkante die beste Zustandsfolge ermittelt und in der zweiten Erkennungsstufe auf die Modellwahrscheinlichkeiten des SLDMs abgebildet wurde (SLDM-M16-FB1word).

In Tab. 8.5 ist die Erkennungsrate für verschiedene Werte von S_α angegeben.

S_α	$0,0025$	$0,005$	$0,01$	$0,02$	$0,04$	$0,08$
WAC	81,09	81,11	81,21	81,17	81,10	80,83

Tabelle 8.5: SLDM-M16-FB1word: Erkennungsrate (WAC) auf Test-Set A der AURORA2 Datenbank abhängig vom akustischen Skalierungsfaktor S_α

Die Erkennungsrate weist ein Maximum bei $S_\alpha \approx 0,01$ auf, ist jedoch in einem großen Bereich von S_α näherungsweise konstant. Tab. 8.7 und Tab. 8.8 zeigen, dass die Ergebnisse durch die Rückkopplung der a posteriori Wahrscheinlichkeiten der HMM-Zustände anstelle der besten Zustandsfolge deutlich verbessert werden können. Durch die exakte Berechnung der a posteriori Wahrscheinlichkeiten der HMM-Zustände mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene wurden auf Test-Set A und B der AURORA2 Datenbank Erkennungsraten von $83,25\%$ und $82,44\%$ mit der Rückkopplungsmethode FB1 erreicht (Tab. 8.7). Mit der Rückkopplungsmethode FB2 ergaben sich auf Test-Set A und B Erkennungsraten von $83,57\%$ und $82,72\%$ (Tab. 8.8). Auf Test-Set A ergab sich, wie in Tab. 8.9 angegeben, durch die Anwendung von Uncertainty Decoding bei der Ermittlung der Zustandswahrscheinlichkeiten und der Decodierung der optimalen Wortsequenz ein Anstieg der Erkennungsrate auf $83,85\%$ für die erste Rückkopplungsmethode (SLDM-M16-FB1state+UD) und auf $84,10\%$ für die zweite Rückkopplungsmethode (SLDM-M16-FB2state+UD). Die Durchführung einer dritten Erkennungsstufe für das SLDM-M16-FB2state+UD, nachdem in den ersten beiden Erkennungsstufen jeweils die a posteriori Wahrscheinlichkeiten der HMM-Zustände mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene berechnet wurden, führte zu einem weiteren Anstieg der Erkennungsrate für einige Sub-Sets, so dass sich, wie aus Tab. 8.10 zu entnehmen ist, mit Uncertainty-Decoding insgesamt Erkennungsraten von $84,62\%$ und $83,52\%$ auf Test Set A und B ergaben (SLDM-M16-FB2state-it+UD).

Im Folgenden wird die Anwendung des in Abschnitt 8.6.2 beschriebenen Vorwärts-Rückwärts-Algorithmus auf Wortebene zur Beschleunigung und Speicherreduktion untersucht. Die beiden Möglichkeiten, den Wortgraphen bei der Berechnung der a posteriori Wahrscheinlichkeiten auszunutzen, werden im Folgenden durch die Endungen "word" und "state-cs" bei der Angabe des untersuchten Verfahrens gekennzeichnet. Die Endung "word" bedeutet, dass zunächst ein Vorwärts-Rückwärts-Algorithmus auf Wortebene angewendet wird, um die a posteriori Wahrscheinlichkeiten der Wörter zu berechnen, und diese anschließend zur Gewichtung der besten Zustandsfolge der einzelnen Wörter verwendet werden (siehe Abschnitt 8.6.2.3, Ansatz 1). In dem An-

satz "state-cs" wird der Wortgraph zur Einschränkung der möglichen Zustandsübergänge im Vorwärts-Rückwärts-Algorithmus auf Zustandsebene verwendet (siehe Abschnitt 8.6.2.3, Ansatz 2). Die Wortgraphfehlerrate (WGE) und Wortgraphdichte (WGD) der extrahierten Wortgraphen auf Test-Set A der AURORA2 Datenbank sind in Tab. 8.6 dargestellt. Die Sprachmerkmale wurden vor der Erzeugung des Wortgraphen mit dem SLDM-M16 entstört. Bei der Extraktion der Wortgraphen wurde kein Wortgraph-Pruning angewendet. Die Wortgraphfehlerrate beträgt, gemittelt über $0dB$ bis $20dB$, für Test-Set A der AURORA2 Datenbank ungefähr $3,09\%$ und liegt somit deutlich unter der Fehlerrate bei der Erkennung mit dem SLDM-M16.

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	0,00	0,00	0,00	0,00	0,00
20dB	1,13	0,57	0,79	0,40	0,72
15dB	1,26	1,27	0,77	0,55	0,96
10dB	1,85	2,53	0,94	1,15	1,62
5dB	3,80	5,59	2,07	2,75	3,55
0dB	8,45	12,75	6,44	6,68	8,58
-5dB	17,51	23,98	20,00	15,81	19,32
Ø	3,30	4,54	2,20	2,31	3,09

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	1,21	1,22	1,20	1,19	1,21
20dB	2,62	3,36	2,29	2,31	2,65
15dB	3,89	5,31	3,17	3,43	3,95
10dB	6,53	9,21	5,35	5,85	6,73
5dB	11,50	15,38	10,77	10,80	12,11
0dB	17,78	22,59	19,36	16,07	18,95
-5dB	20,13	21,10	20,25	18,34	19,96
Ø	8,46	11,17	8,19	7,69	8,88

Tabelle 8.6: Qualitätsmaße der Wortgraphen auf Test-Set A der AURORA2 Datenbank (SLDM-M16, kein Wortgraph-Pruning): a) WGE b) WGD

Die Erkennungsraten für die Wortgraph-Algorithmen sind in Tab. 8.11-Tab. 8.15 dargestellt. Mit dem ersten Wortgraph-Algorithmus, in dem die Viterbi-Approximation auf Wortebene durchgeführt wird, ergaben sich gegenüber dem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene deutliche Einbußen in der Erkennungsrate. Für die Rückkopplungsmethode FB1 wurden Erkennungsraten von $81,43\%$ und $80,57\%$ auf den beiden Test-Sets der AURORA2 Datenbank erzielt (Tab. 8.11). In Tab. 8.14 a) sind zum Vergleich die Ergebnisse für die Rückkopplung der Zustandswahrscheinlichkeiten mit einer N-Best-Liste auf Test-Set A der AURORA2 Datenbank dargestellt (SLDM-M16-FB1nbest). In diesem Ansatz wurden die HMM-Zustände der besten $N_{hyp} = 20$ Wortfolgen berechnet und mit den Likelihoods $p(\mathbf{y}_1^T|(w_1^{N(j)})^{(j)})^{S_\alpha}$ der Wortfolgen $(w_1^{N(j)})^{(j)}$, die mit einem HTK-Tool berechnet wurden, gewichtet. Dabei wurde angenommen, dass die a priori Wahrscheinlichkeiten $p((w_1^{N(j)})^{(j)})$ auf der AURORA2 Zifferndatenbank näherungsweise konstant sind. Dieser Ansatz führte gegenüber dem SLDM-FB1word zu keinem wesentlichen Unterschied in der Erkennungsrate, was darauf zurückzuführen ist, dass auf der AURORA2 Datenbank mit $N_{hyp} = 20$ Wortfolgen eine hinreichende Überlagerung des Wortgraphen erzielt wird, um die a posteriori Wahrscheinlichkeiten der Wortfolgen zuverlässig ermitteln zu können. Mit der zweiten Rückkopplungsmethode ergaben sich für das SLDM-FB2word Erkennungsraten von $81,59\%$ und $80,66\%$ (Tab. 8.12). Auf Test-Set A wurden durch die Anwendung von Uncertainty Decoding Erkennungsraten von $82,27\%$ und $82,75\%$ für die Rückkopplungsmethoden FB1 und FB2 erzielt (Tab. 8.13).

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,49	99,69	99,61
20dB	98,62	97,16	99,08	98,73	98,40
15dB	96,81	93,02	98,66	96,76	96,31
10dB	91,43	84,22	96,12	91,18	90,74
5dB	80,26	66,29	87,71	80,50	78,69
0dB	54,38	36,03	62,06	56,00	52,12
-5dB	24,69	13,69	22,64	26,47	21,87
Ø	84,30	75,34	88,73	84,63	83,25

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	99,66	99,61	99,49	99,69	99,61
20dB	97,64	98,55	98,09	98,55	98,21
15dB	93,40	96,86	96,39	96,88	95,88
10dB	85,60	90,30	89,02	94,05	89,74
5dB	70,56	76,75	75,84	84,29	76,86
0dB	44,34	50,48	51,57	59,73	51,53
-5dB	21,49	18,62	21,71	25,58	21,85
Ø	78,31	82,59	82,18	86,70	82,44

Tabelle 8.7: SLDM-M16-FB1state: Erkennungsrate auf Test-Set A und B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,51	99,67	99,61
20dB	98,80	97,25	99,16	98,43	98,41
15dB	97,02	92,71	98,72	96,85	96,33
10dB	92,57	83,83	96,63	92,69	91,43
5dB	81,21	67,02	88,34	80,84	79,35
0dB	55,76	36,40	62,15	55,04	52,34
-5dB	24,78	14,15	24,13	25,89	22,24
Ø	85,07	75,44	89,00	84,77	83,57

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	99,75	99,55	99,52	99,69	99,63
20dB	97,33	98,85	97,70	98,43	98,08
15dB	92,48	96,98	95,91	96,91	95,57
10dB	85,57	90,54	89,14	93,74	89,75
5dB	70,77	77,51	77,57	84,79	77,66
0dB	45,38	52,12	52,40	60,26	52,54
-5dB	20,51	21,67	22,25	27,24	22,92
Ø	78,31	83,20	82,54	86,83	82,72

Tabelle 8.8: SLDM-M16-FB2state: Erkennungsrate auf Test-Set A und B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

FB1	Sub.	Bab.	Car	Exh.	Ø
Clean	99,60	99,61	99,52	99,69	99,61
20dB	98,71	97,16	99,16	98,64	98,42
15dB	96,96	93,02	98,63	96,45	96,27
10dB	92,14	85,40	96,18	92,01	91,43
5dB	80,78	69,65	87,77	81,33	79,88
0dB	55,76	36,88	63,64	56,77	53,26
-5dB	25,24	11,31	23,11	27,28	21,74
Ø	84,87	76,42	89,08	85,04	83,85

FB2	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,51	99,67	99,61
20dB	98,71	97,25	99,16	98,26	98,35
15dB	97,18	92,71	98,69	96,91	96,37
10dB	92,82	83,83	96,75	92,66	91,52
5dB	82,84	67,17	88,99	82,47	80,37
0dB	57,84	36,40	64,48	56,96	53,92
-5dB	27,26	14,24	27,83	27,34	24,17
Ø	85,88	75,47	89,61	85,45	84,10

Tabelle 8.9: SLDM-M16-FB1/2state+UD: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,52	99,69	99,62
20dB	98,80	96,89	99,16	98,52	98,34
15dB	97,45	92,71	98,75	96,91	96,46
10dB	93,40	85,37	96,84	92,84	92,11
5dB	82,59	70,41	89,17	82,72	81,22
0dB	56,92	41,11	64,39	57,54	54,99
-5dB	24,72	14,41	27,92	27,65	23,68
Ø	85,83	77,30	89,66	85,71	84,62

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	99,66	99,61	99,52	99,69	99,62
20dB	97,18	98,73	97,67	98,25	97,96
15dB	92,85	96,98	96,27	97,35	95,86
10dB	86,43	91,72	90,07	94,42	90,66
5dB	73,20	78,99	79,27	85,99	79,36
0dB	47,74	51,69	55,32	60,35	53,78
-5dB	19,80	20,68	23,29	25,33	22,28
Ø	79,48	83,62	83,72	87,27	83,52

Tabelle 8.10: SLDM-M16-FB2state-it+UD: Erkennungsrate auf Test-Set A und B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Sub.	Bab.	Car	Exh.	Ø
Clean	99,54	99,55	99,46	99,51	99,52
20dB	97,73	95,92	99,14	98,83	97,91
15dB	95,89	90,96	98,63	96,17	95,41
10dB	89,96	80,11	95,17	90,43	88,92
5dB	76,92	62,48	86,25	78,28	75,98
0dB	49,12	31,74	60,42	54,35	48,91
-5dB	21,22	12,21	19,32	27,65	20,10
Ø	81,92	72,24	87,92	83,61	81,43

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	99,54	99,51	99,46	99,57	99,52
20dB	96,19	97,99	97,14	98,30	97,41
15dB	90,94	95,86	94,54	96,70	94,51
10dB	81,79	88,36	86,52	93,34	87,50
5dB	66,29	75,24	73,04	82,60	74,29
0dB	42,55	47,34	49,06	57,70	49,16
-5dB	18,11	18,47	20,64	25,70	20,73
Ø	75,55	80,96	80,06	85,73	80,57

Tabelle 8.11: SLDM-M16-FB1word: Erkennungsrate auf Test-Set A und Test-Set B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Set A	Res.	Str.	Tra.	Air.	Ø
Clean	99,63	99,55	99,46	99,57	99,55
20dB	98,37	95,95	99,14	98,18	97,91
15dB	96,44	90,42	98,81	96,45	89,69
10dB	91,03	80,20	95,97	91,55	89,69
5dB	77,34	63,24	86,52	78,06	76,29
0dB	49,92	32,44	58,48	53,38	48,56
-5dB	21,92	11,58	21,80	24,81	20,03
Ø	82,62	72,45	87,78	83,52	81,59

Set B	Res.	Str.	Tra.	Air.	Ø
Clean	99,63	99,58	99,46	99,57	99,56
20dB	95,82	98,58	96,99	98,21	97,40
15dB	89,90	96,13	94,45	96,42	94,23
10dB	80,96	88,81	86,76	92,50	87,26
5dB	66,44	75,36	73,78	82,57	74,54
0dB	41,11	50,70	49,00	58,78	49,90
-5dB	17,81	22,37	22,10	28,02	22,58
Ø	74,85	81,92	80,20	85,70	80,66

Tabelle 8.12: SLDM-M16-FB2word: Erkennungsrate auf Test-Set A und B der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

FB1	Sub.	Bab.	Car	Exh.	Ø
Clean	99,57	99,43	99,43	99,63	99,52
20dB	98,07	95,98	99,14	98,18	97,84
15dB	96,53	91,41	98,51	96,14	95,65
10dB	91,74	80,96	96,21	90,84	89,94
5dB	79,67	64,33	86,70	79,45	77,54
0dB	52,87	33,62	58,57	56,40	50,37
-5dB	21,61	9,82	13,24	24,90	17,39
Ø	83,78	73,26	87,83	84,20	82,27

FB2	Sub.	Bab.	Car	Exh.	Ø
Clean	99,63	99,37	99,43	99,63	99,52
20dB	98,00	95,65	99,14	98,06	97,71
15dB	96,50	90,54	98,78	96,36	95,55
10dB	91,86	79,90	96,48	91,92	90,04
5dB	80,93	64,06	88,82	80,13	78,49
0dB	54,53	33,98	61,91	57,54	51,99
-5dB	24,32	10,10	15,24	27,43	19,27
Ø	84,36	72,83	89,03	84,80	82,75

Tabelle 8.13: SLDM-M16-FB1/2word+UD: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

a)	Sub.	Bab.	Car	Exh.	Ø
Clean	99,57	99,69	99,55	99,60	99,60
20dB	97,97	96,52	99,14	98,80	98,11
15dB	96,22	90,75	98,48	96,51	95,49
10dB	90,64	79,93	95,29	91,33	89,30
5dB	76,60	61,79	85,42	79,36	75,79
0dB	49,83	31,74	60,42	57,74	49,93
-5dB	22,62	11,79	17,78	25,02	19,30
Ø	82,25	72,15	87,21	84,03	81,41

b)	Sub.	Bab.	Car	Exh.	Ø
Clean	99,57	99,69	99,55	99,29	99,53
20dB	97,67	96,25	98,54	98,36	97,71
15dB	95,39	92,90	97,38	95,93	95,40
10dB	89,81	84,52	93,29	91,82	89,86
5dB	77,68	69,50	83,33	80,96	77,87
0dB	49,46	38,18	53,95	54,37	48,99
-5dB	18,32	11,55	10,11	24,47	16,11
Ø	82,00	76,27	85,30	84,29	81,96

Tabelle 8.14: SLDM-M16-FB1nbest mit a) Rauschätzung aus Sprachrahmen am Anfang und Ende des Satzes b) adaptiver Rauschschätzung: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

	Sub.	Bab.	Car	Exh.	Ø
Clean	99,66	99,61	99,55	99,69	99,63
20dB	98,71	96,88	99,16	98,36	98,28
15dB	97,17	92,43	98,78	96,88	96,32
10dB	92,87	83,41	96,72	92,95	91,49
5dB	82,29	67,43	88,90	82,22	80,21
0dB	56,75	36,56	63,87	56,80	53,50
-5dB	25,42	14,19	24,94	26,93	22,87
Ø	85,56	75,34	89,49	85,44	83,96

Tabelle 8.15: SLDM-M16-FB2state-cs+UD: Erkennungsrate auf Test-Set A der AURORA2 Datenbank bei verschiedenen Umgebungsbedingungen

Mit dem zweiten Wortgraph-Algorithmus, in dem der Wortgraph dazu verwendet wird, die Übergangswahrscheinlichkeiten im Vorwärts-Rückwärts-Algorithmus auf Zustandsebene einzuschränken, wurde auf der AURORA2 Datenbank für $T_{wp} = 0$ und $N_{wp} = 10$ kein signifikanter Rückgang der Erkennungsrate gegenüber dem Vorwärts-Rückwärtsalgorithmus ohne Einschränkungen beobachtet (SLDM+FB1/2state-cs). Für das SLDM+FB2state-cs+UD (Tab. 8.15) wurde auf Test-Set A der AURORA2 Datenbank z.B. eine Erkennungsrate von $83,96\%$ verglichen mit $84,10\%$ ohne die Einschränkung erzielt.

Die Ergebnisse für die in Abschnitt 8.5 beschriebene Rauschschätzung sind in Tab. 8.14 für das SLDM-M16-FB1nbest dargestellt. Durch die adaptive Rauschschätzung wurde eine Erkennungsrate von $81,96\%$ (Tab. 8.14 a)) verglichen mit $81,41\%$ bei einer Rauschschätzung aus den ersten und letzten 10 Sätzen des Satzes erzielt (Tab. 8.14 b)). Dabei ergab sich jedoch eine starke Abhängigkeit von dem Hintergrundrauschen. Während für das instationäre Babble-Rauschen eine Verbesserung von $4,12$ Prozentpunkten erzielt wurde, wurde die Erkennungsrate für Car-Rauschen um $1,91$ Prozentpunkte verringert, während die Ergebnisse sich für Subway- und Exhibition-Rauschen nur leicht änderten.

Tab. 8.16 zeigt für die wichtigsten Verfahren, die in diesem Kapitel untersucht wurden, die Komplexität der Verarbeitung im Front-End und Back-End auf Test-Set A der AURORA2 Datenbank für ein SNR von $5dB$. Die angegebenen Laufzeiten sind auf die Laufzeit des SFEs normiert. In den durchgeführten Experimenten betrug der

Verfahren	Front-End	Back-End
SLDM-M16	523	34
SLDM-M16+FB	1046	68
SLDM-M16+FBword	1046	110
SLDM-M16+FBstate	1046	189

Tabelle 8.16: Laufzeiten der untersuchten Verfahren auf der AURORA2 Datenbank normiert auf die Laufzeit des SFEs

Rechenaufwand der Standarderkennung, die in Anhang A.1 spezifiert ist, ungefähr das 34-fache des Rechenaufwandes für die Merkmalsextraktion mit dem SFE (vgl. Kapitel 4). Der Rechenaufwand für die modellbasierte Merkmalsentstörung mit $M = 16$ schaltenden Modellen (SLDM-M16) aus [DA04] lag, wie in Kapitel 4 angegeben, deutlich über dem Aufwand für die Erkennung. Die Rückkopplungsmethoden SLDM-M16+FB1 und SLDM-M16-FB2, die in diesem Kapitel eingeführt wurden, führten näherungsweise zu einer Verdoppelung der Komplexität, da sowohl die Merkmalsentstörung als auch die Erkennung aufgrund der Iteration zweimal durchgeführt werden. Die Ersetzung der Viterbi-Approximation durch einen Vorwärts-Rückwärtsalgorithmus auf Wort- (SLDM-M16+FBword: Approximationen aus Abschnit 8.6.2.3) oder Zustandsebene (SLDM-M16+FBstate: Algorithmus aus Abschnit 8.6.2.3) führte ungefähr zu einer Verdoppelung bzw. Verdreifachung der Komplexität im Back-End. Dies ist von

Bedeutung, wenn sich beispielsweise aufgrund eines größeren Vokabulars oder der Verwendung komplexerer Akustikmodelle die Komplexität des Back-Ends im Vergleich zu der Gesamtkomplexität des Erkenners erhöht.

Zusammenfassung

Die experimentellen Untersuchungen in diesem Abschnitt zeigen, dass mit den Rückkopplungsmethoden FB1 und FB2, die in diesem Kapitel eingeführt wurden, eine konsistente Erhöhung der Erkennungsrate erzielt werden kann. Die Rückkopplung einer einzelnen Zustandsfolge führte sowohl auf der AURORA2 Datenbank als auch auf der AURORA4 Datenbank bereits zu signifikanten Erkennungsgewinnen. Auf der AURORA2 Datenbank konnten die Erkennungsgewinne durch die Rückkopplung der a posteriori Wahrscheinlichkeiten der HMM-Zustände anstelle einer einzelnen Zustandsfolge deutlich verbessert werden. Eine weitere Verbesserung ergab sich auf beiden Datenbanken durch die Anwendung von Uncertainty Decoding. Insgesamt wurde durch die Kombination des mehrstufigen Erkennungsansatzes mit Uncertainty Decoding auf Test-Set A der AURORA2 Datenbank ein Anstieg der Erkennungsrate von $79,87\%$ auf $84,62\%$ und auf Test-Set B von $79,16\%$ auf $83,52\%$ erreicht. Auf der AURORA4 Datenbank wurde die Fehlerrate von $40,5\%$ auf $32,0\%$ reduziert. Dabei wurden auf der AURORA4 Datenbank größere Gewinne durch die Anwendung von Uncertainty Decoding ($40,5\% \rightarrow 33,0\%$) als durch die Rückkopplung ($40,5\% \rightarrow 36,6\%$) erzielt, während die Erkennungsgewinne auf der AURORA2 Datenbank zum größten Teil auf den mehrstufigen Erkennungsansatz zurückzuführen sind. Bei instationärem Rauschen konnten weiterhin Erkennungsgewinne dadurch erzielt werden, dass die a posteriori Wahrscheinlichkeiten der HMM-Zustände zur Berechnung einer Soft-VAD-Variable verwendet wurden, die bei der Rauschschätzung eingesetzt wurde. Die Laufzeit steigt durch die mehrstufige Verarbeitung näherungsweise proportional zu der Anzahl der Verarbeitungsstufen an.

Kapitel 9

Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde die Erweiterung eines HMM-basierten Spracherkennungsansatzes um die Modellierung statistischer Abhängigkeiten zwischen den Sprachmerkmalen verschiedener Sprachrahmen untersucht. Die optimale Wortsequenz wurde unter der Modellannahme in Abb. 6.1 analog zu [KF02] und [IHU08b] als Lösung eines Optimierungsproblems für verrauschte Sprachmerkmale ermittelt, wodurch sich ein konsistenter statistischer Rahmen für die Entstörung und Decodierung der verrauschten Sprachmerkmale ergab.

Zunächst wurde in Kapitel 4 die Merkmalsentstörung mit schaltenden Modellen im Front-End des Spracherkenners untersucht. Den Ausgangspunkt für die Untersuchungen stellte ein SLDM dar, das in [DA04] eingeführt wurde (siehe Anhang C.3). Im Vergleich zu einem ansonsten identischen GMM, in dem eine a priori Verteilung der Ordnung Null verwendet wird, ergab sich mit dem SLDM auf Test Set A und B der AURORA2 Datenbank ein Anstieg der Erkennungsrate von $76,40\%$ auf $79,87\%$ bzw. von $76,94\%$ auf $79,16\%$, während auf der AURORA4 Datenbank bei der verwendeten Abtastrate von $8kHz$ deutlich bessere Ergebnisse mit dem GMM erzielt wurden (SLDM: $40,5\%$, GMM: $36,1\%$). Wie sich in den weiteren Untersuchungen dieser Arbeit herausgestellt hat, können die Nachteile des SLDMs gegenüber dem GMM auf der AURORA4 Datenbank allerdings durch die Verwendung von Uncertainty Decoding kompensiert werden. In Kapitel 4 wurden verschiedene Alternativen untersucht, um die a posteriori Verteilung der Sprachmerkmale mit schaltenden Modellen zu ermitteln. Durch die Verwendung erweiterter Kalman Filter wurde gemittelt über alle Testdaten und die SNR-Pegel zwischen $0dB$ und $20dB$ auf der AURORA2 Datenbank ungefähr die gleiche Erkennungsrate wie mit den Filtern aus [DA04] erzielt, während auf der AURORA4 Datenbank das SLDM aus [DA04] zu höheren Erkennungsraten führte. Die Glättung der Sprachmerkmale und die Verwendung eines iterativen erweiterten Kalman Filters führte auf beiden Datenbanken jeweils zu einem Anstieg der Erkennungsrate. Außerdem wurde das iterative erweiterte Kalman Filter in Kapitel 4 um dynamische Sprachmerkmale erweitert, deren Dimension mittels einer PCA-Matrix reduziert wurde, um die Robustheit und Effizienz des Verfahrens zu erhöhen. Sowohl auf der AURORA2 Datenbank als auch auf der AURORA4 Datenbank ergaben sich dadurch im Vergleich zu den anderen Ansätzen auf der Basis schaltender Modelle, die in Kapitel 4 untersucht wurden, die insgesamt höchsten Erkennungsraten (AURORA2, Set A: $83,01\%$, AURORA2, Set B: $81,65\%$, Fehlerrate AURORA4: $36,6\%$).

Die Erkennungsergebnisse des AFE wurden jedoch auf beiden Testdatenbanken mit den in Kapitel 4 untersuchten schaltenden Modellen bei Weitem nicht erreicht. Außerdem erfordert der Einsatz schaltender Modelle einen höheren Rechenaufwand als die Verwendung des AFE. Wie in Kapitel 3 ausgeführt wird, wurden die schaltenden Modelle trotz der genannten Nachteile gegenüber dem AFE als Ausgangspunkt für die weiteren Untersuchungen verwendet, da sie die Modellierung von Inter-Frame Korrelationen ermöglichen (vgl. Abschnitt 2.2.2) und Vorteile beim Austausch von Informationen zwischen der Merkmalsentstörung und der Sprachdecodierung aufweisen (vgl. Abschnitt 2.4).

Als mögliche Schwachpunkte der modellbasierten Merkmalsentstörung wurden in Kapitel 5 neben der Nichtlinearität des Beobachtungsmodells die Modellierung des Rauschens und die Schätzung der Rauschparameter identifiziert. Die experimentellen Untersuchungen in Abschnitt 5.3 haben gezeigt, dass die Annahme stationären Rauschens, die in Kapitel 4 getroffen wurde, auch bei einer Parameterschätzung aus den wahren Rauschwerten bei der Merkmalsentstörung mit dem SLDM zu großen Fehlerraten auf den verwendeten Testdatenbanken führt. Aus diesem Grund wurde in Kapitel 5 ein dynamisches Rauschmodell eingeführt. Im Gegensatz zu anderen Literaturansätzen wurde in diesem Ansatz die Dynamik einer versteckten Zustandsvariable mit einem Zustandsmodell beschrieben, während die verbleibende Unsicherheit auf ein Beobachtungsmodell zurückgeführt wurde. Zur Schätzung der Rauschparameter wurden EM-Algorithmen hergeleitet. Die Varianz des Zustandsmodells wurde mit einem blockweisen EM-Algorithmus aus charakteristischen Trainingsdaten bestimmt (Abschnitt 5.1), da die Schätzung zur Laufzeit sich aufgrund der überlagerten Sprache als unzuverlässig herausgestellt hat. Um eine Adaption an veränderte Umgebungsbedingungen zu erreichen, wurde in Abschnitt 5.2 ein sequentieller EM-Algorithmus zur Schätzung der Beobachtungsvarianz hergeleitet. Die Modifikation der Rauschschätzung führte auf der AURORA4 Datenbank zu einer Reduktion der Fehlerrate des EKF-d-M16 mit nicht-iterierten erweiterten Kalman Filtern von $37,6\%$ auf $35,5\%$. Auf der AURORA2 Datenbank wurde durch das neue Rauschmodell keine signifikante Verbesserung der Erkennungsrate erreicht, was wahrscheinlich auf die kürzere mittlere Satzlänge und die damit verbundene kürzere Adaptionszeit zurückzuführen ist.

In Kapitel 6 wurden die theoretischen Grundlagen für die folgenden beiden Kapitel gelegt, indem das eingangs erwähnte Optimierungsproblem für verrauschte Eingangsdaten [KF02] unter der Modellannahme in Abb. 6.1 betrachtet wurde. Die Vereinfachung des Optimierungsproblems führte auf eine Uncertainty-Decodierregel, die in [IHU08b] bei der Kompensation von Paketverlusten eingesetzt wird, wobei sich durch die Modellerweiterung ein Ansatz für ein segmentielles HMM, das in Kapitel 7 eingeführt wurde, ergab. In Abschnitt 6.3 wurden experimentelle Untersuchungen zu den betrachteten Uncertainty-Decodierregeln durchgeführt. Durch die Anwendung von Uncertainty Decoding wurde sowohl auf der AURORA2 Datenbank als auch auf der AURORA4 Datenbank bei der Verwendung unterschiedlicher a posteriori Verteilungen der Sprachmerkmale (GMM: $p(\mathbf{x}_t|\mathbf{y}_t)$, SLDM: $p(\mathbf{x}_t|\mathbf{y}_1^t)$, SLDM-S: $p(\mathbf{x}_t|\mathbf{y}_1^T)$) eine Erhöhung der Erkennungsrate erzielt. Auf der AURORA4 Datenbank waren die Gewinne bei einer

a priori Verteilung erster Ordnung höher als bei einer a priori Verteilung der Ordnung Null, so dass mit dem IEKF-d-M16-UD (32, 5%) bessere Erkennungsergebnisse als mit dem SLDM-M16 (33, 0%) und dem GMM-M16 (33, 2%) erzielt wurden. Durch die Kombination der dynamischen Rauschschätzung, die in Kapitel 5 eingeführt wurde, mit Uncertainty Decoding (EKF-d-M16-dyn-UD) ergab sich eine Fehlerrate von 31, 8%.

Das segmentielle HMM, das in Kapitel 7 eingeführt wurde, basiert auf der Approximation der Emissionsverteilung $p(\mathbf{x}_t|\mathbf{x}_1^{t-1}, q_t)$ des HMMs auf der Ebene der Mischungsgewichte. Durch eine geeignete Suchstrategie, in der der Viterbi-Algorithmus um ein diskretes Filter zur Aktualisierung der Mischungsgewichte erweitert wurde, konnte ein signifikanter Anstieg der Rechenzeit verhindert werden. Der Speicheraufwand wurde durch die Verwendung linearer Zustandsmodelle im Merkmalsraum anstelle der Übergangswahrscheinlichkeiten zwischen den Mischungsgewichten reduziert. Auf der AURORA2 Datenbank wurde mit dem segmentiellen HMM gegenüber einem HMM mit vergleichsweisen Rechenaufwand und um den Speicherplatz für die Zustandsübergangstabelle erhöhten Rechenaufwand ein konsistenter Anstieg der Erkennungsrate von 79, 87% auf 81, 24% auf Test-Set A bzw. von 79, 16% auf 79, 67% auf Test-Set B erzielt. Die Anwendung von Uncertainty-Decoding führte für das segmentielle HMM zu keinem signifikanten Anstieg der Erkennungsrate (Test-Set A: 81, 50%, Test-Set B: 79, 77%).

Die Ausnutzung von Informationen aus den komplexen Sprach- und Akustikmodellen des Spracherkenners bei der Merkmalsentstörung wurde in Kapitel 8 untersucht. Dazu wurden ausgehend von dem SLDM in Abb. 4.2 zusätzlich die Zustände des statistischen Modells in Abb. 6.1 berücksichtigt, wodurch sich das Modell in Abb. 8.3 ergab. Auf der Grundlage dieses Modells wurden zwei Rückkopplungsstrukturen hergeleitet, die es ermöglichen, in einer mehrstufigen Verarbeitung jeweils bei der Merkmalsentstörung die Erkennungsergebnisse aus der vorangehenden Stufe zu berücksichtigen. In der ersten Methode wurden die a posteriori Wahrscheinlichkeiten der HMM-Zustände verwendet, um die Modellwahrscheinlichkeiten des SLDMs zu bestimmen, während in der zweiten Methode unmittelbar die a posteriori Verteilung der Sprachmerkmale durch die Rückkopplung beeinflußt wurde. Beide Ansätze führten im Vergleich zu der Merkmalsentstörung mit schaltenden Modellen, die in Kapitel 4 untersucht wurde, auf der AURORA2 Datenbank und auf der AURORA4 Datenbank zu einer konsistenten Verbesserung der Erkennungsergebnisse. Sowohl auf der AURORA2 Datenbank als auch auf der AURORA4 Datenbank ergab sich durch die Rückkopplung der besten Zustandsfolge bereits ein signifikanter Anstieg der Erkennungsrate. Auf der AURORA2 Datenbank wurden diese Ergebnisse durch die Berechnung der a posteriori Wahrscheinlichkeiten der HMM-Zustände mit einem Vorwärts-Rückwärts-Algorithmus auf Zustandsebene sowie die Durchführung einer dritten Erkennungsstufe, bei der in der ersten und zweiten Erkennung jeweils die a posteriori Wahrscheinlichkeiten der HMM-Zustände berechnet wurden, deutlich verbessert. Durch die Verwendung eines Wortgraphen bei der Berechnung der Zustandswahrscheinlichkeiten konnten Speicher- und Rechenaufwand ohne signifikante Verluste in der Erkennungsrate deutlich reduziert werden. Dabei hat sich die Einschränkung der möglichen Zustände im Vorwärts-Rückwärts-Algorithmus

auf Zustandsebene als bessere Alternative gegenüber der Berechnung der besten Zustandsfolge für die einzelnen Wortgraphkanten herausgestellt. Insgesamt wurde durch die Kombination des mehrstufigen Erkennungsansatzes mit Uncertainty Decoding auf Test-Set A der AURORA2 Datenbank ein Anstieg der Erkennungsrate von $79,87\%$ auf $84,62\%$ und auf Test-Set B von $79,16\%$ auf $83,52\%$ erreicht. Auf der AURORA4 Datenbank wurde die Fehlerrate von $40,5\%$ auf $32,0\%$ reduziert. Dabei wurden auf der AURORA4 Datenbank größere Gewinne durch die Anwendung von Uncertainty Decoding ($40,5\% \rightarrow 33,0\%$) als durch die Rückkopplung ($40,5\% \rightarrow 36,6\%$) erzielt, während die Erkennungsgewinne auf der AURORA2 Datenbank zum größten Teil auf den mehrstufigen Erkennungsansatz zurückzuführen sind. Bei instationärem Rauschen konnten weiterhin Erkennungsgewinne dadurch erzielt werden, dass die a posteriori Wahrscheinlichkeiten der HMM-Zustände zur Berechnung einer Soft-VAD-Variable verwendet wurden, die bei der Rauschschätzung eingesetzt wurde.

Ein möglicher Ansatzpunkt für weiterführende Arbeiten ist die Merkmalsentstörung im Front-End des Spracherkenners. Zwei Schwachpunkte bei der Entrauschung der cepstralen Sprachmerkmale mit schaltenden, linearen Dynamikmodellen, die in Kapitel 4 beschrieben wird, sind das hochgradig nichtlineare Beobachtungsmodell, das zudem einen Phasenterm beinhaltet, sowie die Rauschschätzung. Die Untersuchungen in Kapitel 5 zeigen, dass das unbekannte Rauschen den größten Einfluß auf die Fehlerrate hat. Für das SLDM-M1 wurde die Erkennungsrate auf der AURORA2 Datenbank durch eine genauere Schätzung der Rauschparameter z.B. von $77,48\%$ auf $94,61\%$ erhöht. Allerdings verbleibt auch bei bekanntem Rauschen ein Restfehler von $5,39\%$, der auf der AURORA2 Datenbank vor allem auf das Beobachtungsmodell zurückgeführt werden kann.

Während sich die Vernachlässigung des Phasenterms im Beobachtungsmodell in [DDA02] und [SVhW05a] als unbedeutend erwiesen hat, erscheint vor allem die Nichtlinearität des Beobachtungsmodells als problematisch. Wie in Abschnitt 2.1.1 ausgeführt wurde, wurden in der Literatur verschiedene Ansätze untersucht, um dieses Problem zu lösen, wobei sich das Beobachungsmodell des iterativen erweiterten Kalman Filters als bessere Alternative gegenüber dem EKF und dem UKF erwiesen hat. Eine mögliche Alternative stellen Partikelfilter dar, deren Einsatz in Kombination mit einem Dynamikmodell für das Rauschen in der Literatur bereits untersucht wurde (vgl. Abschnitt 2.3). Wegen des hochdimensionalen Mermalsraumes und der starken Schwankungen der Sprachmerkmale erscheint dieser Ansatz, insbesondere in Kombination mit schaltenden Modellen, jedoch als wenig erfolgversprechend zur Filterung der Sprachmerkmale. Ein anderer Ansatzpunkt ist eine in [IHU08a] vorgeschlagene Quantisierung des Merkmalsraumes, die die Verwendung eines nichtlinearen Beobachtungsmodells ermöglicht, jedoch bei der Merkmalsentstörung bislang zu schlechteren Ergebnissen als das SLDM führte (vgl. Abschnitt 2.2.2).

Die Schätzung der Rauschparameter wurde in Kapitel 5 untersucht. Da das Rauschen unter den gegebenen Testbedingungen additiv hinzugefügt wurde, war die Modellierung von Faltungsrauschen nicht notwendig. Unter realen Testbedingungen ist es jedoch unter Umständen notwendig, das Faltungsrauschen mit einem konstanten Offset

im Beobachtungsmodell zu modellieren, der wie beispielsweise im ALGONQUIN-Verfahren [FDAK03] zusammen mit dem im Zeitbereich additiven Rauschen geschätzt werden kann. Die wesentlich schlechteren Erkennungsraten der modellbasierten Verfahren, die in Kapitel 4 untersucht werden, gegenüber dem AFE können wahrscheinlich teilweise auf die Annahme stationären Rauschens zurückgeführt werden. In Kapitel 5 wurden auf der AURORA4 Datenbank durch die Verwendung eines dynamischen Rauschmodells bereits höhere Erkennungsraten als unter der Annahme stationären Rauschens erzielt. Ein anderer möglicher Ansatz besteht in der Verwendung der zeitvarianten Rauschschätzung des AFEs, die, ähnlich wie in [SVhW06a] angegeben, ins Cepstrum transformiert werden kann. Es ist zu beachten, dass dieser Ansatz neben der Punktschätzung des Rauschens eine zuverlässige Schätzung der Rauschvarianz erfordert.

Eine Kombination des SLDMs mit dem AFE ist auch möglich, indem das SLDM als Postfilter konzipiert wird bzw. das Beobachtungsmodell des SLDMs um die mit dem AFE entstörten Sprachmerkmale erweitert wird.

Ein anderer Ansatzpunkt für weiterführende Untersuchungen ist das segmentielle HMM, das in Kapitel 7 beschrieben wird. Wie in Abschnitt 2.2.1 dargestellt wurde, ist in der Literatur bereits eine große Anzahl an Lösungsansätzen zu finden, in denen Inter-Frame Korrelationen bei der Decodierung der Sprachmerkmale im Back-End des Erkenners berücksichtigt werden, wodurch bislang allerdings noch keine substantiellen Verbesserungen gegenüber einem HMM mit dynamischen Sprachmerkmalen erzielt werden konnten. In Kapitel 7 wurden die Inter-Frame Korrelationen durch die Modellierung von Übergangswahrscheinlichkeiten zwischen den Mischungsgewichten approximiert. Eine andere mögliche Approximation ergibt sich aus der Faktorisierung

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, q_t) \approx \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|q_t)}{p(\mathbf{x}_t)} \tag{9.1}$$

in Gl. (6.6). In informalen Experimenten in dieser Arbeit wurde untersucht, $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ abhängig von q_t zu wählen und $\mathbf{x}_{t-1}$ durch die Punktschätzung $\hat{\mathbf{x}}_{t-1}$ des SLDMs im Front-End für den Merkmalsvektor des letzten Sprachrahmens zu approximieren, wodurch allerdings keine signifikante Erhöhung der Erkennungsrate erzielt werden konnte. Eine andere Möglichkeit besteht in der parallelen Durchführung einer SLDM-Erkennung mit der wortabhängige Likelihood $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ und einer HMM-Erkennung mit der Likelihood $p(\mathbf{x}_t|q_t)$ der HMM-Zustände, wobei die beste Wortfolge aufgrund des Produktes der Wahrscheinlichkeiten, die sich aus der SLDM- und der HMM-Erkennung ergeben, ermittelt werden kann.

Die Rückkopplung von Informationen zur Merkmalsentstörung, die in Kapitel 8 untersucht wird, könnte mit anderen Entstörungsansätzen als dem SLDM kombiniert werden. Möglich ist z.B. die Kombination mit dem VTS-Verfahren, mit dem auf der AURORA2 Datenbank allerdings eine deutlich niedrigere Erkennungsleistung als mit dem SLDM erzielt wird. Da im VTS-Verfahren die statistischen Abhängigkeiten zwischen aufeinander folgenden Sprachrahmen nicht berücksichtigt werden, erfordert dieser Ansatz auf der anderen Seite aber auch nicht die Approximation, die in Gl. (8.24) vorgenommen wird.

Neben den eingesetzten Algorithmen kann auch die Effizienz der Verfahren verbessert werden. Alle in dieser Arbeit untersuchten Online-Verfahren wurden mit einer Simulationssoftware für Sprachsignal- und Spracherkennungsforschung (SPARK: Speech Processing And Recognition Kit) in C++ entwickelt. Eine Laufzeitreduktion kann unter Umständen durch die Parallelisierung kritischer Programmteile erreicht werden.

Anhang A

Testdatenbanken und Konfigurationen des Spracherkenners

Die experimentellen Untersuchungen in der vorliegenden Arbeit wurden, soweit nicht anders angegeben, auf der AURORA2 Datenbank und der AURORA4 Datenbank durchgeführt. In diesem Anhang werden die beiden Testdatenbanken zusammen mit den jeweils verwendeten Konfigurationen des Spracherkennungssystems beschrieben.

A.1 AURORA2 Testdatenbank

Die AURORA2 Datenbank [PH00] wurde für eine sprecherunabhängige Zeichenerkennung mit kleinem Vokabular konzipiert. Sie basiert auf der TI-Digits Datenbank, einer Datenbank mit verbundenen Ziffern in amerikanischem Englisch. Das Vokabular der beiden Datenbanken umfaßt die 11 englischen Ziffern {oh, zero, one, ..., nine}. Die Daten der TI-Digits Datenbank wurden mit einer Abtastrate von $20kHz$ aufgenommen, die bei der Erzeugung der AURORA2 Datenbank auf $8kHz$ reduziert wurde. Die resultierenden $8kHz$-Daten der AURORA2 Datenbank wurden mit einer G.712 Filtercharakteristik gefiltert. Diese ist für einen Frequenzbereich bis $4kHz$ definiert und weist eine flache Charakteristik zwischen $300Hz$ und $3400Hz$ auf.

Die Trainingsdaten bestehen aus Sprachaufnahmen mit 8440 Sätzen von 55 männlichen und 55 weiblichen Sprechern. Im Rahmen dieser Arbeit wurden unverrauschte Trainingsdaten verwendet. Die Spracherkennung wurde auf Test-Set A und Test-Set B der AURORA2 Datenbank durchgeführt. Die Parameter wurden, soweit erforderlich, auf Test-Set A optimiert, während Test-Set B ausschließlich zur Evaluation verwendet wurde. Beide Test-Sets beinhalten jeweils 4004 Sätze von 52 männlichen und 52 weiblichen Sprechern, die in 4 Sub-Sets mit jeweils 1001 Sätzen und 3257 Wörtern unterteilt wurden. Den Sub-Sets wurden die in Tab. A.1 mit ihren englischen Bezeichnungen aufgeführten Rauscharten für die SNR-Level $-5dB$, $0dB$, $5dB$, $10dB$, $15dB$, $20dB$ und ∞(unverrauscht) überlagert, so dass sich insgesamt 56 Sub-Sets mit 28028 Sätzen je Test-Set ergaben.

Die Erkennung wurde, soweit nicht anders angegeben, mit einem Unigram-Sprachmodell durchgeführt, wobei für jede Ziffer ein akustisches Markovmodell mit Links-Rechts-Topologie trainiert wurde. Die HMMs der einzelnen Ziffern besitzen 16 emittierende Zustände mit zehn Gaußverteilungen je Zustand und diagonalen Kovarianzmatri-

Sub-Set	Test-Set A	Test-Set B
1	Subway	Restaurant
2	Babble	Street
3	Car	Airway
4	Exhibition	Train

Tabelle A.1: AURORA2 Datenbank: Rauschsorten in Test-Set A und Test-Set B

zen der Emissionsverteilungen. Daneben werden die Pausen am Anfang und Ende der verarbeiteten Sätze mit einem ergodischen SIL-HMM modelliert, das drei Zustände mit 16 Gaußverteilungen je Zustand aufweist. Die Pausen zwischen den Wörtern werden durch ein SP-HMM repräsentiert, das einen einzelnen emittierenden Zustand besitzt, der dem mittleren Zustand des SIL-Modells entspricht. Im Rahmen dieser Arbeit wurden Merkmalsvektoren mit 39 Komponenten bestehend aus den statischen cepstralen Komponenten $x^{(0)} \dots x^{(12)}$, wobei $x^{(0)}$ die Energiekomponente bezeichnet, und den dynamischen Komponenten erster und zweiter Ordnung für die Erkennung verwendet. Soweit nicht anders angegeben, wurden die experimentellen Ergebnisse über alle Sub-Sets eines Test-Sets und die SNR-Level $0dB - 20dB$ gemittelt, um den Mittelwert Ø der Erkennungsrate zu bestimmen.

A.2 AURORA4 Testdatenbank

Die AURORA4 Datenbank [PPPH04] basiert auf den in [Hir02] spezifizierten Nov'92 Evaluationsdaten des DARPA Wall Street Journal Task (WSJ0). Dieser beinhaltet unverrauschte Aufnahmen kontinuierlich gesprochener Sprache mit einem Vokabular von 5000 Wörtern. Die Daten des WSJ0 Task wurden mit einer Abtastrate von $16kHz$ aufgenommen. Die AURORA4 Datenbank enthält daneben eine unterabgetastete Version dieser Daten, in der die Abtastrate auf $8kHz$ reduziert wurde. Um die Kompatibilität mit dem Front-End, das für die Erkennung auf der AURORA2 Datenbank eingesetzt wurde, zu gewährleisten, wurden die Untersuchungen in der vorliegenden Arbeit mit der unterabgetasteten Version durchgeführt. Die $8kHz$-Daten wurden mit einer G.712 Filtercharakteristik gefiltert (siehe Anhang A.1). Die Trainingsdaten der AURORA4 Datenbank umfassen eine Aufnahmedauer von ca. $14h$ und beinhalten 7138 Sätze von 83 verschiedenen Sprechern. In den experimentellen Untersuchungen dieser Arbeit wurden ausschließlich unverrauschte Trainingsdaten verwendet.

Die Testdaten sind in 14 verschiedene Test-Sets unterteilt, wobei die Hälfte der Test-Sets mit einem Sennheiser HMD414 Mikrophon und die andere Hälfte mit 18 verschiedenen Mikrophonen aufgenommen wurden. In dieser Arbeit wurden die mit dem Sennheiser HMD414 Mikrophon erzeugten sieben Test-Sets eingesetzt. Das erste dieser Test-Sets enthält unverrauschte Sprachdaten, die wie beschrieben durch Unterabtastung und Filterung aus den Daten des WSJ Task erzeugt werden können. Den anderen Test-Sets wurden zufällig verschiedene Rauschsorten mit Signal-zu-Rauschverhältnissen

zwischen $5dB$ und $15dB$ in Schritten von $1dB$ überlagert. Die einzelnen Rauschsorten sind mit ihren englischen Bezeichnungen in Tab. A.2 aufgeführt.

Test-Set	Rauschen
1	Clean
2	Car
3	Babble
4	Restaurant
5	Street
6	Airport
7	Train

Tabelle A.2: AURORA4 Datenbank: Verwendete Rauschsorten

Die Test-Sets umfassen jeweils eine Aufnahmedauer von ungefähr $40min.$ und beinhalteten ursprünglich 330 Sätze von acht Sprechern. Davon werden in der vorliegenden Arbeit jeweils 166 Sätze mit 2715 Wörtern verwendet, die in [Hir02] als offizieller reduzierter Test-Set ("Official AURORA4 Selection Test Set") definiert sind. Zur Erkennung wurden in den durchgeführten Untersuchungen ein Bigram-Sprachmodell mit 5000 Wörtern und ein akustisches Modell mit ungefähr 3240 HMM-Zuständen, 10 Mischungsverteilungen je HMM-Zustand und diagonalen Kovarianzmatrizen eingesetzt. Genauso wie für die experimentellen Untersuchungen auf der AURORA2 Datenbank wurden Merkmalsvektoren mit 39 Komponenten, bestehend aus den statischen und dynamischen cepstralen Komponenten erster und zweiter Ordnung, verwendet. Um den Mittelwert Ø der Fehlerrate zu bestimmen, wurden die Ergebnisse der experimentellen Untersuchungen über die verwendeten sieben Test-Sets gemittelt.

Anhang B

Qualitätsmaße

Die Ergebnisse der Spracherkennung werden, wie in der Literatur üblich, mit der Wortfehlerrate (WER) bzw. der Erkennungsrate (WAC) bewertet (siehe Anhang B.1). Um Aussagen über die Qualität eines Wortgraphen zu ermöglichen, können die Fehlerrate (WGE) und Dichte (WGD) des Wortgraphen herangezogen werden (Anhang B.2).

B.1 Wortfehlerrate (WER)

Die Wortfehlerrate ist definiert als

$$WER = \frac{N_{ins} + N_{del} + N_{sub}}{N_{ref}} 100\%, \tag{B.1}$$

wobei N_{ins}, N_{del} und N_{sub} die Anzahl der eingefügten, gelöschten und ersetzten Wörter für die optimale Ausrichtung der erkannten Wortfolge w_1^N der Länge N an der Referenzwortfolge $w^{(ref)}{}_1^{N_{ref}}$ der Länge N_{ref} bezeichnet [YEH+02]. Die Erkennungsrate WAC beträgt

$$WAC = 100\% - WER. \tag{B.2}$$

Der Unterschied zwischen zwei Erkennungsergebnissen WER_1 und $WER_2 < WER_1$ ist bei einem Konfidenzintervall von 95%, wie in [Man84] angegeben, statistisch signifikant, wenn gilt:

$$WER_2 + 1.96\sigma_{WER_2} < WER_1 - 1.96\sigma_{WER_1} \tag{B.3}$$

mit der Standardabweichung

$$\sigma_{WER_i} = \sqrt{\frac{WER_i(100\% - WER_i)}{N_i}}, \quad i \in \{1, 2\}, \tag{B.4}$$

wobei N_i die Anzahl der erkannten Wörter mit Verfahren $i \in \{1, 2\}$ bezeichnet. Bei der Abschätzung wird vorausgesetzt, dass die erkannten Wörter statistisch unabängig sind und die gleiche Verteilung aufweisen.Die erste Annahme ist bei isolierten Ziffern, d.h. für die AURORA2 Datenbank, näherungsweise erfüllt, während sie bei Vorhandensein eines Sprachmodells, d.h. für die AURORA4 Datenbank, weniger genau ist. Die zweite Annahme ist in der Regel näherungsweise erfüllt, wenn die Anzahl der Einfügungsfehler in der Größenordnung der Löschungsfehler liegt. Die angegebene Abschätzung stellt aufgrund der ungenauen Modellannahme in der Praxis somit nur einen Indikator für

die statistische Signifikanz einer Verbesserung dar. Genauere Konfidenzmaße wie die Varianzanalyse und der Matched-Pairs-Test [GC88] erfordern neben den Wortfehlerraten und der Anzahl der Wörter in der Regel jedoch zusätzliche Daten, die mit den HTK-Tools nicht explizit berechnet werden können.

B.2 Qualitätsmaße für die Bewertung von Wortgraphen

Die Wortgraphdichte (WGD) ist der Quotient

$$WGD = \frac{N_{edges}}{N_{ref}} \tag{B.5}$$

aus der Anzahl N_{edges} der Kanten im Wortgraph und der Länge N_{ref} der Referenzfolge.

Die Wortgraphfehlerrate (WGE) wird als Minimum der WER (siehe Anhang B.1) über alle im Wortgraphen enthaltenen Wortfolgen $w_1^{N_h,(h)}$ mit Länge N_h für $h = 1 \ldots N_{hyp}$ berechnet:

$$WGE = \min_h \left\{ \frac{N_{ins}^{(h)} + N_{del}^{(h)} + N_{sub}^{(h)}}{N_{ref}} 100\% \right\} \quad . \tag{B.6}$$

Die Implementierung von Gl. (B.6) wird in Anhang D.3 beschrieben.

Anhang C

Front-End

C.1 ETSI Standard Front-End (SFE)

Die Merkmalsextraktion wird in der vorliegenden Arbeit, soweit nicht anders angegeben, mit dem ETSI Standard Front-End ES 201 108 V1.1.2 [ETS00] ohne Merkmalsvektorquantisierung durchgeführt. Die Konformität mit dem cepstralen Beobachtungsmodell (siehe Abschnitt 4.1.0.3) wird dabei durch zwei in [DDA03b] eingeführte Modifikationen gewährleistet. Anstelle des Leistungsdichtespektrums, das im SFE dem Standard entsprechend berechnet wird, wird das Energiedichtespektrum ermittelt. Daneben wird die logarithmierte Energiekomponente, die im SFE aus den quadrierten Abtastwerten des Zeitsignals bestimmt wird, durch die Energiekomponente $x^{(0)}$ des cepstralen Merkmalsvektors ersetzt, da diese zur inversen DCT-Transformation der standardmäßig extrahierten cepstralen Komponenten $x^{(1)} \dots x^{(12)}$ benötigt wird.

Bei der Extraktion der Sprachmerkmale aus dem tiefpassgefilterten und digitalisierten Sprachsignal wird zunächst der DC-Offset entfernt und eine Pre-Emphasis durchgeführt. Anschließend werden überlappende Segmente der Länge $25ms$ mit einem Vorschub von $10ms$ mittels einer Hamming-Fensterung herausgeschnitten. Aus den extrahierten Sprachsignalausschnitten wird das Energiedichtespektrum mittels einer Fouriertransformation berechnet. Das Spektrum wird mit einer Mel-Filterbank mit 23 Kanälen an die Frequenzauflösung des menschlichen Gehörs und mit einer nachfolgenden Logarithmierung an die Lautstärkewahrnehmung angepaßt. Anschließend werden die log-spektralen Merkmale durch die Anwendung der inversen diskreten Cosinustransformation dekorreliert und die Anzahl der Koeffizienten reduziert.

Für die AURORA4 Datenbank wird, anders als in [ETS00] spezifiziert, eine Schwellwertbegrenzung der statischen Sprachmerkmale vorgenommen:

$$x^{(0)} \geq 0, \quad x^{(i)} \geq -40,\ i \in 1, \dots 12 \quad . \tag{C.1}$$

Falls eine Merkmalsentstörung durchgeführt wird, erfolgt die Schwellwertbegrenzung nach der Entstörung der Sprachmerkmale.

Aus den resultierenden 13 statischen MFCC-Koeffizienten werden vor der Erkennung, wie in Anhang C.2 beschrieben, zusätzlich 26 dynamische MFCC-Koeffizienten berechnet.

C.2 Dynamische Sprachmerkmale

In [YEH+02] ist die folgende Approximation für die Komponenten $\delta x_t^{(i)}$ und $\delta^2 x_t^{(i)}$, $i = 0, \ldots, 12$, der dynamischen Sprachmerkmale spezifiziert:

$$\begin{aligned} \delta x_t^{(i)} &= \frac{\sum_{l=1}^{L} l(x_{t+l}^{(i)} - x_{t-l}^{(i)})}{2\sum_{l=1}^{L} l^2} \approx \frac{\partial \mathbf{x}_t}{\partial t}, \quad \text{mit} \quad L = 3 \\ \delta^2 x_t^{(i)} &= \frac{\sum_{l=1}^{L_2} l(\delta x_{t+l}^{(i)} - \delta x_{t-l}^{(i)})}{2\sum_{l=1}^{L} l^2} \approx \frac{\partial^2 \mathbf{x}_t}{\partial t^2}, \quad \text{mit} \quad L_2 = 2 \end{aligned} \tag{C.2}$$

Die dynamischen Sprachmerkmale werden in dieser Arbeit, falls eine modellbasierte Merkmalsentstörung durchgeführt wird, aus den entstörten Sprachmerkmalen berechnet. Die Kovarianzmatrix $\mathbf{P}_{t|t}^{(x)}$ der cepstralen Sprachmerkmale $\mathbf{x}_t$, die z.B. für die Anwendung von Uncertainty Decoding berechnet werden muß, wird im Back-End als Diagonalmatrix mit den Elementen $p^{(i,i)}$ modelliert (vgl. Abschnitt 6.1). Bei der Berechnung der dynamischen Komponenten $\delta p^{(i,i)}$ und $\delta^2 p^{(i,i)}$ wird weiterhin die vereinfachende Annahme getroffen, dass die Merkmalsvektoren $\mathbf{x}_t$ in aufeinanderfolgenden Sprachrahmen unkorreliert sind, so dass sich die Beziehungen [IHU06]

$$\delta p^{(i,i)} = \frac{\sum_{l=1}^{L} l^2 (p_{t+l}^{(i,i)} - p_{t-l}^{(i,i)})}{4\left(\sum_{l=1}^{L} l^2\right)^2}, \quad \delta^2 p^{(i,i)} = \frac{\sum_{l=1}^{L_2} l^2 (\delta p_{t+l}^{(i,i)} - \delta p_{t-l}^{(i,i)})}{4\left(\sum_{l=1}^{L} l^2\right)^2} \tag{C.3}$$

ergeben.

C.3 Modellbasierte Ansätze zur Merkmalsentstörung

In diesem Abschnitt werden zwei modellbasierte Entstörungsansätze für die mit dem modifizierten SFE extrahierten Sprachmerkmale beschrieben. Beide Ansätze können in das SLDM, das in Kapitel 4 dargestellt wird, integriert werden.

C.3.1 VTS-Verfahren

In dem VTS-Verfahren von Moreno [Mor96] wird zunächst eine initiale Schätzung $(\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$ der Rauschparameter berechnet. Dazu werden z.B. sprachfreie Signalabschnitte am Anfang und Ende des Satzes verwendet. Diese Schätzung wird iterativ mit einem EM-Algorithmus verbessert, der auf einer Vektortaylorreihenentwicklung des Beobachtungsmodells in Gl. (4.15) basiert. Der Erwartungswert $\boldsymbol{\mu_n}^{(N_{it})}$ des Rauschens, den man nach N_{it} Iterationen des EM-Algorithmus erhält, wird bei der Filterung mit dem SLDM als Erwartungswert

$$\mathbf{n}_{t|1:t-1} = \boldsymbol{\mu_n}^{(N_{it})} \tag{C.4}$$

der a priori Verteilung des Rauschens verwendet. Die a priori Verteilung der Sprache wird im VTS-Verfahren als GMM modelliert, d.h. im SLDM wird beim Training der Modellparameter in Gl. (4.4) die Einschränkung

$$\mathbf{A}(s_t) = \mathbf{0} \tag{C.5}$$

vorgenommen, wodurch sich für die einzelnen Filter s_t jeweils eine gaußförmige a priori Verteilung mit den Momenten

$$\mathbf{x}_{t|1:t-1}(s_t) = \mathbf{b}(s_t), \quad \mathbf{P}_{t|1:t-1}(s_t) = \mathbf{C}(s_t) \tag{C.6}$$

ergibt. Die Meßgleichung (4.19) wird im VTS-Verfahren durch die approximative MMSE-Schätzung

$$\mathbf{x}_{t|1:t}(s_t) \approx \mathbf{y}_t - \mathbf{M}_{DCT} \log(1 + e^{\mathbf{M}_{DCT}^{-1}(\mathbf{n}_{t|1:t-1} - \mathbf{x}_{t|1:t-1}(s_t))}) \tag{C.7}$$

für die Erwartungswerte der a posteriori Verteilungen $p(\mathbf{x}_t|\mathbf{y}_1^t, s_t)$ des Subvektors für die Sprache vereinfacht. Die Erwartungswerte der einzelnen Filter s_t werden abschließend entsprechend Gl. (4.17) zu einem einzelnen MMSE-Schätzwert für die unverrauschten Sprachmerkmale kombiniert.

C.3.2 Iterative Verbesserung einer SNR-Variablen

Neben dem VTS-Verfahren wird in dieser Arbeit eine Methode eingesetzt, in der anstelle der Rauschparameter eine SNR-Variable

$$\mathbf{r}_t(s_t) = \mathbf{x}_t(s_t) - \mathbf{n}_t \tag{C.8}$$

iterativ verbessert wird [DA04]. In diesem Ansatz wird zunächst für jedes Filter s_t die gemeinsame a priori Verteilung von Sprache und Rauschen entsprechend Gl. (4.18) aus der a posteriori Verteilung des letzten Sprachrahmens berechnet, wobei für das Rauschen durch die Festlegung

$$\mathbf{D} = \mathbf{0}, \quad \mathbf{e} = \boldsymbol{\mu}_{\boldsymbol{n}}, \quad \mathbf{V} = \boldsymbol{\Sigma}_{\boldsymbol{n}} \tag{C.9}$$

in Gl. 4.6 eine a priori Verteilung der Ordnung Null verwendet werden kann. Die Iteration der SNR-Variable $\mathbf{r}_t(s_t)$ in den einzelnen Filtern s_t wird anders als der blockweise EM-Algorithmus im VTS-Verfahren für jeden einzelnen Sprachrahmen durchgeführt:

$$\begin{aligned}
(\mathbf{P}^{(r)}_{t|1:t})^{(l)}(s_t)^{-1} &= (\mathbf{G}_0^{(l-1)}(s_t) - \mathbf{I})'(\mathbf{P}^{(x)}_{t|1:t-1}(s_t))^{-1}(\mathbf{G}_0^{(l-1)}(s_t) - \mathbf{I}) \\
&\quad + \mathbf{G}_0^{(l-1)}(s_t)'(\mathbf{P}^{(n)}_{t|1:t-1})^{-1}\mathbf{G}_0^{(l-1)}(s_t) \\
\mathbf{r}^{(l)}_{t|1:t}(s_t) &= (\mathbf{P}^{(r)}_{t|1:t})^{(l)}(s_t)(\mathbf{G}_0(s_t)^{(l-1)} - \mathbf{I})'(\mathbf{P}^{(x)}_{t|1:t-1}(s_t))^{-1} \\
&\qquad (\mathbf{y}_t - \mathbf{g}_0^{(l-1)}(s_t) + \mathbf{G}_0^{(l-1)}(s_t)\mathbf{r}^{(l-1)}_{t|1:t}(s_t) - \mathbf{x}_{t|1:t-1}(s_t)) \\
&\quad + (\mathbf{P}^{(r)}_{t|1:t})^{(l)}(s_t)\mathbf{G}_0^{(l-1)}(s_t)'(\mathbf{P}^{(n)}_{t|1:t-1})^{-1} \\
&\qquad (\mathbf{y}_t - \mathbf{g}_0^{(l-1)}(s_t) + \mathbf{G}_0^{(l-1)}(s_t)\mathbf{r}^{(l-1)}_{t|1:t}(s_t) - \mathbf{n}_{t|1:t-1})).
\end{aligned} \tag{C.10}$$

Dabei bezeichnet $l = 1 \dots N_{it}$ den Iterationsindex. $\mathbf{g}_0^{(l-1)}(s_t)$ und $\mathbf{G}_0^{(l-1)}(s_t)$ sind die ersten beiden Terme in der Vektortaylorreihenentwicklung der Funktion

$$\mathbf{g}(\mathbf{r}_t) = \mathbf{M}_{DCT} \log\left(e^{\mathbf{M}_{DCT}^{-1}\mathbf{r}_t} + 1\right) \tag{C.11}$$

um den Entwicklungspunkt $\mathbf{r}_{t|1:t}^{(l-1)}(s_t)$:

$$\begin{aligned} \mathbf{g}_0^{(l-1)}(s_t) &= \mathbf{M}_{DCT} \log\left(e^{\mathbf{M}_{DCT}^{-1}\mathbf{r}_{t|1:t}^{(l-1)}(s_t)} + 1\right) \\ \mathbf{G}_0^{(l-1)}(s_t) &= \mathbf{M}_{DCT} \mathrm{diag}\left\{\frac{1}{1 + e^{-\mathbf{M}_{DCT}^{-1}\mathbf{r}_{t|1:t}^{(l-1)}(s_t)}}\right\} \mathbf{M}_{DCT}^{-1} \quad . \end{aligned} \tag{C.12}$$

Der Entwicklungspunkt $\mathbf{r}_{t|1:t}^{(l-1)}(s_t)$ der Taylorreihe ergibt sich aus der jeweils vorangehenden Iteration $l-1$ und wird mit

$$\mathbf{r}_{t|1:t}^{(0)}(s_t) = \mathbf{x}_{t|1:t-1}(s_t) - \mathbf{n}_{t|1:t-1} \tag{C.13}$$

initialisiert. Die MMSE-Schätzung, die im VTS-Verfahren in den einzelnen Filtern entsprechend Gl. (C.7) durchgeführt wird, wird in [DA04] als Funktion der SNR-Variable $\mathbf{r}_{t|1:t}^{(N_{it})}(s_t)$ der letzten Iteration geschrieben:

$$\mathbf{x}_{t|1:t}(s_t) \approx \mathbf{y}_t - \mathbf{M}_{DCT} \log\left(1 + e^{\mathbf{M}_{DCT}^{-1}\mathbf{r}_{t|1:t}^{(N_{it})}(s_t)}\right) + \mathbf{r}_{t|1:t}^{(N_{it})}(s_t). \tag{C.14}$$

Die Varianz $\mathbf{P}_{t|1:t}^{(x)}(s_t)$ des Merkmalsvektors $\mathbf{x}_t(s_t)$ wird in diesem Ansatz durch die Varianz der SNR-Variable $\mathbf{r}_t(s_t)$ approximiert:

$$\mathbf{P}_{t|1:t}^{(x)}(s_t) \approx \mathbf{P}_{t|1:t}^{(r)}(s_t). \tag{C.15}$$

Die Likelihood $p(\mathbf{y}_t|s_t)$, mit der die Zustandsschätzungen der Modelle s_t in Gl. (4.17) gewichtet werden, kann über die Beziehung

$$\begin{aligned} p(\mathbf{y}_t|s_t) = \mathcal{N}(&\mathbf{y}_t - \mathbf{g}_0^{(N_{it})}(s_t) + \mathbf{G}_0^{(N_{it})}(s_t)\mathbf{r}_{t|1:t}^{(N_{it})}(s_t); \\ &\mathbf{n}_{t|1:t-1} + \mathbf{G}_0^{(N_{it})}(s_t)(\mathbf{x}_{t|1:t-1}(s_t) - \mathbf{n}_{t|1:t-1}), \\ &\mathbf{G}_0^{(N_{it})}(s_t)\mathbf{P}_{t|1:t-1}^{(x)}(s_t)\mathbf{G}_0^{(N_{it})}(s_t) + (\mathbf{G}_0^{(N_{it})}(s_t) - \mathbf{I})\mathbf{P}_{t|1:t-1}^{(n)}(\mathbf{G}_0^{(N_{it})}(s_t) - \mathbf{I})') \end{aligned} \tag{C.16}$$

berechnet werden.

C.4 ETSI Advanced Front-End (AFE)

Das ETSI Advanced Front-End (AFE) [ETS05] ist aus einer Kooperation von Motorola, France Telecom und Alcatel hervorgegangen und wird aufgrund hoher Erkennungsraten in vielen aktuellen Veröffentlichungen als Referenzverfahren angegeben. Die Grundlage stellt ein 2-stufiges Wiener-Filter zur Rauschreduktion dar. Im AFE wird zunächst

eine Hanning-Fensterung der tiefpassgefilterten und digitalisierten Sprachdaten durchgeführt. Das Energiedichtespektrum wird genauso wie im SFE alle $10ms$ aus einem Fenster der Größe $25ms$ berechnet. Anschließend erfolgt eine Mittelwertbildung über jeweils zwei aufeinanderfolgende Frequenzbins, um die Varianz des Spektrums zu reduzieren. Das Rauschen wird im Spektrum mit Hilfe einer energiebasierten VAD in den Sprachpausen aktualisiert. Daneben werden im Spektrum die Parameter des Wiener-Filters berechnet. Die Impulsantwort des Wiener-Filters wird durch die Anwendung einer MEL-IDCT ermittelt und mit dem verrauschten Eingangssignal gefaltet. Das auf diese Weise entrauschte Sprachsignal dient als Eingangssignal für die zweite Phase der Wiener-Filterung. Diese unterscheidet sich von der ersten Phase in der Rauschschätzung, für die keine VAD mehr eingesetzt wird. Weiterhin wird in der zweiten Phase ein Verstärkungsfaktor vor der MEL IDCT verwendet, um das Ausmaß der Rauschunterdrückung zu kontrollieren. Am Ende der Rauschreduktion wird der Gleichanteil des Ausgangssignals entfernt. Das entrauschte Sprachsignal wird anschließend aufbereitet, indem Signalanteile mit großem SNR verstärkt und Anteile mit geringem SNR abgeschwächt werden. Im nächsten Schritt werden die MFCC-Koeffizienten aus dem entrauschten Sprachsignal berechnet. Dazu wird zunächst eine Vorverstärkung und eine Hamming-Fensterung durchgeführt. Danach wird das Energiedichtespektrum mit einer FFT berechnet und mittels einer MEL-Filterbank, einer Logarithmierung und einer DCT ins Cepstrum transformiert. Parallel wird aus dem entrauschten Sprachsignal auch ein Energiekoeffizient berechnet. Abschließend wird eine blinde Equalisierung der cepstralen Koeffizienten $x^{(0)} \dots x^{(12)}$ durchgeführt.

Anhang D

Implementierungsdetails

Der folgende Anhang beinhaltet Implementierungsdetails zur Berechnung der SLDM-Parameter (Anhang D.1), der numerisch stabilen Durchführung der Vorwärts-Rückwärts-Algorithmen (Anhang D.2) und der Berechnung der Wortgraphfehlerrate (Anhang D.3).

D.1 Initialisierung der SLDM-Parameter

Der EM-Algorithmus, der in Abschnitt 4.1.0.1 zur Schätzung der SLDM-Parameter eingeführt wurde, erfordert eine Initialisierung der Modellparameter. In dieser Arbeit wird zu diesem Zweck ein Splitting-Algorithmus angewendet, in dem die Parameter zunächst für $M = 1$ berechnet werden und M nach einer festgelegten Anzahl von Iterationen des EM-Algorithmus jeweils verdoppelt wird. Dabei ergeben sich aus den Modellen mit den Indizes $m = 0 \ldots M-1$ jeweils zwei neue Modelle mit den Indizes $2m$ und $2m+1$:

$$\begin{aligned} \mathbf{A}(2m+1) &= \mathbf{A}(2m) = \mathbf{A}(m) \\ \mathbf{b}(2m+1) &= \mathbf{b}(m) + \boldsymbol{\epsilon}, \quad \mathbf{b}(2m) = \mathbf{b}(m) - \boldsymbol{\epsilon} \\ \mathbf{C}(2m+1) &= \mathbf{C}(2m) = \gamma \mathbf{C}(m) \\ P(2m+1) &= P(2m) = P(m)/2 \quad . \end{aligned} \tag{D.1}$$

In Gl. (D.1) bezeichnet $\boldsymbol{\epsilon}$ einen konstanten Vektor und γ eine Konstante, für die $\gamma < 1$ gilt. Da einige Klassen für große M zu wenige Trainingsdaten enthalten können, werden Modelle, deren Modellwahrscheinlichkeit $P(m)$ unterhalb eines konstanten Schwellwertes T_{split} liegt, nicht weiter berücksichtigt. Stattdessen wird für jedes der auf diese Weise eliminierten Modelle das Modell mit der größten a priori Wahrscheinlichkeit $P(m) = \max_{\tilde{m}}\{P(\tilde{m})\}$ entsprechend Gl. (D.1) in zwei neue Modelle aufgeteilt.

D.2 Numerische Berechnungen

In den Vorwärts-Rückwärts-Algorithmen auf Zustands- und Wortebene werden, wie es bei der Implementierung des Viterbi-Algorithmus Standard ist, anstelle der Wahrscheinlichkeiten P die logarithmischen Scores $S = \log(P)$ gespeichert, um numerische

Probleme zu vermeiden. Dies hat genauso wie im Viterbi-Algorithmus die Auswirkung, dass anstelle eines Produktes von Wahrscheinlichkeiten P_i die Summe der Scores $S_i = \log(P_i)$ berechnet wird:

$$P = \prod_i P_i \quad \Leftrightarrow \quad S = \sum_i S_i. \tag{D.2}$$

In den Vorwärts-Rückwärts-Algorithmen ist zusätzlich die Berechnung der Summen von Wahrscheinlichkeiten erforderlich. Die numerisch stabile Berechnung dieser Summen erfolgt analog zu [Wes02], wo anstelle der Scores S die negativen Scores $-S$ gespeichert werden, über die Beziehung

$$P = \sum_i P_i \quad \Leftrightarrow \quad S = S_{\max} + \log \sum_i e^{S_i - S_{\max}} \quad \text{mit} \quad S_{\max} = \max_i\{S_i\}. \tag{D.3}$$

D.3 Berechnung der Wortgraphfehlerrate

Die Wortgraphfehlerrate (WGE) kann entsprechend Gl. (B.1) dadurch berechnet werden, dass für jede Wortsequenzhypothese, die im Wortgraphen enthalten ist, die Wortfehlerrate (WER) berechnet wird und die Minimierung anschließend über alle möglichen Wortfolgen durchgeführt wird. In der Praxis kann die Anzahl N_{hyp} der möglichen Wortfolgen, wie in Abschnitt 8.6.2.2 ausgeführt wurde, jedoch sehr groß sein, so dass die Minimierung im Folgenden direkt auf dem Wortgraphen durchgeführt wird ohne einzelne Wortsequenzen isoliert zu betrachten.

Den Ausgangspunkt stellt der Alignment-Algorithmus für eine einzelne Wortsequenz w_1^N und eine Referenzwortfolge ${w^{(ref)}}_1^{N_{ref}}$ dar [HU05a]. Der Alignment-Algorithmus wird zunächst in einer modifizierten Form dargestellt, die die elementare Erweiterung auf einen Wortgraphen erlaubt. Zur Ausrichtung der Wortfolge w_1^N wird eine Sequenz

$$P_L = h_1 \ldots h_l \ldots h_L = (w_1, w_1^{(ref)})_1 \ldots (w_i, w_j^{(ref)})_l \ldots (w_N, w_{N_{ref}}^{(ref)})_L \tag{D.4}$$

von Zuordnungen zwischen den Wörtern w_i aus w_1^N und den Wörtern $w_j^{(ref)}$ aus ${w^{(ref)}}_1^{N_{ref}}$ gesucht, so dass die Kosten für diese Zuordnungen minimal sind:

$$P_L^{opt} = \underset{P_L}{\operatorname{argmin}} \left\{ N_{error}(P_L) \right\}, \tag{D.5}$$

mit dem in Gl. (D.7) definierten Fehlermaß $N_{error}(P_L)$. Die Optimierung wird dabei auch über die Länge L des Pfades P_L durchgeführt. Dies wird in Gl. (D.5), wie in der Literatur üblich, weggelassen. Ein solcher Pfad im Alignment-Gitter ist in Abb. D.1 exemplarisch für die Wortsequenz (w_1, w_2, w_3) des rechts abgebildeten Wortgraphen und die Referenzwortfolge $(w_1^{(ref)}, w_2^{(ref)}, w_3^{(ref)}, w_4^{(ref)}, w_5^{(ref)})$ dargestellt (blaue Linie). Neben der in Gl. (D.4) bereits festgelegten Bedingung über die Anfangs- und Endzeitpunkte des Pfades wird die folgende Randbedingung bzgl. der Monotonie und Kontinuität des Pfades festgelegt:

$$h_l = (w_i, w_j^{(ref)})_l \Rightarrow h_{l-1} \in \{(w_{i-1}, w_j^{(ref)})_{l-1}, (w_{i-1}, w_{j-1}^{(ref)})_{l-1}, (w_i, w_{j-1}^{(ref)})_{l-1}\} \quad . \tag{D.6}$$

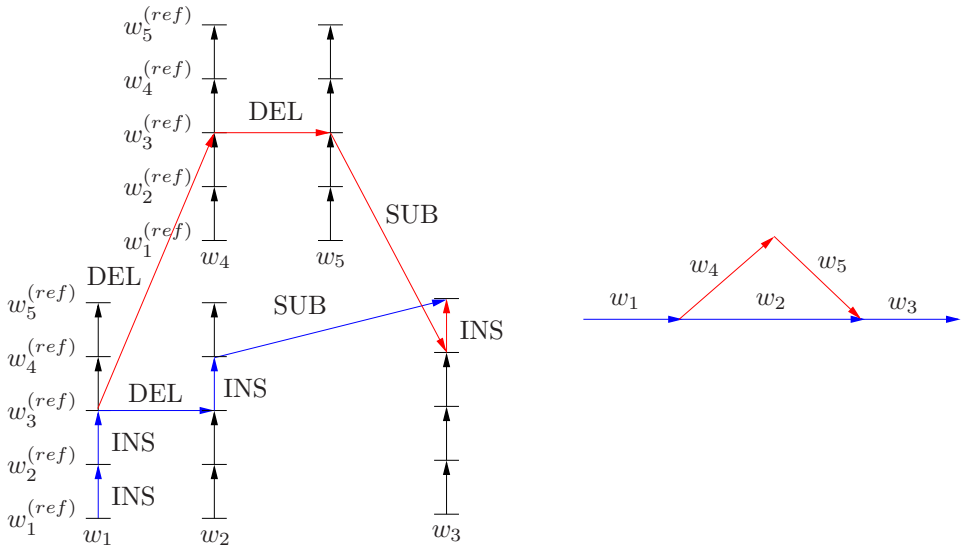

Abbildung D.1: *Verlauf zweier Pfade im Alignment-Gitter*

Für $i = 1$ bzw. $j = 1$ sind die Tupel mit w_{i-1} bzw. $w_{j-1}^{(ref)}$ nicht in der Menge der Vorgänger enthalten. Die drei möglichen Übergänge werden als Einfügung (INS), Ersetzung (SUB) und Löschung (DEL) bezeichnet. In Abb. D.1 sind diese Übergänge für den blauen Musterpfad eingezeichnet.

Die Kosten für eine Einfügung oder Löschung betragen jeweils eins. Bei einer Einfügung sind die Kosten für identische Werte von w_i und $w_j^{(ref)}$ null und für verschiedene Werte eins. Die Kosten eines vorgegebenen Pfades P_L können somit in der Form

$$N_{error}(P_L) = \sum_{l=1}^{L} b(h_l) a(h_{l-1}, h_l) \tag{D.7}$$

mit

$$b(h_l) = 1 \tag{D.8}$$

und

$$a(h_{l-1}, h_l) = \begin{cases} 1 & : \quad h_{l-1} \in \{(w_{i-1}, w_j^{(ref)})\} & \text{(DEL)} \\ 1 & : \quad h_{l-1} \in \{(w_i, w_{j-1}^{(ref)})\} & \text{(INS)} \\ 1 & : \quad h_{l-1} \in \{(w_{i-1}, w_{j-1}^{(ref)})\} \text{ und } w_i \neq w_j^{(ref)} & \text{(SUB)} \\ 0 & : \quad h_{l-1} \in \{(w_{i-1}, w_{j-1}^{(ref)})\} \text{ und } w_i = w_j^{(ref)} & \text{(SUB)} \end{cases} \tag{D.9}$$

geschrieben werden. Zur Lösung des Optimierungsproblems wird ein dynamischer Programmieransatz verwendet [Vit67]. Dabei wird jeder Kombination $h = (w_i, w_j^{(ref)})$ eines detektierten Wortes $w_i \in \{w_1 \dots w_i \dots w_N\}$ und der Referenzwortfolge $w_j^{(ref)} \in \{w_1^{(ref)} \dots w_j^{(ref)} \dots w_{N_{ref}}^{(ref)}\}$ die Fehleranzahl

$$N_{error}(h) = \min_{h'} \{N_{error}(h') + a(h', h)\} \tag{D.10}$$

mit

$$h' \in \{(w_{i-1}, w_j^{(ref)}), (w_{i-1}, w_{j-1}^{(ref)}), (w_i, w_{j-1}^{(ref)})\} \tag{D.11}$$

zugeordnet, wobei h' einen möglichen Vorgänger von h entsprechend Gl. (D.6) bezeichnet. Die Aktualisierung von $N_{error}(h)$ erfolgt iterativ, wobei $N_{error}(h)$ aktualisiert wird, sobald alle Vorgänger h' entsprechend Gl. (D.11) aktualisiert sind. Die Iteration beginnt mit $N_{error}(w_1, w_1^{(ref)}) = 0$ und endet, sobald $N_{error}(w_N, w_{N_{ref}}^{(ref)})$ aktualisiert ist, so dass sich die Wortfehlerrate

$$WER = \frac{N_{error}(w_N, w_{N_{ref}})}{N_{ref}} 100\% \tag{D.12}$$

aus Gl. (B.1) ergibt.

Erweitert man diese Darstellung auf die Berechnung der Wortgraphfehlerrate WGE, ergibt sich das Optimierungsproblem

$$P_L^{opt} = \min_h \left\{ \min_{P_L} \{ N_{error}(P_L^{(h)}) \} \right\} \tag{D.13}$$

mit

$$P_L^{(h)} = h_1^{(h)} \dots h_l^{(h)} \dots h_L^{(h)} = (w_1^{(h)}, w_1^{(ref)})_1 \dots (w_i^{(h)}, w_j^{(ref)})_l \dots (w_N^{(h)}, w_{N_{ref}}^{(ref)})_L. \tag{D.14}$$

Das bedeutet, dass die Minimierung in Gl. (D.10), wie in Abb. D.1 rot dargestellt, statt für einen einzelnen Vorgänger w_{i-1} des Wortes w_i für die Menge der $N_{WG-Arcs}$ Vorgängerknoten $w_{i-1}^{(1)} \dots w_{i-1}^{(N_{WG-Arcs})}$ von w_i im Wortgraphen vorgenommen werden muß. Dies wird dadurch erreicht, dass Gl. (D.11) durch

$$\begin{aligned} h' \in \{ & (w_{i-1}^{(1)}, w_j^{(ref)}), \dots, (w_{i-1}^{(N_{WG-Arcs})}, w_j^{(ref)}), \\ & (w_{i-1}^{(1)}, w_{j-1}^{(ref)}), \dots, (w_{i-1}^{(N_{WG-Arcs})}, w_{j-1}^{(ref)}), (w_i, w_{j-1}^{(ref)}) \} \end{aligned} \tag{D.15}$$

ersetzt wird.

Anhang E

Mathematische Herleitungen

E.1 Ergänzung zum IEKF

Wie im Folgenden gezeigt wird, ist die Iteration in Gl. (4.28) äquivalent zu einem Schritt des ALGONQUIN-Verfahrens [FDAK03]. Durch mehrmalige Anwendung des Matrixinversionslemmas [HU05b]

$$(\mathbf{RS})^{-1} = \mathbf{S}^{-1}\mathbf{R}^{-1} \tag{E.1}$$

erhält man für den Kalman-Gain

$$\begin{aligned}
\mathbf{K}_t^{(i)}(s_t) &= \left(\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)\prime}\right)\left(\mathbf{H}_z^{(i)}\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)\prime} + \mathbf{W}\right)^{-1} \\
&= \left(\mathbf{H}_z^{(i)}\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)\prime}\left(\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)\prime}\right)^{-1} + \mathbf{W}\left(\mathbf{P}_{t|1:t-1}(s_t)\mathbf{H}_z^{(i)\prime}\right)^{-1}\right)^{-1} \\
&= \left(\mathbf{H}_z^{(i)} + \left(\mathbf{W}\mathbf{H}_z^{(i)\prime-1}\right)\mathbf{P}_{t|1:t-1}(s_t)^{-1}\right)^{-1} \\
&= \left(\left(\mathbf{H}_z^{(i)\prime}\mathbf{W}^{-1}\right)\mathbf{H}_z^{(i)} + \mathbf{P}_{t|1:t-1}(s_t)^{-1}\right)^{-1}\left(\mathbf{H}_z^{(i)\prime}\mathbf{W}^{-1}\right).
\end{aligned} \tag{E.2}$$

Durch das Einsetzen der umgeformten Kalman-Verstärkung in die zweite Zeile in Gl. (4.28) ergibt sich die Formel für die Aktualisierung des Erwartungswertes aus [FDAK03]:

$$\mathbf{z}_{t|1:t}^{(i)}(s_t) = \mathbf{z}_{t|1:t}^{(i-1)}(s_t) + \left(\mathbf{P}_{t|1:t-1}(s_t)^{-1} + \mathbf{H}_z^{(i)\prime}\mathbf{W}^{-1}\mathbf{H}_z^{(i)}\right)^{-1}\mathbf{H}_z^{(i)\prime}\mathbf{W}^{-1}\left(\mathbf{y}_t - \mathbf{h}(\mathbf{z}_{t|1:t}^{(i-1)}(s_t))\right). \tag{E.3}$$

Aus der dritten Zeile in Gl. (4.28) erhält man durch die Substitution von

$$\mathbf{Z} = \mathbf{P}_{t|1:t}^{(i)}(s_t), \quad \mathbf{R} = \mathbf{P}_{t|1:t-1}(s_t), \quad \mathbf{S} = \mathbf{H}_z^{(i)\prime}, \quad \mathbf{Y} = \mathbf{W} \tag{E.4}$$

in dem Matrixinversionslemma [HU05b]

$$\begin{aligned}
\mathbf{Z} &= \mathbf{R} - \mathbf{RS}(\mathbf{S}'\mathbf{RS} + \mathbf{Y})^{-1}\mathbf{S}'\mathbf{R} \\
\mathbf{Z}^{-1} &= \mathbf{R}^{-1} + \mathbf{S}\mathbf{Y}^{-1}\mathbf{S}'
\end{aligned} \tag{E.5}$$

die Formel für die Aktualisierung der Varianzen im ALGONQUIN-Verfahren:

$$\mathbf{P}_{t|1:t}^{(i)}(s_t) = \left(\mathbf{P}_{t|1:t-1}(s_t)^{-1} + \mathbf{H}_z^{(i)\prime}\mathbf{W}^{-1}\mathbf{H}_z^{(i)}\right)^{-1}. \tag{E.6}$$

E.2 EM-Algorithmen zur Rauschschätzung

Im Folgenden werden Herleitungen und Ergänzungen zu Kapitel 5 angegeben. In diesem Zusammenhang bezeichet $\mathbf{M}$ eine Matrix mit den Komponenten $m^{(k,l)}$ und $\mathbf{S}$ eine symmetrische Matrix. $\mathbf{a}$, $\mathbf{b}$ und $\mathbf{v}$ bezeichnen Vektoren. $\mathbf{c}_k = [0 \dots 0 \quad 1 \quad 0 \dots 0]$ ist ein Spaltenvektor mit von Null verschiedener k-ter Komponente,

$$\delta_{kl} = \begin{cases} 1 & : \; l = k \\ 0 & : \; l \neq k \end{cases} \tag{E.7}$$

das Kronecker-Symbol und $\mathbf{I}^*_{kl}$ eine Matrix mit Einsen an den Positionen (k,l) und (l,k) sowie Nullen an allen anderen Positionen. Weiterhin gilt: $\mathbf{\Theta} = \mathbf{M}^{-1}\mathbf{S}\mathbf{M}^{-1}$ mit den Komponenten θ_{kl} von $\mathbf{\Theta}$. Für die folgenden Matrixoperationen ist in eckigen Klammern jeweils die Quelle angegeben:

$$\begin{aligned}
\mathbf{v}'\mathbf{M}^{-1}\mathbf{v} &= \mathrm{tr}\{\mathbf{M}^{-1}\mathbf{v}\mathbf{v}'\} \quad [\text{Fuk90}] && \text{(E.8)}\\
\mathbf{a}'\mathbf{M}^{-1}\mathbf{b} &= \mathrm{tr}\{\mathbf{a}'\mathbf{M}^{-1}\mathbf{b}\} = \mathrm{tr}\{\mathbf{M}^{-1}\mathbf{b}\mathbf{a}'\} \quad [\text{BSMM01}] && \text{(E.9)}\\
\frac{\partial \mathrm{tr}\{\mathbf{M}^{-1}\mathbf{S}\}}{\partial m^{(k,l)}} &= -[\theta_{kl} + \theta_{lk} - \delta_{kl}\theta_{kl}] \quad [\text{Fuk90}] && \text{(E.10)}\\
&= -\mathbf{c}_k'\Theta\mathbf{c}_l - \mathbf{c}_l'\Theta\mathbf{c}_k + \delta_{kl}\mathbf{c}_k'\Theta\mathbf{c}_l \quad . && \text{(E.11)}
\end{aligned}$$

Für symmetrische Matrizen $\mathbf{M}$ gilt:

$$\begin{aligned}
\mathbf{c}_k'\mathbf{\Theta}\mathbf{c}_l &= \mathbf{c}_l'\mathbf{\Theta}\mathbf{c}_k && \text{(E.12)}\\
\Rightarrow \frac{\partial \mathrm{tr}\{\mathbf{M}^{-1}\mathbf{S}\}}{\partial m^{(k,l)}} &= -(2-\delta_{kl})\mathbf{c}_k'\mathbf{\Theta}\mathbf{c}_l. && \text{(E.13)}\\
\frac{\partial \log|\mathbf{M}|}{\partial \mathbf{M}} &= [2\mathbf{M}^{-1} - \mathrm{diag}\{\mathbf{M}^{-1}\}] \quad [\text{Fuk90}] && \text{(E.14)}\\
\Rightarrow \frac{\partial \log|\mathbf{M}|}{\partial m^{(k,l)}} &= (2-\delta_{kl})\mathbf{c}_k'\mathbf{M}^{-1}\mathbf{c}_l && \text{(E.15)}\\
\frac{\partial \mathbf{M}^{-1}}{\partial m^{(k,l)}} &= -\mathbf{M}^{-1}\mathbf{I}^*_{kl}\mathbf{M}^{-1} \quad [\text{Fuk90}] && \text{(E.16)}\\
&= -\mathbf{M}^{-1}(\mathbf{c}_k\mathbf{c}_l' + \mathbf{c}_l\mathbf{c}_k' - \delta_{kl}\mathbf{c}_k\mathbf{c}_l')\mathbf{M}^{-1} && \text{(E.17)}\\
&= -(2-\delta_{kl})\mathbf{M}^{-1}\mathbf{c}_k\mathbf{c}_l'\mathbf{M}^{-1}. && \text{(E.18)}
\end{aligned}$$

E.2.1 Maximierung der Log-Likelihood einer Summe multivariater Gaußverteilungen

Der Erwartungswert der Log-Likelihood einer Summe multivariater Gaußverteilungen

$$p(\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^{N_c}|\mathbf{V}|}} e^{\Delta\mathbf{x}_t'\mathbf{V}^{-1}\Delta\mathbf{x}_t} \tag{E.19}$$

kann mit Gl. (E.8) und unter Ausnutzung der Linearität der beteiligten Operatoren als

$$\begin{aligned} Q(\mathbf{V}) = E[\log p(\mathbf{x}_1^T)] &= C - \frac{T}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{t=1}^{T} E[\Delta\mathbf{x}_t'\mathbf{V}^{-1}\Delta\mathbf{x}_t] \\ &= C - \frac{T}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{t=1}^{T} E[\mathrm{tr}(\mathbf{V}^{-1}\Delta\mathbf{x}_t\Delta\mathbf{x}_t')] \\ &= C - \frac{T}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\sum_{t=1}^{T} E[\Delta\mathbf{x}_t\Delta\mathbf{x}_t']) \\ &= C - \frac{T}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{D}) = C + \frac{1}{2}f(\mathbf{V}) \end{aligned} \tag{E.20}$$

mit $\mathbf{D} := \sum_{t=1}^{T} E[\Delta\mathbf{x}_t\Delta\mathbf{x}_t']$. geschrieben werden, wobei C einen konstanten Vorfaktor, $\Delta\mathbf{x}_t = \mathbf{x}_t - \boldsymbol{\mu}_x$, $t = 1 \ldots T$, die Abweichung gegebener Vektoren $\mathbf{x}_t$ der Dimension N_c von einem gegebenen Vektor $\boldsymbol{\mu}_x$ und $\mathbf{V}$ eine unbekannte, konstante Kovarianzmatrix bezeichnen. Entsprechend Lemma 3.2.3 aus [And84] beträgt der eindeutige ML-Schätzwert, der $f(\mathbf{V})$ und damit $Q(\mathbf{V})$ maximiert:

$$\hat{\mathbf{V}} = \frac{1}{T}\mathbf{D} = \frac{1}{T}\sum_{t=1}^{T} E[\Delta\mathbf{x}_t\Delta\mathbf{x}_t']. \tag{E.21}$$

E.2.2 Ergänzung zur Herleitung des sequentiellen EM-Algorithmus

Gegeben:

$$\begin{aligned} f(\mathbf{W}(\mathbf{W}^{(n)})) :=& R_t(\mathbf{W}^{(n)}) \\ =& \frac{N_c}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{W}(\mathbf{W}^{(n)})| \\ &+ \frac{1}{2}\mathrm{tr}\{\mathbf{W}(\mathbf{W}^{(n)})^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)})\} \end{aligned} \tag{E.22}$$

mit der konstanten Matrix $\mathbf{P}_{t|1:t}^{(y)}$ und der Dimension N_c des Vektors $\Delta\mathbf{y}_{t|1:t}$. Die Elemente $w^{(k,l)}$ der Matrix

$$\mathbf{W}(\mathbf{W}^{(n)}) = \mathbf{H}_n(\hat{s}_\tau)'\mathbf{W}^{(n)}\mathbf{H}_n(\hat{s}_\tau) + \mathbf{W}_c \tag{E.23}$$

können als Funktion der Diagonalmatrix $\mathbf{W}^{(n)}$ mit den Komponenten $w_n^{(i)}$ geschrieben werden (Gl. (5.32)). Im Folgenden werden die erste und zweite Ableitung der Funktion $R_t(\mathbf{W}^{(n)})$ berechnet.

1. Ableitung:

$$
\begin{aligned}
&\frac{\partial R_t(\mathbf{W}^{(n)})}{\partial w_n^{(i)}} \\
&= \sum_{k=1}^{N_c}\sum_{l=1}^{N_c}\frac{\partial f(\mathbf{W})}{\partial w^{(k,l)}}\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} && \text{(E.24)}\\
&= \frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}(2-\delta_{kl})(\mathbf{c}_k'\mathbf{W}^{-1}\mathbf{c}_l)\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} \\
&\quad -\frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}(2-\delta_{kl})\mathbf{c}_k'\mathbf{W}^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}'+\mathbf{P}_{t|1:t}^{(y)})\mathbf{W}^{-1}\mathbf{c}_l\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} && \text{(E.25)}\\
&= -\frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}(2-\delta_{kl})\mathbf{c}_k'\mathbf{W}^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}'+\mathbf{P}_{t|1:t}^{(y)}-\mathbf{W})\mathbf{W}^{-1}\mathbf{c}_l\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} && \text{(E.26)}
\end{aligned}
$$

Anmerkung: In Gl. (E.25) wurden die partiellen Ableitungen von $\mathrm{tr}\{\mathbf{W}(\mathbf{W}^{(n)})^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}'+\mathbf{P}_{t|1:t}^{(y)})\}$ mit Gl. (E.13) berechnet. Zur Berechnung der partiellen Ableitungen von $\log|\mathbf{W}(\mathbf{W}^{(n)})|$ wurde Gl. (E.15) angewendet.

2. Ableitung:

Durch die Berechnung der partiellen Ableitungen mit Gl. (E.9) und Gl. (E.13) bzw. die Anwendung der Produktregel erhält man aus den beiden Termen in Gl. (E.25):

$$
\begin{aligned}
&\frac{\partial^2 R_t(\mathbf{W}^{(n)})}{\partial w_n^{(i)}\partial w_n^{(j)}} \\
&= -\frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}\sum_{m=1}^{N_c}\sum_{n=1}^{N_c}(2-\delta_{kl})(2-\delta_{mn})\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}}\frac{\partial w^{(m,n)}}{\partial w_n^{(j)}}\mathbf{c}_m'\mathbf{W}^{-1}\mathbf{c}_k\mathbf{c}_l'\mathbf{W}^{-1}\mathbf{c}_n' \\
&\quad -\frac{1}{2}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c}\sum_{m=1}^{N_c}\sum_{n=1}^{N_c}(2-\delta_{kl})(2-\delta_{mn})\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}}\frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \\
&\quad \cdot\mathbf{c}_k'\left[\frac{\partial\mathbf{W}^{-1}}{\partial w_n^{(j)}}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}'+\mathbf{P}_{t|1:t}^{(y)})\mathbf{W}^{-1}\right. \\
&\qquad \left.+\mathbf{W}^{-1}(\Delta\mathbf{y}_{t|1:t}\Delta\mathbf{y}_{t|1:t}'+\mathbf{P}_{t|1:t}^{(y)})\frac{\partial\mathbf{W}^{-1}}{\partial w_n^{(j)}}\right]\mathbf{c}_l.
\end{aligned}
\qquad \text{(E.27)}
$$

In Gl. (E.27) wird ausgenutzt, dass die partielle Ableitung

$$
\frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} = h_n^{(k,i)}h_n^{(l,i)} \qquad \text{(E.28)}
$$

unabhängig von $\mathbf{W}^{(n)}$ ist (siehe Gl. (5.44)) und somit eine Konstante bzgl. der partiellen Ableitung nach $w_n^{(j)}$ ist. Durch Umformung von Gl. (E.27) ergibt sich:

$$\begin{aligned}
&\frac{\partial^2 R_t(\mathbf{W}^{(n)})}{\partial w_n^{(i)} \partial w_n^{(j)}} \\
= \; & -\frac{1}{2} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} (2 - \delta_{kl}) \frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} \\
& \cdot \Bigg[\sum_{m=1}^{N_c} \sum_{n=1}^{N_c} (2 - \delta_{mn}) \frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \mathbf{c}_k' \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} \mathbf{c}_l \\
& - \sum_{m=1}^{N_c} \sum_{n=1}^{N_c} (2 - \delta_{mn}) \frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \mathbf{c}_k' \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) \mathbf{W}^{-1} \mathbf{c}_l \\
& - \sum_{m=1}^{N_c} \sum_{m=1}^{N_c} (2 - \delta_{mn}) \frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \mathbf{c}_k' \mathbf{W}^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} \mathbf{c}_l \Bigg]
\end{aligned} \tag{E.29}$$

$$\begin{aligned}
= \; & -\frac{1}{2} \sum_{k=1}^{N_c} \sum_{l=1}^{N_c} \sum_{m=1}^{N_c} \sum_{n=1}^{N_c} (2 - \delta_{kl})(2 - \delta_{mn}) \frac{\partial w^{(k,l)}}{\partial w_n^{(i)}} \frac{\partial w^{(m,n)}}{\partial w_n^{(j)}} \\
& \cdot [\mathbf{c}_k' \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} \mathbf{c}_l \\
& - \mathbf{c}_k' \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) \mathbf{W}^{-1} \mathbf{c}_l \\
& - \mathbf{c}_k' \mathbf{W}^{-1} (\Delta \mathbf{y}_{t|1:t} \Delta \mathbf{y}_{t|1:t}' + \mathbf{P}_{t|1:t}^{(y)}) \mathbf{W}^{-1} \mathbf{c}_m \mathbf{c}_n' \mathbf{W}^{-1} \mathbf{c}_l]
\end{aligned} \tag{E.30}$$

Die partiellen Ableitungen in Gl. (E.29) wurden mit Gl. (E.18) berechnet.

E.3 Momente einer Funktion von Gaußverteilung

Im Folgenden bezeichnen $p_i(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i \in \{1, 2, 3\}$ multivariate Gaußverteilungen mit den Momenten $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$ der Dimension $N_c = \dim(\boldsymbol{\mu}_i)$. Gesucht sind die Momente $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$ der Verteilung

$$\begin{aligned}
p(\mathbf{x}_t) &= \frac{1}{\sqrt{(2\pi)^{N_c} |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu})} \propto \frac{p_1(\mathbf{x}_t) p_2(\mathbf{x}_t)}{p_3(\mathbf{x}_t)} \\
&= \sqrt{\frac{1}{(2\pi)^{N_c}} \frac{|\boldsymbol{\Sigma}_3|}{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} e^{-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_2) + \frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_3)' \boldsymbol{\Sigma}_3^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_3)}.
\end{aligned} \tag{E.31}$$

Da die Exponentialfunktion streng monoton steigend ist und der Vorfaktor unabhängig von $\mathbf{x}_t$ ist, können die Momente von $p(\mathbf{x}_t)$ eindeutig durch Koeffizientenvergleich im

Exponenten bestimmt werden:

$$\begin{aligned}
& \mathbf{x}_t'\boldsymbol{\Sigma}^{-1}\mathbf{x}_t - \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{x}_t - \mathbf{x}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
= \;& \mathbf{x}_t'\boldsymbol{\Sigma}_1^{-1}\mathbf{x}_t - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\mathbf{x}_t - \mathbf{x}_t'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\
+ \;& \mathbf{x}_t'\boldsymbol{\Sigma}_2^{-1}\mathbf{x}_t - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\mathbf{x}_t - \mathbf{x}_t'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 \\
- \;& \mathbf{x}_t'\boldsymbol{\Sigma}_3^{-1}\mathbf{x}_t + \boldsymbol{\mu}_3'\boldsymbol{\Sigma}_3^{-1}\mathbf{x}_t + \mathbf{x}_t'\boldsymbol{\Sigma}_3^{-1}\boldsymbol{\mu}_3 - \boldsymbol{\mu}_3'\boldsymbol{\Sigma}_3^{-1}\boldsymbol{\mu}_3 \\
\Rightarrow\; & \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_3^{-1} \\
& \boldsymbol{\mu} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_3^{-1}\boldsymbol{\mu}_3).
\end{aligned} \tag{E.32}$$

Anhang F

Symbolverzeichnis

$P()$	Wahrscheinlichkeit
$p()$	Wahrscheinlichkeitsdichte
$P_\alpha(), p_\alpha()$	Skalierte Wahrscheinlichkeiten (mit akustischem Skalierungsfaktor S_α)
$\mathbf{x}_1^T$	$\mathbf{x}_1 \dots \mathbf{x}_T$
$\mathbf{x}_{t\|1:\tau}$	Erwartungswert einer Zustandsschätzung für den Zustandsvektor $\mathbf{x}_t$ bei gegebenen Messungen $\mathbf{y}_1^\tau$
$\mathbf{P}_{t\|1:\tau}$	Kovarianzmatrix einer Zustandsschätzung für den Zustandsvektor $\mathbf{x}_t$ bei gegebenen Messungen $\mathbf{y}_1^\tau$
$x^{(c)}$	Komponente c des Vektors $\mathbf{x}$
$\mathbf{M}^{(k,l)}$	Komponente in Zeile k und Spalte l der Matrix $\mathbf{M}$
$\mathbf{M}'$	Transponierte der Matrix $\mathbf{M}$
$\mathbf{M}^+$	Pseudo-Inverse der Matrix $\mathbf{M}$
$\mathbf{0}$	Vektor bzw. Matrix mit Nullen
$\mathbf{1}$	Vektor bzw. Matrix mit Einsen
α	Gaußverteilte Zufallsvariable im Phasenterm des cepstralen Beobachtungsmodells
α_t	Vorwärtsvariable im Viterbi- bzw. Vorwärts-Rückwärts-Algorithmus
α_{ij}	Übergangswahrscheinlichkeiten im Vorwärts-Rückwärtsalgorithmus
$A(\tau_e; v, w)$	Akustische Likelihood der Wortgraphkante für Wort w, Endknoten τ_e und Vorgänger v
$\mathbf{A}$	Transitionsmatrix im Zustandsübergangsmodell für die Sprachmerkmale
$\tilde{\mathbf{A}}$	Transitionsmatrix im erweiterten Zustandsübergangsmodell
$\mathbf{A}_{\boldsymbol{\eta}}$	Transitionsmatrix im Zustandsübergangsmodell für $\boldsymbol{\eta}_t$
$\mathbf{A}_z$	Transitionsmatrix des gemeinsamen Zustandsübergangsmodells für Sprache und Rauschen
β_t	Rückwärtsvariable im Vorwärts-Rückwärts-Algorithmus
$\mathbf{b}$	Konstanter Offset im Zustandsübergangsmodell für die Sprachmerkmale
$\tilde{\mathbf{b}}$	Konstanter Offset im erweiterten Zustandsübergangsmodell
$\mathbf{b}_{\boldsymbol{\eta}}$	Konstanter Offset im Zustandsübergangsmodell für $\boldsymbol{\eta}_t$
$\mathbf{b}_z$	Konstanter Offset im Zustandsübergangsmodell für Sprache und Rauschen
B_t	Wortgrenze im Viterbi-Algorithmus

$\mathbf{C}$	Kovarianzmatrix des Zustandsübergangsmodells für die Sprachmerkmale
$\tilde{\mathbf{C}}$	Kovarianzmatrix des erweiterten Zustandsübergangsmodells für die Sprachmerkmale
$\mathbf{C}_{\boldsymbol{\eta}}$	Kovarianzmatrix des Zustandsübergangsmodells
$\mathbf{C}_z$	Kovarianzmatrix des Zustandsübergangsmodells für Sprache und Rauschen
d_x	Mittlerer quadratischer Prädiktionsfehler
$\mathbf{D}$	Transitionsmatrix im Zustandsübergangsmodell für das Rauschen
δ_{ij}	Kronecker-Symbol für die natürlichen Zahlen i und j
$\delta\mathbf{x}_t$	Dynamische Sprachmerkmale
$\delta^2\mathbf{x}_t$	Dynamische Sprachmerkmale zweiter Ordnung
$\Delta\mathbf{y}_t$	Innovationssequenz des Kalman-Filters
e	Wortgraphkante
$\mathbf{e}$	Konstanter Offset im Zustandsübergangsmodell für das Rauschen
$\boldsymbol{\epsilon}$	Kovarianzmatrix des Zustandsübergangsmodells für das Rauschen
$\boldsymbol{\eta}_t$	Zustandsvariable des erweiterten Rauschmodells $[\mathbf{n}_t \quad \mathbf{n}_{t-1}]'$
$f(k)$	Faltungsrauschen im Zeitbereich
$\mathbf{f}$	Additiver Offset im cepstralen Beobachtungsmodell durch Faltungsrauschen
$\mathbf{F}$	Kovarianzmatrix des Zustandsübergangsmodells für das Rauschen
γ_t	Vorwärts-Rückwärtsvariable
γ_t^m	Modellwahrscheinlichkeit des Modells m zum Zeitpunkt t
$\mathbf{h}(\mathbf{z}_t)$	Beobachtungsgleichung
$H(w,t)$	Likelihood des besten Pfades mit Endwort w und Endknoten t
$\mathbf{H}$	Jacobi-Matrix des Beobachtungsmodells für $\mathbf{z}_t$
$\tilde{\mathbf{H}}$	Jacobi-Matrix des Beobachtungsmodells für den erweiterten Zustandsvektor
$\mathbf{H}_{\boldsymbol{\eta}}$	Jacobi-Matrix des Beobachtungsmodells für $\boldsymbol{\eta}_t$
$\mathbf{H}_n$	Jacobi-Matrix des Beobachtungsmodells für $\mathbf{n}_t$
$\mathbf{H}_x$	Jacobi-Matrix des Beobachtungsmodells für $\mathbf{x}_t$
$I(q_0, q_\tau)$	Mittlere wechselseitige Information zwischen q_0 und q_τ
$\mathbf{I}$	Einheitsmatrix
k	Zeitbereichsindex
$\mathbf{K}$	Kalman-Verstärkung
m_t	Mischungskomponente des HMMs
M	Anzahl der schaltenden Modelle
$\mathbf{M}_{DCT}$	Matrix der diskreten Kosinustransformation
$\mathbf{M}_{PCA}$	PCA-Matrix
$n(k)$	Rauschen im Zeitbereich
$n^{(l)}$	Komponente des log-spektralen Rauschvektors
$\mathbf{n}_t$	Cepstrales Rauschen
$\tilde{\mathbf{n}}_t$	Beobachtetes cepstrales Rauschen
$\boldsymbol{\mu}_e$	Offset des Erwartungswertes in der Uncertainty Decodierregel

$\boldsymbol{\mu}_n$	Erwartungswert der a priori Verteilung des Rauschens
$\boldsymbol{\mu}_x$	Erwartungswert der a priori Verteilung der Sprachmerkmale
$\boldsymbol{\mu}_{\delta x}$	Erwartungswert der a priori Verteilung der dynamischen Sprachmerkmale erster Ordnung
$\boldsymbol{\mu}_{\delta^2 x}$	Erwartungswert der a priori Verteilung der dynamischen Sprachmerkmale zweiter Ordnung
N	Anzahl der Wörter in einer Wortsequenz
N_{it}	Anzahl der Iterationen im IEKF
N_{ar}	Ordnung eines autoregressiven Modells
N_c	Dimension des cepstralen Merkmalsvektors
N_l	Dimension des log-spektralen Merkmalsvektors
N_q	Anzahl der HMM-Zustände
N_τ	Anzahl der Knoten eines Wortgraphen
N_w	Anzahl der Wörter einer Datenbank
N_{wp}	Schwellwert für das Histogramm-Pruning der a posteriori Wahrscheinlichkeiten des Wortgraphen
$\emptyset$	Mittelwert
p_t	Phonem
$\mathbf{P}$	Kovarianzmatrix der Zustandsschätzung für $\mathbf{z}_t$
$\mathbf{P}^{(\eta)}$	Kovarianzmatrix der Zustandsschätzung für $\boldsymbol{\eta}_t$
$\mathbf{P}^{(n)}$	Kovarianzmatrix der Zustandsschätzung für $\mathbf{n}_t$
P_{sil}	Soft-VAD-Variable
$\mathbf{P}^{(x)}$	Kovarianzmatrix der Zustandsschätzung für $\mathbf{x}_t$
$\mathbf{P}^{(y)}$	Kovarianzmatrix einer Schätzung für $\mathbf{y}_t$
$\boldsymbol{\varphi}_\alpha$	Phasenterm im cepstralen Beobachtungsmodell
q_t	Versteckte, diskrete Zustandsvariable im Back-End des Spracherkenners (z.B. HMM-Variable)
$\hat{q}_1^T$	Optimale Zustandsfolge
Q_w^e	Endzustand des HMMs für das Wort w
Q	Kostenfunktion
Q_t	Kausale Kostenfunktion
Q_w^s	Startzustand des HMMs für das Wort w
$\mathbf{r}_t$	SNR-Variable
s_t	Versteckte, diskrete Zustandsvariable im Front-End des Spracherkenners (z.B. Modellvariable des SLDMs)
S_α	Akustischer Skalierungsfaktor
S_β	Sprachmodell-Skalierungsfaktor
$\boldsymbol{\Sigma}_e$	Zusätzliche Varianz in der Uncertainty Decodierregel
$\boldsymbol{\Sigma}_n$	Kovarianzmatrix der a priori Verteilung des Rauschens
$\boldsymbol{\Sigma}_x$	Kovarianzmatrix der a priori Verteilung der Sprache
$\tau(t; v, w)$	Startknoten der Wortgraphkante für Wort w und Endknoten t bei Vorgänger v
τ_e	Endknoten einer Wortgraphkante

τ_i	Wortgraphknoten
τ_s	Startknoten einer Wortgraphkante
t	Index eines Sprachrahmens
T	Anzahl der Sprachrahmen in einem Satz
T_α	Schwellwert für die Uncertainty-Decodierregel
T_{PTB}	Anzahl der Sprachrahmen bis ein Partial Trace-Back durchgeführt wird
$\mathbf{T}_{\boldsymbol{\Sigma_n}}$	Schwellwerte für die Kovarianzmatrix des Rauschens
T_{wp}	Schwellwert für die a posteriori Wahrscheinlichkeiten des Wortgraphen
T_{wg}	Schwellwert für das Wortgraphpruning bei der Erstellung des Wortgraphen
$\boldsymbol{\theta}$	Parametervektor
$\hat{\boldsymbol{\theta}}_{ML}$	ML-Schätzwert des Parametervektors
$\mathbf{u}_t$	Zustandsrauschen des Modells für den Zustandsvektor
$\tilde{\mathbf{u}}_t$	Zustandsrauschen des erweiterten Zustandsübergangsmodells
$\mathbf{v}_t$	Zustandsrauschen des Rauschmodells
$v_0(w;t)$	Bester Vorgänger des Wortes w mit Endknoten t
$\mathbf{v}_t^{(\eta)}$	Zustandsrauschen des erweiterten Rauschmodells für $\boldsymbol{\eta}_t$
$\mathbf{W}^{(n)}$	Beobachtungsvarianz des Rauschmodells
$\mathbf{V}$	Varianz des Beobachtungsmodells
$\mathbf{V}_c$	Konstanter Anteil der Varianz des Beobachtungsmodells
$\mathbf{V}^{(n)}$	Zusätzliche Varianz des Beobachtungsmodells aufgrund des Rauschmodells
$\mathcal{V}(w;t)$	Menge der besten Vorgänger des Wortes w mit Endknoten t
w_1^N	Wortsequenz
$\hat{w}_1^N$	Optimale Wortsequenz
$\mathbf{w}_t$	Beobachtungsrauschen
$\mathbf{w}_t^{(n)}$	Zusätzliches Beobachtungsrauschen aufgrund des Rauschmodells
$x(k)$	Unverrauschtes Sprachsignal im Zeitbereich
$x_t^{(l)}$	Komponente des unverrauschten, log-spektralen Merkmalsvektors
$\mathbf{x}_t$	Unverrauschte, cepstrale Sprachmerkmale
$y(k)$	Verrauschtes Sprachsignal im Zeitbereich
$y_t^{(l)}$	Komponente des verrauschten, log-spektralen Merkmalsvektors
$\mathbf{y}_t$	Verrauschte, cepstrale Sprachmerkmale
$\mathbf{z}_t$	Zustandsvektor des gemeinsamen Zustandsübergangsmodells für Sprache und Rauschen

Anhang G

Abkürzungsverzeichnis

Verfahren zur Merkmalsentstörung

AFE	(ETSI) Advanced Front-End zur Merkmalsextraktion und -entrauschung
EKF-Mx	SLDM mit x EKFs
EKF-d-Mx	SLDM mit dynamischen Sprachmerkmalen und x EKFs
IEKF-Mx	SLDM mit x IEKFs
IEKF-d-Mx	SLDM mit dynamischen Sprachmerkmalen und x EKFs
GMM-Mx	GMM mit x Mischungsverteilungen und iterativer Verbesserung einer SNR-Variablen entsprechend Anhang C.3
SFE	(ETSI) Standard Front-End zur Merkmalsextraktion
SLDM-Mx	Baseline SLDM mit x schaltenden Modellen
SLDM-Sx	Baseline SLDM mit x schaltenden Modellen und Glättung
SLDM-Mx-UD	Baseline SLDM mit x schaltenden Modellen und Uncertainty Decoding
SLDM-Mx-FB	Baseline SLDM mit x schaltenden Modellen und Rückkopplung von Informationen
VTS-Mx	GMM mit x Mischungsverteilungen und VTS-Verfahren entsprechend Anhang C.3

Sonstige Abkürzungen

AR	Auto-regressiv
ASR	Automatische Spracherkennung
AURORA2	Sprachdatenbank mit isolierten Ziffern
AURORA4	Sprachdatenbank mit kontinuierlich gesprochenen Sätzen
BEQ	Blinde Equalisierung
BMM	Markovmodell mit verdeckter Variable
CMHMM	Segmentielles Markovmodell basierend auf einer modifizierten Emissionsverteilung
CMN	Cepstrale Mittelwertnormalisierung
DAG	Gerichteter, azyklischer Graph
DARPA	Forschungsbehörde des US-Verteidigungsministeriums
DCT	Diskrete Kosinustransformation
DFG	Deutsche Forschungsgemeinschaft

DFT	Diskrete Fouriertransformation
EKF	Erweitertes Kalman-Filter
EKF-d-Mx	SLDM mit x EKFs und dynamischen Sprachmerkmalen
EM	Expectation-Maximization
ETSI	Europäisches Institut für Telekommunikationsnormen
FFT	Schnelle Fouriertransformation
GMM	Gaußmischungsmodell
GMM-Mx	GMM mit x Mischungskomponenten
GPB-Technik	Generalisierte Pseudo-Bayes'sche Technik
GSF	Gauß'sches Summenfilter
HMM	Markovmodell
HTHMM	Markovmodell mit versteckter Trajektorie
IDCT	Inverse diskrete Kosinustransformation
IEKF	Iteratives erweitertes Kalman-Filter
IMCRA	Iterative, durch Minima kontrollierte, rekusive Glättung
IMM	Interagierende Modelle (Interacting Multiple Models)
MAP	Maximale a posteriori Wahrscheinlichkeit
MCRA	Rekusive, durch Minima kontrollierte Glättung
MFCC	Cepstrale Koeffizienten (Mel Frequency Cepstral Coefficients)
MMSE	Mittlerer quadratischer Fehler
ML	Maximum Likelihood
LDA	Lineare Diskriminantenanalyse
LMMSE	Linearer, mittlerer quadratischer Fehler
LPC	Lineare Prädiktionskoeffizienten
NEC	Nippon Electric Company (japanischer, weltweit agierender Elektronikkonzern)
PaSCo	Paderborner Institut für Wissenschaftliches Rechnen
PCA	Principal Component Analysis
PLP	Lineare Prädiktion, an die menschliche Wahrnehmung angepaßt
PMC	Parallele Modellkombination
RCA	Ehemaliger US-Elektronikkonzern
RTS-Glätter	Rauch-Tung-Striebel-Glätter
SIR	Sampling-Importance-Resampling
SLDM	Schaltende, lineare Dynamikmodelle
SNR	Signal-zu-Rausch-Verhältnis
UKF	Unscented Kalman-Filter
VAD	Sprachaktivitätsdetektion
VTS	Vektortaylorreihe
WAC	Worterkennungsrate
WGD	Wortgraphdichte
WER	Wortfehlerrate
WGE	Wortgraphfehlerrate
WSJ	Wallstreet Journal

ZVM	Zero Variance Model

Literaturverzeichnis

[AC02] Arrowood, J.A. ; Clements, M.A.: Using Observation Uncertainty in HMM Decoding. In: *Proc. ICSLP* Bd. 3 ISCA, 2002, S. 1561–1564

[Ace93] Acero, A.: *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, 1993

[Afi05] Afify, M.: Accurate compensation in the log-spectral domain for noisy speech recognition. In: *IEEE Transactions on Audio, Speech and Language Processing* 13 (2005), Nr. 3, S. 388– 398

[AH04] Ahmed, B. ; Holmes, H.: A voice activity detector using the chi-square test. In: *Proc. ICASSP* Bd. 1 IEEE, 2004, S. 625–628

[And84] Anderson, T.W.: *An Introduction To Multivariate Statistical Analysis.* John Wiley and Sons, Inc., 1984

[AS72] Alspach, D.L. ; Sorenson, H.W.: Nonlinear Bayesian Estimation using Gaussian Sum Approximation. In: *IEEE Transactions on Automatic Control* 17 (1972), Nr. 4, S. 439–448

[Ata74] Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. In: *Journal of the Acoustical Society of America* 55 (1974), Nr. 6, S. 1304–1322

[Bel57] Bellman, R.-E.: *Dynamic programming.* Princeton University Press, 1957

[Bil03] Bilmes, J.A.: Buried markov models: a graphical-modeling approach to automatic speech recognition. In: *Computer Speech and Language* 17 (2003), Nr. 2-3, S. 213–231

[BMC05] Benesty, J. ; Makino, S. ; Chen, J.: *Speech Enhancement.* Springer, 2005

[Bol79] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 27 (1979), Nr. 2, S. 113–120

[BP66] Baum, L.E. ; Petri, T.: Statistical inference for probabilistic functions of finite state Markov chains. In: *Annals of Mathematical Statistics* 37 (1966), Nr. 6, S. 1554–1563

[Bri73] Bridle, J.S.: An efficient elastic template method for detecting keywords in running speech. In: *Proc. British Acoustical Society Meeting* Bd. 1 British Acoustical Society, 1973, S. 1–4

[Bro87] Brown, P.F.: *The Acoustic-Modeling Problem in Automatic Speech Recognition*, Carnegie Mellon University, Diss., 1987

[Bro92] Brown, G.J.: *Computational Auditory Scene Analysis: A Represential Approach.*, University of Sheffield, Diss., 1992

[BS91] Berstein, A. ; Shallom, I.: An hypothesized Wiener filtering approach to noisy speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 1991, S. 913–916

[BSMM01] Bronstein, I.N. ; Semendjajew, K.A. ; Musiol, G. ; Mühlig, H.: *Taschenbuch der Mathematik*. Verlag Harri Deutsch, 2001

[BSRLK01] Bar-Shalom, Y. ; Rong Li, X. ; Kirubarajan, T.: *Estimation with Applications to Tracking and Navigation*. John Wiley and Sons, Inc., 2001

[BST+04] Benitez, M.C. ; Segura, J.C. ; Torre, A. de l. ; Ramirez, J. ; Rubio, A.: Including uncertainty of speech observations in robust speech recognition. In: *Proc. ICSLP* Bd. 1 ISCA, 2004, S. 137–140

[BWJ01] Brown, G.J. ; Wang, D.L. ; J., Barker: A neural oscillator sound separator for missing data speech recognition. In: *IJCNN* Bd. 4 IEEE, 2001, S. 2907–2912

[BWN95] Beulen, K. ; Welling, L. ; Ney, H.: Experiments with linear feature extraction in speech recognition. In: *Proc. EUROSPEECH* Bd. 2 ISCA, 1995, S. 1415–1418

[CB01] Cohen, I. ; Berdugo, B.: Speech enhancement for non-stationary noise environments. In: *IEEE Signal Processing Letters* 81 (2001), Nr. 11, S. 2403–2418

[CB02] Cohen, I. ; Berdugo, B.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. In: *IEEE Signal Processing Letters* 9 (2002), Nr. 1, S. 12–15

[DA04] Droppo, J. ; Acero, A.: Noise robust speech recognition with a switching linear dynamic model. In: *Proc. ICASSP* Bd. 1 IEEE, 2004, S. 953–956

[DAD02a] Droppo, J. ; Acero, A. ; Deng, L.: A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies. In: *Proc. ICSLP* Bd. 1 IEEE, 2002, S. 182–185

[DAD02b] Droppo, J. ; Acero, A. ; Deng, L.: Uncertainty decoding with SPLICE for noise robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2002, S. 57–60

[DAPH00] Deng, L. ; Acero, A. ; Plumpe, M. ; Huang, X.D.: Large vocabulary speech recognition under adverse acoustic environments. In: *Proc. ICSLP* Bd. 1 ISCA, 2000, S. 806–809

[DBB52] Davis, K.-H. ; Biddulph, R. ; Balashek, S: Automatic Recognition of Spoken Digits. In: *Journal of the Acoustical Society of America* 24 (1952), Nr. 6, S. 637–642

[DBY06] Deng, J. ; Bouchard, M. ; Yeap, T.H.: Speech feature estimation under the presence of noise with a switching linear dynamic model. In: *Proc. ICASSP* Bd. 1 IEEE, 2006, S. 497–500

[DDA02] Deng, L. ; Droppo, J. ; Acero, A.: Log-domain speech feature enhancement using sequential map noise estimation and a phase-sensitive model of the acoustic environment. In: *Proc. ICSLP* Bd. 1 ISCA, 2002, S. 192–195

[DDA03a] Deng, L ; Droppo, J. ; Acero, A.: Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. In: *IEEE Transactions on Speech and Audio Processing* 11 (2003), Nr. 6, S. 568–580

[DDA03b] Droppo, J. ; Deng, L. ; Acero, A.: A comparison of three non-linear observation models for noisy speech features. In: *Proc. EUROSPEECH* Bd. 1 ISCA, 2003, S. 681–684

[DDA04a] Deng, L. ; Droppo, J. ; Acero, A.: Enhancement of log-spectra of speech using a phase-sensitive model of the acoustic environment. In: *IEEE Transactions on Speech and Audio Processing* 12 (2004), Nr. 3, S. 133–143

[DDA04b] Deng, L. ; Droppo, J. ; Acero, A.: Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features. In: *IEEE Transactions on Speech and Audio Processing* 12 (2004), Nr. 3, S. 218–233

[DDA05] Deng, L. ; Droppo, J. ; Acero, A.: Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. In: *IEEE Transactions on Speech and Audio Processing* 13 (2005), Nr. 3, S. 412–421

[Dig92] Digalakis, L.: *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*, Boston University, Diss., 1992

[DM80] Davis, S.B. ; Mermelstein, P.: Comparison of parametric representations for monosyllable word recognition in continously spoken sentences. In: *IEEE Transactions on Speech and Audio Processing* 28 (1980), Nr. 4, S. 357–366

[DO03] Deng, L. ; O'Shaughnessy, D.-O.: *Speech Processing: A Dynamic and Optimization-Oriented Approach*. Marcel Dekker, Inc., 2003

[DVCW02] Demuynck, K. ; Van Compernolle, D. ; Wambacq, P.: Doing away with the Viterbi approximation. In: *Proc. ICASSP* Bd. 1 IEEE, 2002, S. 717–720

[EM84] Ephraim, Y. ; Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 32 (1984), Nr. 6, S. 1109–1112

[EMJ89] Ephraim, Y. ; Malah, D. ; Juang, B.H.: On the application of hidden Markov models for enhancing noisy speech. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 37 (1989), Nr. 12, S. 1846–1856

[Eph92] Ephraim, Y.: Speech enhancement using state dependent dynamical system model. In: *Proc. ICASSP* Bd. 1 IEEE, 1992, S. 289–292

[ETS00] ETSI: *ES 201 108, Standard front-end feature extraction algorithm.* V1.1.2. 2000

[ETS05] ETSI: *ES 202 050, Advanced front-end feature extraction algorithm.* V1.1.4. 2005

[EW93a] Erell, A. ; Weintraub, M.: Energy conditioned spectral estimation for recognition of noisy speech. In: *IEEE Transactions on Speech and Audio Processing* 1 (1993), Nr. 1, S. 84–89

[EW93b] Erell, A. ; Weintraub, M.: Filterbank-energy estimation using mixture and Markov models for recognition of noisy speech. In: *IEEE Transactions on Speech and Audio Processing* 1 (1993), Nr. 1, S. 68–76

[EW00] Evermann, G. ; Woodland, P.C.: Large vocabulary decoding and confidence estimation using word posterior probabilities. In: *Proc. ICASSP* Bd. 3 IEEE, 2000, S. 1655–1658

[FA00] Fujimoto, M. ; Ariki, Y.: Noisy speech recognition using noise reduction method based on Kalman filter. In: *Proc. ICASSP* Bd. 1 IEEE, 2000, S. 1727–1730

[FA04] Fujimoto, M. ; Ariki, Y.: Robust speech recognition in additive and channel noise environments using GMM and EM algorithm. In: *Proc. ICASSP* Bd. 1 IEEE, 2004, S. 941–944

[FDAK03] Frey, B. ; Deng, L. ; Acero, A. ; Kristjansson, T.: ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In: *Proc. EUROSPEECH* Bd. 1 ISCA, 2003, S. 901–904

[FF59] Forgie, J.-W. ; Forgie, C.-D.: Results obtained from a vowel recognition computer program. In: *Journal of the Acoustical Society of America* 31 (1959), Nr. 11, S. 1480–1489

[FG01] Fong, W. ; Godsill, S.: Monte Carlo smoothing with application to audio signal enhancement. In: *SSP Workshop* Bd. 1 IEEE, 2001, S. 18–210

[FGDW02] Fong, W. ; Godsill, S. ; Doucet, A. ; West, M.: Monte Carlo smoothing with application to audio signal enhancement. In: *IEEE Transactions on Signal Processing* 50 (2002), Nr. 2, S. 438–449

[FK07] Frankel, J. ; King, S.: Speech recognition using linear dynamic models. In: *IEEE Transactions on Audio Speech and Language Processing* 15 (2007), Nr. 1, S. 213–231

[FN05] Fujimoto, M. ; Nakamura, S.: Particle Filter Based Non-Stationary Noise Tracking For Robust Speech Recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2005, S. 257–260

[FN06] Fujimoto, M. ; Nakamura, S.: A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging. In: *IEICE Transactions on Information and Systems* E89-D (2006), Nr. 3, S. 922–930

[Fry59] Fry, D.-B.: Theoretical Aspects of Mechanical Speech Recognition. In: *British Inst. Radio Engr.* 19 (1959), Nr. 4, S. 211–229

[Fuk90] Fukunaga, K.: *Statistical Pattern Recognition.* Academic Press, Inc., 1990

[FW06] Faubel, F. ; Wölfel, M.: Coupling Particle Filters with Automatic Speech Recognition for Speech Feature Enhancement. In: *Proc. Interspeech* Bd. 1 ISCA, 2006, S. 37–41

[Gag93] Gagnon, L.: A state-based noise reduction approach for non-stationary additive interference. In: *Speech Communication* 12 (1993), Nr. 3, S. 213–219

[Gal95] Gales, M.-J.F.: *Model-based techniques for noise robust speech recognition*, University of Cambridge, Diss., 1995

[GBW98] Gannot, S. ; Burshtein, D. ; Weinstein, E.: Iterative and sequential Kalman filter-based speech enhancement algorithms. In: *IEEE Transactions on Speech and Audio Processing* 6 (1998), Nr. 4, S. 373–385

[GC88] Gillick, L. ; Cox, S.J.: Some statistical issues in the comparison of speech recognition algorithms. In: *Proc. ICASSP* Bd. 1 IEEE, 1988, S. 532–535

[GM03] Gannot, S. ; Moonen, M.: On the application of the unscented Kalman filter to speech processing. In: *IWAENC* Bd. 1 IEEE, 2003, S. 27–30

[Gon03] Gong, Y.: Model-space compensation of microphone and noise for speaker-independent speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2003, S. 660–663

[HE95] Hirsch, H.G. ; Ehrlicher, C.: Noise estimation techniques for robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 1995, S. 153–156

[Her90] Hermansky, H.: Perceptual Linear Predictive (PLP) analysis of speech. In: *Journal of the Acoustical Society of America* 87 (1990), Nr. 4, S. 1738–1752

[Hir02] Hirsch, G.: Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task. In: *STQ AURORA DSR WORKING GROUP* Bd. 1 ETSI, 2002

[HJH07] Hendriks, R. C. ; Jensen, J. ; Heusdens, R.: DFT Domain Subspace Based Noise Tracking for Speech Enhancement. In: *Proc. Interspeech* Bd. 1 ISCA, 2007, S. 830–833

[HL89] Hunt, M. ; Lefebvre, C.: A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In: *Proc. ICASSP* Bd. 1 IEEE, 1989, S. 262–265

[HM94] Hermansky, H. ; Morgan, N.: RASTA processing of speech. In: *IEEE Transactions on Speech and Audio Processing* 2 (1994), Nr. 4, S. 578–589

[HU05a] Haeb-Umbach, R.: *Algorithmen der Spracherkennung*. 2004/05. – Skript zur Vorlesung, Universität Paderborn

[HU05b] Haeb-Umbach, R.: *Optimale und adaptive Filter*. 2004/05. – Skript zur Vorlesung, Universität Paderborn

[HUN92] Haeb-Umbach, R. ; Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 1992, S. 13–16

[HW04] Hermus, K. ; Wambacq, P.: Assessment of signal subspace based speech enhancement for noise robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2004, S. 945–948

[IHU06] Ion, V. ; Haeb-Umbach, R.: Uncertainty decoding for distributed speech recognition over error-prone networks. In: *Speech Communication* 48 (2006), Nr. 11, S. 1435–1446

[IHU08a] Ion, V. ; Haeb-Umbach, R.: *Investigations into uncertainty decoding employing a discrete feature space for noise robust ASR.* 2008. – accepted for ITG Fachtagung Sprachkommunikation

[IHU08b] Ion, V. ; Haeb-Umbach, R.: *A novel uncertainty decoding rule with applications to transmission error robust speech recognition.* 2008. – accepted for IEEE Transactions on Acoustics, Speech and Signal Processing

[Ita75] Itakura, F.: Minimum Prediction Residual Applied to Speech Recognition. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 23 (1975), Nr. 1, S. 67–72

[JBM75] Jelinek, F. ; Bahl, L.R. ; Mercer, R.L.: Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech. In: *IEEE Transactions on Information Theory* 21 (1975), Nr. 3, S. 250–256

[Jel76] Jelinek, F.: Continuous speech recognition by statistical methods. In: *Proc. IEEE* 64 (1976), Nr. 10, S. 532–556

[Jel85] Jelinek, F.: The Development of an Experimental Discrete Dictation Recognizer. In: *Proc. IEEE* 73 (1985), Nr. 11, S. 1616–1624

[JH96] Junqua, J.-C. ; Haton, J.-P.: *Robustness in automatic speech recognition.* Kluwer Academic Publishers, 1996

[KF02] Kristjansson, T.T. ; Frey, B.J.: Accounting for uncertainty in observations: A new paradigm for robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2002, S. 61–64

[Kim98a] Kim, N.S.: IMM-based estimation for slowly evolving environments. In: *IEEE Signal Processing Letters* 5 (1998), Nr. 6, S. 146–149

[Kim98b] Kim, N.S.: Nonstationary environment compensation based on sequential estimation. In: *IEEE Signal Processing Letters* 5 (1998), Nr. 3, S. 57–59

[Kim98c] Kim, N.S.: Statistical linear approximation for environment compensation with noise statistics. In: *IEEE Signal Processing Letters* 5 (1998), Nr. 1, S. 8–10

[Kim02] Kim, N.S.: Feature domain compensation of nonstationary noise for robust speech recognition. In: *Speech Communication* 37 (2002), Nr. 3-4, S. 231–248

[KKKK97] Kim, N.S. ; Kim, D.Y. ; Kong, B.G. ; Kim, S.R.: Application of VTS to environment compensation with noise statistics. In: *Workshop on Robust Speech Recognition for Unknown Communication Channels* Bd. 1 ESCA, 1997, S. 99–102

[KLL00] Kim, J.B. ; Lee, K.Y. ; Lee, C.W.: On the applications of the interacting multiple model algorithm for enhancing noisy speech. In: *IEEE Transactions on Speech and Audio Processing* 8 (2000), Nr. 3, S. 349–352

[KLS05] Kim, N.S. ; Lim, W. ; Stern, R.M.: Feature compensation based on switching linear dynamic model. In: *IEEE Signal Processing Letters* 12 (2005), Nr. 6, S. 473–476

[Koo89] Koo, B.: Filtering of colored noise for speech enhancement and coding. In: *Proc. ICASSP* Bd. 1 IEEE, 1989, S. 349–352

[KSS06] Kato, M. ; Sugiyama, A. ; Serizawa, M.: Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA. In: *Electronics and Communications in Japan* 89 (2006), Nr. 2, S. 43–53

[KUK98] Kim, D.Y. ; Un, C.K. ; Kim, N.S.: Speech recognition in noisy environments using first-order vector Taylor series. In: *Speech Communication* 24 (1998), Nr. 1, S. 39–49

[LG04] Liao, H. ; Gales, M.J.F.: Issues with Uncertainty Decoding for Noise Robust Speech Recognition / Cambridge University Engineering Department. 2004 (499). – Forschungsbericht

[LG06] Liao, H. ; Gales, M.J.F.: Issues with Uncertainty Decoding for Noise Robust Speech Recognition. In: *Proc. Interspeech* Bd. 1 ISCA, 2006, S. 1627–1630

[LG08] Liao, H. ; Gales, M.J.F.: Issues with Uncertainty Decoding for Noise Robust Automatic Speech Recognition. In: *Speech Communication* 50 (2008), Nr. 4, S. 265–277

[Lim78] Lim, J.: Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 26 (1978), Nr. 5, S. 471–472

[Lip82] Liporace, L.: Maximum likelihood estimation for multi-variate observations of Markov sources. In: *IEEE Transactions on Information Theory* 28 (1982), Nr. 5, S. 729–734

[LO78] Lim, J.S. ; Oppenheim, A.V.: All-pole modeling of degraded speech. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 26 (1978), Nr. 3, S. 197–210

[LRS83] Levinson, S.E. ; Rabiner, L.R. ; Sondhi, M.M.: An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. In: *Bell System Technical Journal* 62 (1983), Nr. 4, S. 1035–1074

[LS96] Lee, K.Y. ; Shirai, K.: Efficient recursive estimation for speech enhancement in colored noise. In: *IEEE Signal Processing Letters* 3 (1996), Nr. 7, S. 196–199

[LW95] Leggetter, C.J. ; Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In: *Computer Speech and Language* 9 (1995), Nr. 2, S. 171–185

[MAC03] Morris, R.W. ; Arrowood, J.A. ; Clements, M.A.: Markov chain Monte Carlo methods for noise robust feature extraction using autoregressive model. In: *Proc. EUROSPEECH* Bd. 1 ISCA, 2003, S. 3097–3100

[Man84] Mandel, J.: *The Statistical Analysis of Experimental Data.* Springer, 1984

[Mar94] Martin, R.: Spectral subtraction based on minimum statistics. In: *EUSIPCO* Bd. 1 EURASIP, 1994, S. 1182–1185

[Mar96] Markowitz, J.-A.: *Using speech recognition.* Prentice Hall PTR, 1996

[Mar01] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. In: *IEEE Transactions on Speech and Audio Processing* 9 (2001), Nr. 5, S. 504–512

[Mar02] Martin, R.: Speech enhancement using MMSE short time spectral estimation with gamma distributed priors. In: *Proc. ICASSP* Bd. 1 IEEE, 2002, S. 253–256

[MBB01] Morris, A. ; Barker, J. ; Bourland, H.: From missing data to maybe useful data: Soft data modelling for noise robust ASR. In: *WISP* Bd. 6 IEEE, 2001, S. 153–164

[MCA99] Malah, D. ; Cox, R.V. ; Accardi, A.J.: Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In: *Proc. ICASSP* Bd. 1 IEEE, 1999, S. 17–20

[MD04] Ma, J. ; Deng, L.: Target-directed mixture linear dynamic models for spontaneous speech recognition. In: *IEEE Transactions on Speech Audio Processing* 12 (2004), Nr. 1, S. 47–58

[MHN03] Molau, S. ; Hilger, F. ; Ney, H.: Feature space normalization in adverse acoustic conditions. In: *Proc. ICASSP* Bd. 1 IEEE, 2003, S. 656–659

[MMC+02] Macho, D. ; Mauuary, L. ; Cheng, Y.M. ; Ealey, D. ; Jouvet, D. ; Kelleher, H. ; Pearce, D. ; Saadoun, F.: Evaluation of a noise-robust DSR front-end on Aurora databases. In: *Proc. ICSLP* Bd. 1 ISCA, 2002, S. 17–20

[MN03] Myrvoll, T.A. ; Nakamura, S.: Optimal Filtering of Noisy Cepstral Coefficients for Robust ASR. In: *ASRU* Bd. 1 IEEE, 2003, S. 381–386

[MNZ64] Martin, T.-B. ; Nelson, A.-L. ; Zadell, H.J.: Speech Recognition by Feature Abstraction Techniques / Air Force Avionics Lab. 1964 (AL-TDR-64-176). – Forschungsbericht

[Mor96] Moreno, P.J.: *Speech recognition in noisy environments*, Carnegie Mellon University, Diss., 1996

[MRGS95] Moreno, P.J. ; Raj, B. ; Gouvea, E. ; Stern, R.M.: Multivariate-Gaussian-based cepstral normalization for robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 1995, S. 137–140

[MW97] McKinley, B.L. ; Whipple, G.H.: Model based speech pause detection. In: *Proc. ICASSP* Bd. 1 IEEE, 1997, S. 1179–1182

[NKC63] Nakata, K ; Kato, Y. ; Chiba, S.: Spoken Digit Recognizer for Japanese Language / NEC Res. Develop. 1963 (6/63). – Forschungsbericht

[NN88] Ney, H. ; Noll, A.: Phoneme modeling using continuous mixture densities. In: *Proc. ICASSP* Bd. 1 IEEE, 1988, S. 437–440

[OB56] Olson, H.-F. ; Belar, H.: Phonetic Typewriter. In: *Journal of the Acoustical Society of America* 28 (1956), Nr. 6, S. 1072–1081

[ONA97] Ortmanns, S. ; Ney, H. ; Aubert, X.: A word graph algorithm for large vocabulary continuous speech recognition. In: *Computer Speech and Language* 11 (1997), Nr. 1, S. 43–72

[OR89] Ostendorf, M. ; Roukos, S.: A stochastic segment model for phoneme-based continuous speech recognition. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 37 (1989), Nr. 12, S. 1857–1869

[PB87] Paliwal, K.K. ; Basu, A.: A speech enhancement method based on Kalman filtering. In: *Proc. ICASSP* Bd. 1 IEEE, 1987, S. 177–180

[PH00] Pearce, D. ; Hirsch, H.-G.: The Aurora Experimental Framework for the performance evaluation of speech recognition systems under noisy conditions. In: *Proc. ICSLP* Bd. 1 ISCA, 2000, S. 29–32

[PPPH04] Parihar, N. ; Picone, J. ; Pearce, D. ; Hirsch, G.: Performance analysis of the Aurora large vocabulary baseline system. In: *Proc. EUSIPCO* Bd. 1 EURASIP, 2004, S. 553–556

[RD01] Ris, C. ; Dupont, S.: Assessing local noise level estimation methods: Application to noise robust ASR. In: *Speech Communication* 34 (2001), Nr. 1-2, S. 141–158

[RDVGS97] Rubio, A.J. ; Diaz-Verdejo, J.E. ; Garcia, P. ; Segura, J.C.: On the influence of frame-asynchronous grammar scoring in a CSR system. In: *Proc. ICASSP* Bd. 2 IEEE, 1997, S. 895–898

[Red66] Reddy, D.-R.: An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave / Stanford University. 1966 (C549). – Forschungsbericht

[RGMS96] Raj, B. ; Gouvea, E.B. ; Moreno, P.J. ; Stern, R.M.: Cepstral compensation by polynomial approximation for environment-independent speech recognition. In: *Proc. ICSLP* Bd. 4 ISCA, 1996, S. 2340–2343

[RH97] Russel, M.-J. ; Holmes, W.-J.: Linear Trajectory Segmental HMMs. In: *IEEE signal processing letters* 4 (1997), Nr. 3, S. 72–74

[RJ93] Rabiner, L. ; Juang, B.-H.: *Fundamentals of speech recognition*. Prentice Hall Signal Processing Series, 1993

[RL06] Rangachari, S. ; Loizou, P.C.: A noise-estimation algorithm for highly non-stationary environments. In: *Speech Communication* 48 (2006), Nr. 2, S. 220–231

[RLRW79] Rabiner, L.-R. ; Levinson, S.-E. ; Rosenberg, A.-E. ; Wilpon, J.-G.: Speaker independent recognition of isolated words using clustering techniques. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 27 (1979), Nr. 4, S. 336–349

[Ros04] Rosti, A.-V.I.: *Linear Gaussian Models for Speech Recognition*, University of Cambridge, Diss., 2004

[RSG04] Ristic, B. ; S., Arulampalam ; Gordon, N.: *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Radar Library, 2004

[RSS04] Raj, B. ; Singh, R. ; Stern, R.: On tracking noise with linear dynamical system models. In: *Proc. ICASSP* Bd. 1 IEEE, 2004, S. 965–968

[SA91] Schwartz, R. ; Austin, S.: A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses. In: *ICASSP* Bd. 1 IEEE, 1991, S. 701–704

[SC71] Sakoe, H. ; Chiba, S.: Recognition of continuously spoken words based on time-normalization by dynamic programming. In: *Journal of the Acoustical Society of America* 27 (1971), Nr. 9, S. 483–490

[SC78] Sakoe, H. ; Chiba, S.: Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In: *IEEE Transactions on Acoustics, Speech And Signal Processing* 26 (1978), Nr. 1, S. 43–49

[SD62] Sakai, J. ; Doshita, S.: The Phonetic Typewriter. In: *Proc. of the International Federation for Information Processing Congress* Bd. 1 IFIP, 1962, S. 445–450

[SG07] Sim, K.C. ; Gales, M.J.F.: Discriminative semi-parametric trajectory model for speech recognition. In: *Computer Speech and Language* 21 (2007), Nr. 4, S. 669–687

[SHU05] Schmalenstroer, J. ; Haeb-Umbach, R.: A comparison of particle filtering variants for speech feature enhancement. In: *Proc. Interspeech* Bd. 1 ISCA, 2005, S. 913–916

[SN61] Suzuki, J. ; Nakata, K.: Recognition of Japanese Vowels-Preliminary to the Recognition of Speech. In: *Radio Research Lab* 37 (1961), Nr. 8, S. 193–212

[SN94] Seymour, C.W. ; Niranjan, M.: An HMM-based cepstral domain speech enhancement system. In: *Proc. ICSLP* Bd. 1 ISCA, 1994, S. 1595–1598

[SP04] Schlegel, C.B. ; Perez, L.C.: *Trellis and Turbo Coding*. IEEE Press Series on Digital and Mobile Communication, 2004

[SR03] Singh, R. ; Raj, B.: Tracking noise via dynamical systems with a continuum of states. In: *Proc. ICASSP* Bd. 1 IEEE, 2003, S. 396–399

[SS98] Sohn, J. ; Sung, W.: A voice activity detector employing soft decision based noise spectrum adaptation. In: *Proc. ICASSP* Bd. 1 IEEE, 1998, S. 365–399

[STBP01] Segura, J.C. ; Torre, A. de l. ; Benitez, M.C. ; Peinado, A.M.: Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks. In: *Proc. EUROSPEECH* Bd. 1 ISCA, 2001, S. 221–224

[Sto06] Stouten, V.: *Robust Automatic Speech Recognition in Time-varying Environments*, K.U. Leuven, Diss., 2006

[SVhDW03] Stouten, V. ; Van hamme, H. ; Demuynck, K. ; Wambacq, P.: Robust speech recognition using model-based feature enhancement. In: *Proc. EUROSPEECH* Bd. 1 ISCA, 2003, S. 17–20

[SVhW04] Stouten, V. ; Van hamme, H. ; Wambacq, P.: Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In: *Proc. ICSLP* Bd. 1 ISCA, 2004, S. 105–108

[SVhW05a] Stouten, V. ; Van hamme, H. ; Wambacq, P.: Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement. In: *Proc. ICASSP* Bd. 1 IEEE, 2005, S. 433–436

[SVhW05b] Stouten, V. ; Van hamme, H. ; Wambacq, P.: Kalman and unscented Kalman filter feature enhancement for noise robust ASR. In: *Proc. Interspeech* Bd. 1 ISCA, 2005, S. 953–956

[SVhW06a] Stouten, V. ; Van hamme, H. ; Wambacq, P.: Application of minimum statistics and minima controlled recursive averaging methods to estimate a cepstral noise model for robust ASR. In: *Proc. ICASSP* Bd. 1 IEEE, 2006, S. 765–768

[SVhW06b] Stouten, V. ; Van hamme, H. ; Wambacq, P.: Model-based feature enhancement with uncertainty decoding for noise robust ASR. In: *Speech Communication* 48 (2006), Nr. 11, S. 1616–1624

[SZD03] Seide, F. ; Zhou, J. ; Deng, L.: Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM-MAP decoding and evaluation. In: *Proc. ICASSP* Bd. 1 IEEE, 2003, S. 748–751

[TDRC71] Tappert, C.-C. ; Dixon, N.-R. ; Rabinowitz, A.-S. ; Chapman, W.-D.: Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery / Rome Air Dev. Cen. 1971 (TR-71-146). – Forschungsbericht

[Tit84] Titterington, D.M.: Recursive parameter estimation using incomplete data. In: *Journal of the Royal Statistical Society* 46 (1984), Nr. 2, S. 257–267

[TSB+02] Torre, A. De l. ; Segura, J.C. ; Benitez, C. ; Peinado, A.M. ; Rubio, A.J.: Non-linear transformations of the feature space for robust speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 2002, S. 401–404

[TZK03] Tokuda, K. ; Zen, H. ; Kitamura, T.: Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In: *EUROSPEECH* Bd. 1 ISCA, 2003, S. 865–868

[VADG02] Vermaak, J. ; Andrieu, C. ; Doucet, A. ; Godsill, S.: Particle methods for bayesian modeling and enhancement of speech signals. In: *IEEE Transactions on Speech and Audio Processing* 10 (2002), Nr. 3, S. 173–185

[VC89] Van Compernolle, D.: Spectral estimation using a log-distance error criterion applied to speech recognition. In: *Proc. ICASSP* Bd. 1 IEEE, 1989, S. 258–261

[Vin68] Vintsjuk, S.: Speech discrimination by dynamic programming. In: *Kibernetika* 4 (1968), Nr. 1, S. 81–88

[Vin69] Vincens, P.: *Aspects of Speech Recognition by Computers*, Stanford University, Diss., 1969

[Vit67] Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. In: *IEEE Transactions on Information Theory* 13 (1967), Nr. 1, S. 260–269

[VM90] Varga, A.P. ; Moore, R.K.: Hidden Markov model decomposition of speech and noise. In: *Proc. ICASSP* Bd. 1 IEEE, 1990, S. 845–848

[VZ70] Velichko, V.-M. ; Zagoruyko, N.-G.: Automatic Recognition of 200 Words. In: *International Journal of Man-Machine Studies* 2 (1970), Nr. 3, S. 223–234

[WAP74] Weiss, M. ; Aschkenasy, E. ; Parsons, T.: Processing speech signals to attenuate interference. In: *Symposium of Speech Recognition* Bd. 1 IEEE, 1974, S. 292–293

[Wes02] Wessel, F.: *Word Posterior Probabilities for Large Vocabulary Continuous Speech Recognition*, RWTH Aachen, Diss., 2002

[WHU06a] Windmann, S. ; Haeb-Umbach, R.: Einkanalige Sprachsignalverbesserung mit Hilfe eines marginalisierten Partikelfilters. In: *7. Fachtagung Sprach-Kommunikation* Bd. 1 ITG, 2006, S. 133–136

[WHU06b] Windmann, S. ; Haeb-Umbach, R.: Iterative speech enhancement with a non-linear model of speech and its parameters. In: *Proc. ICASSP* Bd. 1 IEEE, 2006, S. 465–468

[WHU07] Windmann, S. ; Haeb-Umbach, R.: An iterative approach to speech feature enhancement and recognition. In: *INTERSPEECH* Bd. 1 ISCA, 2007, S. 1086–1089

[WHU08a] Windmann, S. ; Haeb-Umbach, R.: Modelling the Dynamics of Speech and Noise for Speech Feature Enhancement in ASR. In: *Proc. ICASSP* Bd. 1 IEEE, 2008, S. 4409–4412

[WHU08b] Windmann, S. ; Haeb-Umbach, R.: *A novel approach to noise estimation in model-based speech feature enhancement*. 2008. – accepted for 8. ITG Fachtagung Sprach-Kommunikation

[WHUL08] Windmann, S. ; Haeb-Umbach, R. ; Leutnant, V.: *A segmental HMM based on a modified emission probability.* 2008. – accepted for 8. ITG Fachtagung Sprach-Kommunikation

[XRK06] Xu, H. ; Rigazio, L. ; Kryze, D.: Vector Taylor Series based Joint Uncertainty Decoding. In: *Proc. Interspeech* Bd. 1 ISCA, 2006, S. 1688–1691

[YEH+02] Young, S. ; Evermann, G. ; Hain, T. ; Kershaw, D. ; Moore, G. ; Odell, J. ; Ollason, D. ; Povey, D. ; Valtchev, V. ; Woodland, P.: *The HTK Book (for HTK Version 3.2.1).* Cambridge University Engineering Department, 2002

[YL04] Yao, K. ; Lee, T.W.: Time-varying noise estimation for speech enhancement and recognition using sequential Monte Carlo methods. In: *EURASIP Journal on Applied Signal Processing* 15 (2004), Nr. 1, S. 2366–2384

[YSFK07] Yang, C. ; Soong ; F.-K., T.: Static and Dynamic Spectral Features: Their Noise Robustness and Optimal Weights for ASR. In: *IEEE Transactions on Information Theory* 15 (2007), Nr. 3, S. 1087–1097

[YSW07] Yan, Z.-J. ; Soong, F.-K. ; Wang, R.-H.: Word graph based feature enhancement for noisy speech recognition. In: *Proc. ICASSP* Bd. 4 IEEE, 2007, S. 373–376

[ZVY06] Zavarehei, E. ; Vaseghi, S. ; Yan, Q.: Inter-frame modeling of DFT trajectories of speech and noise for speech enhancement using Kalman filters. In: *Speech Communication* 48 (2006), Nr. 11, S. 1545–1555

Publikationsliste

Windmann, S. ; Haeb-Umbach, R.: Einkanalige Sprachsignalverbesserung mit Hilfe eines marginalisierten Partikelfilters. In: *7. Fachtagung Sprach-Kommunikation Bd. 1 ITG*, 2006, S. 133-136

Windmann, S. ; Haeb-Umbach, R.: Iterative speech enhancement with a non-linear model of speech and its parameters. In: *Proc. ICASSP Bd. 1 IEEE*, 2006, S. 465-468

Windmann, S. ; Haeb-Umbach, R.: An iterative approach to speech feature enhancement and recognition. In: *Proc. Interspeech* Bd. 1 ISCA, 2007, S. 1086-1089

Windmann, S. ; Haeb-Umbach, R.: Modelling the Dynamics of Speech and Noise for Speech Feature Enhancement in ASR. In: *Proc. ICASSP* Bd. 1 IEEE, 2008, S.4409-4412

Windmann, S. ; Haeb-Umbach, R.: *A novel approach to noise estimation in model-based speech feature enhancement.* 2008. - accepted for 8. ITG Fachtagung Sprach-Kommunikation

Windmann, S. ; Haeb-Umbach, R. ; Leutnant, V.: *A segmental HMM based on a modified emission probability.* 2008. - accepted for 8. ITG Fachtagung Sprach-Kommunikation

Windmann, S. ; Haeb-Umbach, R.: *Approaches to iterative speech feature enhancement and recognition.* 2008. - accepted for IEEE Transactions on speech and language processing

Windmann, S. ; Haeb-Umbach, R.: *Parameter Estimation of a State Space Model of Noise for Robust Speech Recognition.* 2008. - accepted for IEEE Transactions on speech and language processing